International Journal of Language, Translation and Intercultural Communication

Vol 11 (2026)

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


## Understanding AI interpreting in context

*Kayo Matsushita*

doi: 10.12681/ijltic.44192



**To cite this article:**

Matsushita, K. (2026). Understanding AI interpreting in context: A comprehension-based evaluation of human vs. machine-generated interpretations in a real-world setting. *International Journal of Language, Translation and Intercultural Communication*, *11*, 71–85. https://doi.org/10.12681/ijltic.44192

# Understanding AI interpreting in context: A comprehension-based evaluation of human vs. machine-generated interpretations in a real-world setting.

**Kayo Matsushita**
Rikkyo University
kayo.matsushita@rikkyo.ac.jp

**Abstract (150 words)**

*The rise of AI in the interpreting industry poses pressing questions about the sustainability of interpreting as a profession. While commercial platforms promise real-time multilingual communication at scale, their functional effectiveness in high-stakes professional contexts remains underexplored. This study presents a comprehension-based evaluation comparing human and AI interpreting of a climate-related press conference. Following Reithofer's (2013, 2014) methodology, 56 journalists were divided into two groups: one listening to professional human interpretation and the other to a cutting-edge AI service (KUDO AI Speech Translator). Results showed that the human group achieved higher comprehension scores (mean 4.5/10) than the AI group (mean 3.7/10), with the latter exhibiting a 17.9% "Don't Know" rate. Qualitative feedback highlighted that AI's lack of prosodic salience increased cognitive load, hindering deep information synthesis. These findings suggest that human intervention remains essential for ensuring semantic adequacy and effective information transfer in professional journalistic settings.*

***Keywords****: artificial intelligence (AI), machine interpreting, Speech-to-Speech (S2S) translation, comprehension test, press conference*

## 1 Introduction

The field of simultaneous interpreting is currently facing a transformative era, driven by the rapid integration of Large Language Models (LLMs) and Speech-to-Speech (S2S) translation technologies. As commercial AI platforms gain traction (Slator, 2024), the discourse has shifted toward the concept of human parity (Hassan et al., 2018)—the idea that machine-generated output can functionally replace human experts. However, much of this debate relies on automated linguistic metrics (e.g., BLEU scores) that measure surface-level correspondence rather than the communicative effect on the listener. For professional audiences such as journalists who attend press conferences for the purpose of news reporting, the value of interpretation lies in its ability to accurately reconstruct the speaker's intent, nuance, and logical structure under time pressure. If an interpretation is literally accurate but prosodically flat or structurally fragmented, the cognitive load on the listener increases, potentially compromising information transfer.

This study seeks to empirically compare the level of information transfer between human and AI interpreting. By conducting a controlled experiment with professional journalists as subjects, we measure the end product of interpretation, i.e., the depth of comprehension, following Reithofer (2013, 2014). The findings aim to contribute to the ongoing discussion on the occupational implications of AI and the necessity of human-centric evaluation frameworks in the age of machine translation and interpreting. In short, this research asks the question: How sustainable is professional interpreting in the face of rising AI parity claims? Can comprehension-based evaluation help differentiate human from AI performance? Through a controlled experiment utilizing authentic press conference materials and a cohort of active and former

journalists, this study aims to provide empirical evidence to inform the ongoing practical, pedagogical, and academic discourse around the questions within interpreting and translation studies.

## 2 Literature review

### 2.1 The rise of AI interpreting and industry shifts

The interpreting landscape is undergoing a paradigm shift due to the rapid advancement of AI and LLMs. According to Slator (2024), the industry is witnessing a convergence of technologies where traditional interpreting platforms are increasingly integrating AI-powered Speech-to-Speech (S2S) translation. This technological surge is driven by the demand for scalable, real-time multilingual communication in settings ranging from international conferences to community interpreting. However, as Slator highlights, while the speed and cost-efficiency of AI are undisputed, the industry's focus is shifting toward fit-for-purpose quality—questioning whether AI can meet the rigorous demands of high-stakes professional environments.

The proliferation of generative AI has also sparked intense debate regarding the future of the interpreting profession. Tomlinson et al. (2025) examine the occupational implications of working with AI, emphasizing that generative models are increasingly integrated into core professional activities such as information gathering and writing. While their research highlights the high applicability of AI across knowledge work, the rapid integration of these systems also sparks debate over potential risks. Specifically, scholars caution that while AI can automate routine linguistic conversions, it may unintentionally introduce new cognitive burdens or risk "deskilling" professionals. In the context of interpreting, this raises a critical question: does the presence of AI assist the listener, or does it complicate the communicative act by putting the cognitive burden on the audience to make sense of non-human output and filter inaccuracies?

### 2.2 Beyond human parity: The shift toward user-centric assessment

A central theme in recent discourse is whether AI can achieve human parity, a performance threshold where the quality of machine-generated output is statistically indistinguishable from that produced by a human professional (Hassan et al., 2018). In a practical context, it suggests that AI has reached a level of accuracy and fluency sufficient to functionally replace human experts in specific tasks.

However, recent studies such as Fantinuoli (2025) and Lu and Fantinuoli (2025) challenge the simplistic notion of parity often implied by commercial metrics. They argue that quality should not be measured solely by string-based metrics such as BLEU or NIST, which focus on surface-level correspondence and fail to account for the situated nature of interpreting. While acknowledging that newer methodologies—including Large Language Model (LLM)-based evaluations—offer a more granular analysis of semantic errors, they maintain that these still fall short of capturing the full complexity of the interpreting ecosystem (Lu & Fantinuoli, 2025).

Instead, they emphasize that true parity must be evaluated based on communicative effectiveness and user reception (Lu & Fantinuoli, 2025). They posit that even if AI achieves high literal accuracy, it may still lack the pragmatic competence such as managing prosody, emotional resonance, and cultural nuance that human interpreters provide. Without these

elements, machine-generated output may fail to ensure the listener's deep understanding or could increase the cognitive load required for the audience to process the information.

Alongside these evolving perspectives on interpreting quality, the methodology for assessment is also advancing. As AI interpreting becomes more prevalent, the methodology for assessing quality is also evolving. Shafiei (2024) proposes a refined analytic rubric for interpreting assessment, shifting the focus from mere error counting to a more holistic evaluation of communicative effectiveness. This work categorizes performance into content, form, and delivery, underscoring the importance of a structured approach in professional contexts. This reflects a broader trend in interpreting studies—a move away from source-text-centric evaluations toward user-centric models that measure how much information is actually retained and processed by the audience—which this study intends to explore further.

## 3 Theoretical framework

### 3.1 Reithofer's functional approach

The theoretical foundation of this study is rooted in Reithofer's (2013, 2014) reconceptualization of interpreting quality through the lens of "equivalent effect." Moving beyond traditional, source-text-centric models that prioritize formal correspondence, Reithofer argues that the success of an interpretation should be judged by its impact on the listener. This perspective aligns with the user-centric paradigm proposed by Shafiei (2024). It shifts the evaluative focus toward semantic adequacy—the degree to which the core meaning of the speaker's message is successfully reconstructed in the listener's mind.

For professional audiences such as journalists, semantic adequacy represents the most critical functional metric. Professionals in news reporting are trained to extract the central message from complex rhetoric; thus, an interpretation is deemed successful if it enables the listener to accurately identify and summarize a speaker's primary intent. Adopting Reithofer's teleological perspective, this study evaluates whether machine-generated output can achieve the professional utility threshold of the journalists, ensuring that the communicative purpose is preserved regardless of any minor linguistic deviations.

### 3.2 The application of idea units (IUs)

To measure the semantic adequacy of the participants' responses to the open-ended questions this study utilizes the concept of idea units (IUs). Originally developed in cognitive psychology and linguistics to analyze information processing (Chafe, 1980; Kintsch, 1998), the IU framework was adapted by scholars of interpreting studies to quantify the transfer of meaning in interpreting. An idea unit is typically defined as a minimal segment of information conveying a single propositional thought or semantic concept (e.g., Liu et al., 2004). By deconstructing a source speech into a comprehensive list of IUs, researchers can objectively measure how much of the original propositional content is retained by the audience.

This methodology serves as a robust alternative to the simplistic notion of parity criticized by Fantinuoli (2025). Unlike automated, string-based metrics (e.g., BLEU) that tend to focus on surface-level matches, IU analysis captures the holistic communicative effectiveness by evaluating participants' open-ended responses based on the presence of specific semantic

segments. This granular approach reveals not only the raw recall of facts but also how the listener synthesizes those facts into a logical narrative.

### 3.3 Cognitive load and schematic knowledge

The final pillar of this framework focuses on cognitive load, with particular attention paid to the auditory challenges inherent in AI-mediated communication. Reithofer (2013) posits that the efficiency of information transfer depends on minimizing the cognitive effort required to decode the speech. In professional contexts, an interpretation that lacks prosodic salience—including natural rhythm, emphasis, and intent-driven pausing—imposes an extrinsic cognitive load on the listener. This exhaustion of working memory can prevent the audience from achieving deep understanding, a concern echoed by Fantinuoli's (2025) argument regarding the lack of pragmatic competence in current AI systems.

