Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-02055-6

Evaluating human-in-the-loop strategies for artificial intelligence-enabled translation of patient discharge instructions: a multidisciplinary analysis

Check for updates

Ryan CL Brewster[1,2,3,4] ✉, Gabe Tse[5], Angela L. Fan[1], Marwa Elborki[1], Maiah Newell[1], Priscilla Gonzalez[1,2,4], Amitra Hoq[6], Crystal Chang[7], Maksud Chowdhury[8], Adiba Geeti[9], Marlin Hana[10], Hoda Hassan[11], Osama Ibrahim[1,2], Lucine Keseyan[12], Qing Li[13], Md Mamoon[14], Maymona Nageye[15], Arthur Ohannessian[16], Ilan Rozen Eisenberg[17], Mohammad Sallam[18], Giordano Sosa Soto[2,19], Christina Su[1,2,4], Raffi Tachdjian[20,21], Mondira Ray[1,2], Hannah Lev[22], Jonathan D. Hron[1,2], Nate Shaar[23], Nicholas Kuzma[24,25] & Alisa Khan[1,2,25]

Machine translation supported by artificial intelligence (AI) may enhance linguistically-concordant care for patients speaking languages other than English. This assessment of free-text inpatient discharge instructions in Arabic, Armenian, Bengali, simplified Chinese, Somali, and Spanish compared linguist, clinician, and family caregiver evaluations of translations generated by (1) ChatGPT-4o, (2) professional linguists, and (3) human-in-the-loop (AI-generated, professional linguist post-edited). Likert scales (1–5; higher is better) evaluated linguistic and clinical characteristics of each translation. ChatGPT-4o exhibited variable performance relative to professional translations, with poorest ratings for digitally underrepresented languages (Armenian and Somali). Conversely, human-in-the-loop translations achieved comparable, often better, outcomes to professional translations for all languages, (e.g., Armenian mean overall quality: 3.9 [95% CI 3.7–4.2] vs. professional 3.6 [3.4–3.9], $p = 0.01$), were most frequently preferred (46.5% vs. 28.4%) and had shorter mean translation time (7.1 [5.4–8.8] vs. 16.8 [13.7–19.9] min, $p < 0.001$). Human-in-the-loop strategies may enable safe, efficient, equitable machine translation application in clinical practice.

The U.S. Census estimates that more than 25 million individuals in the United States (U.S.) speak English less than "very well," with a growing proportion speaking a language other than English at home[1]. Patients who use languages other than English for care have poorer clinical and patient-centered outcomes, including medication adherence, healthcare utilization, and adverse events[2–5]. Barriers to high-quality translation are a key driver of these inequities[6,7]. Written discharge instructions and portal messages, among other forms of written materials, are essential modes of patient education and care delivery. Yet, they are frequently only made available in English[8]. Hospital in-house and third-party translation services can be costly and often require multiple days for completion, if utilized at all, making them impractical for time-sensitive communications[9,10]. Importantly, these gaps persist despite federal policies that stipulate non-discrimination on the basis of national origin and language access provisions among institutions receiving funding from the Centers for Medicare and Medicaid Services[11].

In the past decade, advances in artificial intelligence (AI) have generated interest in developing novel solutions to fill translation access challenges. The quality of neural machine translation systems like Google Translate and large language models like ChatGPT vary across languages[12–17]. The most significant deficits in accuracy and quality have been reported in digitally underrepresented languages (e.g., Haitian Creole), where relatively sparse linguistic datasets limits AI model training and validation, regardless of the number of global speakers[18,19]. Within this context, Section 1557 of the Affordable Care Act requires that machine translations must be reviewed by qualified translators[20]. The integration of human oversight—known as human-in-the-loop approaches—emphasizes the need for collaborative workflows with AI systems[21,22]. Manually reviewing AI-generated material may safeguard against potential harm; however, such benefits may be offset by added cognitive burden or over-reliance on automated outputs[23,24]. To date, the human-in-the-loop

A full list of affiliations appears at the end of the paper. ✉e-mail: ryan.brewster@childrens.harvard.edu

paradigm in the context of machine translation has not been systematically examined, limiting thoughtful implementation that balances quality, equity, and operational efficiency. Furthermore, no prior investigation has sought to capture the diverse perspectives of key partners—including those of linguists, clinicians, and family caregivers—commensurate with the nuanced and multidisciplinary nature of medical translation.

Therefore, we conducted a prospective study to compare the quality of free-text hospital discharge instructions translations generated by ChatGPT-4o and human-in-the-loop relative to professional translations across six languages, as assessed by linguist, clinician, and family caregiver evaluators.

Results

The study compared free-text pediatric inpatient discharge instructions translated into Arabic, Armenian, Bengali, simplified Chinese, Somali, and Spanish by three translation modalities: (1) ChatGPT-4o (Version 2024-11-20), (2) human-in-the-loop, involving postediting of ChatGPT-4o translations by professional linguists, and (3) professional linguist, the current reference standard.

Evaluator and source text characteristics

Translation performance was assessed by linguists, pediatric clinicians, and family caregivers. Overall, 42 evaluators participated in the study ($n = 12$ linguists, $n = 16$ clinicians, $n = 14$ caregivers), ranging between 6 and 8 evaluators per study language. The majority had either full professional (45%) or native/bilingual (35%) proficiency in English with a mean of 16.4 years (SD 12.4) of residence in the US (Table 1).

Evaluators most commonly reported using English and their native language equally when speaking and reading (60%), communicating with friends (43%), and thinking (43%). The majority more often (45.2%) or exclusively (21.4%) spoke their native language at home.

The mean text length of original English texts was 74.6 words (SD 28.5) and the mean Flesch-Kincaid readability score, a measure of ease of understanding a written text using sentence and syllable length along a 0–100 scale (higher values indicate easier to read), was 50.3 (SD 16.7), consistent with a 10–12th grade reading level.

Translation ratings

Evaluators measured translation performance using a validated instrument assessing adequacy (preservation of information), fluency (preservation of grammar and readability), meaning (preservation of connotation or intent), severity (risk of clinical harm or error) and overall quality. Each domain was measured along a 5-point Likert scale (1–5; higher is better).

Interrater reliability was moderate for Arabic (ICC 0.50 [95% CI 0.42–0.58]), Bengali (0.53 [0.47–0.65]), Chinese (0.58 [0.44–0.69]), Somali (0.61 [0.55–0.66]), and Spanish (0.55 [0.36–0.64]) and was good for (Armenian (0.77 [0.74–0.80]). The Supplemental Online Materials present numerical values (Supplementary Table 1) and bar plots (Supplementary Fig. 1) of aggregate mean domain-level ratings; mean domain-level ratings stratified by evaluator type (Supplementary Fig. 2); and types of translation errors with illustrative examples (Supplementary Table 2).

ChatGPT-4o vs. professional translation ratings. ChatGPT-4o exhibited differential performance across languages (Fig. 1). Relative to professional translations, evaluators rated all domains for ChatGPT-4o most poorly for Armenian and Somali. For example, ChatGPT-4o translations for Armenian received a mean rating for the overall quality domain of 2.4 (95% CI 2.1–2.7), which was 1.2 (1–1.4) lower than professional translations ratings of 3.6 (3.4–3.9) ($p < 0.001$). Albeit to a slightly lesser degree, simplified Chinese and Arabic ChatGPT-4o translations were also generally rated lower than professional translations. Domain ratings were comparable between ChatGPT-4o and professional translations for Bengali and Spanish. For example, the mean overall quality scores for Bengali were 3.6 (3.4–3.8) for ChatGPT-4o and 3.5 (3.2–3.7) for professional translations, corresponding to a difference of 0.1 (0–0.3) ($p = 0.15$).

Table 1 | Self-reported translation evaluator characteristics

	Linguist	Clinician	Family caregiver	Overall
	N = 12	*N* = 16	*N* = 14	*N* = 42
Native language, *n* (%)				
Arabic	2 (16.7)	3 (18.8)	2 (14.3)	7 (16.7)
Armenian	2 (16.7)	3 (18.8)	2 (14.3)	7 (16.7)
Bengali	2 (16.7)	3 (18.8)	2 (14.3)	7 (16.7)
Simplified Chinese	2 (16.7)	2 (12.5)	3 (21.4)	7 (16.7)
Somali	2 (16.7)	2 (12.5)	2 (14.3)	6 (14.3)
Spanish	2 (16.7)	3 (18.8)	3 (21.4)	8 (19.0)
English proficiency, *n* (%)				
No proficiency	0 (0)	0 (0)	0 (0)	0 (0)
Elementary proficiency	0 (0)	0 (0)	0 (0)	0 (0)
Limited working proficiency	0 (0)	0 (0)	1 (7.1)	1 (2.4)
Professional working proficiency	4 (33.3)	1 (6.3)	2 (14.3)	7 (16.7)
Full professional proficiency	3 (25.0)	6 (37.5)	9 (64.3)	18 (42.9)
Native or bilingual proficiency	5 (41.7)	9 (56.3)	2 (14.3)	16 (38.1)
Years of residence in United States, mean (SD)	3.00 (4.54)	22.8 (12.5)	19.8 (7.50)	16.4 (12.4)
Language(s) read and spoken, *n* (%)				
Only English	1 (8.3)	0 (0)	0 (0)	1 (2.4)
More English than native language	1 (8.3)	5 (31.3)	3 (21.4)	9 (21.4)
Both equally	9 (75.0)	8 (50.0)	8 (57.1)	25 (59.5)
More native language than English	1 (8.3)	3 (18.8)	2 (14.3)	6 (14.3)
Only native language	0 (0)	0 (0)	1 (7.1)	1 (2.4)
Language(s) spoken with friends, *n* (%)				
Only English	1 (8.3)	2 (12.5)	0 (0)	3 (7.1)
More English than native language	0 (0)	5 (31.3)	5 (35.7)	10 (23.8)
Both equally	6 (50.0)	6 (37.5)	6 (42.9)	18 (42.9)
More native language than English	5 (41.7)	3 (18.8)	2 (14.3)	10 (23.8)
Only native language	0 (0)	0 (0)	1 (7.1)	1 (2.4)
Language(s) spoken at home, *n* (%)				
Only English	2 (0)	1 (0)	0 (0)	3 (0)
More English than native language	2 (16.7)	1 (6.3)	3 (21.4)	6 (14.3)
Both equally	2 (16.7)	4 (25.0)	2 (14.3)	8 (19.0)
More native language than English	7 (58.3)	6 (37.5)	6 (42.9)	19 (45.2)
Only native language	1 (8.3)	5 (31.3)	3 (21.4)	9 (21.4)
Language(s) think in, *n* (%)				
Only English	1 (8.3)	2 (12.5)	0 (0)	3 (7.1)
More English than native language	0 (0)	4 (25.0)	4 (28.6)	8 (19.0)
Both equally	8 (66.7)	5 (31.3)	5 (35.7)	18 (42.9)
More native language than English	3 (25.0)	5 (31.3)	4 (28.6)	12 (28.6)
Only native language	0 (0)	0 (0)	1 (7.1)	1 (2.4)

Human-in-the-loop vs. professional translation ratings. Overall quality ratings for human-in-the-loop translations were similar to professional translations for Arabic, simplified Chinese, and Somali.

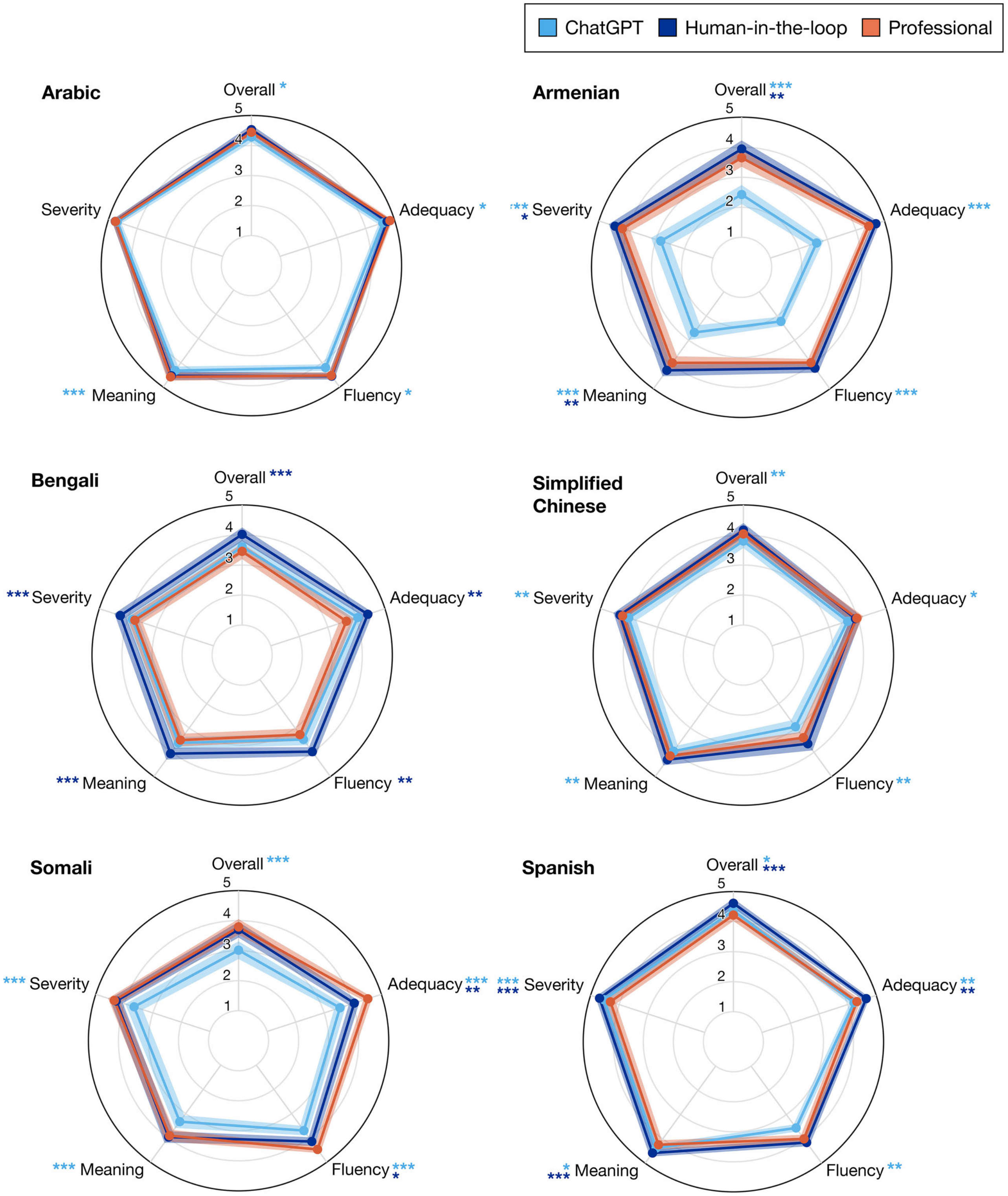
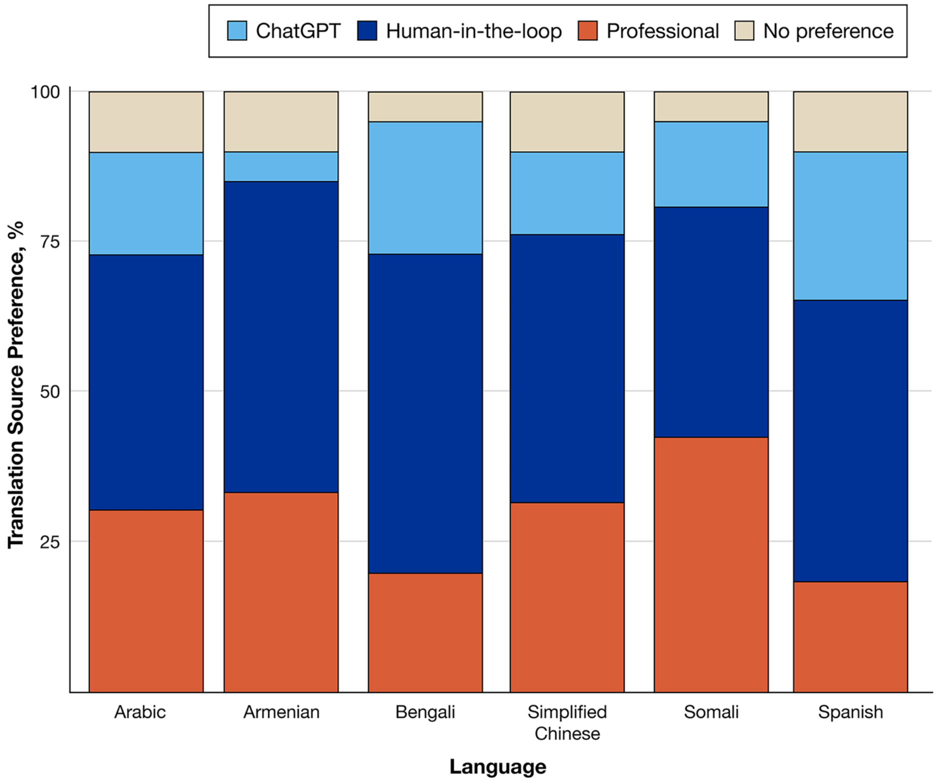


Fig. 1 | Mean (with 95% confidence intervals) domain-level ratings. Each domain on the radar chart is represented on an axis and mean ratings extend radially from the center from 1 to 5, with 5 being the best. Color-coded asterisks (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$) correspond to the Friedman test with post hoc Wilcoxon signed-rank test for statistical significance relative to professional translations. Adequacy and fluency domains only completed by linguist evaluators.

Evaluators scored overall quality for human-in-the-loop translations higher than professional translations for Armenian, Bengali, and Spanish. For example, the mean overall domain rating for Armenian human-in-the-loop translations was 3.9 (3.7–4.2) for human-in-the-loop and 3.6 (3.4–3.9) for professional translations, a difference of 0.3 (0.1–0.5) ($p = 0.01$).

Human-in-the-loop translations also received comparable, if not greater, ratings to professional translations for the adequacy, fluency, meaning, and severity domains across most languages, with the highest scores for Bengali and Spanish. For example, Spanish human-in-the-loop translations achieved a mean adequacy score of 4.7 (4.5–4.8), which was 0.4 (0.2–0.5) greater than professional translations, which had mean ratings of 4.3 (4.2–4.5)

Fig. 2 | Preferred translation source.



(p =0.01). For Somali human-in-the-loop translations, adequacy and fluency ratings were poorer than professional translations, though still high, while all other domains ratings were similar. There was no significant differences in mean meaning scores, for example, with human-in-the-loop translations rated 4 (3.7–4.2) and professional translations rated 3.9 (3.7–4.1) (p = 0.7).

Translation source preference

For most study languages, human-in-the-loop was the most highly preferred translation modality, ranging from 42.9% (95% CI 34.7–51.1%) for Arabic to 53.6% (45.1–61.7%) for Bengali (Fig. 2). Somali human-in-the-loop translations (38.3% [9.6–46.9%]) were similarly preferred to professional translations (42.5% [33.7–51.3%]).

Except for Spanish, evaluators least preferred ChatGPT-4o translations for all languages, including Arabic (17.1% [10.9–23.3%]), Armenian (5.0% [1.1–8.9%]), Bengali (22.1% [15.2–28.9%]), simplified Chinese (13.6% [7.9–19.2%]), and Somali (14.2% [7.9–20.4%]). Preferred translation sources for individual linguist, clinician, and caregiver evaluator groups are presented in Supplementary Fig. 3.

Completion time

Across all languages, the mean time required to produce each translation was significantly faster for human-in-the-loop than professional translations (7.1 [95% CI 5.4–8.8] vs 16.8 [13.7–19.9] min, $p < 0.001$). This held true for each individual language but Armenian, for which human-in-the-loop and professional translations took similar times to produce (human-in-the-loop 11.2 [9.3–13.1] vs. professional 13.0 [10.8–15.2] min) (Fig. 3). Otherwise, human-in-the-loop translations were completed significantly faster than professional translations for Arabic (3.5 [2.9–4.1] vs. 18 [15–21] min), Bengali (7.5 [6.2–8.8] vs 13.5 [11.2–15.8] min), simplified Chinese (7.4 [7.5–6-9.1] vs. 18.8 [15.6–21.9] min), Somali (3.9 [3.2–4.5] vs. 21.0 [17.5–24.5] min), and Spanish (9 [7.5–10.5] vs. 16.5 [13.7–19.3] min), all $p < 0.001$.

Discussion

In this evaluation of machine translation quality across six languages by diverse evaluators, ChatGPT-4o demonstrated variable performance in translating free-text patient discharge instructions, with generally poorer domain-level ratings and preference for digitally underrepresented languages, namely Armenian and Somali. Incorporating a human-in-the-loop yielded similar, if not better, outcomes relative to professional translations for most study languages, substantially reduced time to completion, and was the most preferred translation modality. To our knowledge, this is among the first studies to assess the use of human-in-the-loop approaches for machine translation and to incorporate multidisciplinary evaluator groups, including family caregivers.

Our findings advance an emerging body of literature showing imbalances in machine translation capabilities for patient-facing communications. An assessment of standardized discharge instructions, for example, found that Spanish and Portuguese translations generated by ChatGPT and Google Translate were comparable in quality to professional versions, but shortcomings and safety concerns were noted for Haitian Creole[12]. Such variation is likely due to inherent feature of all generative AI systems: the quality of the output directly depends on the composition of underlying training and validation datasets. Languages with a relatively large digital footprint, such as Spanish, tend to yield better translation performance.

Conversely, machine translations have consistently shown to be worse for languages, such as Haitian Creole, Somali, and Armenian, that are comparatively underrepresented. Differences in sentence composition, grammar, terminology, and other linguistic features relative to English and other Western European languages may also impact translation quality. This may in part explain the unexpectedly poor performance for some aspects of Arabic and simplified Chinese machine translations[19]. Meanwhile, Bengali has been classically considered a digitally underrepresented language, yet ChatGPT-4o received similar domain-level ratings to professional translations. Drawing generalizable conclusions about the safety and efficacy of machine translation may thus be challenging without individualized validation for each language of interest.

