

Article

Human-in-the-Loop AI Feedback in Interpreter Training: An ASR-Based Platform Analysis of Instructor Annotations, Comment Functions, and System Constraints

Juriae Lee

Graduate School of Translation and Interpretation, Ewha Womans University, Seoul 03760, Republic of Korea; julietlee@ewha.ac.kr

Abstract

The use of AI-assisted feedback systems has steadily increased in higher education, yet there has been limited research as to which components of expert feedback can be supported by AI and which require continued human judgment. This issue is particularly important in professional training domains where feedback must rely upon contextual interpretation, source–output comparison, and domain expertise. Accordingly, this study analyzes instructor-generated feedback in TalkTrack, an automatic speech recognition (ASR)-supported digital platform for interpreter training, as a basis for human-in-the-loop AI feedback design. The dataset consists of 44 Korean–Japanese, Japanese–Korean, and Chinese–Korean interpreting performances produced by 12 graduate students and includes 931 feedback tags and 756 substantive qualitative comments manually assigned by 10 instructors. The categories most frequently assigned were expression and translation-error feedback, while corrective feedback was the dominant instructional function. Further, feedback behavior varied substantially by instructor and feedback category, indicating that different types of feedback require different levels of human intervention and system support. Accordingly, this study proposes a human-in-the-loop feedback architecture employing ASR and large language models to support detection, organization, retrieval, and suggestion generation, while instructors retain responsibility for source-text fidelity, discourse coherence, pedagogical judgment, and final feedback validation, from which this study proposes empirical design requirements for scalable AI-supported feedback systems in professional e-learning environments.

Keywords: human-in-the-loop AI; AI-assisted feedback; automatic speech recognition; e-learning; computer-assisted interpreter training; instructor annotation; feedback taxonomy; learning analytics; interpreter training



Academic Editor: César A. Collazos

Received: 8 June 2026

Revised: 1 July 2026

Accepted: 11 July 2026

Published: 15 July 2026

Copyright: © 2026 by the author.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

AI-enhanced e-learning platforms support an increasing repertoire of tools for automated assessment, feedback generation, learner analytics, and adaptive learning pathways, with notable improvements reported in the scalability and rapid feedback delivery of automated feedback systems [1,2]. More recently, research has examined the changing role of instructors in AI-assisted educational environments [3,4]. However, in the complex circumstances of professional training, content-level feedback, which often requires expert human judgment, contextual interpretation, and domain-specific evaluation, cannot be reduced to automatic scoring. Interpreter training is a prime example of this problem be-

cause instructors must evaluate meaning transfer, target-language formulation, delivery, fluency, and fidelity to the source discourse of the interpreted output simultaneously [5,6].

Affordances to integrate source-text upload, learner recording, ASR-based transcription, annotation, and visualization functions are being developed on digital interpreter-training platforms like TalkTrack [7–9], allowing the automatic detection of selected surface fluency phenomena, such as pauses, fillers, and backtracking, but feedback on meaning accuracy and target-language expression still depends on instructors' manual annotation and comments. This creates a human-in-the-loop AI problem: how can a system automate what is reliably automatable while supporting, rather than replacing, expert instructor judgment?

The issue is not merely whether AI can provide feedback, but how AI-assisted feedback infrastructure should be designed around actual instructor behavior. Instructors differ in the number of tags they assign, the likelihood that they will attach comments, the types of comments they write, and how they handle ASR errors or ambiguous feedback categories. If these behaviors are not quantified, AI-supported feedback platforms might automate the wrong components or fail to reduce the most burdensome parts of the feedback workflow.

This study analyzes feedback data generated in TalkTrack, an ASR-supported e-learning platform for interpreter training. Rather than evaluating learner performance directly, the study models instructor feedback as an annotation and comment-generation process within a human-in-the-loop educational AI system and investigates which feedback categories are likely to trigger comments, which types of feedback most require human judgment, where ASR errors constrain annotation reliability, and the design requirements that follow for AI-assisted feedback systems.

To this end, instructor annotations are treated as empirical traces of the human judgment that future AI-assisted feedback systems must support, structure, or partially automate. This study thus focuses on instructor annotation behavior as a design object for human-in-the-loop AI feedback, rather than on learner achievement or AI feedback effectiveness, and addresses the following research questions:

RQ1. What types of instructor annotations are generated in an ASR-supported e-learning platform for interpreter training, and how are they distributed?

RQ2. Which feedback categories and comment functions are least amenable to automation?

RQ3. How do tag type, performance layer, and instructor identity affect the likelihood that a manual tag will include a qualitative comment?

RQ4. How do ASR transcription constraints and residual feedback categories reveal gaps in platform taxonomy and human–AI workflow?

RQ5. What human-in-the-loop AI design requirements can be adduced for scalable feedback systems in professional e-learning?

The remainder of this paper is organized as follows. Section 2 reviews research on interpreter feedback, AI-assisted language learning, and human-in-the-loop educational systems. Section 3 describes the TalkTrack platform, the dataset, and the methodological framework of the analysis. Section 4 presents our findings on instructor feedback patterns and educational functions. Section 5 discusses the implications for AI-assisted interpreter feedback and proposes a human-in-the-loop system architecture. Finally, Sections 6–8 present the conclusions of this study, summarizing its major findings as they pertain to the research questions, the limitations of this study, and future directions for the development and validation of the proposed human-in-the-loop AI feedback framework, respectively.

2. Related Work

2.1. Instructor Feedback in Interpreter Training

Interpreter training requires feedback much more intricate than general language correction, as instructors evaluate whether the interpreted output preserves source meaning, maintains discourse logic, and uses appropriate target-language expression while remaining fluent under real-time processing constraints [10,11]. Researchers have emphasized the role of instructor feedback in identifying performance gaps, guiding strategy development, and supporting professional competence formation [12,13]. Learners also tend to value instructor feedback because it carries expert authority and provides concrete alternatives for improvement [10,14,15].

Research on interpreter pedagogy has examined learners' engagement with instructor feedback and corrective feedback practices [16], disfluency and interpreter performance [17], technology-supported interpreter training [18], listening comprehension [19], strategy acquisition [20], aptitude–treatment interactions [21], and reflective practice through peer feedback [22].

At the same time, interpreter feedback is hard to standardize. The instructor must often listen to an audio recording, compare it with the source text, inspect transcripts, and identify whether a given problem stems from comprehension, reformulation, omission, delivery, or system transcription error. Feedback is thus often labor-intensive and uneven across instructors [23,24], which has motivated the development of digital feedback environments that can store, structure, and reuse feedback data.

2.2. AI-Assisted Feedback and E-Learning

The use of automated feedback systems in online learning, writing instruction, and digital assessment has been widely studied. Systematic reviews have shown that automated feedback can improve scalability and immediacy but also raise concerns about accuracy, explainability, and the division of labor between the system and the teacher [25,26]. Recent work on AI-assisted teaching similarly stresses that automation does not supplant instructors, who instead shift toward the interpretation, validation, and orchestration of AI-generated outputs [27,28].

Accordingly, the key design question in professional e-learning is not full automation but human-in-the-loop automation. Feedback systems must identify which components can be suggested or prestructured by AI, which require instructor verification, and which must remain human-dominant. This distinction is particularly important for feedback on source-text fidelity, discourse coherence, and domain-specific target-language formulation.

2.3. ASR, CAIT, and Speech-Based Learning Analytics

Early computer-aided interpreter training (CAIT) research established the role of computer technology in interpreter training and self-assessment [29,30]. More recent studies have explored AI-supported interpreting tools and online feedback environments [31] and the technical accuracy and evaluation of ASR systems [32–34]. Other studies have highlighted the issues of error-preserving transcription, language-specific recognition problems, and orthographic ambiguity that directly affect its educational use in interpreter training [35,36].

For interpreter training, transcription should be pedagogically faithful rather than merely readable by sufficiently preserving learner errors and disfluencies for diagnosis. In particular, as Japanese kanji conversion errors, accent-related recognition problems, and orthographic ambiguity can make it difficult to distinguish learner errors from sys-

tem errors, ASR-supported interpreter training is an appropriate domain for investigating human-in-the-loop feedback design.

3. Platform, Dataset, and Analytical Methods

This section presents the platform, dataset, and analytical procedures used to examine instructor-generated feedback. The analysis focuses on feedback tags, qualitative comments, comment provision, and their implications for human-in-the-loop AI feedback design.

3.1. TalkTrack as an ASR-Supported Feedback Platform

TalkTrack is a digital feedback platform developed to support computer-assisted interpreter training. It allows instructors to upload source texts and audio, collect recordings of student interpreting, generate ASR transcripts, inspect visual fluency indicators, assign feedback tags, and write segment-level comments [7–9]. Speech transcription was performed using commercially available ASR services selected according to the target language as providing the most reliable transcription of spontaneous interpreting speech while preserving disfluency phenomena relevant to feedback analysis. The Korean and Japanese interpreting outputs were transcribed using Clova Note and Microsoft Azure Speech Service, respectively. The ASR outputs generated by these commercial services were used as part of the established TalkTrack workflow. In interpreter training, instructors must evaluate multiple dimensions simultaneously, including fluency features, such as pauses, fillers, and backtracking; content-level accuracy through comparison with the source text; and delivery-related features, such as intonation and speech rate. By automatically supporting fluency analysis, TalkTrack can substantially reduce the instructor's evaluation burden. However, content-level feedback has not yet been automated; it must be manually generated by instructors as they consult the source text, transcript, and audio together.

Figure 1 shows the TalkTrack feedback annotation interface, which displays the source text alongside the ASR transcript of the student's interpretation, while the feedback panel allows instructors to select a predefined tag and enter comments for a selected segment. This interface allows for the analysis of feedback data as a structured human annotation process.

Although the platform can accept source texts in any language, the current ASR-based automatic transcription and annotation modules operate only for interpreted output in Korean or Japanese. This limitation reflects the current availability of interpreter-oriented ASR models, which have been developed and validated only for these two target languages. In the future, the platform will incorporate multilingual ASR models able to support a broader range of target interpreting languages.

Figure 2 illustrates how TalkTrack transforms instructor annotations into learner-facing feedback visualizations. As shown in Panel (a), the feedback tag distribution for an individual student is displayed as a radar profile, allowing instructors to identify that student's error profile and the categories where feedback is concentrated, while Panel (b) shows how the same learner's feedback profile can be compared with the class average to support relative diagnosis within a cohort. These visualizations are important in AI-enhanced e-learning because they convert segment-level human annotations into analyzable feedback data. They also indicate how instructor-generated tags can serve as inputs for future learning analytics, adaptive feedback dashboards, and human-in-the-loop AI feedback support.

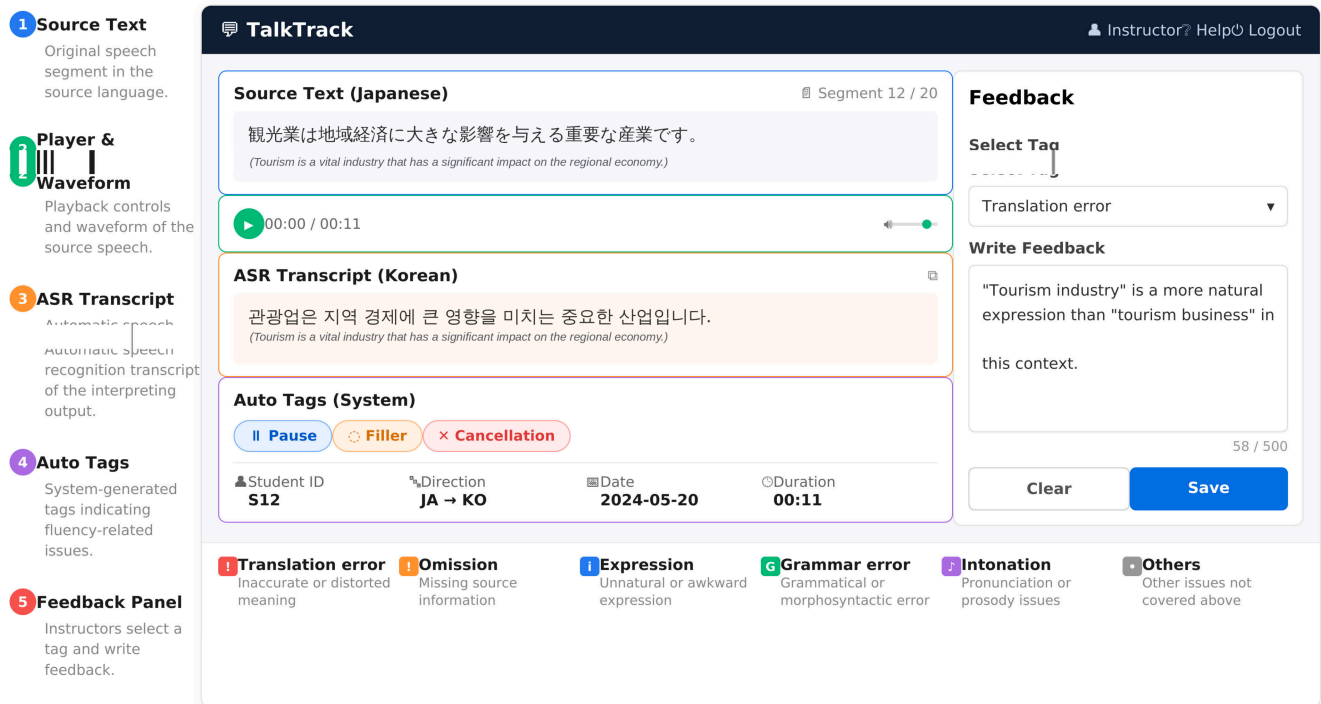


Figure 1. The TalkTrack feedback annotation interface.

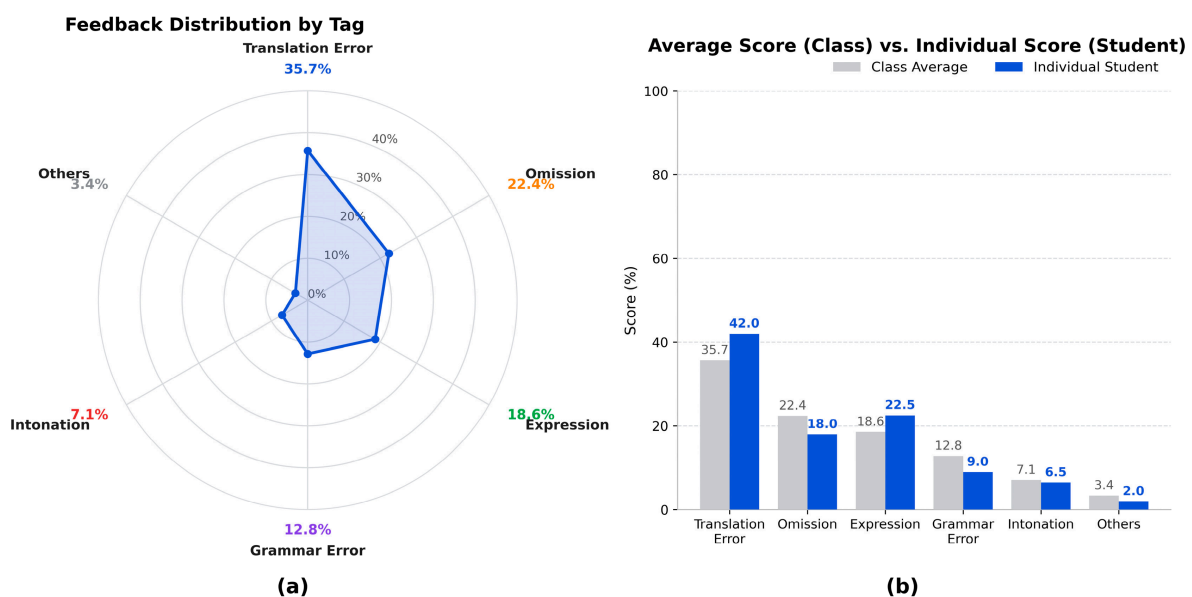


Figure 2. Examples of feedback visualization generated by TalkTrack: (a) feedback tag distribution for an individual student; (b) comparison of an individual student's feedback profile with the class average.

3.2. Dataset

The dataset comprises 44 interpreting performance files produced by 12 master's students enrolled in a graduate school of translation and interpretation in Korea, 11 of whom were native speakers of Korean, and 1 a native speaker of Japanese. Ten instructors (P1–P10) provided feedback through TalkTrack. All instructors had completed graduate-level professional training in interpretation and translation, were active professional interpreters and translators, and served as faculty members or lecturers teaching graduate-level interpreting and translation courses, with teaching experience ranging from fewer than

10 years ($n = 4$), to between 10 and 20 years, ($n = 4$) and 20 or more years ($n = 2$). All instructors used the same TalkTrack platform and the same predefined feedback taxonomy during routine instructional activities. The observed instructor-level variations thus reflect authentic pedagogical practice within a shared instructional environment rather than differences in annotation procedures. The dataset covers consecutive interpreting (CI) and simultaneous interpreting (SI) in the Korean-to-Japanese (KO-JA), Japanese-to-Korean (JA-KO), and Chinese-to-Korean (CN-KO) directions (Table 1).

Table 1. Overview of the dataset.

Item	Description
Participants	Twelve master's students in a graduate school of translation and interpretation
Native language	Eleven Korean native speakers and one Japanese native speaker
Number of interpreting performances	44 files total: 20 Korean–Japanese/Japanese–Korean files and 24 Chinese–Korean files
Korean–Japanese/Japanese–Korean dataset	CI/KO-JA: 7; CI/JA-KO: 7; SI/KO-JA: 3; SI/JA-KO: 3
Chinese–Korean dataset	CI/CN-KO: 12; SI/CN-KO: 12
Number of instructors providing feedback	Ten instructors total: four instructors for the Korean–Japanese/Japanese–Korean dataset and six instructors for the Chinese–Korean dataset
Instructors' teaching experience	Fewer than 10 years: 4; 10 to fewer than 20 years: 4; 20 years or more: 2
Interpreting mode	Consecutive interpreting (CI) and simultaneous interpreting (SI)
Language direction	Korean-to-Japanese (KO-JA), Japanese-to-Korean (JA-KO), and Chinese-to-Korean (CN-KO)
Platform	TalkTrack

The unit of analysis in this study is not learner scores or interpreting quality but instructor-generated annotation. The analyzed data include manually assigned feedback tags, presence or absence of a comment, qualitative comment content, instructor identity, performance file count, and performance-layer classification.