Furthermore, this study considers the role of schematic knowledge (Reithofer, 2014) unique to journalists. While professional listeners can leverage their expertise to mitigate some cognitive strain, the machine-like delivery typical of commercial AI platforms available at the time of writing may still hinder their ability to decode nuanced diplomatic rhetoric (e.g., specific references to international treaties). This study hypothesizes that the human interpreter's ability to manage nuance and prosody is essential for minimizing cognitive load, thereby enabling a level of semantic adequacy that current AI platforms may struggle to achieve.

## 4 Methodology

### 4.1 Research design

The present study employs a quasi-experimental, between-subjects design to compare the cognitive impact and semantic adequacy of human interpreting (hereafter referred to as HI) and AI-powered simultaneous interpreting (hereafter AI). The primary objective is to measure the transfer of meaning to a professional audience in a real-world setting. Following the methodological framework established by Reithofer (2013, 2014) but adapted to include open-ended questions and a "Don't Know" metric, this study utilizes a comprehension-based evaluation, focusing on the listener's ability to comprehend and retain information delivered through each interpreting mode.

### 4.2 Stimulus material

The stimulus was a video recording of an online press conference held at the Japan National Press Club (JNPC) on December 12th, 2024. The speaker was Simon Stiell, Executive Secretary of the United Nations Framework Convention on Climate Change (UNFCCC) who participated online from Bonn, Germany (see Stiell, 2024, for detailed descriptions about the press conference). The material was selected based on several criteria to ensure the appropriateness for an interpreting comprehension test:

1. **Linguistic clarity:** The speaker is a native English speaker from Grenada, an English-speaking Caribbean nation, with a widely intelligible accent and deliberate delivery style, making the source text suitable for both human and AI processing.

2. **Standard difficulty:** The duration of the monologic opening speech was 11 minutes and 48 seconds, containing 1,316 words. This results in an average delivery speed of 111.5 words per minute (WPM), which is considered a standard and manageable pace for simultaneous interpreting.

3. **Thematic relevance:** The topic (climate change policy and the role of Japan) was chosen for its universality and familiarity to the Japan-based journalists, while being specific enough to allow for the creation of questions that test information acquired during the speech rather than prior general knowledge.

### 4.3 Participants

A total of 56 participants were recruited for this study. To ensure a highly homogeneous and professionally relevant sample, recruitment was restricted to current and former journalists based on their responses to the pre-experiment questionnaire. Given that the source material selected was a press conference held at JNPC whose role is to facilitate news coverage by Japan-based media outlets, journalists residing in Japan were deemed the ideal target group. They were also appropriate subjects for this study because they are well trained to extract accurate and meaningful information and reproduce them under time pressure.

The participants were recruited through a combination of snowball and purposive sampling, leveraging the author's professional network established through her 14 years of experience as a newspaper reporter in Japan. Although the criterion for participation was simply having professional journalistic experience, participants were asked to answer basic questions about their career backgrounds. The resulting sample included both active and former journalists, with a wide range of experience both in terms of medium and tenure.

Participants were divided into two groups ($n$ = 28 each). To ensure the validity of the comparison, a professional research firm (Trust One Co., Ltd.) was commissioned to balance the groups based on key attributes including age, gender, and years of journalistic experience. While two of the 56 participants identified themselves as non-native Japanese speakers, follow-up email interviews confirmed that their Japanese proficiency was sufficient for professional journalistic work in Japan. To maintain balance, these two individuals were assigned to different experimental groups.

Table1: Participant Demographics and Backgrounds

| **Category** | **Sub-category** | **Group 1: HI ($n$ = 28)** | **Group 2: AI ($n$ = 28)** |
|---|---|---|---|
| **Gender** | Male | 17 (60.7%) | 17 (60.7%) |
| | Female | 11 (39.3%) | 11 (39.3%) |
| **Age** | 30s | 7 (25.0%) | 5 (17.9%) |
| | 40s | 6 (21.4%) | 11 (39.3%) |
| | 50s | 10 (35.7%) | 6 (21.4%) |
| | 60s | 3 (10.7%) | 5 (17.9%) |
| | 70s+ | 2 (7.1%) | 1 (3.6%) |
| **Journalism Experience** | 1–5 years | 1 (3.6%) | 1 (3.6%) |
| | 6–10 years | 4 (14.3%) | 2 (7.1%) |
| | 11–20 years | 9 (32.1%) | 9 (32.1%) |

| | | | |
|---|---|---|---|
| | 21–30 years | 9 (32.1%) | 9 (32.1%) |
| | 31+ years | 5 (17.9%) | 7 (25.0%) |
| **Climate Reporting Experience** | Yes | 10 (35.7%) | 9 (32.1%) |
| | No | 18 (64.3%) | 19 (67.9%) |

### 4.4 Interpreting conditions

Participants in Group 1 (HI) listened to the Japanese interpretation provided by one of the two professional simultaneous interpreters assigned for the event. The audio was sourced from the recording available on the JNPC YouTube channel (Stiell, 2024) and used under a Memorandum of Understanding (MOU) between the author's research group and the JNPC. The interpreter for the opening speech identified by the JNPC was one of the top interpreters in Japan, with decades of experience in high-level diplomatic and journalistic settings.

