Review

Exploring Artificial Intelligence Systems for Language Translation of Health Information: An Environmental Scan

Sarah Elliott , Samantha Guitard , Emily Vandermeer , Lisa Hartling *

Posted Date: 27 February 2026

doi: 10.20944/preprints202602.1740.v1

Keywords: artificial intelligence; large language models; machine translation; health information

 doi:10.20944/preprints202602.1740.v1

Review

Exploring Artificial Intelligence Systems for Language Translation of Health Information: An Environmental Scan

Sarah Elliot [1], Samantha Guitard [1], Emily Vandermeer [2] and Lisa Hartling [1,*]

[1] Alberta Research Centre for Health Evidence, Department of Pediatrics, University of Alberta

[2] Department of Computing Science, University of Alberta

* Correspondence: hartling@ualberta.ca; Tel.: 1-780-492-6124

Abstract

The proliferation of artificial intelligence (AI) systems presents an opportunity for impacts in the automated translation of health information. Several types of systems are available, including machine translation tools, large language models, computer-assisted translation platforms, and specialized/localized platforms. Cost, usability, specific language coverage, and performance (e.g., accuracy) need to be considered when selecting an AI system for language translation. We conducted an environmental scan to identify studies that evaluated the performance of AI for language translation of patient-facing health information. We identified 19 studies that met our inclusion criteria, all of which translated resources from English to other languages and evaluated a variety of outcomes using various AI systems. Most authors concluded that AI systems have strong potential but recommended post-editing by humans, particularly for high stakes situations. Only one study involved the public, and none involved patients, which should be a priority for future research in this area.

Keywords: artificial intelligence; large language models; machine translation; health information

Introduction

Artificial intelligence (AI), particularly in the form of machine learning and natural language processing, has emerged as a transformative tool in the context of language translation. AI-powered systems like Google Translate, DeepL, and other large language models (LLMs) have significantly improved the speed, scalability and accessibility of multilingual communication (1, 2). Google Translate was publicly launched in 2006 (3). Since then, extensive research has focused on improving Machine Translation (MT) tools in terms of linguistic accuracy. In 2017, DeepL was released which relied on more complex AI systems to perform translations; when released it outperformed all other competitors (4).

In subsequent years the landscape of MT tools has changed once again with the release of a multitude of LLMs. LLMs are next word predictors, meaning that these models are trained on incredibly large corpuses of data to be able to string together output by predicting the next word in a sequence (5). However, translating health-related content presents unique challenges, such as navigating complex medical terminology, maintaining cultural relevance, and ensuring compliance with legal standards for privacy, clarity and safety (6-8). While existing AI translation tools demonstrate promise, their accuracy and reliability in the healthcare domain remain variable (9). This highlights a critical gap in current research: the need to systematically explore, evaluate, and optimize AI-based translation tools specifically for health information to ensure equity and clarity in patient communication across languages and regions.

An essential part of advancing research in the area of AI for language translation of health information is selecting an appropriate AI system. Table 1 provides examples of different AI systems

 doi:10.20944/preprints202602.1740.v1

that are currently commercially-available for language translation. The table groups the AI systems based on how they work: non LLM (MT tools), LLM, computer assisted, and translation services. The examples are "off-the-shelf" (i.e., ready-made, immediately available, standardized format) that most users could access (pending costs) to support language translations. Non LLM systems (e.g., Google Translate) involve models that do phrase matching and are trained on bilingual data sets in order to accurately translate information. LLMs (e.g., ChatGPT) are general models that are trained on massive amounts of data in order to understand and replicate natural language in the form of text. This allows for these models to be able to perform many tasks like text generation, question answering, and language translation. There are many LLMs that exist; however, we have presented examples that are most common and accessible. Some LLMs have been designed specifically to complete the task of translation (e.g., DeepL) while others (e.g., ChatGPT) perform multiple functions. Computer assisted translation (CAT; e.g., SmartCAT) are tools designed to support human translators in their tasks of translating texts, such as use of consistent vocabulary, memories of previous corrections, and quality assurance checks. These tools are not designed to replace human translators. Localization platforms (e.g., Lokalise) are designed to ensure websites and other content can be localized to different regions in their language. Many of these professional services that include products like CATs emphasize that humans are involved for part of the translation process.