Heterogeneity of machine translation quality reinforces the need for ongoing oversight in clinical settings[25]. In our analysis, the utilization of human-in-the-loop approaches improved outputs from ChatGPT-4o alone, approximating and often exceeding the level of quality and

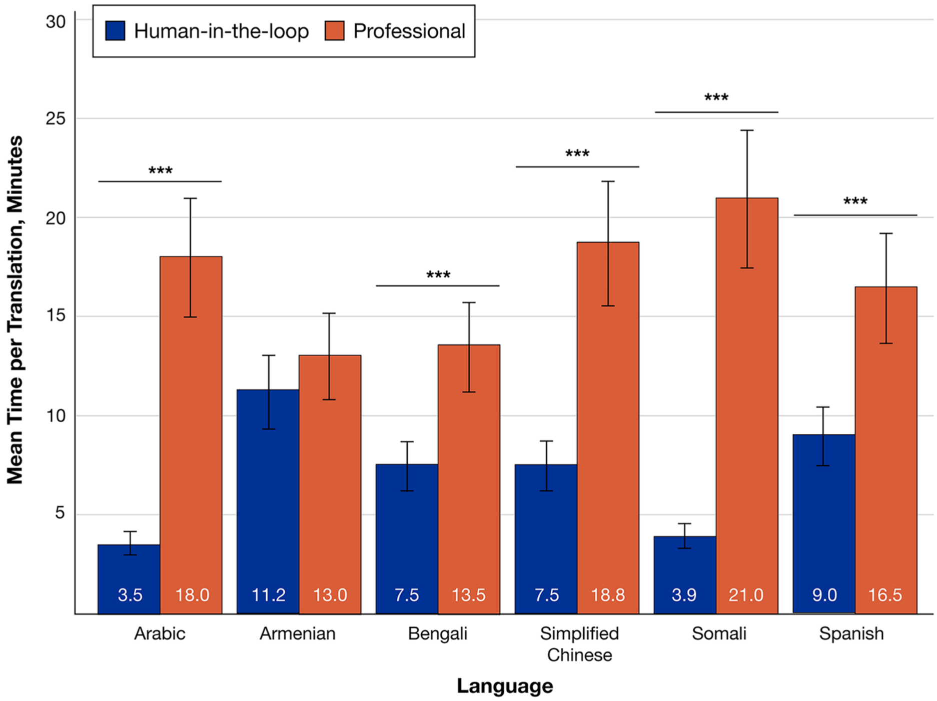


Fig. 3 | Mean (with 95% confidence interval) time to translation completion for human-in-the-loop. Asterisks (**P* < 0.05, ***P* < 0.01, ****P* < 0.001) correspond to statistically significant differences by two-tailed t-tests.

preferability to that of professional translations. Clinicians have increasingly been integrated into AI-enabled clinical workflows as human safeguards against potential machine-generated errors leading to patient harm[26–28]. The same imperatives should apply to machine translation. Linguists may be able to navigate the nuances of cultural and clinical context in ways not yet feasible for AI systems, particularly for digitally underrepresented languages.

That human-in-the-loop and professional translations were more preferred in our study than ChatGPT-4o despite comparable domain-level ratings for some languages illustrates the importance of continued human involvement. Our results altogether indicate that hybrid workflows combining machine translation with manual proofreading could enhance efficiency while maintaining high quality. Discharge instructions, portal messages, and other time-sensitive written communication may be more conducive to human-in-the-loop strategies than conventional translation services that often require advance notice and long turnaround times. However, these benefits may diminish if machine translation outputs warrant considerable revision for a particular language, as was appreciated for Armenian translations.

Optimizing efficiency, accuracy, and equity when operationalizing machine translation into health systems remains an active area of inquiry. Any technology intended for translation would be governed by federal provisions that mandate a qualified translator "when accuracy is essential, or when the source documents or materials contain complex, non-literal or non-technical languages."[20] Based on our findings, fully automated translation for certain low-risk, non-clinical activities, such as appointment scheduling, may be permissible, however, it should be strictly limited to languages that meet rigorous performance standards. Cautious introduction of human-in-the-loop strategies may be appropriate along a broader range of languages and applications.

Within this framework, institutional and national policies must prioritize the perspectives and needs of multidisciplinary partners, particularly of patients and family caregivers[25]. We aimed to capture a more comprehensive assessment than prior studies, which have exclusively relied on a single evaluator type[12,17]. Linguists, clinicians, and family caregivers each provide distinct and complementary perspectives towards machine translation quality. For example, linguists may place more weight on adequacy and fluency while clinicians and family caregivers may prioritize clinical meaning independent of linguistic structure. These groups should be engaged within a responsible AI framework to not only define acceptable clinical workflows and use cases, but also to protect patient privacy and safety[29].

Strengths of this study include the representative sampling of multiple different target languages and evaluator groups and masked and randomized study design to minimize bias. In addition, we characterized actual free-text pediatric patient discharge instructions with variable readability, as opposed to standardized or vendor-supplied materials. That said, important limitations exist. Firstly, although our findings carry immediate practical implications, they reflect a small number of source texts and may not be generalizable to all languages, machine translation platforms, patient-facing communications, or translation vendors. Secondly, we only evaluated the first output from ChatGPT-4o with a relatively simple prompt to approximate real-world user behavior. Translations may have been improved with iterative prompt engineering and in-context learning, approaches that been increasingly leveraged in machine translation applications[30,31]. Herein lies a fundamental challenge in the era of AI research—Innovations in technology and associated workflows are rapidly outpacing traditional academic and regulatory processes[32]. Future work should consider more resource-efficient methods that facilitate larger-scale and timely assessment; for example, establishing benchmarking datasets or standardized quality metrics[33,34]. Thirdly, while we pre-specified our pairwise comparisons and treated domain-level outcomes as independent, the risk of Type 1 error cannot be excluded. Lastly, despite receiving comprehensive training on study procedures, evaluators demonstrated only moderate interrater reliability for most languages. Cultural and dialect-specific differences within languages and subjectivity inherent to the evaluation framework may have contributed to observed variation.

AI has created novel opportunities to improve linguistically-appropriate care and clinical outcomes among patients who use languages other than English. Yet, persistent disparities in machine translation may place already vulnerable populations at risk of excess harm[35]. Human-in-the-loop techniques were able to overcome performance gaps, producing comparable results to professional linguists with greater efficiency. Our findings suggest a viable strategy for implementing machine translation into clinical practice safely and equitably, provided efforts are guided by linguists, clinicians, caregivers, and patients themselves.

Methods

Source text selection and translation

We conducted a comparative analysis to evaluate quality and translation time of three different translation modalities: (1) ChatGPT-4o, (2) human-in-the-loop, and (3) professional translation. First, we extracted customized, patient-specific free text discharge instructions from inpatient hospital admissions between 1/1/2023 and 1/1/2024 to non-surgical, non-intensive care services at an urban, quaternary care children's hospital. Discharge instructions were originally prepared in English by physicians. Using purposive sampling, we selected 20 source texts to represent common content areas including summaries of hospitalization, return precautions, and medication instructions. Prior to translation, we manually removed patient and family identifiers, including names, addresses, contact information, and medical record numbers. The study received approval from the Boston Children's Hospital Institutional Review Board (IRB-P00042876).

Subsequently, we translated de-identified source texts from English into Arabic, Armenian, Bengali, simplified Chinese, Somali, and Spanish (Fig. 4). We selected these languages based on available resources for translation and evaluation and to reflect a representative sample of language families, scripts, syntactic structures, and representation in computational linguistics. Armenian, Bengali, and Somali are generally considered digitally underrepresented languages. We used three translation modalities: (1) ChatGPT-4o, (2) human-in-the-loop, wherein ChatGPT-4o translations were manually post-edited by a professional linguist, and (3) professional translation by a linguist (the reference standard).

ChatGPT-4o (Version 2024-11-20), the latest large language model released by OpenAI at the time of the study, was accessed with a personal subscription[36]. We used the following prompt for the ChatGPT-4o translation, adapted from a prior study: "Imagine you are a translator at a children's hospital. Please translate the following information into [target language], which will be provided to patients and their families."[12]. The prompt was developed in consultation with clinical informaticists and language access researchers with the intention of mirroring real-world clinician usage. We entered individual prompts for each source text and target language on December 1, 2024. We used a certified third-party translation service (LanguageLine; Monterey, CA) for the human-in-the-loop and professional translations. Of note, human-in-the-loop and professional translations were produced independently by separate linguists to minimize contamination bias. Linguists recorded the total time (in minutes) required to complete translations, excluding internal administrative procedures, like pre-processing and production queues.

Evaluator types

We recruited three evaluator types to rate translation performance: (1) bilingual clinicians, (2) medically-certified linguists (i.e., professional translators), and (3) bilingual family caregivers. All evaluators had native fluency in the target language and were at least professionally fluent in English. Clinician evaluators were practicing physicians and nurses with backgrounds in pediatric specialties. Linguist evaluators were contracted from a different third-party translation service (Multilingual Connections; Evanston, IL) than the vendor generating the professional and human-in-the-loop translations. We partnered with the research and hospital family advisory committee to identify caregiver evaluators. We assessed evaluator demographics, language use, and proficiency of evaluators with an adaptation of the Short Acculturation Scale for Hispanics (SASH)[37].

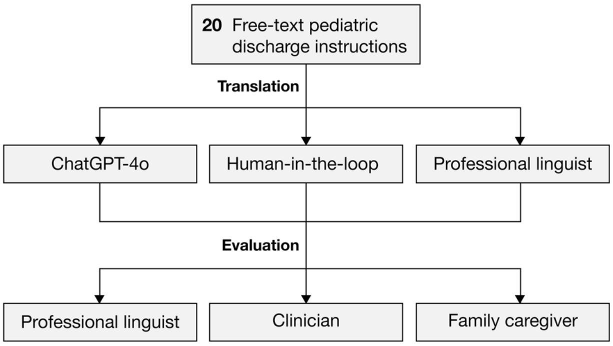


Fig. 4 | Study design and evaluation methodology for each language.

Translation evaluations

Each evaluator rated translations using a validated framework for machine translation performance. The instrument consists of five linguistic domains: (1) *adequacy*, or preservation of information, (2) *fluency*, or preservation of grammar and readability, (3) *meaning*, or preservation of connotation or intent, (4) *severity*, or risk of clinical harm or error, and (5) *overall quality*[12,13,16]. Evaluators scored domains along a 5-point Likert scale (1–5; higher values indicate better performance) (Supplementary Table 3). Linguists scored all 5 domains while caregivers and clinicians only scored the meaning, severity, and overall quality domains, which we deemed to be the most relevant for comprehension and clinical impact. In addition, all evaluators indicated their preferred translation modality, if any, for each set of discharge instruction translations.

Individual evaluator groups participated in a comprehensive training prior to study completion. During the training, study investigators described the objectives of the study and evaluation metrics. Evaluators reviewed and scored example translations as a group until internal consensus was reached.

Evaluators entered ratings into a secure, web-based data collection platform (REDCap; Nashville, TN) with anonymized survey links. We masked the translation source to evaluators and randomized the order in which translations were presented. Translations were accompanied by the original English text for comparison. Domain-level ratings and translation source preferences were reported independently without internal discussion or arbitration.

Statistical analysis

Evaluator and source text characteristics were summarized with descriptive statistics. For each language, we aggregated the mean Likert rating (1–5) across evaluator groups for the adequacy (linguist evaluator only), fluency (linguistic evaluator only), meaning, severity, and overall quality domains. We used the non-parametric Friedman test to assess differences in mean domain-level ratings across translation modalities to account for the within-subjects, repeated measures design. Where the Friedman test was statistically significant, we conducted post hoc pairwise comparisons using the Wilcoxon signed-rank test to compare ChatGPT-4o and human-in-the-loop translations with professional translations. Each domain was treated as an independent outcome. We did not apply a correction for multiple comparisons given the relatively small number of distinct, pre-specified comparisons and a priori hypotheses regarding machine translation performance. Translation source preference was described with proportions (with 95% confidence intervals [CI]). We used two-tailed t-tests to compare time to completion between human-in-the-loop and professional translations. Interrater reliability was measured with the intraclass correlation coefficient (ICC) with a two-way mixed effects model. We interpreted the ICC in accordance with accepted classifications: Less than 0.5 indicates poor reliability; between 0.5 and 0.75, moderate reliability; between 0.75 and 0.9, good reliability; and greater than 0.9, excellent reliability[38]. The threshold for statistical significance was $p < 0.05$. All analyses were conducted in R statistical software, version 4.3.1 (R Project for Statistical Computing) using the following packages: dplyr, rstatix, tidyverse, and interpretCI.