Participants in Group 2 (AI) listened to the same speech interpreted by KUDO AI Speech Translator, a commercial AI simultaneous interpreting service. The specific settings used were:

- **Mode:** Speech-to-Speech (direct audio-to-audio translation).
- **Voice output:** A female Text-to-Speech (TTS) engine was selected to match the gender of the human interpreter in Group 1, thereby controlling for potential gender-based bias in auditory perception.
- **Technology level:** The video was generated using the latest proprietary engine available from KUDO's technical team at the time of the experiment (February 2025), representing cutting−edge AI interpreting technology.

### 4.5 Comprehension test

Immediately after viewing the video, participants completed an online comprehension test consisting of 10 items. The test was designed following established pedagogical standards for listening comprehension and Reithofer's (2013, 2014) focus on semantic adequacy. The test items and scoring criteria were finalized following a pilot study.

- **Q1–Q5 (Single-Choice):** Measured factual recall of key statements and specific terms.
- **Q6–Q8 (Multiple-Response):** Tested the ability to retain and identify multiple pieces of information within a single argument.
- **Q9–Q10 (Open-Ended):** Required participants to synthesize the speaker's intent and describe complex logical connections in their own words. These were scored based on the presence of predefined idea units (IUs).

The test also included a "Don't Know" option for every question to discourage guessing and to measure the participants' subjective sense of uncertainty, establishing a distinct category not present in Reithofer's (2013, 2014) experiments. The full list of questions and answers as well as the IUs are provided in the Appendix.

### 4.6 Procedure and environmental control

The study was conducted online by the research firm Trust One Co., Ltd under specific instructions provided by the author. In order to accommodate the participants' professional schedules, multiple dates were set between late February and early March 2025. The following protocols were enforced to ensure data integrity:

- **Environment:** Participants were instructed to watch the video in a quiet environment using headphones or earphones.
- **Restrictions:** The video was restricted to a single viewing. To prevent repeated viewings or the use of external devices, the duration of access was monitored and measured against the video's runtime.
- **Note-taking:** Participants were told to follow their usual professional habits—if they would typically take notes during a press conference, they were permitted to do so.

### 4.7 Data analysis

Quantitative data from the multiple-choice items were analyzed for mean scores and accuracy rates. Qualitative data from the open-ended questions were scored based on the presence of predefined IUs. Finally, the feedback from the self-reported section was qualitatively analyzed to correlate the objective scores with the participants' subjective experience of the interpreting quality.

Following the methodological framework established by Reithofer (2013, 2014), the data analysis in this study prioritizes the nature and direction of the communicative effect. Reithofer argues that in exploratory experimental research focused on effect equivalence, the application of hypothesis testing (p-values) may be useful but inadequate, since the primary objective is not merely to detect a statistical difference, but to evaluate the quality and equivalence of information transfer across different modalities.

The data were analyzed using a combination of descriptive quantitative scoring and detailed qualitative analysis:

1. **Descriptive statistical analysis:** For the majority of the assessment, the analysis focused on the calculation of mean scores and accuracy rates to identify overall trends and the distribution of scores between the HI and AI groups, emphasizing the communicative impact rather than statistical generalization.

2. **Granular IU analysis (Q9 and Q10):** For the open-ended questions (Q9 and Q10), the responses were evaluated against a set of predefined IUs derived from the source text (10 IUs for Q9 and 9 for Q10) to determine what information was successfully reconstructed in the listeners' minds. This allowed for a more detailed analysis of semantic adequacy in high-stakes professional contexts.

3. **Qualitative synthesis:** By combining the quantitative scores with the more detailed IU analysis of Q9 and Q10, this study aims for a qualitative judgment of equivalence. This approach evaluates the degree to which AI interpretation meets the professional utility threshold required by journalists.

## 5 Key findings

## 5.1 Overview of participant performance

The primary objective of this study was to evaluate the semantic adequacy of HI and AI in an authentic professional context. A total of 56 participants, all with professional journalism backgrounds, completed a ten-item comprehension test immediately following the stimulus. The maximum possible score was 10.0.

The quantitative analysis revealed a distinct performance gap between the two groups (see Table 2). The mean score was 4.50 in the HI group ($n$ = 28) and 3.71 in the AI group ($n$ = 28). This disparity, while reflecting the inherent difficulty of a one-time listening task without any preparation, underscores the advantage of human intervention in preserving the propositional content and logical coherence of the source text. Notably, the HI group outperformed the AI group on 8 out of 10 questions, including all three multiple response questions and two open-ended questions, which require a higher degree of cognitive processing and precise information retrieval compared to single-choice items.

Table 2 presents a comparative analysis of the accuracy rates and the frequency of "Don't Know" (DK) responses for all 10 items. For the open-ended questions (Q9 and Q10), correct responses were determined by the presence of predefined IUs. This ensured that the qualitative depth of the participants' comprehension was quantified alongside the objective factual recall measured in the earlier items (Q1–Q8).

Table2: Comprehension Scores and "Don't Know" Rates by Item and Group

| **Item** | **Theme** | **HI Group ($n$ = 28) Correct** | **HI Group ($n$ = 28) Don't Know** | **AI Group ($n$ = 28) Correct** | **AI Group ($n$ = 28) Don't Know** |
|---|---|---|---|---|---|
| **Q1** | Urgency of Climate Action (single) | 17 (60.7%) | 3 (10.7%) | **24 (85.7%)** | 0 (0.0%) |
| **Q2** | Specific Impacts on Japan (single) | **23 (82.1%)** | 1 (3.6%) | 20 (71.4%) | 5 (17.9%) |
| **Q3** | Negative Impacts on GDP (single) | **6 (21.4%)** | 5 (17.9%) | 2 (7.1%) | 9 (32.1%) |
| **Q4** | G7 Communiqué Commitments (single) | **19 (67.9%)** | 4 (14.3%) | 8 (28.6%) | 6 (21.4%) |
| **Q5** | Economic Impacts (single) | 7 (25.0%) | 1 (3.6%) | **18 (64.3%)** | 1 (3.6%) |
| **Q6** | Mentions of COP29 (multiple) | **6 (21.4%)** | 2 (7.1%) | 2 (7.1%) | 1 (3.6%) |
| **Q7** | Reasons for Clean Energy (multiple) | **8 (28.6%)** | 1 (3.6%) | 3 (10.7%) | 2 (7.1%) |
| **Q8** | Japan's Strengths (multiple) | 8 (28.6%) | 1 (3.6%) | 8 (28.6%) | 2 (7.1%) |
| **Q9** | Challenges Mentioned by Business Leaders (open) | **11 (39.3%)** | 13 (46.4%) | 4 (14.3%) | 18 (64.3%) |
| **Q10** | Motivation for Future Action (open) | **21 (75.0%)** | 4 (14.3%) | 15 (53.6%) | 6 (21.4%) |
| **Mean** | **Total Score (out of 10.0)** | **4.50** | **12.5%** | **3.71** | **17.9%** |

## 5.2 Quantitative analysis of comprehension (Q1–Q8)

### 5.2.1 Single-choice items (Q1–Q5)

In the single-choice category, the HI group demonstrated superior performance in capturing nuanced policy details and specific impacts, particularly in questions requiring the distinction of complex clauses. In Q4, which focused on the commitments in the G7 communiqué, the HI group showed considerably higher retention (HI 67.9% vs. AI 28.6%). This question required participants to distinguish between general environmental concern and a specific policy commitment—namely, "further limits on greenhouse gas emissions." The AI group's lower performance here suggests that while the system can translate individual words, it often fails to preserve the logical weight of specific clauses. This leads to a flattening of the message, where crucial policy details are lost in a sea of generalities, whereas the human interpreter successfully conveyed the pragmatic intent and emphasis of the speaker. Similar tendencies were observed in Q2 and Q3 as well.

On the other hand, the results for Q1 (the reason for the urgency of climate action by Japan) and Q5 (the economic impact of climate actions) showed that the AI group outperformed the HI group (Q1: AI 85.7% vs. HI 60.7%; Q5: AI 64.3% vs. HI 25.0%). To understand this discrepancy, a comparative analysis was conducted between the source speech, the HI transcript, and the AI transcript.

Regarding Q5, when the speaker repeated the phrase "clean energy boom" four times, the AI rendered it literally as "クリーンエネルギーブーム" (clean energy boom) in every instance. In contrast, the human interpreter replaced all four occurrences with more sophisticated Japanese expressions, likely to avoid the informal or non-diplomatic nuance associated with the loanword "boom" in Japanese. However, this professional refinement may have made it more difficult for participants to map the interpretation onto the specific answer choice, "Growth of the clean energy industry."

For Q1, which concerned the urgency of climate action mentioned at the beginning of the speech, the AI consistently voiced the full phrase "national climate plans." The human interpreter, however, employed a time-saving technique by abbreviating the term to "NDC (Nationally Determined Contributions)" from the second mention onward. This technical efficiency may have inadvertently weakened the verbal cues needed for participants to recall the full phrase in the answer choice. Furthermore, as the answer to Q1 appeared at the very start of the speech, the cascaded speech-to-speech translation system was at its most stable, with minimal delays or unnatural pauses—issues that typically compound as a session progresses.

### 5.2.2 Multiple-response items (Q6–Q8)

The performance gap widened significantly in the multiple-response section (Q6, Q7, and Q8), which required participants to identify all correct statements from a list of four (excluding DK). These items were intended to measure the information density that a listener could successfully process, since multiple response questions demand that participants verify each individual proposition against their mental reconstruction of the speech.

In Q6 (details regarding COP29) and Q7 (reasons for promoting clean energy), the HI group showed a markedly higher success rate. For Q7, which required recollection of four distinct socio-economic benefits (international competitiveness, living standards, economic growth, and productivity), 28.6% of the HI group identified the full set of correct responses, compared to only

10.7% in the AI group. It could be argued that the prosodic cues provided by the human interpreter—such as pausing, emphasis, and intonational grouping—played a vital role in helping journalists organize the information when storing it in working memory. In contrast, the AI's relatively monotonic and occasionally erratic delivery pace appeared to increase the cognitive load, preventing listeners from capturing the full list of items. This argument was generally supported by the qualitative feedback in the metadata, part of which will be described in the following sections.

## 5.3 Qualitative analysis of open-ended responses (Q9–Q10)

Responses to the two open-ended questions were evaluated based on how effectively the semantic content of the original message was reconstructed. While Reithofer's (2013, 2014) semi-open format facilitates scoring, an open-ended design was adopted to better reveal differences in semantic adequacy between HI and AI, prioritizing communicative effect over binary correctness.

### 5.3.1 Q9: Specific challenges highlighted by business leaders

Q9 required participants to synthesize a complex argument regarding the "urgent need for acceleration" in Japan's climate transition. While the raw response rate was low for both groups, the HI group achieved a significantly higher IU-based accuracy rate of 39.3%, compared to just 14.3% in the AI group.

The qualitative difference was stark: HI participants correctly identified the nuanced link between overcoming the "misconception that climate action harms the economy" and the formulation of "more ambitious NDCs." In contrast, AI group participants exhibited a high "Don't Know" rate of 64.3%. Successful AI responses were typically limited to isolated keywords such as "clean energy" without capturing the underlying logical connection to Japan's economic security or specific policy recommendations. This suggests that the AI's rendering lacked the logical connectors necessary for the audience to construct a coherent mental model of the business leaders' specific advocacy.

### 5.3.2 Q10: Reasons for advancing to the next stage

Q10 focused on the speech's conclusion, which linked Japan's historical role (e.g., the Kyoto Protocol) to its future national interest in the $2 trillion clean energy market. The HI group demonstrated robust performance with a 75.0% accuracy rate, whereas the AI group scored 53.6%.

Most HI participants successfully integrated the speaker's rhetorical arc, linking past achievements to future economic survival. While the AI group performed relatively well on this item due to the high frequency of the "2 trillion dollar" keyword in the AI transcript, they often failed to capture the pragmatic force of the "national interest" argument. The 21.4% "Don't Know" rate in the AI group (compared to 14.3% in HI) indicates that even at the conclusion of the speech, the AI's delivery continued to cause cognitive fatigue, preventing a segment of the audience from fully grasping the final call to action.

## 5.4 The "Don't Know" factor and subjective uncertainty

A critical metric in this study was the frequency of "Don't Know" (DK) responses, which serves as a proxy for participant uncertainty and perceived communicative failure. Across all 10 questions, the AI group recorded a higher DK rate (17.9%) compared to the HI group (12.5%).

This gap is particularly pronounced in the open-ended section, where the AI's DK rate reached 64.3% for Q9. This disparity suggests that the AI's failure is not merely a matter of lexical mistranslation, but a failure of prosodic salience and semantic weighting. As indicated by participant feedback, the "flat" and "mechanical" nature of the AI output made it difficult to distinguish between peripheral information and core arguments. While the human interpreter used techniques such as professional abbreviation (e.g., using "NDC") and sophisticated synonymy to maintain engagement, the AI's often verbatim and repetitive output may have forced the listener to dedicate excessive mental resources simply to decoding the syntax and rhythm of the speech. This left insufficient cognitive capacity for the higher-order tasks of meaning-making or long-term retention. The significantly higher rate of DK responses in the AI group serves as a proxy for this communicative failure.