Some key factors (described in Table 1) influencing the selection of an AI system for translation of health information are cost, ease of use (e.g., whether technical computing knowledge is needed), and the ability of the system to translate specific languages of interest. Another essential factor in selecting an AI system for translation of health information is evidence of its performance for this purpose (e.g., accuracy, validity, reliability, etc.). To inform our own evaluation in terms of choice of AI system, we conducted an environmental scan to identify previous studies that evaluated the functionality, performance and limitations of AI systems for language translation of health information, including recommended approaches for effective use.

Table 1. Characteristics of AI tools available for written language translation.

AI tool	**Cost*†**	**Usability**	**Designed specifically for language translation**	**Languages**
Machine Translation Tools (non-large language models, LLMs)				
Google Translate	Free	App, Webpage, does not need an account	Yes	249 languages (29)
No-Language-Left-Behind (NLLB) by Meta	Free	Requires download of code and computing knowledge to operate	Yes	204 languages (30)
LLMs				
ChatGPT	Free - $200	App/ Web Page	No	85 languages (31)

 doi:10.20944/preprints202602.1740.v1

	monthly (or billed per token)	Need an account		
Claude	Free - $100 monthly (or billed per token)	App/Web Page, need an account	No	Over 100 (not all are listed, however, mentions support for most major languages) (32)
Gemini	Free - $340 per month (CA)	App/Web Page	No	Claims similar language performance to Google Translate with less support for certain languages (33)
Meta AI	Free (or billed per token)	App/ Webpage Do not need an account to test	No	Official list not available, however, Meta AI's other models like seamless communication models translate over 100 languages (34)
Language LLMs				
DeepL	Free - $81 monthly (CDN) (or $7.50 monthly + token pricing)	Need to have an account, App/Website	Yes	32 Languages (35)
Aya	Free	Download the model and run it locally on the computer, has a small playground for testing	Yes	101 Languages (36)
Computer Assisted Translation and Localized Platforms				
SmartCAT	Paid Professional Service	Paid Professional Service	Yes	280+ Languages (37)
Lokalise	Paid	Paid Professional	Yes	677 Languages (38)

 doi:10.20944/preprints202602.1740.v1

	Professional Service	Service		
Text United	Paid Professional Service	Paid Professional Service	Yes	150+ Languages (39)

* Costs in USD unless otherwise noted. † Billed per token means that instead of a flat fee, a person (usually a corporation) is billed per token, which are usually words or parts of words. This is often used when the model is being used as part of a service such as a website that uses an LLM to be their chatbot online.

Methods

Searches

We used three main approaches to searching for relevant studies. First, Google Scholar was searched (in July and August 2025) utilizing combinations of the following key concepts (see Box): AI Translation, Translation, Medical Translation, LLM, Large Language Models, ChatGPT OR Gemini OR Llama OR Claude, Patient Education, Pediatrics, Medicine, Public Health. Given the substantial volume of search results retrieved from Google Scholar, only the first 10 results were reviewed for each combination of terms (n=90 search results). This subset was selected based on Google Scholar's default relevance-ranking algorithm, which prioritizes the most pertinent studies. Relevance was observed to decline markedly beyond this point. This approach ensured a manageable scope for analysis while preserving transparency and reproducibility in the selection process (9). Search results were exported to Excel and duplicates were removed.

Box. Combinations of terms used to search Google Scholar

AI Translation + Public Health

LLM + Translation + Medicine

LLM + Translation + Pediatrics

ChatGPT OR Gemini OR Llama OR Claude + Translation + Patient Education

Large Language Models + Medical Translation

("AI Translation" + "Health") OR ("AI Translation" + "Health Information")

LLM + Translation + Health

LLM + