Research Square

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

Evaluating LingualAI: A Prospective Validation of AI-Based Real- Time Translation Against Certified Human Interpreters

Uday P. Singh

uday.pratapsingh@uth.tmc.edu

The University of Texas Health Science Center at Houston

Carlos A. Jaimes Garcia

The University of Texas Health Science Center at Houston

Gabriel M. Aisenberg

The University of Texas Health Science Center at Houston

Javier Barreda Garcia

The University of Texas Health Science Center at Houston

Jessica A. Hernandez-Chilatra

The University of Texas Health Science Center at Houston

Cecilia Wang

The University of Texas Health Science Center at Houston

Dalilah Reyes Jesus

The University of Texas Health Science Center at Houston

Eileen Whalen

The University of Texas Health Science Center at Houston

Veronica Santos Canellas

The University of Texas Health Science Center at Houston

Amanda R. Falk Vargas

The University of Texas Health Science Center at Houston

Brian O. Rodriguez Echevarria

The University of Texas Health Science Center at Houston

Martin J. Citardi

The University of Texas Health Science Center at Houston

Babatope O. Fatuyi

The University of Texas Health Science Center at Houston

Xiaoqian Jiang

The University of Texas Health Science Center at Houston

Article

Keywords: Artificial intelligence, Medical translation, Clinical communication, Limited English proficiency, Interpreter services, Otorhinolaryngology

Posted Date: December 5th, 2025

DOI: https://doi.org/10.21203/rs.3.rs-8205991/v1

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

Additional Declarations: No competing interests reported.

Abstract

Background: Limited English proficiency (LEP) affects >25 million people in the United States and is linked to health disparities in safety, quality, and outcomes. While professional interpreters remain the standard, access is often constrained. Real-time AI translation systems are increasingly available, yet their clinical performance relative to certified interpreters is uncertain.

Objective: To evaluate whether an in-house AI application (LingualAI) achieves non-inferior translation quality compared with certified interpreters in English–Spanish otorhinolaryngology encounters.

Design, Setting, and Participants: Prospective, within-subject comparison using three standardized outpatient scenarios (33 lines: 18 clinicians, 15 patients) enacted by two pairs of native speakers. Each line was translated by LingualAI and by two certified medical interpreters. Nine bilingual clinicians, blinded to source but given scenario context, independently rated anonymized audio clips.

Main Measures: Twelve domains on 5-point Likert scales: primary (*terminology accuracy*, *adequacy of meaning*), secondary (*completeness*, *grammar*, *vocabulary*, *cultural appropriateness*), voice-related (*fluency*, *clarity*, *prosody*, *pacing*), and conclusive (*overall quality*, *clinical confidence*). Non-inferiority margin prespecified at 0.30 points (Δ = Human − AI). Analyses used paired tests and mixed-effects models with random intercepts for line; inter-rater reliability via Krippendorff's α.

Results: Across models, LingualAI was non-inferior for adequacy of meaning and terminology accuracy; completeness also met the criterion. Human interpreters scored higher on delivery-related and linguistic-mechanics domains, including clarity/intelligibility (Δ≈0.50), fluency (Δ≈1.1), prosody (Δ≈0.6), pacing (Δ≈0.4), grammar, vocabulary, and cultural appropriateness. Conclusive ratings favored humans for overall quality (Δ≈0.6) and clinical confidence (Δ≈0.6). Findings were consistent in direction-specific contrasts (English→Spanish clinician lines; Spanish→English patient lines). Inter-rater reliability was modest (α=0.31), reflecting first-impression scoring. In exploratory system metrics, mean end-to-end translation latency was ~9.7s with substantially lower estimated per-session costs than phone/video interpreter services.

Conclusions: LingualAI preserves core meaning and terminology at near-interpreter levels but lags in speech naturalness and delivery (*fluency*, *prosody*, *pacing*), leading to lower overall quality and clinical confidence. AI translation may serve as a useful aid when interpreters are unavailable; however, its use today should remain aligned with professional standards and ideally follow an *interpreter-in-the-loop model* rather than replacement. Continued refinement of voice and delivery features potentially will improve perceived speech naturalness and delivery and thus, in the long-run applications such as Lingual AI, will more closely approximate the performance of human interpreters on all measures. Technical work, plus clinical validation is necessary for the safe and effective deployment of applications such as Lingual AI in real-world settings (Fig. 1).

INTRODUCTION

Effective communication is fundamental to safe and equitable healthcare. In countries with substantial linguistic diversity, such as the United States, a growing proportion of clinical encounters involve patients with limited English proficiency (LEP), estimated at approximately 25 million people, or about 8% of the population aged 5 years and older who report speaking English less than “very well”. (1, 2) LEP patients face barriers to understanding diagnoses, treatment instructions, and follow-up care, contributing to disparities in safety, quality, and health outcomes. (3, 4) A recent systematic review found that when LEP patients receive language-concordant care, clinical outcomes are improved in most settings. (1) Although professional medical interpreters are the standard of care, access is often limited, particularly in primary care, rural settings, and time-sensitive encounters, leaving many patients without reliable language support. (5, 6)

Recent advances in artificial intelligence (AI) have enabled speech- and language-model–based systems capable of real-time multilingual translation. (7, 8) Several mobile or app-based tools are now available at the point of care; yet systematic evidence regarding their performance, accuracy, and clinical appropriateness remains scarce. In particular, little is known about how AI-based translations compare with certified human interpreters when assessed against clinical communication standards such as terminology accuracy, adequacy of meaning, cultural appropriateness, and speech fluency. Notably, in a recent study evaluating three commercially available applications for bidirectional clinician–patient dialogue, none were found to be suitable for safe two-way clinical communication when compared with professional interpreters. (9) This evidence gap complicates safe integration into health systems.

The LingualAI application was created to reduce language barriers in clinical settings through real-time voice translation. Developed at UTHealth Houston, its purpose it help clinical team members and patients communicate seamlessly through a continuous flow of voice capture, transcription, translation, and synthesis. Its interface is designed to be intuitive: users can select patient and clinician languages, record conversations linked to each patient, and view interactive bubbles on screen with real-time transcription and translation. In addition, the app offers translation quality feedback, conversation history, secure device synchronization, and an offline mode that ensures reliability even in low-connectivity environments. However, its effectiveness relative to certified interpreters has not been formally evaluated.

We conducted a prospective, within-subject comparative study using scripted bilingual dialogues to evaluate LingualAI’s performance in English–Spanish interactions in an outpatient otorhinolaryngology setting. We assess whether LingualAI’s audio translations are non-inferior to certified human interpreters across multiple domains of translation quality, including terminology accuracy, adequacy of meaning, clarity/fluency, cultural appropriateness, and overall confidence for clinical use. By applying a structured validation framework, this work provides early evidence on the role of AI-based translation tools in supporting equitable, multilingual healthcare delivery.

METHODS

We conducted a prospective, within-subject comparison of translations generated by the LingualAI application versus certified human interpreters (Fig. 2). Two translation directions were evaluated: English → Spanish (clinician utterances) and Spanish → English (patient utterances). Each input phrase was translated by both arms, enabling paired comparisons with evaluators blinded to translation source but not scenario context. Similar prospective, paired or randomized designs have been used in prior translation and interpreter studies that have compared Machine or AI Translation with human translators and bilingual clinical evaluators (10, 11).

Clinical scenarios and audio collection

Three standardized otorhinolaryngology scenarios were developed to represent typical care interactions. Together, the scripts included 18 dialogue paragraphs (33 lines) covering history-taking, clinical instructions, and empathic statements. Each scenario was enacted by two pairs of native English and Spanish speakers to reduce bias from voice quality or pronunciation differences. As a result, each phrase was recorded twice, yielding two independent input sets. Use of standardized scripted encounters with native speakers mirrors prior interpreter research and simulation-based translation evaluations (12, 13). The full text of the three standardized scenarios is provided in Supplementary Material SM1.

Translation pipelines

In the AI arm, recordings were processed through LingualAI, which integrates automatic speech recognition (ASR), neural machine translation (NMT), and text-to-speech synthesis.

Each phrase was submitted twice using the two independent source recordings, generating two AI-translated audio files per phrase in each direction. Outputs were pooled and randomized in blocks of 2 using a fixed random seed to ensure reproducibility, and evaluators were blinded to the translation source (AI vs. human). They were, however, provided with scenario context, line order, and speaker role (clinician vs. patient) so that ratings reflected the intended clinical meaning. Modular pipelines combining ASR, neural MT, and synthesis are well described in biomedical translation research (7, 14, 15).

LingualAI's pipeline operates in five real-time stages: (i) voice capture, ii) text conversion via ASR, (iii) refinement using prior dialogue and domain-specific vocabulary, (iv) translation into the target language, and (v) speech synthesis for natural-sounding delivery. Each stage is designed for real-time performance with fallback strategies in case of network interruptions. The most recent update includes support for custom vocabularies, allowing clinicians to register specialized or uncommon medical terms to improve transcription and translation accuracy in clinical contexts. Comparable staged architectures with domain adaptation have been reported for ambient clinical transcription and multilingual MT systems (14).

In the human arm, two certified medical interpreters independently translated the scenarios, producing two human-translated audio files per phrase. Outputs were similarly pooled and randomized to reduce interpreter-specific bias. All translations were recorded under uniform audio conditions.

Evaluators

Nine fluent bilingual clinicians (English–Spanish) served as evaluators. Participation in the evaluation was voluntary and completed outside protected clinical time. Evaluators submitted ratings anonymously; unique rater identifiers were not collected. Audio clips were randomized and presented blinded to translation source (AI vs. human), but with scenario context provided to support intended meaning. Although blinding was applied, it is possible that evaluators could have inferred AI outputs due to the use of the same synthetic voice and a slightly mechanistic tone compared with human translations. We did not conduct a calibration round, as the study was designed to capture first-impression ratings rather than consensus. All evaluators provided Consent to Participate prior to beginning the study procedures. Blinded bilingual rater designs have been applied in comparable translation and medical communication studies (10, 16).

Rating instrument

Translation quality was assessed across 12 domains: (i) primary: accuracy of meaning, terminology accuracy; (ii) secondary: completeness, cultural appropriateness, grammar, vocabulary; (iii) voice-related: fluency, clarity, prosody, pacing; and (iv) conclusive: overall quality and clinical confidence. Ratings used 5-point Likert scales (1 = poor, 5 = excellent), and evaluators were provided with a guide containing definitions and examples. Evaluators rated each clip independently using Google Forms, a secure, web-based platform containing domain definitions and illustrative examples. Similar multidomain rubrics have been applied in medical machine translation and LLM evaluation studies (10, 11). Established frameworks such as the Multidimensional Quality Metrics (MQM) cover accuracy, fluency, terminology, and adequacy, with error taxonomies also informing clarity, grammar, and overall quality (17). In line with broader recommendations, our rubric emphasized context awareness, distinguishing accuracy from fluency, and incorporating completeness, cultural appropriateness, and clinical relevance into evaluation. (18, 19). Additionally, because of this, we did not conduct a calibration round, as the study was designed to capture first-impression ratings rather than consensus. The complete rating instrument and evaluator instructions are provided in Supplementary Material SM2.

Statistical analysis

Normality of scores was assessed using the Shapiro–Wilk test. For normally distributed outcomes, paired t-tests were applied; for non-normal distributions, Wilcoxon signed-rank tests were used. Effect sizes were reported as Cohen's *d* (t-test) or correlation coefficient (*r*) (Wilcoxon). Inter-rater reliability was estimated using Krippendorff's α, which accommodates missing data and unequal numbers of raters per

item, a necessity given that evaluators submitted ratings anonymously without unique identifiers. Mixed-effects linear models with random intercepts for item (line) were applied to account for repeated ratings of the same line across multiple evaluators.

A non-inferiority margin of 0.30 points on the 5-point scale was prespecified. This margin was chosen a priori based on three considerations: (i) the half-SD rule commonly applied to Likert ratings, where typical standard deviations of ~ 0.5–0.7 correspond to a minimal meaningful difference of ~ 0.25–0.35; (ii) prior studies that used similar small margins for clinically relevant differences on 5-point rating scales (20); and (iii) expert consensus among bilingual clinician-evaluators that a 0.30 shift would begin to alter clinical confidence without changing literal meaning. For each domain, one-sided 95% confidence intervals (CIs) were calculated for the mean difference between human and AI ratings. LingualAI was considered non-inferior if the upper bound of the CI was less than the margin. All analyses were conducted in Python (Ver 3.12.5 using Pandas, Numpy, StatsModels and SciPy libraries) and R (Ver 4.4.2 using ggplot2, dplyr, tibble libraries), with statistical significance set at two-sided $p < 0.05$. This non-inferiority framework is consistent with prior comparative translation studies (9, 20, 21).

For a priori power, we planned for 33 lines, each rated by ~ 7 bilingual clinicians (≈ 231 ratings). Because ratings on the same line are correlated, we adjusted for clustering using an assumed intraclass correlation (ICC) of 0.25, giving an effective sample size of $n_{eff} = \frac{33 \times 7}{1+(7-1)\times 0.25} \approx 92$. With a one-sided non-inferiority test ($\alpha = 0.025$), a prespecified margin of $\Delta = 0.30$, and an expected true mean difference of 0.10, the projected power was ~89% (Supplementary Material SM3). Sensitivity checks varying ICC (0.20–0.30) and score variability (SD 0.60–0.70) yielded power between ~74% and 93%, supporting the adequacy of the planned sample size.

Exploratory System Metrics

In addition to translation quality, we recorded system-level performance metrics, including end-to-end translation latency and estimated operational cost. Latency was computed from system logs as the total time between audio input and synthesized output, encompassing automatic speech recognition, neural translation, and text-to-speech synthesis. Operating cost was estimated based on API usage and computational overhead for a 10-minute bidirectional conversation, using unit prices of deployed components. These exploratory analyses aimed to contextualize LingualAI's technical efficiency and cost-effectiveness relative to conventional interpreter services.

RESULTS

Results are organized in three analytical layers. Differences are reported as Δ = Human − AI (positive values favor human), with a prespecified non-inferiority margin of 0.30 points on the 5-point scale. Analyses include both paired and mixed-effects models to account for evaluator clustering and confirm robustness. Findings are presented as follows: (i) evaluator participation and reliability, (ii) domain-level performance evaluated using non-inferiority testing and direction-specific contrasts, and (iii) secondary analyses, including preference patterns, robustness checks, and representative error types.

Evaluator and Data Overview

Nine bilingual clinicians contributed ratings across three standardized scenarios comprising 18 clinician and 15 patient lines. Evaluator participation varied by scenario (8, 7, and 4 raters for Scenarios 1–3, respectively), resulting in an unbalanced rater–item panel. Participation declined across scenarios due to evaluator fatigue and the substantial time required, leading to fewer ratings in Scenario 3 (Table 1). Because raters were anonymous, inter-rater reliability was estimated using Krippendorff's α, which accommodates missing data and unequal numbers of raters per item. Across all domains, α = 0.31, reflecting fair agreement. This modest reliability likely reflects our design choice to capture first-impression ratings rather than consensus. Domain-specific α values are shown in Table 2. Score distributions were approximately normal except for fluency and prosody, which violated Shapiro–Wilk assumptions and were analyzed with non-parametric tests.

Primary Analyses (Fixed and Mixed Effects)

Results from paired and mixed-effects models were largely consistent (Table 3). Adjusting for evaluator clustering did not alter the direction or significance of findings, confirming the robustness of observed differences between human and AI translations.

Primary factors:

Adequacy of meaning (human 4.82 vs. AI 4.70; 95% CI, 0.00 to 0.24; $p = 0.050$) and terminology accuracy (4.81 vs. 4.75; CI, −0.05 to 0.19; $p = 0.282$) remained statistically comparable across both models, indicating semantic equivalence between human and AI outputs. In contrast, clarity/intelligibility showed a consistent and large human advantage (4.88 vs. 4.36; $p < 0.001$; $r = 0.76$).

Secondary factors:

Human interpreters achieved higher scores in completeness (4.87 vs. 4.73; $p = 0.022$; 95% CI, 0.02–0.26), grammar/syntax (4.88 vs. 4.67; $p = 0.007$; CI, 0.09–0.32), vocabulary (4.85 vs. 4.67; $p = 0.060$; CI, 0.07–0.29), and cultural appropriateness (4.89 vs. 4.49; $p < 0.001$; CI, 0.27–0.52). Effect sizes ranged from moderate for vocabulary to large for grammar/syntax and cultural appropriateness.

Voice-quality factors:

The most pronounced differences were observed in fluency (4.86 vs. 3.72; $p < 0.001$; CI, 0.97–1.31), prosody (4.89 vs. 4.30; $p < 0.001$; CI, 0.47–0.72), and pacing (4.88 vs. 4.46; $p = 0.002$; CI, 0.31–0.53), all demonstrating large, statistically robust advantages for human interpreters.

Conclusive factors:

Overall quality (4.81 vs. 4.23; $p < 0.001$; CI, 0.44–0.72) and confidence for clinical use (4.82 vs. 4.22; $p < 0.001$; CI, 0.47–0.75) were also significantly higher for human translations.

Direction-specific contrasts:

Stratified analyses confirmed these trends (Supplementary Material SM4). For clinician statements (English → Spanish), human interpreters outperformed the AI system in clarity, fluency, prosody, pacing, overall quality, and clinical confidence, while terminology and adequacy of meaning were equivalent. For patient statements (Spanish → English), human interpreters again scored higher across delivery-related and linguistic-mechanics domains (grammar, vocabulary, prosody), whereas accuracy and meaning remained non-inferior.

Non-Inferiority Testing

A prespecified non-inferiority margin of 0.30 points on the 5-point scale was applied (Fig. 3). Only a subset of domains met this threshold, indicating that while LingualAI preserved semantic accuracy, delivery-related aspects remained inferior to human interpreters.

Primary factors: Terminology accuracy (mean difference 0.07; 95% CI, 0.18) and adequacy of meaning (0.13; CI, 0.23) satisfied the non-inferiority criterion, confirming comparable semantic performance. Clarity exceeded the margin (0.50; CI, 0.72), reflecting a perceptible decline in intelligibility relative to human translation.

Secondary factors: Completeness (0.14; CI, 0.25) met the criterion, but vocabulary (0.18; CI, 0.32), grammar (0.21; CI, 0.34), and cultural appropriateness (0.39; CI, 0.64) exceeded the margin, suggesting subtle but consistent linguistic differences.

Voice-quality factors: None met the non-inferiority threshold, differences were largest for fluency (1.13; CI, 1.49), followed by prosody (0.59; CI, 0.81) and pacing (0.40; CI, 0.55), indicating that human interpreters retained a clear advantage in speech naturalness and rhythm.

Conclusive factors: Both overall quality (0.58; CI, 0.81) and confidence for clinical use (0.61; CI, 0.84) exceeded the margin, underscoring the evaluators' preference and higher perceived reliability for human translations. The mean differences and 95% confidence intervals for all domains are visualized in Table 4 and Figure 3, plotted against the prespecified non-inferiority boundary.

When a stricter non-inferiority margin of 0.20 was applied, limited to Scenarios 1–2 due to smaller sample size in Scenario 3, terminology accuracy remained non-inferior while all other domains exceeded the margin. Detailed results are provided in Supplementary Material SM5.

Preference Analysis

Pairwise preference testing compared individual evaluator choices between human and AI audio clips for each domain (Table 5). Overall, preference patterns closely mirrored the quantitative scoring results: evaluators perceived both systems as equivalent for meaning and terminology but favored human interpreters for delivery quality and naturalness.

Primary factors: Terminology accuracy (tie rate, 66%) and adequacy of meaning (65%) showed high equivalence, with minimal net bias toward human translations. Clarity revealed a stronger human preference (40% human wins vs. 4% AI wins; net bias of 0.36).

Secondary factors: Completeness also demonstrated a high tie rate (73%), whereas grammar (29% vs. 8%; net bias of 0.21) and cultural appropriateness (31% vs. 8%; net bias of 0.23) favored human interpreters. Vocabulary showed a smaller but consistent human advantage (net bias of 0.12).