Data availability

The data that support the findings of this study are available and may be provided from the authors upon reasonable request.

Received: 3 June 2025; Accepted: 2 October 2025;
Published online: 24 October 2025

References

1. United States Census Bureau. *What Languages Do We Speak in the United States?* https://www.census.gov/library/stories/2022/12/languages-we-speak-in-united-states.html (2022).
2. Khan, A. et al. Association between parent comfort with English and adverse events among hospitalized children. *JAMA Pediatr.* **174**, e203215 (2020).
3. Castro, M. R. H. et al. The association of limited English proficiency with morbidity and mortality after trauma. *J. Surg. Res.* **280**, 326–332 (2022).
4. Manuel, S. P., Nguyen, K., Karliner, L. S., Ward, D. T. & Fernandez, A. Association of Eenglish language proficiency with hospitalization cost, length of stay, disposition location, and readmission following total joint arthroplasty. *JAMA Netw. Open* **5**, e221842 (2022).
5. Lion, K. C. et al. Association between language, serious adverse events, and length of stay among hospitalized children. *Hosp. Pediatr.* **3**, 219–225 (2013).
6. Choe, A. Y. et al. Improving discharge instructions for hospitalized children with limited English proficiency. *Hosp. Pediatr.* **11**, 1213–1222 (2021).
7. Platter, E. et al. Completeness of written discharge guidance for English- and Spanish-speaking patient families. *Hosp. Pediatr.* **9**, 516–522 (2019).
8. Diamond, L. C., Wilson-Stronks, A. & Jacobs, E. A. Do hospitals measure up to the national culturally and linguistically appropriate services standards?. *Med. Care* **48**, 1080–1087 (2010).
9. Davis, S. H. et al. Translating discharge instructions for limited English-proficient families: strategies and barriers. *Hosp. Pediatr.* **9**, 779–787 (2019).
10. Regenstein, M. & Andres, E. Hospital language service programs: a closer look at translation practices. *J. Health Care Poor Underserved* **25**, 2003–2018 (2014).
11. Office for Civil Rights, Office of the Secretary, Department of Health and Human Services & Centers for Medicare and Medicaid Services. *Nondiscrimination in Health Programs and Activities* (2024).
12. Brewster, R. C. L. et al. Performance of ChatGPT and Google translate for pediatric discharge instruction translation. *Pediatrics* **154**, e2023065573 (2024).
13. Khanna, R. R. et al. Performance of an online translation tool when applied to patient educational material. *J. Hosp. Med.* **6**, 519–525 (2011).
14. Chen, X., Acosta, S. & Barry, A. E. Evaluating the accuracy of Google translate for diabetes education material. *JMIR Diab.* **1**, e3 (2016).
15. Rodriguez, J. A., Fossa, A., Mishuris, R. & Herrick, B. Bridging the language gap in patient portals: an evaluation of Google translate. *J. Gen. Intern. Med.* **36**, 567–569 (2021).
16. Taira, B. R., Kreger, V., Orue, A. & Diamond, L. C. A pragmatic assessment of Google Translate for emergency department instructions. *J. Gen. Intern. Med.* **36**, 3361–3365 (2021).
17. Khoong, E. C., Steinbrook, E., Brown, C. & Fernandez, A. Assessing the use of Google translate for Spanish and Chinese translations of emergency department discharge instructions. *JAMA Intern. Med.* **179**, 580–582 (2019).
18. Magueresse, A., Carles, V. & Heetderks, E. Low-resource languages: a review of past work and future challenges. Preprint at https://doi.org/10.48550/ARXIV.2006.07264 (2020).
19. Fan, A. et al. Beyond English-centric multilingual machine translation. Preprint at https://dl.acm.org/doi/abs/10.5555/3546258.3546365 (2020).
20. Health and Human Services Office for Civil Rights. *Language Access Annual Report*. https://www.hhs.gov/sites/default/files/language-access-report-2023.pdf (2023).
21. Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J. & Fernández-Leal, Á Human-in-the-loop machine learning: a state of the art. *Artif. Intell. Rev.* **56**, 3005–3054 (2023).
22. Wu, X. et al. A survey of human-in-the-loop for machine learning. *Future Gener. Comput. Syst.* **135**, 364–381 (2022).
23. Biro, J. M. et al. Opportunities and risks of artificial intelligence in patient portal messaging in primary care. *NPJ Digit. Med.* **8**, 222 (2025).
24. Ohde, J. W., Rost, L. M. & Overgaard, J. D. The burden of reviewing LLM-generated content. *NEJM AI* **2**, (2025).
25. Lion, K. C., Lin, Y.-H. & Kim, T. Artificial intelligence for language translation: the equity is in the details. *JAMA* https://doi.org/10.1001/jama.2024.15296 (2024).
26. Bakken, S. AI in health: keeping the human in the loop. *J. Am. Med. Inform. Assoc.* **30**, 1225–1226 (2023).
27. Mello, M. M. & Guha, N. Understanding liability risk from using health care artificial intelligence tools. *N. Engl. J. Med.* **390**, 271–278 (2024).
28. Goh, E. et al. GPT-4 assistance for improvement of physician performance on patient care tasks: a randomized controlled trial. *Nat. Med.* https://doi.org/10.1038/s41591-024-03456-y (2025).
29. Embí, P. J., Rhew, D. C., Peterson, E. D. & Pencina, M. ichaelJ. Launching the Trustworthy and Responsible AI Network (TRAIN): a consortium to facilitate safe and effective AI adoption. *JAMA* https://doi.org/10.1001/jama.2025.1331 (2025).
30. Meskó, B. Prompt engineering as an important emerging skill for medical professionals: tutorial. *J. Med. Internet Res.* **25**, e50638 (2023).
31. Ray, M. et al. Evaluating a large language model in translating patient instructions to Spanish using a standardized framework. *JAMA Pediatr.* https://doi.org/10.1001/jamapediatrics.2025.1729 (2025).
32. Warraich, H. J., Tazbaz, T. & Califf, R. M. FDA perspective on the regulation of artificial intelligence in health care and biomedicine. *JAMA* **333**, 241 (2025).
33. NLLB Team et al. Scaling neural machine translation to 200 languages. *Nature* **630**, 841–846 (2024).
34. Hicks, S. A. et al. On evaluation metrics for medical applications of artificial intelligence. *Sci. Rep.* **12**, 5979 (2022).
35. Vieira, L. N., O'Hagan, M. & O'Sullivan, C. Understanding the societal impacts of machine translation: a critical review of the literature on medical and legal use cases. *Inf., Commun. Soc.* **24**, 1515–1532 (2021).
36. OpenAI. ChatGPT. https://chatgpt.com/
37. Ellison, J., Jandorf, L. & Duhamel, K. Assessment of the Short Acculturation Scale for Hispanics (SASH) among low-income, immigrant Hispanics. *J. Cancer Educ.* **26**, 478–483 (2011).
38. Portney, L. & Watkins, M. *Foundations of Clinical Research: Applications to Practice* (*3rd edn*) (Prentice Hall, 2009).

Acknowledgements

We are grateful for the support of the linguists, clinicians, and family caregivers who participated in the study, with particular thanks to the Armenian American Medical Society, Bangladesh Medical Association of North America, Somali American Medical Society, and thes I-PASS LISTEN Patient and Family Centered I-PASS LISTEN Study. This project was supported by the Patient-Centered Outcomes Research Institute (PCORI), AD-2021C3-24848 (PI Khan) and by the Lucile Packard Foundation for Children's Health (Tse). The views expressed herein are those of the authors and do not necessarily represent those of the funding sources.

Author contributions

R.B. contributed to study conceptualization, methodology, formal analysis, investigation, and writing and revision of the manuscript. G.T., A.F., M.E., M.N., P.G., and A.H. contributed to study methodology, formal analysis, data curation, project administration, and critical revision of the manuscript. C.C., M.H., A.G., M.H., H.H., O.I., L.K., Q.L., M.D., M.N., A.O., I.E., M.S., G.S., C.S., and R.T. contributed subject matter expertise and critical revision of the manuscript. M.R., H.L., J.H., and N.S. contributed to study methodology, formal analysis, and critical revision of the manuscript. N.K. and A.K.

contributed to study conceptualization, methodology, formal analysis, supervision, and writing and revision of the manuscript. All authors reviewed and approved the submitted version of the manuscript and agreed to be accountable for their contributions.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-025-02055-6.

Correspondence and requests for materials should be addressed to Ryan CL Brewster.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

[1]Division of General Pediatrics, Boston Children's Hospital, Boston, MA, USA. [2]Department of Pediatrics, Harvard Medical School, Boston, MA, USA. [3]Department of Neonatology, Beth Israel Deaconess Medical Center, Boston, MA, USA. [4]Department of Pediatrics, Boston Medical Center, Boston, MA, USA. [5]Department of Pediatrics, Stanford University School of Medicine, Stanford, CA, USA. [6]New York University College of Global Public Health, New York City, NY, USA. [7]Kaiser Permanente Bernard J. Tyson School of Medicine, Pasadena, CA, USA. [8]Department of Pediatrics, State University of New York Downstate Medical Center, Brooklyn, NY, USA. [9]Division of Pulmonary Medicine, New York University Langone Grossman School of Medicine, New York City, NY, USA. [10]Department of Allergy and Immunology, Children's Hospital of Richmond at Virginia Commonwealth University, Richmond, VA, USA. [11]Department of Psychiatry and Behavioral Health, The Ohio State University Wexner Medical Center, Columbus, OH, USA. [12]Department of Hematology and Oncology, Cedars-Sinai Medical Center, Los Angeles, CA, USA. [13]School of Nursing, University of Mississippi Medical Center, Jackson, MS, USA. [14]Department of Medicine, Icahn School of Medicine at Mount Sinai, New York City, NY, USA. [15]Avalon University School of Medicine, Willemstad, CW, USA. [16]Department of Family Medicine, University of California Los Angeles School of Medicine, Los Angeles, CA, USA. [17]Division of Pediatric Infectious Diseases, Department of Pediatrics, Boston Medical Center, Boston, MA, USA. [18]Department of Pediatric Critical Care, MercyOne Des Moines Medical Center, Des Moines, IA, USA. [19]Division of Infectious Diseases, Department of Pediatrics, Boston Children's Hospital, Boston, MA, USA. [20]Division of Allergy, Immunology and Rheumatology, University of California Los Angeles School of Medicine, Los Angeles, CA, USA. [21]Providence St. John Medical Center, Santa Monica, CA, USA. [22]Biostatistics and Research Design (BARD) Center, Boston Children's Hospital, Boston, MA, USA. [23]Interpreter Services, Boston Children's Hospital, Boston, MA, USA. [24]Division of Hospital Medicine, Department of Pediatrics, St. Christopher's Hospital for Children, Drexel University College of Medicine, Philadelphia, PA, USA. [25]These authors contributed equally: Nicholas Kuzma, Alisa Khan. ✉e-mail: ryan.brewster@childrens.harvard.edu