## 6 Conclusion

This study conducted a comprehensive evaluation of the functional equivalence between human interpreting (HI) and AI interpreting (AI) by measuring the comprehension of professional journalists. By building on Reithofer's (2013, 2014) framework of measuring semantic adequacy, the research moved beyond automated linguistic metrics to assess the communicative impact of these two modes. The empirical results demonstrate a clear disparity: while AI has reached a level of lexical competence capable of capturing high-frequency keywords and specific terms, it still lags behind human experts in facilitating the deep, structural understanding required in high-stakes professional settings.

The quantitative findings, showing a mean score of 4.50 for HI versus 3.71 for AI, are particularly telling when viewed alongside the "Don't Know" (DK) rates. The AI group's significantly higher DK rate in open-ended questions, such as 64.3% in Q9, suggests that machine-generated output often leaves professional listeners in a state of cognitive uncertainty. This supports the theoretical contention that semantic adequacy is not merely a product of word-for-word accuracy, but a result of prosodic salience and logical segmentation (elements that were notably absent in the AI's flat, mechanical delivery). Furthermore, the higher rates of correct answers in Q1 and Q5 for AI reveals that while AI's literal repetition can serve as a mnemonic anchor for specific terms, it still struggles to convey the pragmatic intent of the speaker. The human interpreter's professional interventions, such as stylistic refinements and strategic abbreviations, prioritize professional register and temporal efficiency, providing a more comprehensible and often convincing narrative that current S2S technologies cannot replicate.

From a professional standpoint, the results underscore a critical risk: the extrinsic cognitive load imposed by AI. For journalists, whose work involves the rapid synthesis of complex information under deadline pressure, the need to manually filter and repair fragmented AI output remains a barrier to professional utility. However, the AI's success in specific keyword-heavy segments indicates its potential as a powerful supplementary tool, pointing toward a hybrid model where AI manages data-heavy segments while humans retain control over the rhetorical and context-dependent dimensions of the discourse.

Despite these insights, the study has obvious limitations. The experimental setting utilized a single, albeit highly representative, press conference. Future research should explore different genres—such as highly technical lectures or emotionally charged negotiations—and investigate whether the gap between HI and AI varies by content type. Additionally, as AI voices become more naturalistic, further investigation into the specific impact of synthetic prosody on listener comprehension and long-term fatigue is warranted.

In conclusion, the quest for human parity in AI must shift its focus from the machine's output to the listener's intake. As this study demonstrates, for professional communication where the transfer of intent, nuance, and logical coherence is paramount, human expertise remains indispensable, as the true measure of interpreting quality lies in the successful and efficient reconstruction of meaning within the human mind.

## References (in alphabetical order)


Chafe, W. L. (ed.). (1980). *The Pear Stories: Cognitive, Cultural, and Linguistic Aspects of Narrative Production*. Norwood, NJ: Ablex.

Fantinuoli, C. (2025). Machine interpreting. In E. Davitti, T. Korybski & S. Braun (eds.), *The Routledge Handbook of Interpreting, Technology and AI* (pp. 209–228). Abingdon, Oxon: Routledge.

Hassan, H. et al. (2018). Achieving human parity on automatic Chinese to English news translation. *arXiv.* Retrieved 18/01/2026 from https://arxiv.org/abs/1803.05567

Kintsch, W. (1998). *Comprehension: A Paradigm for Cognition*. Cambridge: Cambridge University Press.

KUDO. (2025). *KUDO AI Speech Translator*. Retrieved 18/01/2026 from https://kudo.ai/solutions/kudo-ai-speech-translator/

Liu, M., Schallert, D.L. and Carroll, P.J. (2004) Working memory and expertise in simultaneous interpreting. *Interpreting*, 6(1), 19–42.

Lu, X. & Fantinuoli, C. (2025). Machine and computer-assisted interpreting: Innovations in and implications for interpreting practice, pedagogy and research. *Linguistica Antverpiensia, New Series: Themes in Translation Studies*, 24*,* 1-22. Retrieved 18/01/2026 from https://lans-tts.uantwerpen.be/index.php/LANS-TTS/article/view/869

Reithofer, K. (2013). Comparing modes of communication: The effect of English as a lingua franca vs. interpreting on information processing in conference situations. *Interpreting*, 15(1), 48–73.

Reithofer, K. (2014). Englisch als Lingua Franca und Dolmetschen: Ein Vergleich zweier Kommunikationsmodi unter dem Aspekt der Wirkungsäquivalenz. Tübingen: Narr Francke Attempto Verlag.

Shafiei, S. (2024). A proposed analytic rubric for consecutive interpreting assessment: Implications for similar contexts. *Language Testing in Asia*, 14, Article 13.