Because raw comment counts can be affected by the length of the interpreted output, source-text length and ASR-transcript length were calculated by interpreting type. After removing system markers like pause tags and segment separators, comment density was calculated as the number of function-coded comments per 1000 output characters, exclusive of blank spaces. The detailed source/output length and density values are reported in Appendix A; this normalization supports the interpretation of file-level comment counts across interpreting modes and language directions.

3.3. Variables, Quantitative Procedures, and Functional Coding

The original instructor feedback taxonomy consisted of six manually assigned tags: expression, translation error, omission, grammar error, intonation, and others. For each manually assigned tag, the presence of a comment was first identified from the JSON attribute field. Because some attribute fields were empty or nonsubstantive, the final binary variable for quantitative analysis was coded as *Comment present* = 1 and *No Comment* = 0. Automated fluency annotations like pauses, fillers, and cancelations were excluded from this analysis. The dataset thus comprised 931 manually assigned feedback tags. To quan-

tify the proportion of annotations that contained substantive instructional comments, the comment rate was calculated as follows:

$$\text{Comment Rate} = \frac{\text{Number of qualitative comments}}{\text{Number of manually assigned tags}} \times 100, \quad (1)$$

where the number of qualitative comments refers to comments containing substantive instructional information, excluding empty or nonsubstantive entries.

Of the 931 manually assigned feedback tags, 756 (81.2%) contained sufficient qualitative information for functional coding. The six feedback tags were then analytically grouped into four performance layers: meaning transfer, target-language formulation, delivery/prosody, and others.

Figure 3 summarizes the overall methodological workflow of the study from data collection and ASR-assisted annotation through statistical analysis to the identification of implications for the design of the proposed human-in-the-loop AI feedback framework.

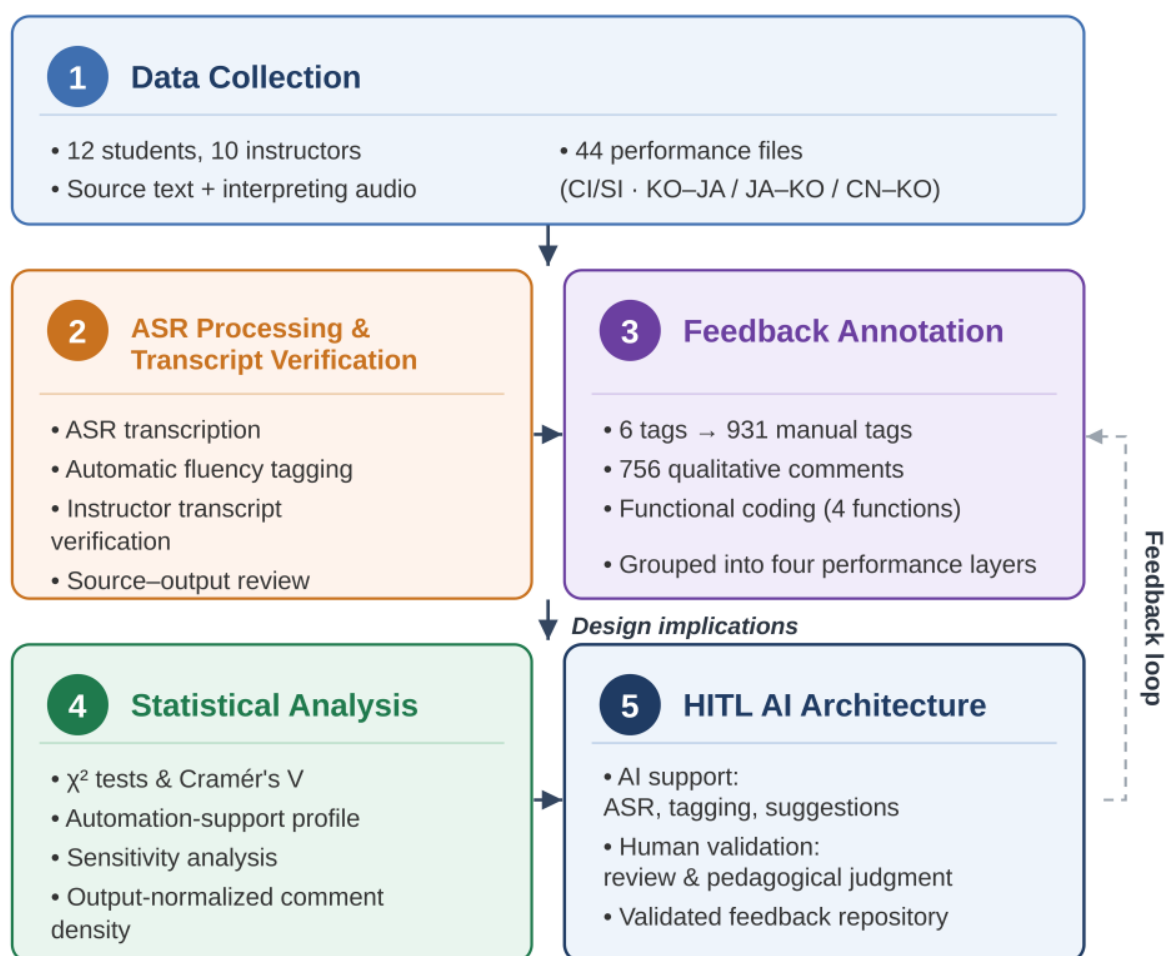


Figure 3. Overall methodological workflow of the study.

To better suit the analysis to AI/e-learning system design, three quantitative procedures were employed. The principal quantitative measures used in these analyses are defined below for methodological transparency and to facilitate replication. Feedback frequency (n) denotes the absolute number of observations in each feedback category and serves as the basis for descriptive frequency analyses. First, descriptive frequency analyses and automation-support profiling were performed to identify the distribution of feedback tags and classify them by their potential for AI-assisted or human-dominant review. The automation-support profile was a design-oriented heuristic, rather than an empirical

measure of automation accuracy, based upon the informational dependency of each feedback type; namely, whether reliable feedback could be generated primarily from the target-language output, required source–output comparison, or depended on acoustic–prosodic features of the speech signal. Second, Pearson’s chi-square tests of independence were employed to examine the association between substantive comment provision and tag type, performance layer, and instructor, given the categorical nature of these variables. For statistically significant associations, Cramér’s V was calculated as an effect-size measure to quantify the strength of the observed associations. Effect sizes were interpreted following Cohen [37]:

$$\chi^2 = \sum \frac{(O - E)^2}{E}, \quad (2)$$

where O and E denote the observed and expected frequencies, respectively; the expected frequencies were calculated under the assumption of independence between the row and column variables.

Cramér’s V was calculated as an effect-size measure:

$$V = \sqrt{\frac{\chi^2}{N(k-1)}}, \quad (3)$$

where k denotes the smaller of the number of rows or columns of the contingency table.

Third, sensitivity and output-normalized density analyses were conducted to strengthen interpretation. The sensitivity analysis compared the full dataset with a dataset excluding P3, whose tag-centered feedback pattern strongly affected the “Others” category. Output-normalized comment density (comments per 1000 whitespace-excluded characters) was calculated using Equation (4):

$$D = \frac{C}{L} \times 1000, \quad (4)$$

where D denotes the comment density, C is the number of function-coded comments, and L is the total number of whitespace-excluded characters in the interpreted output.

Statistical significance was evaluated using Pearson’s chi-square test ($p < 0.05$). For significant associations, Cramér’s V was calculated as an effect-size measure. Following Cohen [37], values of approximately 0.10, 0.30, and 0.50 were interpreted as the thresholds for small, medium, and large effect sizes; see also Agresti [38] for the interpretation of the contingency table analyses.

Each qualitative comment was treated as the unit of analysis and assigned to one educational function according to its primary instructional purpose. Coding was performed following predefined rules based upon explicit category definitions. Following the initial coding, all comments were systematically reviewed against the coding criteria, and comments that did not fully match the predefined definitions were recoded. Throughout the process, the associated feedback tag, source segment, transcript, and interpreting context were reviewed to ensure consistent application of the coding criteria. Table 2 provides representative examples of the application in practice of the coding criteria.

Each comment was reviewed with respect to its associated tag, source segment, transcript, and performance context. As this is an exploratory study, the analysis emphasized transparency, traceability, and the consistent application of the coding criteria rather than generalizable inter-coder agreement. Accordingly, the functional coding was intended to serve as an analytical framework for interpreting instructor feedback patterns rather than to establish a new annotation standard. Independent multi-coder validation should therefore be considered in future studies involving larger multilingual datasets. The analyses were intended to identify annotation patterns relevant to AI-supported feedback system

design rather than to make causal claims about learning outcomes. Because multiple feedback tags were nested within instructors and performance files, the chi-square analyses are interpreted descriptively as indicators of structured variation in instructor feedback behavior rather than as evidence of causal relationships.

Table 2. Representative examples of educational function coding.