Voice-quality factors: Preferences were most pronounced in these domains. Fluency had the lowest tie rate (25%) and the highest human win rate (67%; net bias of 0.59), followed by pacing and prosody, both exhibiting substantial net biases (>0.38).

Conclusive factors: For overall quality and confidence for clinical use, human translations won in roughly half of all comparisons (~49%), with average net biases near 0.39. These findings emphasize that, while evaluators recognized comparable semantic accuracy, they consistently preferred the tone, rhythm, and expressiveness of human interpretations. Detailed tie rates for each domain can be found in Supplementary Material SM6.

Error Analysis

Qualitative review of evaluator comments and audio outputs revealed that most residual errors arose from *delivery* rather than *meaning*. Specifically, deviations in *fluency* and *prosody*, such as monotone intonation, abrupt phrasing, and irregular pacing occasionally disrupted conversational flow, even when lexical and semantic accuracy were preserved.

These patterns were observed across all scenarios. For instance, certain clinician prompts (e.g., “Are you still doing the Budesonide irrigations?”) exhibited misplaced pauses or unnatural emphasis, while reassurance statements (e.g., “That’s expected. The swelling should go down over the next few weeks.”) were rendered with rigid or mechanical rhythm. Importantly, semantic fidelity remained high, confirming that LingualAI’s primary limitation lies in *speech synthesis naturalness* rather than translation accuracy.

Representative examples with linked audio clips are provided in Supplementary Material SM7, illustrating how subtle deviations in tone and pacing can alter perceived empathy and conversational realism.

Translation Latency and Cost

Based on system log analyses, LingualAI achieved an average end-to-end translation latency of 9.7 seconds per message, encompassing automatic speech recognition, neural translation, and speech synthesis. This turnaround time falls within the typical conversational pause window, allowing smooth, turn-based clinical dialogue without noticeable interruption.

The estimated operating cost for a 10-minute bilingual conversation was approximately $0.03–$0.04 (USD), compared with $6.90–$10.60 for phone or video interpreter services, representing a cost

reduction exceeding 99% per session. These results highlight LingualAI's potential for scalable, low-cost deployment in time-sensitive or resource-limited clinical settings.

Detailed latency distributions and cost-calculation parameters are provided in Supplementary Material SM8.

Summary of Findings

Overall, LingualAI preserved semantic accuracy at levels comparable to certified interpreters but showed limitations in delivery quality and expressiveness. Among the primary factors, both terminology accuracy and adequacy of meaning met the non-inferiority thresholds, confirming that the AI system conveyed medical content reliably.

Within the secondary factors, only completeness met the non-inferiority criterion, while grammar, vocabulary, and cultural appropriateness favored human interpreters. Notably, differences in vocabulary and cultural appropriateness reached significance primarily in patient statements (Spanish → English).

All voice-quality factors – fluency, prosody, and pacing showed large and consistent human advantages, underscoring the gap in naturalness and conversational rhythm.

Finally, both conclusive factors – overall quality and confidence for clinical use were significantly higher for human translations. Together, these findings suggest that while LingualAI effectively preserves meaning and accuracy, its synthesized speech still lacks the nuance and expressivity essential for fully natural clinical communication.

DISCUSSION

Key Findings

In this prospective, within-subject comparison of an AI-based translation tool (LingualAI) with certified medical interpreters, we found that LingualAI preserved two primary factors, terminology accuracy and adequacy of meaning, at levels that met non-inferiority thresholds, ensuring reliable conveyance of core medical content. LingualAI also achieved non-inferiority for completeness, supporting its ability to maintain the overall flow of dialogue. However, human interpreters consistently outperformed LingualAI in various secondary and voice-related factors (grammar, vocabulary, cultural appropriateness, fluency, prosody, and pacing), underscoring the continued importance of delivery quality on clinical communications. These findings suggest that while semantic integrity was maintained, elements crucial for patient trust and nuanced interaction remain stronger with professional interpreters. Importantly, the preservation of meaning and terminology, the dimensions most directly linked to clinical safety suggests that LingualAI can reliably convey essential medical information, even if its spoken delivery remains less natural than human interpreters.

Clinical Implications

The findings indicate that AI-based translation tools such as LingualAI are beginning to demonstrate credible performance in the critical clinical safety dimensions, preserving meaning and terminology accuracy. These strengths suggest a role for AI in supporting basic comprehension and instruction delivery, particularly in settings where interpreter services are unavailable, such as urgent encounters, rural clinics, or after-hours consultations. In such scenarios, LingualAI may be an alternative to delaying care or rendering care without translation. From a practical standpoint, accuracy in conveying clinical meaning is often of higher value than perfectly natural prosody, especially in urgent or resource-limited settings. While human interpreters provide richer tone and empathy, LingualAI's consistent accuracy and ultra-low operational cost may make it a viable adjunct when access or affordability is constrained.

Moving forward, it will be essential to work closely with certified medical translation bodies, since the acceptability of AI-supported translation depends on alignment with professional standards. The objective is not to replace certified medical translators but to augment their role and provide coverage in situations where they are absent. In practice, this could also include hybrid workflows where certified translators remotely monitor AI-facilitated conversations, ensuring they remain in the loop and can intervene when needed. AI translation should be viewed as an adjunct that extends access while maintaining the primacy of professional interpreters to ensure safe, nuanced, and culturally appropriate communication.

The baseline capabilities of AI translation models are expected to continue improving, which may steadily raise overall performance on core tasks. At the same time, the inherent risk of errors could also increase as these systems are deployed more widely and in higher-stakes settings. The broader deployment of these systems in critical settings highlights the importance of continuously monitoring primary factors such as meaning and terminology accuracy, alongside steady technological improvements in secondary and voice-related factors. To ensure safe and equitable integration, such progress should be documented through well-structured, real-world studies that evaluate not only technical performance but also clinical usability and patient experience.

Under Section 1557 of the Affordable Care Act and related regulations and guidance from the Office for Civil Rights (OCR), AI-based translation tools may be used to support language access in healthcare, but they cannot serve as the sole method for providing meaningful access to individuals with limited English proficiency (22, 23). OCR's regulatory framework generally requires that most substantive clinical communications such as informed consent, treatment instructions, and discharge summaries be reviewed or provided by qualified human interpreters or translators. While AI tools may improve efficiency and accuracy in medical terminology, they must be integrated into workflows that include human oversight to ensure compliance with nondiscrimination standards.

Comparison to literature

Our findings are consistent with prior evaluations of general-purpose MT tools, which often demonstrate strengths in medical terminology and adequacy of meaning, but show persistent weakness in contextual

interpretation, cultural appropriateness, fluency, and other delivery-related aspects. For example, studies report that tools such as Google Translate perform well in terms of adequacy of meaning and basic clinical content, but often struggle with cultural context and have dialect sensitivity issues (24, 25). Work on LLMs has further highlighted the limitations of literal translation approaches that miss nuances essential for medical communication (26, 27). Evidence also suggests that AI-based translations lack the cultural awareness to support trust and rapport in diverse patient populations, potentially leading to misunderstandings (28). While some studies indicate that AI can generate fluent output, syntactic and semantic inconsistencies remain common, particularly in languages with complex grammatical structures (29).

Our results extend this body of evidence by demonstrating that even a purpose-built, domain-specific system like LingualAI, optimized for clinical communication, continues to mirror these patterns. It performed comparably to certified interpreters on meaning and terminology accuracy but lagged in prosody, fluency, and clinical confidence. Unlike prior studies, however, our evaluation used a prospective, within-subject design with blinded bilingual clinician raters, multidomain scoring, and a non-inferiority framework. This methodological rigor strengthens confidence in the findings and provides a more granular understanding of where AI translation tools may safely support clinical communication and where human oversight remains indispensable.

Strengths & Limitations

This study has several strengths. The prospective, within-subject paired design ensured direct comparability between AI and human translators, while blinded evaluation by bilingual clinicians reduced bias. A structured, multidomain rubric and a pre-specified non-inferiority framework provided a rigorous translation quality assessment across clinically relevant dimensions. In addition, combining paired statistical tests and mixed-effects modeling strengthened the robustness of findings and enabled a nuanced interpretation of where AI tools achieved equivalence and where differences persisted.

Several limitations should be acknowledged. The scenarios were scripted and restricted to otorhinolaryngology, which may not capture the variability of free-flowing clinical dialogue, including patient emotions, interruptions, or background noise. The study evaluated audio translations only, without considering LingualAI's text-based interface, which could also influence comprehension. The evaluation occurred in a non-live setting, outside of real patient encounters, which may not fully represent the pressures and dynamics of clinical care.

Evaluator-related factors add further constraints. The clinical evaluators were anonymized, and an unequal number of ratings were available across scenarios, leading to an unbalanced dataset. Inter-rater reliability was modest (Krippendorff's $\alpha = 0.31$), and no calibration was performed to align evaluators before the study; this was an intentional decision to capture first impressions rather than consensus, but one that limited agreement. Additionally, the raters were clinicians rather than patients, meaning domains such as fluency, rapport, or cultural appropriateness may not reflect patient experience directly.

Finally, the findings were restricted to English–Spanish translation, and results may not generalize to other language pairs or dialects. These limitations should be considered when interpreting the scope of our results.

Future Directions

Future development of LingualAI should prioritize enhancements in speech synthesis, particularly prosody, pacing, and fluency, which accounted for the most significant performance gaps compared with certified interpreters. Beyond these technical refinements, validation in live, unscripted clinical encounters will be essential to test performance under real-world conditions such as interruptions, emotional tone, and background noise. Incorporating patient perspectives and expanding evaluation to additional languages and dialects will strengthen generalizability. As language models and translation technologies evolve, monitoring new error models and working closely with certified medical translation bodies to ensure alignment with professional standards will be critical. Therefore, well-structured, real-world studies should assess accuracy, delivery quality, patient-centered outcomes, and clinical usability, aiming to develop evidence-based, reliable, and trusted AI translation tools that extend access to multilingual care while preserving the central role of professional interpreters. Another important consideration is consistency. While human interpreter quality can vary substantially across sessions and providers, automated systems like LingualAI offer the potential for stable, repeatable performance once adequately validated. Future improvements in voice synthesis, particularly expressiveness and conversational pacing, will further enhance user experience and make AI-assisted translation increasingly compelling in real-world use.

Conclusion

LingualAI demonstrated non-inferior performance in preserving meaning and terminology – the core dimensions of translation accuracy while certified interpreters remained superior in delivery-related qualities such as fluency, prosody, and clinical confidence. These findings suggest that AI translation tools can safely support essential communication when interpreters are unavailable, provided their use remains aligned with professional standards. Consistency and scalability may further enhance their utility, particularly as voice synthesis continues to improve. In practice, the most responsible pathway forward is an interpreter-in-the-loop model, in which professional interpreters oversee or supplement AI-mediated communication to ensure contextual accuracy and patient trust. Professional interpreters therefore remain the gold standard, yet AI systems like LingualAI can serve as valuable supplementary tools that extend access to multilingual care. Continued technical refinement and clinical validation will be essential for ensuring safe, effective, and trusted deployment in real-world healthcare settings.

Declarations

Ethics Approval: This study has been approved by the Committee for the Protection of Human Subjects (the UTHSC-H IRB) under protocol HSC-SBMI- 13-0549.

ACKNOWLEDGEMENTS

We gratefully acknowledge Drs. German Martinez-Gamba and Amalia Guardiola for their contributions as clinical evaluators. We also thank William B. McElhiney for his valuable guidance on legal and regulatory considerations.

FUNDING

This work was supported by institutional funds from The University of Texas Health Science Center at Houston (UTHealth Houston) and by discretionary research funds provided by Dr. Xiaoqian Jiang.

DATA AVAILABILITY STATEMENT

De-identified audio clips and raw evaluator rating data underlying this study are available from the corresponding author upon reasonable request. Access will require completion of a data-use agreement and approval by the appropriate Institutional Review Board. Summary-level data (means, standard deviations, and confidence intervals) are included in the main manuscript and Supplementary Materials. Due to file size, PHI redaction requirements, and copyright considerations, raw audio files and individual rating sheets are not publicly posted.

HUMAN ETHICS AND CONSENT TO PARTICIPATE

All clinician evaluators participated on a voluntary basis and provided informed consent before initiating any study-related activities. The study was conducted in accordance with the ethical principles outlined in the Declaration of Helsinki.

No patients or members of the public were recruited or enrolled as participants in this study. Certified translators provided translation services as compensated professionals and were not considered study participants.

AUTHOR CONTRIBUTIONS STATEMENT

UPS led the study conceptualization, methodology, formal analysis, data curation, visualization, and drafting of the original manuscript. CJG contributed to software development, methodology, data curation, validation, and visualization. GMA, JBG, DRJ, EW, VSC, AFV, and BRE contributed to investigation, validation, and manuscript review. JHC and CW supported project administration, with JHC additionally contributing to manuscript review and editing. BOF contributed resources, investigation, project administration, and manuscript review. MJC provided supervision, conceptual guidance, and manuscript review. XJ provided overall conceptualization, supervision, methodological guidance, and manuscript review, editing and funding support.

All authors reviewed and approved the final manuscript.

References

1. Diamond L, Izquierdo K, Canfield D, Matsoukas K, Gany F. A Systematic Review of the Impact of Patient–Physician Non-English Language Concordance on Quality of Care and Outcomes. J Gen Intern Med. 2019 Aug;34(8):1591–606.
2. U.S. Census Bureau. People that speak English less than very well [Internet]. Washington, DC: U.S. Department of Commerce; n.d. Available from: https://www.census.gov/library/visualizations/interactive/people-that-speak-english-less-than-very-well.html
3. Twersky SE, Jefferson R, Garcia-Ortiz L, Williams E, Pina C. The Impact of Limited English Proficiency on Healthcare Access and Outcomes in the U.S.: A Scoping Review. Healthcare. 2024 Jan 31;12(3):364.
4. Williams KM, Dougherty D, Plagens C, Shah NR, Tubbs D, Ehrlich PF. Limited English Proficiency can Negatively Impact Disease/Treatment in Children With Cancer Compared to Those Who are English Proficient-an Institutional Study. J Pediatr Surg. 2024 May;59(5):800–3.
5. Tran AV, Roberts KP. Language Accommodations for Limited English Proficient Patients in Rural Health Care. J Immigr Minor Health. 2023 June;25(3):674–9.
6. Habhab M, Pham MT. Medical Interpretation Services: Challenges for LEP Communities. Am J Bioeth. 2024 Nov;24(11):72–4.
7. Hudelson P, Chappuis F. Using Voice-to-Voice Machine Translation to Overcome Language Barriers in Clinical Communication: An Exploratory Study. J Gen Intern Med. 2024 May;39(7):1095–102.
8. Genovese A, Borna S, Gomez-Cabello CA, Haider SA, Prabha S, Forte AJ, et al. Artificial intelligence in clinical settings: a systematic review of its role in language translation and interpretation. Ann Transl Med. 2024 Dec;12(6):117–117.
9. Lee W, Khoong EC, Zeng B, Rios-Fetchko F, Ma Y, Liu K, et al. Evaluation of Commercially Available Machine Interpretation Applications for Simple Clinical Communication. J Gen Intern Med. 2023 Aug;38(10):2333–9.
10. Kong M, Fernandez A, Bains J, Milisavljevic A, Brooks KC, Shanmugam A, et al. Evaluation of the accuracy and safety of machine translation of patient-specific discharge instructions: a comparative analysis. BMJ Qual Saf. 2025 July 9;bmjqs-2024-018384.
11. Taira BR, Kreger V, Orue A, Diamond LC. A Pragmatic Assessment of Google Translate for Emergency Department Instructions. J Gen Intern Med. 2021 Nov;36(11):3361–5.
12. Herrmann-Werner A, Loda T, Zipfel S, Holderried M, Holderried F, Erschens R. Evaluation of a Language Translation App in an Undergraduate Medical Communication Course: Proof-of-Concept and Usability Study. JMIR MHealth UHealth. 2021 Dec 2;9(12):e31559.
13. Pinto Taylor E, Mulenos A, Chatterjee A, Talwalkar JS. Partnering With Interpreter Services: Standardized Patient Cases to Improve Communication With Limited English Proficiency Patients. MedEdPORTAL. 2019 May 20;10826.
14. Han L, Gladkoff S, Erofeev G, Sorokina I, Galiano B, Nenadic G. Neural machine translation of clinical text: an empirical investigation into multilingual pre-trained language models and transfer-learning.

Front Digit Health. 2024 Feb 26;6:1211564.

15. Zolnoori M, Vergez S, Xu Z, Esmaeili E, Zolnour A, Anne Briggs K, et al. Decoding disparities: evaluating automatic speech recognition system performance in transcribing Black and White patient verbal communication with nurses in home healthcare. JAMIA Open. 2024 Oct 8;7(4):ooae130.
16. Chen CL, Dong Y, Castillo-Zambrano C, Bencheqroun H, Barwise A, Hoffman A, et al. A systematic multimodal assessment of AI machine translation tools for enhancing access to critical care education internationally. BMC Med Educ. 2025 July 8;25(1):1022.
17. Freitag M, Foster G, Grangier D, Ratnakar V, Tan Q, Macherey W. Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation. Trans Assoc Comput Linguist. 2021 Dec 17;9:1460–74.
18. Läubli S, Castilho S, Neubig G, Sennrich R, Shen Q, Toral A. A Set of Recommendations for Assessing Human-Machine Parity in Language Translation. J Artif Intell Res [Internet]. 2020 Mar 23 [cited 2025 Aug 29];67. Available from: http://arxiv.org/abs/2004.01694
19. Liu T, Lo C kiu, Marshman E, Knowles R. Evaluation Briefs: Drawing on Translation Studies for Human Evaluation of MT. 2024;
20. Gerlinger C, Schmelter T. Determining the non-inferiority margin for patient reported outcomes. Pharm Stat. 2011 Sept;10(5):410–3.
21. Do Campo Bayón M, Sánchez-Gijón P. Evaluating NMT using the non-inferiority principle. Nat Lang Process. 2025 July;31(4):1042–61.
22. U.S. Department of Health & Human Services, Office for Civil Rights. Dear Colleague Letter: Language Access Provisions of the Final Rule Implementing Section 1557 of the Affordable Care Act [Internet]. 2024 Dec. Available from: https://www.hhs.gov/sites/default/files/ocr-dcl-section-1557-language-access.pdf
23. National Health Law Program. Title VI and Section 1557 Language Access Requirements [Internet]. Washington, DC; 2024 May. Available from: https://healthlaw.org/wp-content/uploads/2024/05/T-VI-and-Sec-1557-explainer-2024.pdf
24. Brandenberger J, Stedman I, Stancati N, Sappleton K, Kanathasan S, Fayyaz J, et al. Using artificial intelligence based language interpretation in non-urgent paediatric emergency consultations: a clinical performance test and legal evaluation. BMC Health Serv Res. 2025 Jan 24;25(1):138.
25. Brewster RCL, Gonzalez P, Khazanchi R, Butler A, Selcer R, Chu D, et al. Performance of ChatGPT and Google Translate for Pediatric Discharge Instruction Translation. Pediatrics. 2024 July 1;154(1):e2023065573.
26. Yan J, Yan P, Chen Y, Li J, Zhu X, Zhang Y. Benchmarking GPT-4 against Human Translators: A Comprehensive Evaluation Across Languages, Domains, and Expertise Levels [Internet]. arXiv; 2024 [cited 2025 Sept 16]. Available from: http://arxiv.org/abs/2411.13775
27. Keles B, Gunay M, Caglar SI. LLMs-in-the-loop Part-1: Expert Small AI Models for Bio-Medical Text Translation [Internet]. arXiv; 2024 [cited 2025 Sept 16]. Available from:

http://arxiv.org/abs/2407.12126

28. Merx R, Phillips C, Suominen H. Machine Translation Technology in Health: A Scoping Review. In: Bichel-Findlay J, editor. Studies in Health Technology and Informatics [Internet]. IOS Press; 2024 [cited 2025 Sept 16]. Available from: https://ebooks.iospress.nl/doi/10.3233/SHTI240895
29. Slosarek T, Paeschke D, Sivtsev I, Böttinger EP. Comparison of Machine Translation Services in the Biomedical Context. In: 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) [Internet]. Lisbon, Portugal: IEEE; 2024 [cited 2025 Sept 16]. p. 5539–45. Available from: https://ieeexplore.ieee.org/document/10822680/

