InterpreTutor: Using Large Language Models for Interpreter Assessment

Cihan ÜNLÜ[1][0000-0001-7857-2031]

[1] Istanbul Yeni Yüzyıl University, Istanbul, Turkiye
cihan.unlu@yeniyuzyil.edu.tr

Abstract. The recent development of large language models (LLMs) has shown remarkable natural language processing capabilities and created new possibilities for applications in various fields. With their advanced NLP features, such as text comparison, semantic analysis, text summarization, text classification, and text completion, this study aims to investigate whether these models have the potential to be used as tools for evaluating translation/interpretation output in a textual representation. In this paper, we question the capabilities of an LLM and propose InterpreTutor[1], an LLM-powered application with simple UI that utilizes Generative Pre-trained Transformer (GPT-3.5 turbo and GPT-4 models) and speech recognition (OpenAI Whisper) to offer detailed feedback on interpreters' performances based on automatic analysis of the transcriptions of their interpreting practice. The tool's primary focus is to act as an easy-to-use self-tutoring tool, offering feedback based on four human evaluation criteria. While InterpreTutor may not cover all aspects of interpreting performance, such as prosodic features of the delivery, it still provides insights into aspects that can be assessed through textual representation. In this paper, we will discuss the potential of LLMs, the development of InterpreTutor, its underlying methodology, and provide examples of its application with a small-scale experiment.

Keywords: large language models, interpreter training, artificial intelligence, speech recognition, automatic interpreter assessment, gpt-3.5, gpt-4

1 Introduction

The rise of large language models (LLMs) and LLM-powered applications, such as ChatGPT[2] and DALL-E 2[3] by OpenAI, Claude[4] by Anthropic, PaLM[5] by Google, LLaMA[6] by Meta, have brought about new avenues for the full and semi-automatization of the complex natural language tasks. The wide-ranging impacts of

[1] Publicly available at https://www.interpretutor.app
[2] https://openai.com/blog/chatgpt
[3] https://openai.com/product/dall-e-2
[4] https://www.anthropic.com/index/introducing-claude
[5] https://ai.googleblog.com/2022/04/pathways-language-model-palm-scaling-to.html
[6] https://ai.facebook.com/blog/large-language-model-llama-meta-ai/

Proceedings of the International Conference HiT-IT 2023, pages 78–96,
7-9 July 2023, Naples, Italy
https://doi.org/10.26615/issn.2683-0078.2023_007

LLM and their advanced natural language processing (NLP) capabilities hold significant promise for future derivative LLM-powered end-user applications like support agent assist, personalized tutoring, grammar correction and style, frontend/website generation, database query optimization, code generation and autocomplete, copywriting, etc. Github Copilot, Bing Search, ZoomIQ, Khanmigo, Duolingo Max, InworldAI, and many other end-user software are commercial examples on the market. The field of translation and interpreting is set to experience such transformative advancements, too. Currently, such advancements are limited to possible benefits such as document-level machine translation (MT) [1], producing synthetic data for training better MT engines, term extraction, glossary creation, and conducting quality estimation. Still, LLMs have the potential to be an integral part of the hybrid workflow of language service providers soon.

This study aims to investigate the capabilities of two of the most advanced LLMs of Open AI: GPT-3.5 and GPT-4, which are the base models used in the publicly available ChatGPT[7], with a particular focus on interpreter assessment. To see the generative features of an LLM in interpreting evaluation, we introduce an LLM-powered application *InterpreTutor* designed to transcribe the audio recordings and offer comprehensive feedback for performances based on various human assessment criteria. Details and usage scenarios for LLMs in the literature are briefly discussed in 1.1. The architecture of InterpreTutor is outlined in section two. On the other hand, section three presents the methodology of the small-scale experiment comparing the InterpreTutor's performance in GPT-3.5 and GPT-4. Further, an experimental setup was designed to include two trainee interpreters whose performances were recorded in audio and used as input for our tool. InterpreTutor's automatic performance analysis was conducted by leveraging GPT-3.5 and GPT-4. In its analysis, InterpreTutor gave scorings for each of the four criteria. Afterward, the scorings were compared to human evaluators' scorings. We aimed to understand its effectiveness in a restricted task: understanding the context, feedback generation on the completeness of a text, evaluating coherence, handling terminological consistency across texts, and providing an accurate report.

1.1 Large Language Models

LLMs refer to large-scale deep learning models trained on massive amounts of text data to perform various natural language processing tasks. Among these models, the Generative Pre-trained Transformer (GPT) models [2][3] have gained considerable attention due to their ability to generate coherent and context-aware text. Two of the widely known LLMs, GPT-3.5 and GPT-4 are the third and fourth iteration, respectively, in the GPT series of US-based AI research laboratory OpenAI. As a deep learning-based language model, LLMs generate human-like text using an autoregressive approach. In simpler terms, it's a system that creates persuasive, semantically coherent words, code, or data sequences based on a given input, known as the prompt.

[7] https://openai.com/blog/introducing-chatgpt-and-whisper-apis

GPTs, with their traits of general-purpose technologies, have not only opened up new possibilities for advanced natural language tasks but also raised concerns regarding their broader implications, including the economy, social dynamics, and policy development. In terms of the economy, a recent study reported that the rise of LLMs could affect a substantial portion of the U.S. workforce, with around 80% of workers seeing at least 10% of their tasks influenced by these technologies and approximately 19% experiencing impacts on at least 50% of their tasks [4]. In politics, governments are grappling with the rapid advancement of AI and call for regulation, as large language models enable AI to create realistic art, write essays, and generate code within seconds. In education, institutions have started taking measures to address the potential misuse due to concerns about students using it for assignments and the potential for unreliable or inappropriate content [5]. On the other hand, LLMs have a positive potential in creating educational content, enhancing student engagement, and personalizing learning experiences.

The works evaluating the performance of GPT and other LLM models indicate outperforming results in some domains compared to state-of-the-art NLP methods. For instance, GPT-3.5 has shown remarkable performance in text classification [6][7] and has proven to be more precise in MT evaluation compared to statistical metrics [8].

Among many capabilities, GPTs can be employed for the following linguistic tasks as they generate promising results:

- **Machine translation:** GPTs are capable of producing fluent and competitive machine translation outputs that perform closer to state-of-the-art neural machine translation models (e.g., Google) on high-resource European Languages [9][10][11]. However, there is the risk of generating hallucinations for non-English-centric MT tasks [12].
- **Text summarization:** LLMs can generate concise summaries of long pieces of text by understanding the main points and condensing the information into a shorter version [7] [13] [14].
- **Text generation:** LLMs can generate coherent, context-aware, and human-like text in various formats, such as stories, articles, or conversations for any given purpose. GPTs can adapt their text generation to specific writing styles, formats, or tones, allowing for tailored content production based on user preferences [15].
- **Dependency parsing:** LLMs can analyze the grammatical structure of a sentence and determine the relationships between words and phrases.
- **Textual similarity analysis:** GPTs can understand complex semantic patterns and relationships between words and phrases [16].

The list may be extended further for other specific linguistic tasks, such as named entity recognition, part-of-speech tagging, sentiment analysis, etc. The textual analysis and context understanding capabilities of GPTs make them particularly relevant for our research, where we are focused on their performance for analysing the interpreting output. In this context, this study has the following goals:

a. to explore and evaluate the potential of an LLM (GPT-3.5 / GPT-4) in assessing interpreting performance through zero-shot prompting based on human evaluation criteria while demonstrating the accuracy, comprehensiveness, and constructiveness of the feedback generated.
b. to launch a system that will provide an accessible tool for trainees that could be used as supplementary feedback in deliberate practice.

1.2 Automatic Evaluation and Scoring in Interpreter Assessment

Testing and assessing spoken-language interpreting has been gaining attention over the past decade, particularly in the fields of interpreter education, professional certification, and interpreting research. Han [17] outlines the assessment procedures for interpreting in three major aspects: (1) specificities of interpreting assessment (modes and directionality), (2) assessment design (difficulty, variety, number, and length of tasks and evaluation criteria), and (3) scoring and rater training. For scoring, different scoring methods are drafted by scholars, like the error-analytic method [18][19], questionnaire-based scoring [20], rubric scoring [21], and comparative judgment [22]. Based on these established views on assessment and scoring, automatic evaluation of spoken-language interpreting focuses on predicting human-assigned scores for interpreting performance. Automatic assessment and scoring techniques embark on a method that aims to minimize human involvement in analysing, assessing, and quantifying interpreting quality. It basically employs specifically designed algorithms to automatically calculate metrics and indices related to interpretation performance. While this approach holds promise in transforming the evaluation of interpreting quality, it remains in its nascent stage and necessitates further comprehensive research [23]. Research-wise, it has gained traction as an important area of research, with numerous studies exploring various methods for predicting interpreting quality. In general, there are three strands of automatic evaluation research [17] (a) researchers evaluate speech fluency using objectively measured temporal variables [24] [25]; (b) analyse linguistic features using corpus linguistic tools [26] [27]; (c) use algorithmic quality evaluation metrics that are already deployed for MT [23] [28].

In the first strand of research, researchers generally focus on analysing the temporal variables relating to fluency [29], like speech rate, phonation time ratio, and mean length of the run, which was proven to correlate with human rater's judgment [24] [25]. On the other hand, the second strand, led by interpreting studies, computer science, and natural language processing researchers, utilizes large corpora and sophisticated machine-learning techniques to analyse diverse surface features and predict human assessment results. For instance, Le et al. [30] built an interpreting corpus of 6,700 utterances and employed word confidence estimation systems integrating automatic speech recognition and machine translation features. The results of the experiments indicate that MT features have the most significant impact on the quality assessment of speech translation, while ASR features provide valuable supplementary

information. Meanwhile, Stewart et al. [31] enhanced the Quest++ quality estimation model with four interpreting-specific features to assess interpreting quality. Li et al., [32] propose a neural-based automatic scoring model for Chinese-English interpretation using a multi-indicator assessment. Three improved attention-based BiLSTM neural models are developed to learn the text of transcribed responses in terms of keywords, content, and grammar. BERT pre-training technique is used for keyword and content vectorization, while grammar is initialized randomly. Additionally, fluency is assessed through speech tempo analysis. An integrated score is derived by combining the four metrics using a random forest regression approach. Experimental results demonstrate that the proposed scoring method is effective and performs comparably to human assessment (manual scoring).

Using MT evaluation metrics for automatic assessment of interpreting constitutes the third strand of research where researchers adapt the translation-based MT evaluation metrics into spoken-language interpreting. For instance, Han and Lu [28] investigated the relationship between machine translation metrics, such as BLEU, TER, BERT, NIST and METEOR, and human scores for interpreting assessment, finding moderate-to-strong correlations that suggest the potential for automation. The study was conducted for English-to-Chinese interpreting tasks within language learning classrooms. The results indicated strong correlations, particularly for METEOR, with sentence-level evaluation showing closer correlations than text-level assessments [28]. Similarly, Lu and Han [23] explored the relationship between automated metric scores and human-assigned scores in assessing English-Chinese bidirectional interpretations performed by 56 students. Five machine translation metrics were used alongside various rater types and scoring methods, such as analytic and holistic rubric scoring and ranking. The study observed a relatively strong correlation between automated metric scores and human scores when BLEU, NIST, and METEOR are in question. These studies demonstrate the possibilities of using machine translation metrics and natural language processing techniques for automatic interpreting assessment.

Indeed, an automatic assessment of an interpreted text using the transcripts only would pose a challenge and be bound to some limitations. Being on a textual level, it will not be able to consider non-verbal cues, fluency, pace, and prosodic features (intonation, stress, rhythm) of the delivery, which are essential components of interpreting assessment. Other non-textual assessment units can be voice modulation, the skill of maintaining calm under pressure, memory retention, etc. In general, a comparative assessment based on source and target text would only provide an overall quality evaluation. GPT models stand out with their ability to handle many NLP tasks which can generate practical output with zero-shot prompting. By making the most of the generative capabilities of GPT-3.5 and 4, we attempt to investigate whether integrating LLMs into the assessment process could provide more accurate and nuanced evaluations, as they can capture linguistic features and semantic understanding beyond traditional metrics. The prompting and evaluation criteria that we embark upon in GPT-3.5 and GPT-4 are outlined in section 2.3.

2 Architecture and Workflow

In this section, we will detail the design principles and objectives behind InterpreTutor, including the rationale for its focus on textual feedback and the considerations made to ensure its usability and effectiveness in interpreting assessment scenarios.