Slator. (2024). *Slator 2024 Interpreting Technology and AI Report*. Retrieved 18/01/2026 from https://slator.com/2024-interpreting-technology-and-ai-report/

Stiell, S. (2024). *Press Conference by Simon Stiell, Executive Secretary of the UNFCCC* [Video]. YouTube. Retrieved 18/01/2026 from https://www.youtube.com/watch?v=ZGcQVmiE1Ec

Tomlinson, K., Jaffe, S., Wang, W., Counts, S. & Suri, S. (2025). Working with AI: Measuring the occupational implications of generative AI. *arXiv*. Retrieved 18/01/2026 from https://arxiv.org/abs/2507.07935

## Appendix: Comprehension test items and correct answers

The following items were used to assess the participants' comprehension of the press conference given by Simon Stiell, Executive Secretary of the UNFCCC. The test was administered in Japanese; the items below are translated for reference. The correct answers are underlined. The Idea Units (IUs) used for the analysis of Q9 and Q10 are listed.

### Part 1: Single-Choice Items

Q1. At the beginning of the speech, it was mentioned that this is a critical moment for Japan's climate action. What is the reason for this?

A) A new industrial revolution is about to begin.

B) All countries are required to submit new national climate plans.

C) International pressure for climate action has intensified.

D) Japan's initiatives for renewable energy have fully commenced.

E) I don't know.

Q2. Which of the following was mentioned in the speech as an impact of global warming that Japan is currently facing?

A) Sea-level rise.

B) Increasing temperature fluctuations.

C) Intense storms and flooding.

D) Poor crop growth.

E) I don't know.

Q3. Which of the following was NOT mentioned in the speech as a negative impact of climate change that could lower the GDP of Asia, including Japan?

A) Water shortages.

B) Extreme heat.

C) Pollution.

D) Increased mortality rates.

E) I don't know.

Q4. Which of the following was explicitly promised by all G7 countries in the recently released communiqué?

A) Expansion of renewable energy use.

B) Resetting international climate goals.

C) Strengthening environmental protection measures.

D) Further reduction of greenhouse gas emissions.

E) I don’t know.

Q5. According to the speech, what is the impact of climate change measures on the Japanese economy?

A) Reorganization of the business community.

B) Growth of the clean energy industry.

C) Emergence of venture capital firms.

D) Stabilization of exchange rates.

E) I don’t know.

**Part 2: Multiple-Response Items**

Q6. Regarding COP29 held in Azerbaijan, select all the statements mentioned in the speech.

A) The G20 decided on the phase-out of fossil fuels.

B) The UK announced a bold emission reduction target.

C) Brazil set a cap on emissions.

D) Switzerland revealed a new national climate plan.

E) I don’t know.

Q7. Select all the correct reasons mentioned in the speech for why clean energy should be promoted.

A) Maintaining international competitiveness.

B) Improving standards of living.

C) Economic growth.

D) Improving productivity.

E) I don’t know.

Q8. Select all of Japan’s strengths mentioned in the speech.

A) High technical capabilities.

B) Low taxes.

C) Highly skilled human resources.

D) Robust legal systems.

**Part 3: Open-Ended Items**

Q9. Explain the challenges regarding climate change measures that Japanese business leaders emphasized during their meeting with Executive Secretary Stiell.

Model Answer:

Japanese business leaders emphasized that accelerating the clean energy transition and climate resilience is essential. Referring to the significant business opportunities in both domestic and overseas markets, they argued that climate action is the only path to the prosperity and security of the Japanese economy. Specifically, they proposed that in order for Japan to achieve sustainable economic growth, it is indispensable to formulate a more ambitious Nationally Determined Contribution (NDC) and to strengthen the policies that support it.

IUs: Japanese business leaders emphasized the need; Accelerating clean energy transition; Accelerating climate resilience; Essential/Urgent requirement; Large business opportunities (Domestic/Overseas); Climate action is the only path; To prosperity and security of Japan's economy; Formulation of more ambitious NDC; Strengthening supporting policies; To achieve sustainable economic growth (10)

Q10. In the conclusion of the speech, which factor was most emphasized as the reason for Japan to advance its climate change measures to the next stage?

Model Answer:

Japan has played an important international role in climate change measures to date. However, the current progression of climate change remains serious, and if it continues at this rate, global temperatures will rise, leading to catastrophic impacts on all economies and people, including those in Japan. Therefore, for Japan to take further climate action is directly linked to its own national interest. In particular, proactive efforts are required now for Japanese companies and citizens to participate in and enjoy the benefits of the $2 trillion clean energy market.

IUs: Japan's long history/important international role; Current progression remains serious; Global temperatures will rise (if unchecked); Catastrophic/devastating impacts; Affects all economies/people (including Japan); Taking further action is in Japan's national interest; $2 trillion clean energy market; Participation and enjoying benefits; Proactive efforts are required now (9)

---