Educational Function	Representative Example (Original)	English Translation
Diagnostic	해당 문장의 주어인 내각부'가 생략되었습니다.	The subject “Cabinet Office” has been omitted from this sentence.
Corrective	日本のニュース → 日本で報道されているニュース	Replace “Japanese news” with “news reported in Japan.”
Reflective	원문에서 말하는 '정조준하다'의 의미를 다시 한번 생각해 보기를 바랍니다.	Please reconsider the intended meaning of “take precise aim” in the source text.
Positive	원문을 충분히 이해하여 문장을 재구성하고 역접의 의미를 가진 접속사를 적절하게 활용한 점은 매우 좋습니다.	You demonstrated a good understanding of the source text by restructuring the sentence appropriately and effectively using a contrastive conjunction.

4. Results

4.1. Feedback Tag Distribution and Automation Support Profile

A total of 931 instructor-generated manual feedback tags were identified across the 44 performance files. Expression was the most frequent tag, followed by translation error, omission, grammar error, intonation, and others (Table 3). Of these tags, 756 (81.2%) contained substantive qualitative comments. The high comment rate indicates that the system generated both categorical annotation data and a substantial corpus of instructor-authored pedagogical feedback comments.

Table 3. Distribution of instructor-generated feedback tags and comment rates.

Feedback Tag	Definition	Freq.	%	Comments	Comment Rate
Expression	Naturalness of the target language, lexical choice, and style	376	40.4%	313	83.2%
Translation error	Inaccurate or distorted meaning	285	30.6%	244	85.6%
Omission	Omission of source-text information	147	15.8%	124	84.4%
Grammar error	Morphosyntactic error in the target language	46	4.9%	25	54.3%
Intonation	Feedback on accent and pronunciation, including vowel length	27	2.9%	26	96.3%
Subtotal		881	94.6%	732	83.1%
Others	Category not covered by the above tags	50	5.4%	24	48.0%
Total		931	100.0%	756	81.2%

To develop an automation-support profile for AI system design, the six instructor feedback tags were reclassified by the type of information necessary for reliable feedback generation based upon three criteria: whether the feedback can be generated primarily from the target-language output, whether it requires comparison between the source text and the interpreted output, and whether it depends on acoustic–prosodic features of the speech signal. This classification was intentionally developed to reflect the current capabilities of AI-assisted language learning technologies rather than theoretical task complexity, and the automation-support profile was designed according to the type of information required for reliable feedback generation. Specifically, feedback was classified based on whether it can be evaluated primarily from the interpreted output itself, requires explicit comparison between the source text and the interpreted output, or depends primarily on acoustic–prosodic information. Accordingly, the proposed framework distinguishes AI-supported functions according to the information required for reliable automation rather than the pedagogical importance of the feedback category.

Translation errors and omissions were classified as human-dominant source–output comparison because they require an explicit comparison between the source text and the interpreted output. Recent LLM-based systems can assist in identifying candidate omissions, mistranslations, numerical inconsistencies, and proper-noun mismatches, but the final evaluation of semantic equivalence and contextual meaning transfer requires instructor validation [39–41]. In the proposed TalkTrack framework, these AI-supported analyses serve as preliminary annotations, while instructors make the final pedagogical judgment and determine the feedback provided to learners. Expression and grammar errors were classified as candidates for LLM-assisted target-language support because they can be evaluated primarily from the interpreted output without explicit reference to the source text [42,43]. This reflects the functionality already implemented in contemporary AI-assisted writing feedback systems, where grammatical correctness, spelling, and many aspects of linguistic well-formedness are automatically detected, while instructors remain responsible for validating stylistic appropriateness and pedagogical relevance.

Intonation was classified as a candidate for acoustic–prosodic AI support because it depends on acoustic features like pitch movement, speech rate, rhythm, and intensity rather than on text alone [44,45].

The “Others” category was treated as a taxonomy-management constraint to handle feedback that cannot be reliably assigned to predefined categories or that reflects ASR-related ambiguity. In future AI-assisted workflows, such instances may be temporarily assigned to the residual category for subsequent instructor review and taxonomy refinement.

Based on this classification, meaning-transfer feedback, consisting of translation error and omission, accounted for 432 tags (46.4%) and was dominant in the human judgment domain because it requires comparison between the source text and the interpreted output. Target-language formulation, consisting of expression and grammar errors, accounted for 422 tags (45.3%) and is a candidate for LLM-assisted suggestion under instructor review. Delivery/prosody accounted for only 27 tags (2.9%), likely reflecting the content-oriented nature of instructor feedback in interpreter education, where semantic accuracy and meaning transfer are typically prioritized over pronunciation or prosodic features. Nevertheless, individualized feedback on pronunciation, stress, rhythm, and intonation is highly labor-intensive because instructors must repeatedly listen to and evaluate the fine-grained acoustic features of learners’ recordings. Given that automatic pronunciation assessment and computer-assisted pronunciation training systems have already demonstrated reliable AI-supported analysis of these speech characteristics [44,45], delivery/prosody represents a promising area for automation support within a human-in-the-loop feedback framework.

The Others category (50 tags, 5.4%) is discussed separately because it primarily concerns taxonomy and system-handling issues rather than interpreter performance itself (Table 4).

The automation-support profile should not be interpreted as indicating a hierarchy of full automation. Rather, it serves to identify where AI can support prestructuring, suggestion generation, or measurement while preserving instructor validation as the final decision layer.

Table 4. Automation-support profile of feedback tags.

Automation-Support Category	Included Tags	Freq.	%	AI Design Implications
Human-dominant source–output comparison	Translation error, omission	432	46.4%	Requires source–output comparison and expert review
LLM-assisted target-language support	Expression, grammar error	422	45.3%	Candidate for AI-generated suggestions with instructor validation
Acoustic–prosodic AI support	Intonation	27	2.9%	Potential support through acoustic and prosodic analysis
Taxonomy and system-handling constraint	Others	50	5.4%	Requires clearer tag design, subcategories, or ASR-error handling
Total		931	100.0%	

4.2. Comment Provision as a Statistical Annotation Pattern

The likelihood of Comment provision varied significantly by tag type. A chi-square test (Table 5) showed a significant association between tag type and Comment presence, $\chi^2(5, N = 931) = 67.49, p < 0.001$, Cramér’s $V = 0.269$. At the performance-layer level, the association between layer and Comment provision was also significant, $\chi^2(3, N = 931) = 44.97, p < 0.001$, Cramér’s $V = 0.220$. These results indicate that comment generation is not random but structurally linked to the type of feedback being assigned.

Table 5. Summary of categorical association tests based on Comment presence.

Analysis	χ^2	df	N	p -Value	Cramér’s V	Interpretation
Tag type $\times$ Comment presence	67.49	5	931	<0.001	0.269	Moderate association; Comment provision differs by tag type
Performance layer $\times$ Comment presence	44.97	3	931	<0.001	0.220	Layer-level differences in the comment provision
Instructor $\times$ Comment presence	389.67	9	931	<0.001	0.647	Strong instructor-style effect
Instructor $\times$ Comment presence, excluding P3	28.81	8	833	<0.001	0.186	Instructor effect remains after the sensitivity check.

Instructor identity had an even stronger association with Comment provision. The instructor $\times$ Comment-presence association was significant, $\chi^2(9, N = 931) = 389.67, p < 0.001$, Cramér’s $V = 0.647$. This large effect was driven partly by P3, who generated 98 manual tags but only 9 comments. However, even after excluding P3, instructor effects remained significant, $\chi^2(8, N = 833) = 28.81, p < 0.001$, Cramér’s $V = 0.186$. This shows that instructor feedback style remains an important system-design variable even when the most deviant case is removed. Detailed instructor-level counts are provided in Appendix B.

The expected frequencies for each chi-square test were calculated under the assumption that Comment presence and the corresponding categorical variable were independent. The resulting statistics therefore test whether the observed substantive-comment-present and nonsubstantive/tag-only counts deviate from what would be expected by chance. Cramér’s V was reported because the large number of tags could lead to small differences being statistically significant; the effect-size measure indicates the strength of the association and shows that instructor identity underlay the strongest variation in Comment provision.

Instructor identity showed the strongest association with comment provision, and this pattern was not attributable to a single outlier; as reported above, the association held even after P3 was excluded. This indicates that instructor-level variation cannot be explained by one atypical case alone. Reporting results both for the full dataset and with P3 excluded therefore guards against inflating the instructor effect and the apparent lack of explicability in the “Others” category.

Taken together, these results show that Comment provision should be treated as a structured annotation behavior rather than a random by-product of instructor feedback. This finding is important for human-in-the-loop AI feedback design because it indicates that the feedback categories do not require the same level or type of human intervention. Categories with high Comment demand require stronger support for comment drafting, retrieval, and instructor validation, whereas categories with low or inconsistent Comment provision require clearer interface guidance. The significant instructor effect further suggests that instructor style is not merely a reflection of noise in the data but a system-design variable that should be accommodated through configurable feedback workflows and minimum explicability constraints.

4.3. Residual Category Audit: The “Others” Tag as a System-Design Signal

The analysis showed that the “Others” tag functioned not as a single category but as a residual category that reflected a variety of annotation decisions. All 50 instances of the “Others” tag were subjected to content analysis to identify where the current tag system failed to accommodate instructors’ actual annotation decisions, revealing that the “Others” category did not represent a single residual phenomenon, but rather ASR transcription ambiguity, delivery-related problems, unnecessary additions, cases that could have been coded under existing tags, and discourse-structural feedback (Appendix C, Table A3).

Because one instructor accounted for a large share of the “Others” tags, Table 6 presents the statistics for the full dataset, P3 only, and the dataset excluding P3 separately to determine whether the low comment rate for the “Others” tag reflected a general problem with the residual category or was mainly driven by one instructor’s tag-centered annotation strategy.

Table 6. Comment rate of the “Others” tag with and without P3.

Dataset	Manual Tags	Comments	Comment Rate	“Others” Tags	Comments on Others	“Others” Comment Rate
Full dataset	931	756	81.20%	50	24	48.00%
P3 only	98	9	9.20%	26	3	11.50%
Excluding P3	833	747	89.70%	24	21	87.50%

These findings show that the “Others” category is not merely a minor taxonomic issue, for in a human-in-the-loop AI feedback system, they mark points at which the platform fails to provide instructors with an adequate decision structure. If ASR errors, delivery-related problems, discourse-level issues, and unclassifiable cases are absorbed into “Others,” information is lost that could otherwise support AI-assisted detec-

tion, instructor review, and learner-facing feedback. The audit therefore supports separating ASR/transcription issues from learner-performance feedback, adding pronunciation/delivery and discourse-level categories, and requiring a subcategory or explanatory comment when a residual tag is selected.

4.4. Functional Modeling of Instructor Comments

Qualitative analysis of the 756 instructor comments identified four educational functions: diagnostic, corrective, reflective, and positive feedback. Corrective feedback was dominant, accounting for 591 comments. Diagnostic feedback accounted for 96 comments, reflective feedback for 66 comments, and positive feedback for 3 comments (Table 7).

Table 7. Educational functions of instructor comments.

Function	Definition	Representative Example	Count	%
Diagnostic	Identifies only the error type or problematic aspect without specifying a correction or revision direction	"Does not match the source content."	96	12.7%
Corrective	Provides information that directly enables revision, including corrected forms, alternative expressions, specific omitted source elements, or explicit revision directions	"The omitted source element is 'in the field of security.'"	591	78.2%
Reflective	Prompts learners to reconsider source meaning, expression choice, tense, or discourse structure without directly specifying the corrected form	"Please reconsider the source meaning."	66	8.7%
Positive feedback	Acknowledges a well-handled part or confirms a strength, often combined with a suggestion for improvement	"Good interpretation, but pronunciation needs attention."	3	0.4%
Total			756	100.0%

Corrective feedback is the most important category for AI-assisted design because it requires the instructor not only to detect a problem but also to generate revision-oriented information, including alternative expressions, explanations of meaning distortions, the specification of omitted source elements, and guidance on discourse structure. The dominance of corrective comments indicates that an AI-assisted platform should prioritize suggestion, retrieval, and template support for recurring corrective patterns while preserving instructor validation.

Positive feedback appeared in only three cases. These comments were not purely praise-oriented, but rather combined acknowledgment of a well-handled element with an additional suggestion for improvement, indicating that the current taxonomy is mainly error-oriented and does not sufficiently structure strengths-based or motivational feedback. Because effective feedback can support both error correction and learner self-monitoring [46,47], positive or motivational feedback should be indicated more explicitly.

4.5. Human Feedback Load and Automation-Related Relevance Metrics

To translate the annotation results into design requirements, the study calculated workload and automation-relevance metrics. The manual tag and comment densities were 21.2 and 17.2 comments per file, respectively. The corrective feedback load was 13.4 corrective comments per file (591 corrective comments across 44 files; Table 7), meaning that the most revision-oriented and judgment-intensive function accounted for the largest share of comment work. Because file-level comment counts may be influenced by output length,

these metrics were interpreted together with the output-normalized densities reported in Appendix A. The highest density was observed in simultaneous Chinese–Korean interpreting (15.18 comments per 1000 output characters), indicating that its high comment volume was not solely attributable to longer output length. The AI-assistable target-formulation rate was 45.3%, whereas the human-dominant source-comparison rate was 46.4%. The near-equality of these figures shows that AI support should be hybrid: LLMs may help generate target-language suggestions, but source fidelity and meaning transfer require human-in-the-loop review.

5. Discussion

5.1. What Can Be Automated and What Should Remain Human-in-the-Loop?

The primary implication of these findings is that AI support in interpreter training must rely upon an analysis of what instructors actually do when providing feedback. The present data show where feedback work is repetitive and partially prestructurable and where it remains dependent on expert source–output comparison.

Instructor feedback in interpreter training is not uniform: some components are suitable for AI-assisted prestructuring, whereas others remain human-dominant. Target-language formulation, which accounted for 45.3% of tags, is the most plausible area for LLM-assisted suggestions. Instructors could receive candidate expressions, reformulations, or grammar explanations, but these should be treated as suggestions rather than final feedback, because the appropriate formulation depends on source meaning, discourse function, and interpreting context.

By contrast, meaning-transfer feedback accounted for 46.4% of tags and required a comparison between source content and interpreted output. Full automation of this category is risky because translation errors and omissions involve semantic alignment, pragmatic judgment, and contextual interpretation. AI may assist by highlighting possible source–output mismatches, but the instructor must make the final decision. This supports a human-in-the-loop approach rather than a fully automated one.

Partial automation is also technically feasible for delivery-related feedback, including speech rate, pause distribution, pronunciation, and intonation. In the current version of TalkTrack, fillers and cancelations are detected mainly from STT-derived text patterns, and pauses are detected from the audio signal when silence exceeds a predefined threshold of one second. A realistic next step is to add an acoustic–prosodic analysis module and an audio–text alignment layer that extract speech rate, articulation rate, pause ratio, fundamental frequency (F0) movement, pitch range, and intensity variation from the audio signal. These measurements would capture prosodic features related to rhythm, emphasis, and intonation, and would be linked to specific transcript segments. Importantly, such outputs should not be treated as final judgments of delivery or intonation errors but presented as candidate flags for instructor review. In this way, AI can mechanize the measurement of delivery-related features, while instructors remain responsible for determining whether the flagged patterns are pedagogically relevant in relation to the source text, the interpreted output, and the interpreting task.

5.2. Instructor Variation as a System Variable

The instructor-level analysis shows that feedback behavior is strongly associated with instructor identity. The chi-square effect remained significant even after P3 was excluded, indicating that instructor style is not simply noise but a real variable in human annotation systems. The “Others”-tag sensitivity analysis further clarifies this point: the low overall comment rate for the residual category was largely driven by P3, whereas most other instructors attached comments when selecting “Others.” In a human-in-the-loop AI plat-

form, provision should be made for this variation rather than treating it as a problem to be eliminated.

The system should therefore support multiple feedback strategies while constraining uninformative inputs. Tag-centered feedback can be pedagogically useful when learners understand the tag meanings, but without comment or indication of a subcategory, an “Others” tag has limited instructional value. Platform design should preserve instructor autonomy while ensuring minimum explicability through subcategories, required comments for residual tags, and examples embedded in the interface.

5.3. Design Requirements for AI-Assisted Feedback Systems

Based upon these findings, this study proposes a human-in-the-loop AI feedback architecture for professional e-learning (Figure 4), in which it translates instructor annotation patterns into an implementation pathway rather than a list of isolated functions. This architecture may be fruitfully applied in other professional domains where automated feedback must be combined with expert judgment.

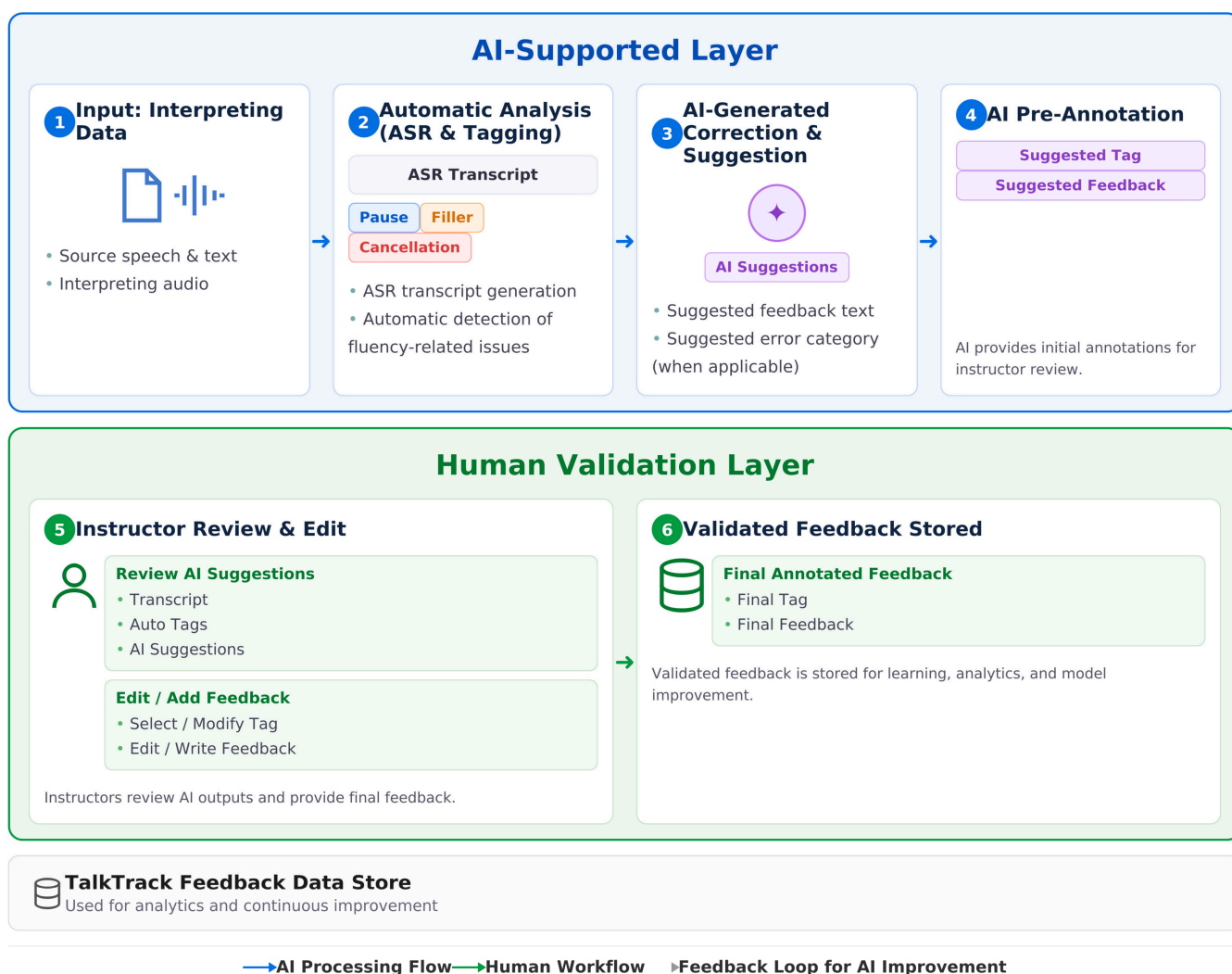


Figure 4. Human-in-the-loop AI feedback architecture in TalkTrack.

As shown in Figure 4, the proposed architecture separates AI-supported prestructuring from human-validated feedback. The automated layer processes speech input, evaluates ASR-based fluency, generates transcripts, performs auto-tagging, and offers candidate suggestions for grammar, expression, and pronunciation/intonation. The human valida-

tion layer includes source–output review, pedagogical judgment, instructor-authored comment writing, the revision or rejection of AI-generated suggestions, feedback memory, and learner–response loops. In this structure, AI supports detection, suggestion, retrieval, and organization, while instructors remain responsible for evaluating source–text fidelity and discourse coherence, rendering pedagogical judgment, and delivering the final feedback.

This architecture extends the current TalkTrack workflow by adding AI-supported suggestions and review functions while preserving instructor validation as the final decision layer. Within this architecture, the acoustic–prosodic measurements discussed above would be incorporated as candidate flags for instructor review rather than as final judgments, ensuring that pedagogical interpretation remains within the human-in-the-loop process.

Another extension of the proposed TalkTrack framework is the incorporation of an AI-generated correction layer. After identifying potential translation, expression, grammar, or delivery-related issues, the system may generate candidate corrections or alternative formulations using large language models to support, rather than replace, instructor feedback. Instructors would review, revise, approve, or reject each suggestion before it is presented to learners, and the validated feedback would then be accumulated in the TalkTrack repository. Through these repeated cycles of AI suggestion and instructor validation, the framework defines a human-in-the-loop improvement cycle in which validated instructor feedback serves as a high-quality resource for refining AI-assisted feedback functions, rather than relying on continuous model retraining, while keeping pedagogical control with instructors.

It should be noted that the current implementation of TalkTrack employs ASR transcription as the basis for transcript generation and instructor feedback. The LLM-assisted feedback functions and acoustic–prosodic analysis described in the proposed architecture are conceptual extensions for future implementation rather than components of the current system.

6. Conclusions

This study reframed the provision of instructor feedback in interpreter training as a human-in-the-loop AI feedback problem. Using 44 interpreting performance files, 931 manual tags, and 756 substantive qualitative comments, this study showed that comment provision differed significantly by tag type, performance layer, and instructor. Output-length normalization showed that high comment volume in simultaneous Chinese–Korean interpreting remained visible even after adjusting for interpreted output length. The “Others”-tag analysis further showed that residual categories must be interpreted with sensitivity to instructor behavior, because P3’s tag-centered strategy strongly affected the full-sample comment rate of the overall dataset. The dominance of corrective feedback suggests that this function represents the most revision-oriented and judgment-intensive component of instructor feedback. The automation-support profile further showed that almost half of the feedback tags were candidates for AI-assisted target-language support, whereas almost half required human-dominant source–output comparison. These findings position instructor-generated feedback data as an essential empirical layer between conventional teacher feedback and future AI-generated feedback. In professional training contexts, AI feedback systems should first be trained to support instructor judgment rather than to replace it.

This study contributes to AI and e-learning research by showing how actual instructor annotations can be transformed into system-design requirements. The findings also offer empirical design guidance for human-in-the-loop AI feedback systems in professional education, particularly in domains where expert source–output comparison and pedagogical

ical judgment cannot be fully automated. As AI models continue to improve, future feedback systems may provide increasingly detailed and pedagogically useful feedback with reduced human intervention. However, in professional training domains, the appropriate goal is not to fully replace instructor judgment, but to support adaptive human oversight, in which instructor intervention is adjusted according to feedback type, task complexity, and system reliability. Accordingly, the proposed framework is intended to reduce repetitive annotation workload while preserving instructor authority over pedagogical interpretation and final feedback decisions. The findings of this study directly address the research questions posed in the Introduction as follows:

RQ1. What characteristics distinguish instructor-generated feedback in interpreter training? The instructor feedback was highly uneven across feedback tags, performance layers, and instructors. Corrective feedback predominated, while pronunciation-related and positive feedback was comparatively rare. These patterns reflect the pedagogical priorities of professional interpreter training and reveal the labor-intensive nature of content-oriented feedback.

RQ2. Which components of instructor feedback are suitable for AI assistance, and which continue to require human judgment? The proposed automation-support profile showed that target-language formulation and acoustic-prosodic features represent promising candidates for AI-assisted support, whereas meaning-transfer feedback requiring explicit source–target comparison remains dependent on expert human judgment. This distinction provides a practical design principle for future AI-assisted interpreter feedback systems.

RQ3. How can authentic instructor feedback inform the design of a human-in-the-loop AI feedback framework? The analysis demonstrated that the design requirements for AI-assisted feedback systems can be systematically derived from authentic instructor-generated feedback. Rather than replacing instructors, AI should provide them support in annotation, suggestion generation, organization, and retrieval, while instructors remain responsible for pedagogical interpretation, source-text fidelity, and final feedback decisions. This human-in-the-loop architecture provides a practical foundation for future AI-assisted interpreter education.

RQ4. How do ASR transcription constraints and residual feedback categories reveal gaps in platform taxonomy and human–AI workflow? The analyses showed that ASR transcription constraints and the residual “Others” category provide valuable diagnostic information for improving the TalkTrack platform rather than indicating isolated technical limitations. ASR-related constraints highlight the need for language-specific transcription support and continued instructor verification, while the concentration of residual tags suggests opportunities to refine the feedback taxonomy and improve the interface design. Together, these findings demonstrate that human-generated annotations can serve not only as learner feedback but also as empirical evidence for iterative system refinement within a human-in-the-loop framework.

RQ5. What human-in-the-loop AI design requirements can be derived for scalable feedback systems in professional e-learning?

The findings demonstrate that scalable AI-assisted feedback systems should be designed as collaborative human-in-the-loop environments rather than fully automated assessment systems. AI can efficiently support transcript generation, feedback detection, categorization, retrieval, and the provision of candidate suggestions, whereas instructors remain responsible for source–output meaning comparison, pedagogical interpretation, and final feedback decisions. The proposed framework provides practical design requirements for future AI-assisted feedback systems, including adaptive human oversight, validated feedback repositories, taxonomy refinement, multilingual ASR integration,

strengths-based feedback, and continuous system improvement based upon authentic instructor feedback.

7. Limitations

This study has several limitations that should be noted. It was conducted in one institutional interpreter-training context and involved interpretation in three language directions, and the functional classification was conducted as an interpretive analysis rather than through multi-coder validation. In particular, the dataset did not include a Korean-to-Chinese (KO-CN) interpreting direction, and the pool of native-speaking instructors and students was not evenly balanced across Korean, Japanese, and Chinese backgrounds; generalizability of the present findings to other language pairs and rater populations should therefore be interpreted with caution. All data were also collected from a single institution within an Asian-language interpreter-training context, so the findings should be read as a design-oriented case study rather than a cross-institutionally or cross-regionally validated model. The study also analyzes existing instructor-generated feedback data rather than experimentally evaluating AI-assisted feedback workflows or AI-generated comment suggestions.

A further consideration is the categorical imbalance in the instructor feedback dataset, as expression-related feedback occurred substantially more frequently than categories such as grammar or intonation. Although this imbalance does not affect the descriptive analyses reported in the present study, it may influence the performance of future AI models trained on instructor annotations by biasing predictions toward dominant feedback categories. However, both this imbalance and the predominance of corrective feedback reflect authentic instructor practice rather than a sampling artifact or a recommended pedagogical model.

Finally, the present study did not quantitatively evaluate the influence of ASR transcription accuracy on subsequent feedback annotation. Because instructors simultaneously consulted the source text, ASR transcript, and audio recordings, the final feedback was not solely dependent on transcription quality.

This study also does not directly examine the ethical implications of AI-assisted feedback systems. Because the proposed architecture is designed to support rather than replace instructor judgment, the immediate risk of instructor replacement is considered low; future work should nonetheless examine safeguards against overreliance on AI-generated suggestions and the potential deskilling of instructors.

8. Future Work

Future work should implement the proposed design requirements in TalkTrack and compare instructor-only and AI-assisted feedback workflows while measuring changes in instructor time and learner uptake. Where comparable systems are available, this evaluation should also benchmark the proposed framework's effectiveness against existing automated or AI-assisted feedback tools used in translator and interpreter training. In addition, the TalkTrack platform will soon be extended to support additional target languages through the integration of multilingual interpreter-oriented ASR models. Future data collection should also incorporate a Korean-to-Chinese (KO-CN) interpreting direction and a more balanced recruitment of native-speaking instructors and students across Korean, Japanese, and Chinese backgrounds, to improve the linguistic diversity and generalizability of the proposed framework. Future work should also extend this validation to interpreter-training programs in other institutional and regional contexts, including European translation and interpreting programs and additional language combinations, to assess the cross-institutional generalizability of the proposed framework. This

extension would enable automatic transcription and annotation for a broader range of interpreting directions, facilitate the construction of larger and more linguistically diverse datasets, and improve the scalability and generalizability of the proposed human-in-the-loop feedback framework.

To address the categorical imbalance noted above, future development should consider targeted data collection for underrepresented categories, data augmentation, class-weighted learning, and continued instructor validation within a human-in-the-loop framework to ensure reliable feedback across all pedagogically important categories while preserving the ecological validity of authentic instructor behavior.

Future versions of TalkTrack may also incorporate AI-assisted detection of interpreting strengths, such as effective terminology choices, fluent delivery, or well-structured target-language production, to generate candidate positive feedback for instructor review. Future interface design should also prompt instructors to specify a reason when selecting the “Others” tag, or to reclassify the feedback into a more specific category where possible, supporting continued refinement of the feedback taxonomy.

Funding: This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2021S1A5A2A03062819).

Institutional Review Board Statement: The study was conducted following the Declaration of Helsinki and approved by the Korean Institutional Review Board (ewha-202104-0019-06) on 21 January 2024.

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

Data Availability Statement: The dataset analyzed in this study was collected through TalkTrack, an LMS-based digital feedback platform developed by the research team to support computer-aided interpreter training (CAIT). The dataset consists of Korean–Japanese, Japanese–Korean, and Chinese–Korean interpreting performances, ASR transcripts, manually assigned instructor feedback tags, and qualitative feedback comments provided through the TalkTrack interface. The dataset generated and analyzed during the current study is publicly available at https://drive.google.com/drive/folders/1rv9rsz604n96zuvKfY830cesVGyAE8t1?usp=drive_link (accessed on 31 May 2026).

Acknowledgments: During the preparation of this manuscript, the author used OpenAI’s ChatGPT (GPT-5.5 Thinking) and Anthropic’s Claude (Sonnet 4.6) for language editing, English translation support, manuscript restructuring, table formatting, consistency checking of statistical descriptions, and improving the resolution of figures. The author reviewed and edited the output and takes full responsibility for the content of this publication.

Conflicts of Interest: The author declares no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
ASR	Automatic Speech Recognition
CAIT	Computer-Assisted Interpreter Training
CI	Consecutive Interpreting
CN-KO	Chinese-to-Korean
F0	Fundamental Frequency
JA-KO	Japanese-to-Korean
KO-JA	Korean-to-Japanese
LLM	Large Language Model
SI	Simultaneous Interpreting

Appendix A.

Table A1. Source/output length and comment density.

Interpreting Type	Source Chars	Output Files	Mean Output Chars	Function-Coded Comments	Comments/1000 Output Chars
CI KO-JA	798	7	1211.4	68	8.02
CI JA-KO	819	7	908.3	73	11.48
CI CN-KO	791	12	1197.4	160	11.14
SI KO-JA	1345	3	1447.3	27	6.22
SI JA-KO	1452	3	1528.3	62	13.52
SI CN-KO	1490	12	2009.2	366	15.18

Appendix B.

Table A2. Instructor-level feedback behavior.

Instructor	Files	Manual Tags	Comments	Comment Rate	Tags per File	Comments per File	Others Tags
P1	7	119	98	82.4%	17.0	14.0	1
P2	7	97	87	89.7%	13.9	12.4	0
P3	3	98	9	9.2%	32.7	3.0	26
P4	3	48	36	75.0%	16.0	12.0	1
P5	4	95	94	98.9%	23.8	23.5	2
P6	2	69	63	91.3%	34.5	31.5	4
P7	4	33	31	93.9%	8.2	7.8	0
P8	4	49	45	91.8%	12.2	11.2	0
P9	4	82	76	92.7%	20.5	19.0	8
P10	6	241	217	90.0%	40.2	36.2	8
Total	44	931	756	81.2%	21.2	17.2	50

Appendix C. Detailed Qualitative Patterns of the “Others” Tag

Table A3. Detailed qualitative patterns of the “Others” tag.

Pattern	Evidence in the Dataset	Interpretation	Proposed Treatment
Japanese ASR kanji conversion errors	Mainly P3 in the Japanese–Korean dataset	Transcription uncertainty rather than learner-performance error	Separate ASR/transcription issue from instructor feedback tags
Pronunciation, speed, or delivery problems	Pronunciation, articulation, speed, and nonverbal utterance comments	Delivery-related issues not fully captured by the intonation tag	Add or integrate Pronunciation/Delivery
Unnecessary addition or nonsource-based content	Added information, unnecessary supplementation, or content not in the source	Problem of interpreting economy and source fidelity	Add unnecessary addition or clarify the relation to the translation error
Cases classifiable under existing tags	Meaning distortion, expression, or omission cases marked as others	Inconsistent tag application	Provide clearer tag definitions and examples
Sentence structure or argument coherence	Comments on sentence continuation, logical connection, or discourse coherence	Discourse-level problems not sufficiently captured by the current taxonomy	Add Sentence structure/argument coherence or allow multi-labeling
Unspecified others without comment	Others selected without sufficient explanation	Pedagogical intention is unclear	Require a comment or subcategory when selected