2.1 User Interface

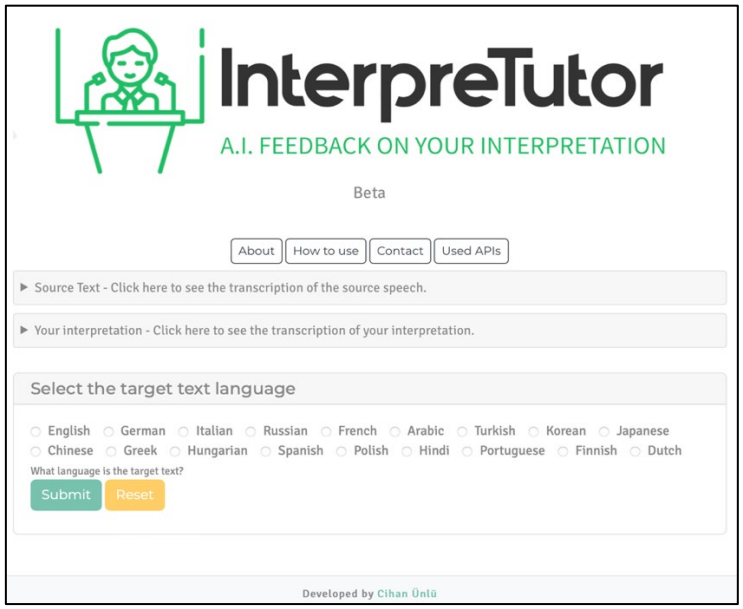


Fig. 1. A screenshot from the main interface of InterpreTutor

The user interface (UI) of the tool provides users with a step-by-step process to facilitate the evaluation of the given inputs. The user begins by uploading the source speech, which is then transcribed using OpenAI Whisper's Automatic Speech Recognition (ASR) system. Upon completion of the transcription, the user is prompted to select the source language. Subsequently, the user is required to upload their interpretation, which is again transcribed into target text using Whisper ASR API. The user must then choose the target language (as shown in Fig. 1)[8]. To generate a reference text in the target language, the user is given the option to either upload the source speech once more, which is then translated automatically via the Whisper Translate API[9], or manually input a pre-determined reference text, also known as the golden standard. With all inputs in place, the source and target texts, as well as the reference

[8] Language selections are done after the transcription/translation processes.

[9] Whisper's speech translation engine currently only works for translations into English. For other directions, the user is asked to enter the reference text manually.

text, are combined to form the general prompt for the model. The evaluation process commences, and upon completion, the user is presented with a concise evaluation summary, offering insights into their interpreting performance based on the predetermined evaluation criteria (see 2.3.).

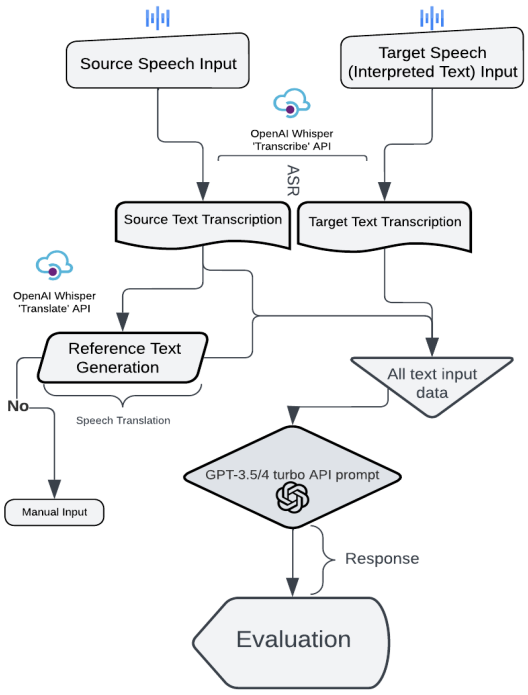


Fig. 2. The workflow applied in InterpreTutor.

2.2 Models Used

In the development of InterpreTutor, we employed two state-of-the-art AI models to facilitate the conversion of speech into text and the generation of assessments, respectively. Both models have been developed by OpenAI and have demonstrated impressive performance in their respective domains.

OpenAI Whisper.

Whisper [33] is an Automatic Speech Recognition (ASR) system that has been trained on a large dataset of multilingual and multitask supervised data collected from the web. Its primary function within InterpreTutor is to convert both the source speech (source text) and the user's spoken input (interpretation) into text to be processed by the LLM. Whisper is an example of the application of transformer models in speech processing is a versatile model designed for speech recognition in challenging environments, such as noisy or low-resource settings. It is capable of multilingual

speech recognition, speech translation, and language identification. In InterpreTutor, we use Whisper API for both transcription and translation. The transcription function is utilized for both source text capturing and target text generation (the user's interpretation), whereas the speech translation (ST) function of Whisper is used for the automatic generation of the reference text. The application relies on the accuracy of Whisper's speech-to-text conversion, which gives promising results with a low word-error rate (WER), especially for the languages English, Italian, German, and Spanish [34]. One thing that Whisper can do is annotate the filled pauses. It is possible to use a prompt to improve the quality of the transcripts generated by the Whisper API. The model may leave out or include common filler words in the audio. We created an example prompt to keep the filler words (e.g., 'uhm(s)' 'hmm' etc.) in the transcript, thereby evaluation of the disfluency is facilitated when the LLM analyses the occurrences of filled pauses.

OpenAI GPT-3.5 & GPT-4.
OpenAI's GPT is a large-scale language model known for its exceptional natural language processing capabilities. In InterpreTutor, we specifically used the model gpt-3.5-turbo and GPT-4[10] (latest iteration) from the GPT family to generate responses and feedback on the user's interpreting performance. Based on the textual input received from the OpenAI Whisper API, GPT-3.5 and GPT-4 are responsible for analysing the transcriptions, comparing the source and target texts, and providing feedback on the quality of the interpretation based on the criteria outlined in section 2.3.

2.3 Evaluation Criteria and Prompt Generation

This section will explain the mechanisms by which InterpreTutor prompts the GPT-3.5 and GPT-4 models to generate feedback, as well as the evaluation criteria used to assess the interpreters' performance. It will also cover the development of these criteria and the considerations made in order to ensure their relevance and fairness.