Tables

Tables 1 to 5 are available in the Supplementary Files section

Figures

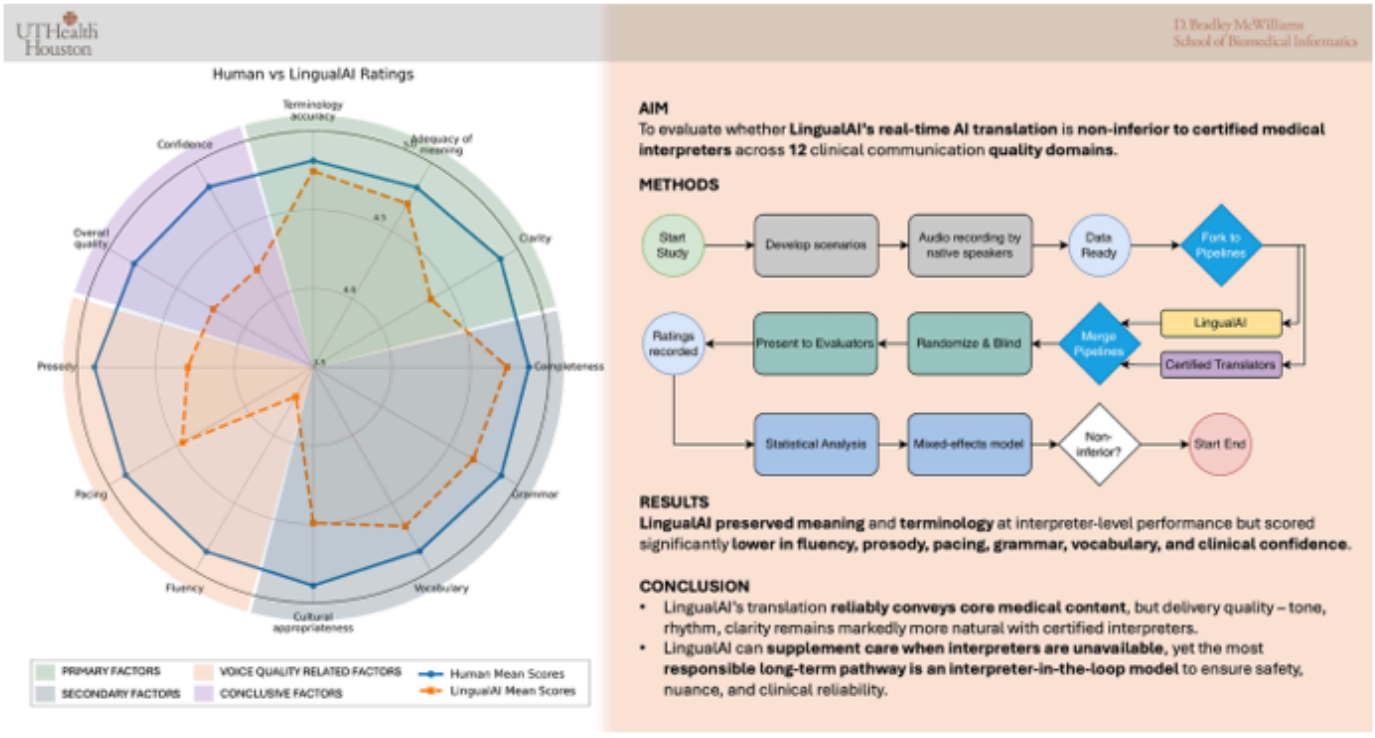

Figure 1

Visual Abstract.

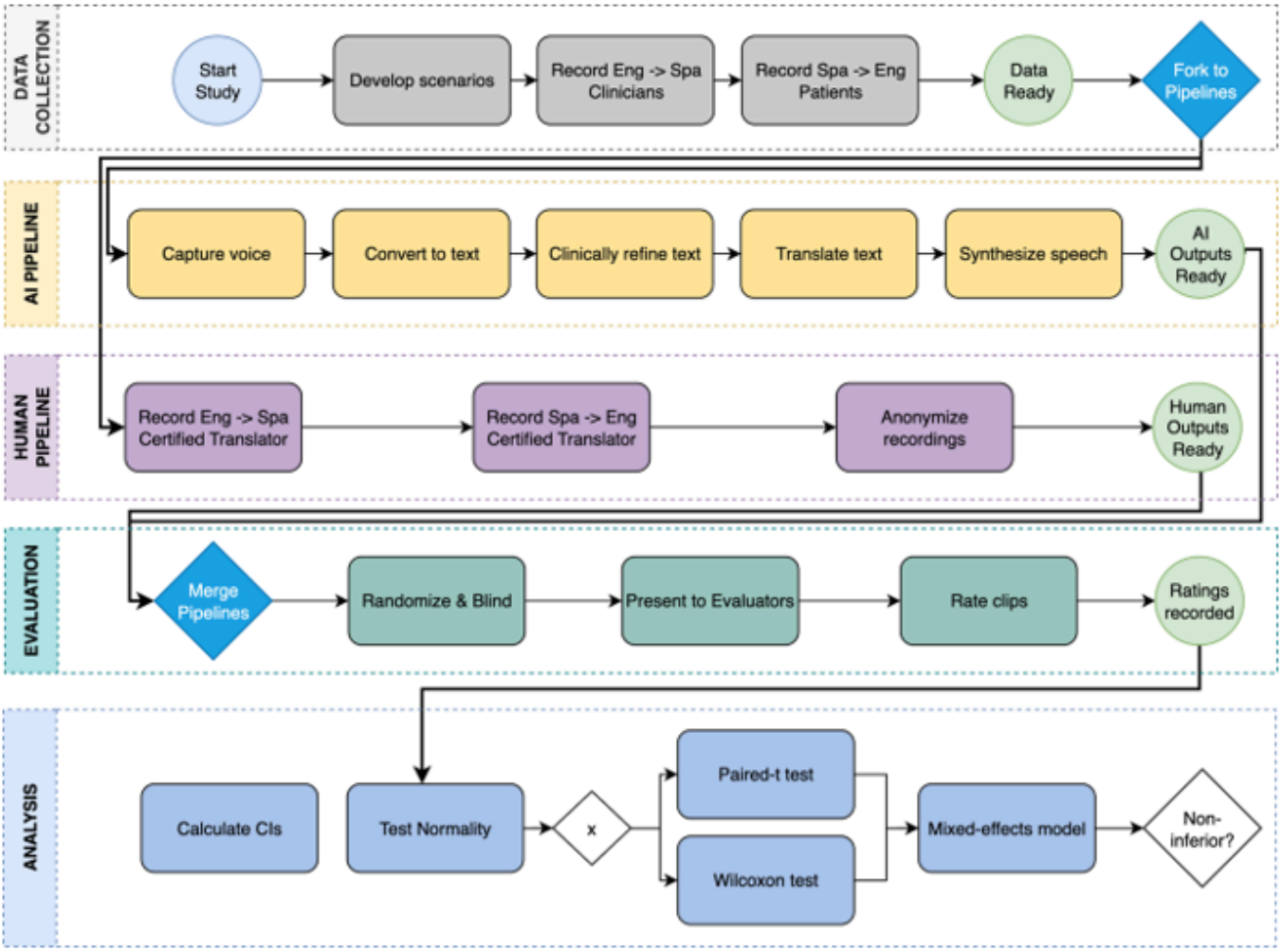

Figure 2

Study design and analytic workflow: The workflow outlines the sequential stages of data collection, translation, evaluation, and analysis. Three standardized otorhinolaryngology scenarios were enacted by native English- and Spanish-speaking pairs and forked into two translation pipelines. The AI pipeline captured voice, converted it to text, refined clinical content, translated, and synthesized speech. The human pipeline involved two certified interpreters (English→Spanish, Spanish→English) producing parallel recordings that were anonymized. All four outputs per line (two AI, two human) were merged, randomized, and presented to blinded bilingual clinicians for multidomain quality rating. Statistical analysis included paired and mixed-effects models with a prespecified non-inferiority margin of 0.30 points on 5-point Likert scales.

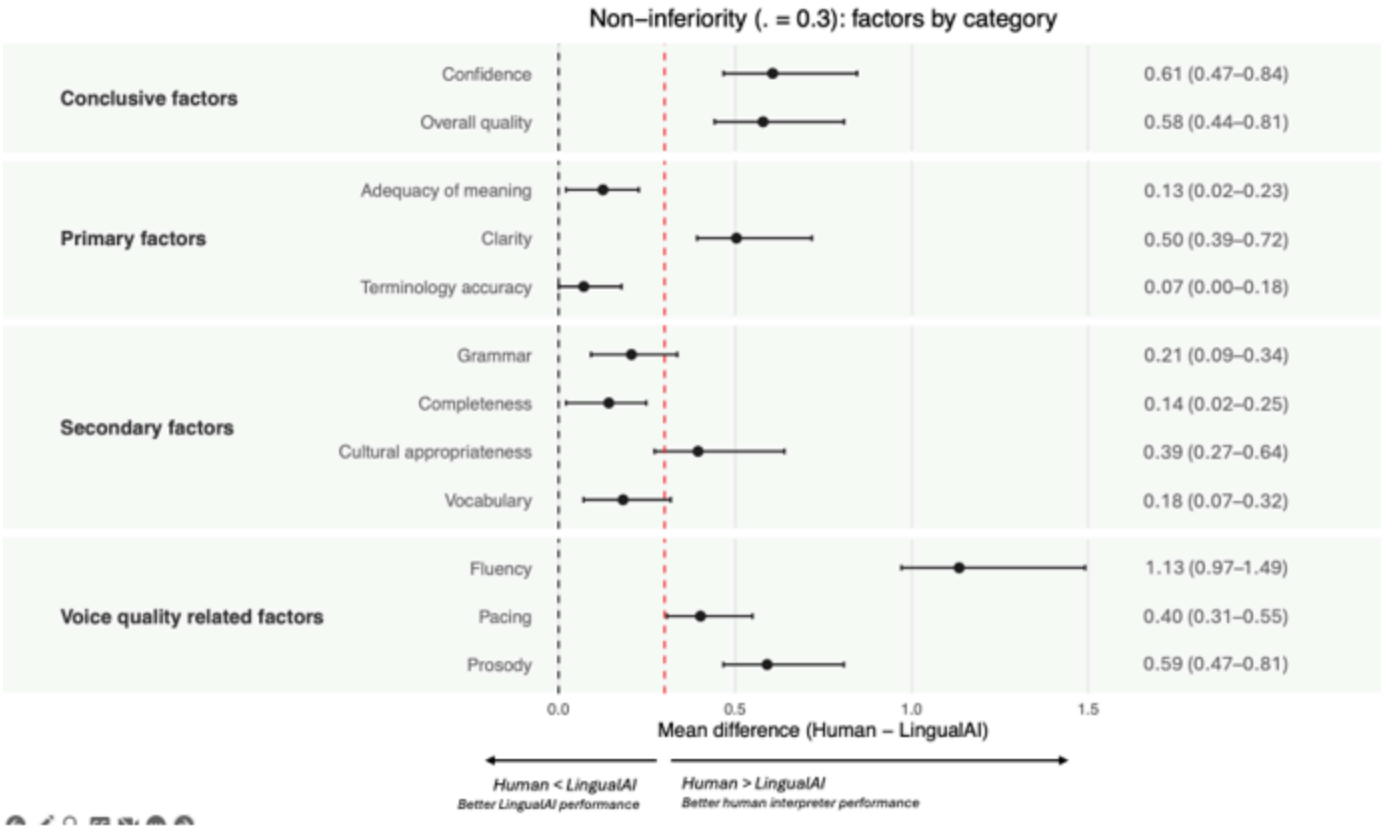

Figure 3

Non-inferiority analysis of AI versus human translation across quality domains: Mean differences (Human − LingualAI) with 95% confidence intervals are displayed for all 12 translation quality domains, grouped by category. The vertical dashed black line indicates no difference (Δ = 0), and the red dashed line marks the prespecified non-inferiority margin (Δ = 0.30). Points to the left of the red line represent domains in which LingualAI achieved non-inferior performance relative to human interpreters. Terminology accuracy, adequacy of meaning, and completeness met this criterion, while all voice-related and delivery factors (fluency, prosody, pacing) and conclusive factors (overall quality, clinical confidence) exceeded the margin, indicating superior human performance.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SM1.Scenarios13usedforvalidationofLingualAI.docx
- SM2.RatingInstrument.docx
- SM3.Aprioripowercalculation.docx
- SM4.StratifiedResultsforClinicianandPatientStatements.docx

- SM5.Noninferioritywithstrictermargin.docx
- SM6.TieRateforeachdomain.docx
- SM7.ErrorAnalysis.docx
- SM8.TranslationLatencyandEstimatedcostper10minuteconversation.docx
- Tables.docx