References

1. Cavalcanti, A.P.; Barbosa, A.; Carvalho, R.; Freitas, F.; Tsai, Y.-S.; Gašević, D.; Mello, R.F. Automatic feedback in online learning environments: A systematic literature review. *Comput. Educ. Artif. Intell.* **2021**, *2*, 100027. [CrossRef]
2. Hahn, M.G.; Navarro, S.M.B.; Valentin, L.D.L.F.; Burgos, D. A systematic review of the effects of automatic scoring and automatic feedback in educational settings. *IEEE Access* **2021**, *9*, 108190–108198. [CrossRef]
3. Alghamdi, L.H.; Alghizzi, T.M. Educators' reflections on AI-automated feedback in higher education: A structured integrative review of potentials, pitfalls, and ethical dimensions. *Front. Educ.* **2025**, *10*, 1704820. [CrossRef]
4. Buckingham Shum, S.; Lim, L.A.; Boud, D.; Bearman, M.; Dawson, P. A comparative analysis of the skilled use of automated feedback tools through the lens of teacher feedback literacy. *Int. J. Educ. Technol. High. Educ.* **2023**, *20*, 40. [CrossRef]
5. Han, C. Interpreting testing and assessment: A state-of-the-art review. *Lang. Test.* **2021**, *39*, 30–55. [CrossRef]
6. Gile, D. *Basic Concepts and Models for Interpreter and Translator Training*, Revised ed.; John Benjamins: Amsterdam, The Netherlands, 2009. [CrossRef]
7. Lee, J.; Kim, J.-D.; Park, H.S.; Park, H.K.; Son, J.-B.; Oh, U.; Sang, W.; Kim, S.; Lim, J.; Jo, H.; et al. Development of an LMS-based automatic annotation authoring tool for computer-assisted interpreter training (CAIT). *Interpret. Transl. Stud.* **2025**, *29*, 79–113. [CrossRef]
8. Sang, W.; Lee, J.R.A.; Kim, S.; Park, H.K.; Son, J.-B. A study on the feasibility of a hybrid feedback platform utilizing AI in interpretation classes. *Interpret. Transl. Stud.* **2025**, *29*, 79–117. [CrossRef]
9. Kang, H.; Kim, J.D.; Lee, J.; Nam, H.J.; Kim, K.W.; Lim, J.; Park, H.S. Speech-adaptive detection of unnatural intra-sentential pauses using contextual anomaly modeling for interpreter training. *Appl. Sci.* **2026**, *16*, 3492. [CrossRef]
10. Lee, J. Feedback on feedback: Guiding student interpreter performance. *Transl. Interpret.* **2018**, *10*, 152–170. [CrossRef]
11. Chen, L. Lost in translation: Decoding the errors in consecutive interpreting by Chinese EFL learners. *PLoS ONE* **2025**, *20*, e0337758. [CrossRef] [PubMed]
12. Hong, J. Feedback in interpreter training. *Interpret. Transl.* **2011**, *13*, 161–174. Available online: <https://www.kci.go.kr/kciportal/ci/sereArticleSearch/ciSereArtiView.kci?sereArticleSearchBean.artiId=ART001569006> (accessed on 3 May 2026).
13. Balaman, S. Exploring undergraduate students' viewpoints on corrective feedback implementations in interpreting. *Korkut Ata Türkiyat Araştırmaları Derg.* **2024**, *15*, 994–1011. [CrossRef]
14. Yıldırım, Ş. Feedback from peers or teachers: The preferences and perceptions of pre-service EFL teachers. *Mehmet Akif Ersoy Univ. J. Educ. Fac.* **2026**, *77*, 53–70. [CrossRef]
15. Zheng, Y.; Zhong, Q.; Yu, S.; Li, X. Examining students' responses to teacher translation feedback: Insights from the perspective of student engagement. *SAGE Open* **2020**, *10*, 2158244020932536. [CrossRef]
16. Lee, J. A discourse analytic study of consecutive interpreting pedagogy with a focus on trainers' corrective feedback. *Interpret. Transl.* **2021**, *23*, 219–244. Available online: <https://www.earticle.net/Article/A404580> (accessed on 3 May 2026). [CrossRef]
17. Kim, W. An analysis of disfluency in trainee interpreters' English-Korean consecutive interpreting. *J. Linguist. Stud.* **2018**, *23*, 27–48. Available online: <https://www.earticle.net/Article/A339011> (accessed on 3 May 2026). [CrossRef]
18. Krouglov, A. Emergency remote teaching and learning in simultaneous interpreting: Capturing experiences of teachers and students. *Train. Lang. Cult.* **2021**, *5*, 41–56. [CrossRef]
19. Cerezo Herrero, E. A critical review of listening comprehension in interpreter training: The case of Spanish translation and interpreting degrees. *Porta Linguarum* **2017**, *28*, 7–22. [CrossRef]
20. Dong, Y.; Li, Y.; Zhao, N. Acquisition of interpreting strategies by student interpreters. *Interpret. Transl. Train.* **2019**, *13*, 408–425. [CrossRef]
21. Xu, Y. Exploring individual differences in the prediction of awareness and improvement in trainee interpreters: An aptitude-treatment interaction approach. *Interpret. Transl. Train.* **2023**, *17*, 133–154. [CrossRef]
22. Holewik, K. Peer feedback and reflective practice in public service interpreter training. *Theory Pract. Second Lang. Acquis.* **2020**, *6*, 133–159. Available online: <https://www.cceol.com/search/article-detail?id=944893> (accessed on 3 May 2026). [CrossRef]
23. Song, Y. A survey of translator trainers on teaching practices in graduate-level classrooms. *Interpret. Transl.* **2016**, *18*, 47–71. Available online: <https://www.earticle.net/Article/A281211> (accessed on 3 May 2026). [CrossRef]
24. Creagh, S.; Thompson, G.; Mockler, N.; Stacey, M.; Hogan, A. Workload, work intensification and time poverty for teachers and school leaders: A systematic research synthesis. *Educ. Rev.* **2023**, *77*, 661–680. [CrossRef]
25. Deeva, G.; Bogdanova, D.; Serral, E.; Snoeck, M.; De Weerd, J. A review of automated feedback systems for learners: Classification framework, challenges and opportunities. *Comput. Educ.* **2021**, *162*, 104094. [CrossRef]
26. Wang, Z.; Han, F. The effects of teacher feedback and automated feedback on cognitive and psychological aspects of foreign language writing: A mixed-methods research. *Front. Psychol.* **2022**, *13*, 909802. [CrossRef] [PubMed]
27. Nolan, H.G.; Coenraad, M.; Young, V. *Teaching Partner, Grading Assistant, Substitute Teacher: Three Ways Teachers Positioned an Artificial Intelligence Tool in Writing Instruction*; Digital Promise: Washington, DC, USA, 2024. [CrossRef] [PubMed]

28. Huang, Z.; Te Winkel, W. Automated feedback: An AI-powered tool to scale microlevel feedback for better academic writing. In Proceedings of the EUNIS 2022 Annual Congress: Good for All in the Digital World, Göttingen, Germany, 1–3 June 2022. Available online: https://eunis.org/download/2022/EUNIS_2022_paper_19.pdf (accessed on 3 May 2026).
29. Sandrelli, A. Becoming an interpreter: The role of computer technology. *MonTI* **2015**, *7*, 111–138. [\[CrossRef\]](#)
30. Sandrelli, A.; de Manuel Jerez, J. The impact of information and communication technology on interpreter training. *Interpret. Transl. Train.* **2007**, *1*, 269–303. [\[CrossRef\]](#)
31. Wang, X.; Wang, C. Can computer-assisted interpreting tools assist interpreting? *Transletters* **2019**, *3*, 109–139. Available online: <https://journals.uco.es/tl/article/view/11575> (accessed on 3 May 2026).
32. Kuhn, K.; Kersken, V.; Reuter, B.; Egger, N.; Zimmermann, G. Measuring the accuracy of automatic speech recognition solutions. *ACM Trans. Access. Comput.* **2023**, *16*, 4. [\[CrossRef\]](#)
33. Prinos, K.; Patwari, N.; Power, C.A. Speaking of accent: A content analysis of accent misconceptions in ASR research. In Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, Rio de Janeiro, Brazil, 3–6 June 2024. [\[CrossRef\]](#)
34. Wein, S.; Te, I.; Cherry, C.; Juraska, J.; Padfield, D.; Macherey, W. Barriers to effective evaluation of simultaneous interpretation. In Proceedings of the Findings of the Association for Computational Linguistics: EACL 2024, St. Julian's, Malta, 17–22 March 2024. [\[CrossRef\]](#)
35. Tian, M. Design and performance optimization of Japanese speech recognition system based on deep learning. In Proceedings of the 2025 International Conference on Educational Technology and Artificial Intelligence, Wuhan, China, 25–27 July 2025. [\[CrossRef\]](#)
36. Karita, S.; Sproat, R.; Ishikawa, H. Lenient evaluation of Japanese speech recognition: Modeling naturally occurring spelling inconsistency. In Proceedings of the Workshop on Computation and Written Language (CAWL 2023), Toronto, ON, Canada, 14 July 2023. [\[CrossRef\]](#)
37. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed.; Routledge: New York, NY, USA, 1988. [\[CrossRef\]](#)
38. Agresti, A. *An Introduction to Categorical Data Analysis*, 3rd ed.; Wiley: Hoboken, NJ, USA, 2019.
39. Kocmi, T.; Federmann, C. Large language models are state-of-the-art evaluators of translation quality. In Proceedings of the 24th Annual Conference of the European Association for Machine Translation, Tampere, Finland, 12–15 June 2023; pp. 193–203. Available online: <https://aclanthology.org/2023.eamt-1.19/> (accessed on 3 May 2026).
40. Minder, J.; Wisniewski, G.; Kübler, N. Testing LLMs' capabilities in annotating translations based on an error typology designed for LSP translation: First experiments with ChatGPT. In Proceedings of the Machine Translation Summit XX: Volume 1, Geneva, Switzerland, 23–27 June 2025; pp. 190–203. Available online: https://arxiv.org/html/2504.15052v1?utm_source=chatgpt.com (accessed on 3 May 2026).
41. Karpinska, M.; Iyyer, M. Large language models effectively leverage document-level context for literary translation, but critical errors persist. In Proceedings of the Eighth Conference on Machine Translation, Singapore, 6–7 December 2023; pp. 419–451. [\[CrossRef\]](#)
42. Dodigovic, M.; Tovmasyan, A. Automated writing evaluation: The accuracy of Grammarly's feedback on form. *Int. J. TESOL Stud.* **2021**, *3*, 71–87. [\[CrossRef\]](#)
43. Wang, I.X.; Wu, X.; Coates, E.; Zeng, M.; Kuang, J.; Liu, S.; Park, J. Neural automated writing evaluation with corrective feedback. *arXiv* **2024**, arXiv:2402.17613. [\[CrossRef\]](#)
44. Neri, A.; Mich, O.; Gerosa, M.; Giuliani, D. The effectiveness of computer assisted pronunciation training for foreign language learning by children. *Comput. Assist. Lang. Learn.* **2008**, *21*, 393–408. [\[CrossRef\]](#)
45. Witt, S.M. Automatic error detection in pronunciation training: Where we are and where we need to go. In Proceedings of the International Symposium on Automatic Detection of Errors in Pronunciation Training (IS ADEPT), Stockholm, Sweden, 6–8 June 2012; pp. 1–8. Available online: <https://www.researchgate.net/publication/250306074> (accessed on 3 May 2026).
46. Lipnevich, A.A.; Panadero, E. A review of feedback models and theories: Descriptions, definitions, and conclusions. *Front. Educ.* **2021**, *6*, 720195. [\[CrossRef\]](#)
47. Morris, R.; Perry, T.; Wardle, L. Formative assessment and feedback for learning in higher education: A systematic review. *Rev. Educ.* **2021**, *9*, e3292. [\[CrossRef\]](#)

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.