Evaluation Criteria.
In evaluation criteria, various quality criteria have emerged in the literature with different viewpoints. In their research, Schjoldager [20] identifies four primary factors to evaluate student interpretation performance, encompassing aspects such as comprehensibility, linguistic accuracy, coherence, and fidelity to the source. On the other hand, Riccardi [35] delves into a more granular assessment approach by proposing 17 specific criteria, which cover areas like phonetic and prosodic accuracy, speech production, non-verbal communication, and lexical variations for both general and specialized vocabulary. The key criteria identified by subsequent research include content, delivery, and language quality as the three primary components of interpreting quality [21] [36]. Drawing on Lee's [21] framework for assessing interpreting perfor-

[10] Fort the technical report, see https://arxiv.org/abs/2303.08774.

mance in the classroom, we adopted a similar three-dimensional approach to interpreting quality, incorporating additional sub-criteria. Our evaluation framework includes 1) accuracy, fidelity and completeness (content); 2) textual integrity, focusing on cohesion and coherence; 3) appropriate usage of terminology; and 4) the presence or absence of disfluency markers (only repetitions and disfluency markers).

Prompting.
The GPT-3.5 and GPT-4 API requests two prompts to generate the answer: system and the user. The system prompt sets the context and the behaviour of the assistant by providing specific information or defining the character of the AI. In our case it is: "You are a conference interpreter trainer who makes detailed and objective evaluations on spoken-language interpreting". The user prompt on the other hand is the main instruction that the GPT handles. Our prompt is designed to encompass three essential criteria as specified above: accuracy, fidelity, and completeness; cohesion and coherence; terminology; disfluency markers. Specifically, we create four responsibilities for the language model and draft the prompt accordingly. 1) Comparing the interpreted content with the source content to check for correctness and completeness. 2) Analysing the logical organization and connectedness of the interpreted content, including transitions. 3) Evaluating the terminological accuracy and consistency. 4) Detecting disfluency markers: repetitions and filled pauses. In the prompt, we elaborated on the aspects and the model is guided accordingly. Consequently, the prompt asks the following questions for the criteria and includes the instructions accordingly:

Accuracy & Fidelity & Completeness:
- **a)** Does the target text accurately convey the meaning of the source text in general?
- **b)** Are there any deviations, such as omissions, additions and unjustifiable changes or misinterpretations of the meaning and intention of the speaker of the source text? Are any important details omitted? Are there grammatical mistakes? Specify in detail.

Cohesion and coherence:
- **a)** Is the target text logically connected and coherent in general?
- **b)** Does the text make sense as a whole in general and make a narrative in the language of the target text despite the minor errors? Specify in detail and give examples if there are any coherence and fragmentation issues in this regard.

Terminology:
- Was correct and consistent terminology used in general? If there are any misusages, point out all major errors, and elaborate.

Disfluency markers:
- Detect the occurrences of repetitions (like saying the same word repeatedly and unnecessarily) and filled pauses.

When all inputs are received for the creation of the prompt, the prompt then guides the model to get the assessment based on these criteria. The model is given the task without any prior examples or training data. Therefore, the model is asked to com-

plete the task using only its pre-existing knowledge, without any fine-tuning or contextual clues[11].

2.4 Output Example

Based on the evaluation criteria and the tasks given in the prompt, the tool generates the general evaluation. An exemplary short and concise output using GPT-3.5 (without scorings) is as follows:

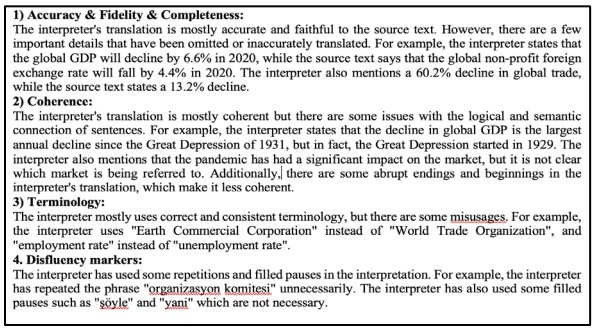

1) Accuracy & Fidelity & Completeness:
The interpreter's translation is mostly accurate and faithful to the source text. However, there are a few important details that have been omitted or inaccurately translated. For example, the interpreter states that the global GDP will decline by 6.6% in 2020, while the source text says that the global non-profit foreign exchange rate will fall by 4.4% in 2020. The interpreter also mentions a 60.2% decline in global trade, while the source text states a 13.2% decline.
2) Coherence:
The interpreter's translation is mostly coherent but there are some issues with the logical and semantic connection of sentences. For example, the interpreter states that the decline in global GDP is the largest annual decline since the Great Depression of 1931, but in fact, the Great Depression started in 1929. The interpreter also mentions that the pandemic has had a significant impact on the market, but it is not clear which market is being referred to. Additionally, there are some abrupt endings and beginnings in the interpreter's translation, which make it less coherent.
3) Terminology:
The interpreter mostly uses correct and consistent terminology, but there are some misusages. For example, the interpreter uses "Earth Commercial Corporation" instead of "World Trade Organization", and "employment rate" instead of "unemployment rate".
4. Disfluency markers:
The interpreter has used some repetitions and filled pauses in the interpretation. For example, the interpreter has repeated the phrase "organizasyon komitesi" unnecessarily. The interpreter has also used some filled pauses such as "söyle" and "yani" which are not necessary.

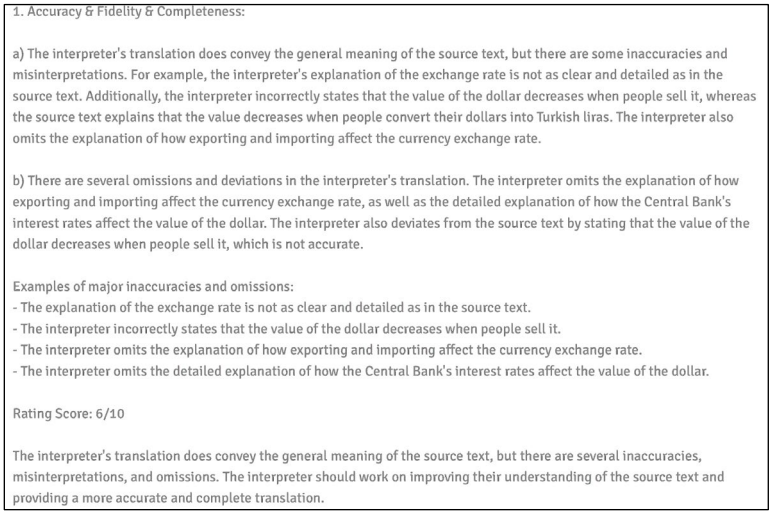
Fig. 3. An example output (Whisper+GPT-3.5) in InterpreTutor

Due to a larger model size than GPT-3, GPT-4 demonstrates superior comprehension and reasoning capabilities[12], which was also observed while experimenting with them in InterpreTutor. Its expanded token limit of 32,000[13], equating to approximately 25,000 words, notably enhances report generation in our tool, providing more extensive and detailed outputs. While utilizing GPT-3.5, text inputs through ASR and reference text make the prompt too long, giving little token size for the answer generation. However, the max token size attributed in the model reinforces the advantage of employing GPT-4, which is set to 4000. Figure 4 shows another example output but only for the first criteria: Accuracy, Fidelity, and Completeness. In addition to model comparison based on size and capabilities, this study also aims to compare the relevancy of the scorings of each model to that of human judgement. However, it's important to note that the focus here is primarily on a score-based comparison. Despite this limitation, GPT-4 generates more detailed and accurate feedback points, which underlines its potential superiority over GPT-3 in the context of producing logical and accurate output. Yet, both models carry the risk of hallucinations and inaccurate information. They can also produce certain biases inherent in the training data, which can possibly affect the feedback provided. In order to mitigate this, a low temperature (0.1) was set to make the model more deterministic and less random.

[11] Apart from the primary system prompt which includes “You are a conference interpreter trainer who gives detailed feedbacks and evaluations on spoken-language interpreting.”

[12] https://openai.com/research/gpt-4

[13] This marks a substantial increase from GPT-3.5's 4,000-token limit, equivalent to around 3,125 words.


1. Accuracy & Fidelity & Completeness:

a) The interpreter's translation does convey the general meaning of the source text, but there are some inaccuracies and misinterpretations. For example, the interpreter's explanation of the exchange rate is not as clear and detailed as in the source text. Additionally, the interpreter incorrectly states that the value of the dollar decreases when people sell it, whereas the source text explains that the value decreases when people convert their dollars into Turkish liras. The interpreter also omits the explanation of how exporting and importing affect the currency exchange rate.

b) There are several omissions and deviations in the interpreter's translation. The interpreter omits the explanation of how exporting and importing affect the currency exchange rate, as well as the detailed explanation of how the Central Bank's interest rates affect the value of the dollar. The interpreter also deviates from the source text by stating that the value of the dollar decreases when people sell it, which is not accurate.

Examples of major inaccuracies and omissions:
- The explanation of the exchange rate is not as clear and detailed as in the source text.
- The interpreter incorrectly states that the value of the dollar decreases when people sell it.
- The interpreter omits the explanation of how exporting and importing affect the currency exchange rate.
- The interpreter omits the detailed explanation of how the Central Bank's interest rates affect the value of the dollar.

Rating Score: 6/10

The interpreter's translation does convey the general meaning of the source text, but there are several inaccuracies, misinterpretations, and omissions. The interpreter should work on improving their understanding of the source text and providing a more accurate and complete translation.

Fig. 4. An example output (Whisper+GPT-4) for the content accuracy and completeness generated in InterpreTutor

3 Pilot test

3.1 Methodology

Our methodology for the pilot test involves a speech to be interpreted by two interpreters and later to be evaluated by InterpreTutor and human evaluators. A number-dense speech (duration: 03:01) was chosen to be interpreted from English into Turkish in consecutive mode by two trainee interpreters. The topic of the speech is FIFA World Cup and it consists of 17 named entities and 15 numerical items which makes it challenging and prone to omissions and/or substitutions in interpreting. The source speech and the record of the interpretation (in mp3/wav/mp4 formats) were used as inputs for our LLM-powered tool InterpreTutor utilizing both GPT-3.5 and GPT-4 separately. Apart from the specific instructions in the prompting, the LLM models were instructed to give a score out of ten for each criterion. For the human judgement, both source and target (interpretation) data were transcribed through Whisper ASR (word-error-rate: 0% in the source text and between 1% and 3% in the target texts). Three human evaluators (interpreter trainers) from two universities were provided with the transcription of the interpreter's performance, the source text and the reference text. The same criteria used in the prompting were also used as a guideline and provided for human annotators beforehand. The evaluators were asked to evaluate the performance based on the given criteria and asked to give scores out of 10 for each criterion specified in 2.3.

Given the limited number of evaluators and the small sample size, the application of statistical models such as the Intraclass Correlation Coefficient (ICC), Pear-

son's correlation coefficient, and Fleiss Kappa would be inappropriate and result in a lack of generalizability. Given the limited number of evaluators and evaluation criteria, we think that, at the first stage, a qualitative analysis of the results based on scoring may be a more appropriate approach to gain insights into the performance of both human and LLM-based evaluation methods. Future research with a larger sample size and a broader range of evaluation criteria could benefit from using these statistical models to obtain more robust and comprehensive results. Though the pilot study employed a score-based design for the evaluation as well as a comparison with human judgement, the results section touches upon a rough analysis of the remarks. Perhaps, a detailed error identification analysis of the models can yield a better understanding. Further research can compare the detailed critics and points in the generated feedback and those of human evaluators by thematic classifications of the remarks.

3.2 Results

An overall overview of the remarks provided in the two outputs shows that GPT-4 produces more accurate feedback with more examples for both evaluations compared to GPT-3.5-based feedback (see the appendix). Scoring-wise, GPT-4 surpasses GPT-3.5 in its alignment with human judgment. Table 1 showcases the scores assigned by human evaluators and those produced by the two models from InterpreTutor.

Table 1. The scorings of each evaluator

	Criteria	Evaluator 1	Evaluator 2	Evaluator 3	InterpreTutor GPT-3.5	InterpreTutor GPT-4
Interpreter I	Accuracy, Fidelity & Completeness	7	6	5	6	6
	Cohesion and Coherence	8	7	6	2	6
	Terminology	6	7	6	5	7
	Disfluency Markers	6	6	7	5	6
Interpreter II	Accuracy, Fidelity & Completeness	4	5	2	2	6
	Cohesion and Coherence	5	3	3	2	6
	Terminology	3	5	4	4	7
	Disfluency Markers	5	8	8	6	8

GPT-3.5 Results

a) **Accuracy, Fidelity & Completeness:** Both the GPT 3.5 model's score and the GPT-4's score (6) are in the same range as the human evaluator scores (5, 6, and 7) for the performance of Interpreter I. This indicates that the LLM evaluation is closely aligned with the human evaluators' assessment of accuracy. However, a rough human evaluation of the outputs indicates the output of GPT-3.5 does not fully catalog all inaccuracies and omissions but gives a couple of accurate exam-

ples.[14] The less number of examples of inaccuracies make the report less persuasive and missing. Repetitive evaluation sessions with GPT-3.5 keeps the score in the same range, but the examples differ each time, though they are correct takeaways. On the other hand, GPT-4-based output constituted more detailed feedback with correct points, making the evaluation fairer given such an average score. For the second interpreter, It's noticeable that the GPT-4 version consistently rated higher than the GPT-3.5 version and gave higher scores than human evaluators. This could suggest an overestimation by the LLM or a difference in the standards used for evaluation between the LLM and human evaluators.

b) **Cohesion and coherence:** For the Interpreter I, the GPT-3.5 score (2) is significantly lower than the human evaluator's scores while the GPT-4 score matches well with the human judgment. However, for the interpreter II, neither the human evaluators nor the LLMs seem to have a common ground. The GPT-3.5 score remains quite lower than the human judgment. On the contrary, the GPT-4 score keeps relatively high for this criterion (6).

c) **Terminology:** For the interpreter I, the GPT-3.5 score (5) is slightly lower than the human evaluator scores (6 and 7), while GPT-4 has given a closer score (7). This result indicates that, for the interpreter I, the GPT-4 evaluation is somewhat aligned with the human evaluators' assessment of terminology consistency, but it may not be as accurate as in the accuracy criterion. For the second interpreter, however, GPT-4 appears to evaluate the use of terminology more favourably than either the human evaluators or GPT-3.5 and GPT-3.5 score is more aligned with those of human evaluators.

d) **Disfluency Markers (filled pauses and repetitions only):** For the first interpreter performance, the GPT-3.5 score (5) falls within the range of the human evaluator scores (6 and 7) while the GPT-4-based evaluation matches the scores of two human evaluators (6). For the second performance, human evaluators have some divergence in their evaluations. The GPT-4 evaluation is reasonably more aligned with the two human evaluators' assessment of disfluency markers (8). The written feedback-wise, it produced more examples of filled pause annotations, especially for the Interpreter II, which makes the score look more reliable. However, the GPT-3.5-based evaluation indicated some hallucinations. This conclusion is based on the output (Interpreter I) that indicates that "there were several instances of filled pauses". While remarks on filled pauses held true, there was only one case of repetition in the interpretation.

4 Conclusion and Future Developments

In this preliminary work, we aimed to investigate the capabilities of a pre-trained LLM, the GPT-4 and GPT-3.5 in the assessment and evaluation of interpreting performance on a textual level based on human criteria. In doing so, we designed a prompt based on the assessment criteria views on the relevant literature and proposed

[14] The examples can be seen in the Appendix 1.

an all-in-one LLM-powered application "InterpreTutor" which streamlines the process by using speech-to-text technology (OpenAI Whisper) and speech translation (OpenAI Whisper Translate). Whisper ASR, thanks to its state-of-the-art architecture and extensive training data, is successful at providing robust and clean ASR outputs[15] which makes it easy to process by the GPT API. In our small-scale pilot test, we designed a prompt and made two interpreters' performances evaluated by the LLM in InterpreTutor and three human evaluators.[16] We observed that for the first interpreting performance, the GPT-4 model performed closely to human evaluators in terms of scoring compared to GPT-3.5. However, for the second interpreter, GPT-4 tended to assign higher scores than GPT-3.5 and sometimes even higher than human evaluators. GPT-3.5 generally aligned more closely with the lower range of human evaluator scores. Though the first test has shown a similarity in scoring between human evaluators and LLMs (particularly GPT-4), the second test provided an evident discrepancy not only between different models but also between models and human judgement, which gives no room for a definite conclusion. Despite the good alignment between human judgement and the InterpreTutor, the usability of such a tool for scoring is still questionable because of the inconsistency and the reliability of the scores. Nevertheless, InterpreTutor with GPT-4 performed well at analysing and commenting more persuasively and elaboratively on the audio inputs transcribed by the Whisper. Detailed feedback generation and accurate error detection of the models hold promise. Further optimization in prompting and fine-tuning of the model can generate useful output. LLMs, at this stage, have the potential to offer in-depth, partly constructive evaluation reports on the content accuracy with their sophisticated natural language understanding capabilities. Applications like InterpreTutor can be used by trainees in their deliberate-practice sessions when they want to independently assess themselves in some aspects.

To advance the understanding and application of LLMs in the assessment of translation and interpretation, the following areas can be recommended for future research. First, more empirical research with fine-tuned LLM models using larger datasets is recommended, as this could potentially lead to improved performance and enhanced evaluation capabilities. Second, this study focused on its feasibility in interpreter assessment and evaluation, while a similar study can be adapted for written translation assessment for scoring, grading or other pedagogical purposes. Even so, with the combination of LLMs' good reasoning capacities and real-time speech-to-text/text-to-speech technologies, similar applications can be developed for the automatic evaluation of standardized speaking exams or similar simulations. This could be done by instructing the LLM to conduct a semantic, syntactic analysis and analyse the argumentation in the speech. However, a lot of development and testing would surely be

[15] The study does not employ detailed benchmarking for ASR outputs except Word-error-rate calculation (which was shown to be very low). It should be acknowledged that ASR errors can also affect the output. Though most of the points made are correct, the model may occasionally misinterpret the text as a result of errors introduced by the ASR. Any lexical error in the ASR results can make the tool detect it as a major error.

[16] This result must be interpreted in a nuanced manner, as it is subject to certain limitations.

required. Third, the current study focused on the Turkish-English language pair, and it would be beneficial to investigate the performance of LLMs in assessing translations and interpretations in other, particularly high-resource, languages. Fourth, the present study utilized a low-temperature setting for the GPT-3.5 due to the 4,000-token limitation imposed by the API at the time of writing, whereas GPT-4 has the capability of using more than 4k tokens which allows us to make it generate logical and longer evaluation reports. Future research should experiment with various temperature levels, different base models, few-shot prompting methods and larger token capacities to determine the optimal configuration for achieving the best results. Fifth, our study did not reveal the superiority of GPT-4 over GPT-3.5. Still, we observe that the GPT-4-based system for the workflow performs the best results in automatic assessment in the context of qualitative feedback generation. That's why, rather than just comparing numerical scores, analyzing the specific content of the feedback could offer a more nuanced understanding of how these two types of evaluators differ in their assessments and how InterpreTutor can use the instructional capabilities of the pre-trained LLMs.

Certain restrictions and disadvantages are present since GPT-4 is slower and more costly. On top of that, ethical considerations on trustworthy usage, privacy, etc. do apply for InterpreTutor. Task-specific prompting keeps the model away from hallucinating and hazardous content. However, it is also true that misleading information or misevaluations can rarely happen in the feedback output, particularly when using low-resource languages for source and target languages. This is why using a gold standard as a reference helps the model to report well-reasoned points in the evaluation. Language variations pose a challenge to the robustness of the transcribed content. Experimenting with different language combinations and even with more and more large-scale language models would undoubtedly unveil more about the practicality of such secondary training-purposed tools like InterpreTutor.

References

1. Wang, L., et al.: Document-Level Machine Translation with Large Language Models. arXiv preprint arXiv:2304.02210 (2023).
2. Vaswani, A., et al.: Attention is all you need. Advances in Neural Information Processing Systems 30 (2017).
3. Brown, T., et al.: Language models are few-shot learners. Advances in Neural Information Processing Systems 33, pp. 1877–1901 (2020).
4. Eloundou, T., Manning, S., Mishkin, P., & Rock, D.: GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models. arXiv preprint arXiv:2303.10130 (2023).
5. Elsen-Rooney, M.: NYC education department blocks ChatGPT on school devices, and networks. Chalkbeat New York January 3, 2023. https://ny.chalkbeat.org/2023/1/3/23537987/nyc-schools-ban-chatgpt-writing-artificial-intelligence (2023).

6. Kuzman, T., Mozetic, I., & Ljubešic, N.: ChatGPT: Beginning of an End of Manual Linguistic Data Annotation? Use Case of Automatic Genre Identification. ArXiv abs/2303.03953 (2023).
7. Luo, Z., Xie, Q., & Ananiadou, S.: ChatGPT as a Factual Inconsistency Evaluator for Abstractive Text Summarization. arXiv preprint arXiv:2303.15621 (2023).
8. Kocmi, T., & Federmann, C.: Large language models are state-of-the-art evaluators of translation quality. arXiv preprint arXiv:2302.14520 (2023).
9. Vilar, D., et al.: Prompting PaLM for Translation: Assessing Strategies and Performance. arXiv preprint arXiv:2211.09102 (2022).
10. Hendy, A., et al.: How good are gpt models at machine translation? A comprehensive evaluation. arXiv preprint arXiv:2302.09210 (2023).
11. Jiao, W., et al.: Is ChatGPT a good translator? A preliminary study. arXiv preprint arXiv:2301.08745 (2023).
12. Peng, K., et al.: Towards Making the Most of ChatGPT for Machine Translation. arXiv preprint arXiv:2303.13780 (2023).
13. Goyal, T., Li, J.J., & Durrett, G.: News summarization and evaluation in the era of gpt-3. arXiv preprint arXiv:2209.12356 (2022).
14. Wang, J., et al.: Is chatgpt a good NLG evaluator? A preliminary study. arXiv preprint arXiv:2303.04048 (2023).
15. Hua, M., & Raley, R.: Playing With Unicorns: AI Dungeon and Citizen NLP. DHQ: Digital Humanities Quarterly 14(4) (2020).
16. Gao, C.A., et al.: Comparing Scientific Abstracts Generated by CHATGPT to Original Abstracts Using an Artificial Intelligence Output Detector, Plagiarism Detector, and Blinded Human Reviewers. bioRxiv 2022.12.23.521610. (2022).
17. Han, C.: Interpreting testing and assessment: A state-of-the-art review. Language Testing 39(1), 30–55 (2021a).
18. Barik, H.C.: A Description of Various Types of Omissions, Additions and Errors of Translation Encountered in Simultaneous Interpretation. Meta: Journal Des Traducteurs 16(4), 199–210 (1971).
19. Setton, R., & Motta, M.: Syntacrobatics: Quality and reformulation in simultaneous-with-text. Interpreting 9(2),199–230 (2007).
20. Schjoldager, A.: Assessment of Simultaneous Interpreting. Teaching Translation and Interpreting 3, p. 187– (1996).
21. Lee, J.: Rating Scales for Interpreting Performance Assessment. The Interpreter and Translator Trainer 2(2), 165–184 (2008).
22. Han, C.: Analytic rubric scoring versus comparative judgment: A comparison of two approaches to assessing spoken-language interpreting. Meta 66(2), 337–361 (2021b).
23. Lu, X., & Han, C.: Automatic assessment of spoken-language interpreting based on machine-translation evaluation metrics: A multi-scenario exploratory study. Interpreting 25(1), 109–143 (2022).
24. Yu, W., & Van Heuven, V.J.: Predicting judged fluency of consecutive interpreting from acoustic measures: Potential for automatic assessment and pedagogic implications. Interpreting 19(1), 47–68 (2017).
25. Han, C., et al.: Modeling the relationship between utterance fluency and raters' perceived fluency of consecutive interpreting. Interpreting 22(2), 211–237 (2020).
26. Ouyang, L., Lv, Q., & Liang, J.: Coh-Metrix model-based automatic assessment of interpreting quality. In: Chen, J., Han, C. (eds) Testing and assessment of interpreting: Recent developments in China, pp. 179–200. Springer, Singapore (2021).

27. Stewart, C., et al.: Automatic estimation of simultaneous interpreter performance. arXiv preprint arXiv:1805.04016 (2018).
28. Han, C., & Lu, X.: Can automated machine translation evaluation metrics be used to assess students' interpretation in the language learning classroom? Computer Assisted Language Learning, 1–24. (2021).
29. Wu, Z.: Chasing the unicorn? The feasibility of automatic assessment of interpreting fluency. In: Chen, J., Han, C. (eds) Testing and assessment of interpreting: Recent developments in China, pp. 143–158. Springer, Singapore (2021).
30. Le, N.-T., Lecouteux, B., & Besacier, L.: Automatic quality assessment for speech translation using joint ASR and MT features. arXiv preprint arXiv:1609.06049 (2016).
31. Stewart, C., et al.: Automatic estimation of simultaneous interpreter performance. arXiv preprint arXiv:1805.04016 (2018).
32. Li, X., et al.: Neural-based automatic scoring model for Chinese-English interpretation with a multi-indicator assessment. Connection Science 34(1),1638–1653 (2022).
33. Radford, A., et al.: Robust speech recognition via large-scale weak supervision. arXiv preprint arXiv:2212.04356 (2022).
34. Heikinheimo, H.: Analyzing Open AI's Whisper ASR Accuracy: Word Error Rates Across Languages and Model Sizes. Speechly, 4 Apr. 2023, www.speechly.com/blog/analyzing-open-ais-whisper-asr-models-word-error-rates-across-languages, last accessed 2023/04/05
35. Riccardi, A.: Evaluation in interpretation. In: Hung, E., (eds.) Teaching translation and interpreting 4, pp. 115–126. John Benjamins Publishing Company, Netherlands (2002).
36. Dawrant, A., & Setton, R.: Conference interpreting: A trainer's guide. John Benjamins Publishing Company, Netherlands (2016).

APPENDIX I

GPT-4 Evaluation (Interpreter I)

Accuracy, Fidelity & Completeness:

a) *In general, the interpreter's translation conveys the meaning of the source text, but there are some misinterpretations and inaccuracies. Another example is "Avrupa ve Güney Afrika'da... Avrupa Birliği... Avrupa'da ve Güney Amerika'da en çok kazanılmıştır." which is a confusing and inaccurate representation of the source text. The source text states that only teams from Europe and South America have won the tournament, but the interpreter's translation mentions Europe and South Africa, and also brings up the European Union, which is not relevant to the source text.*

b) *There are several omissions and deviations in the interpreter's translation. Some examples include:*

- *The interpreter did not mention the fact that the FIFA World Cup is the most widely viewed sporting event.*
- *- The interpreter omitted the specific number of titles won by Argentina, France, and Uruguay, as well as the fact that England and Spain have won the tournament once each.*
- *- The interpreter did not mention the specific years and achievements of the African Federation teams (Cameroon, Ghana, and Senegal) reaching the quarterfinals.*
- *- The interpreter did not mention the Women's World Cup following the same format as the Men's Tournament, and incorrectly stated that England has won the Women's World Cup twice, while the source text states that Germany has won twice.*
- *The interpreter did not mention the FIFA Under-20s and Under-17s World Cups being held every two years and allowing the best youngster players and national teams to compete against each other by age group.*

Cohesion and coherence: *The interpreter's translation is generally coherent, but there are some issues with cohesion and fragmentation. For example, the interpreter mentioned "Güney Afrika'da..."*

instead of "Güney Amerika'da...". There are also some filled pauses like "..." which disrupt the flow of the interpretation.

Terminology: *In general, the interpreter used correct terminology, but there were some inconsistencies and errors. For example, the interpreter mentioned "Güney Afrika'da" instead of "Güney Amerika'da" and "Amerika Birliği" instead of "Avrupa Birliği." Additionally, the interpreter referred to the Women's World Cup as "Kadınların şampiyonası," which is not the exact term for the event.*

Disfluency Markers: *There are several instances of filled pauses in the interpreter's translation, such as "Avrupa ve Güney Afrika'da...", "Bu prestijli turnuva 3 milyar kişi tarafından izlendiği zaman...", and "Bir çok skandala imza atılsa da bu turnuvalar yine futbol sevenler tarafından...". These pauses disrupt the low of the translation and make it less coherent. Additionally, there are repetitions like "Amerika kıtasından katılan ülkeler olarak"*

"32 takım... 32 takım yarışmaktadır." "Avrupa ve Güney Afrika'da... Avrupa Birliği... Avrupa'da ve Güney Amerika'da en çok kazanılmıştır." These disfluency markers indicate that the interpreter had difficulties in maintaining a smooth flow during the interpretation.

GPT-4 Evaluation (Interpreter II)

Accuracy, Fidelity & Completeness:

a) *In general, the target text conveys the meaning of the source text, but there are some misinterpretations and inaccuracies. Some examples of major inaccuracies are:*

- *- The interpreter incorrectly states that the first tournament took place in 1910, while the source text says it was in 1930.*
- *- The interpreter mentions that Argentina has 3 championships, while the source text states that Argentina has 2 championships.*
- *- The interpreter incorrectly states that the USA's best result was in 2013, while the source text says it was in 1930.*
- *- The interpreter mentions Kemerun as an African country, which is not mentioned in the source text and is not a real country. The correct countries mentioned in the source text are Cameroon, Ghana, and Senegal.*

b) *There are some deviations and omissions in the interpreter's translation. Some examples are:*

- *- The interpreter omits the mention of Italy and Germany having four championships each.*
- *- The interpreter omits the mention of France and Uruguay having two championships each.*
- *- The interpreter omits the mention of England and Spain having one championship each.*
- *- The interpreter omits the mention of South Korea's fourth-place finish in 2002 being the best performance by an Asian nation.*
- *- The interpreter omits the mention of the Women's World Cup following the same format as the Men's Tournament.*
- *- The interpreter omits the mention of the FIFA Under-20s and Under-17s World Cups being held every two years.*

Overall Rating Score: 6/10

Cohesion and coherence: *Overall, the target text is logically connected and coherent, but there are some fragmentation issues and minor errors. For example, the interpreter mentioned the first World Cup took place in 1910, which is incorrect; it was in 1930. Also, the interpreter mentioned Argentina having three titles, while the correct number is two. Despite these errors, the target text still makes sense as a whole and maintains a narrative in Turkish.*

Score: 7/10

Terminology: *In general, the interpreter used correct and consistent terminology. However, there are some major errors:*

- *- The interpreter mentioned the first World Cup took place in 1910 instead of 1930.*
- *- The interpreter mentioned Argentina having three titles instead of two.*
- *- The interpreter mentioned the USA's best result was in 2013 instead of 1930.*
- *- The interpreter mentioned Kemerun instead of Cameroon.*
- *- The interpreter mentioned the best result for African countries was in 2006 instead of the correct information about Cameroon, Ghana, and Senegal reaching the quarterfinals.*

Score: 6/10

Disfluency Markers: *There are several occurrences of repetitions and filled pauses in the interpreter's performance. Examples include:*

- *- "Ve ilmiden fazla da türü var." (repetition of "ve")*
- *- "Sonrasında Almanya ve İtalya takip ediyor onu 4 şampiyonlukla." (repetition of "sonrasında")*

- *- "Sonrasında Kemerungan'a ve Senegal gibi Afrika ülkelerinde de en iyisi 2006 yılında gerçekleşmişti." (repetition of "sonrasında")*
- *- "Sonrasında Avustralya için de durum bu şekilde devam ediyor." (repetition of "sonrasında")*
- *- "Tabii ki erkek futbol turnuvası olan FIFA'daki bu erkek futbol turnuvasına çok fazla ilgi var." (repetition of "erkek futbol turnuvası")*

Score: 8/10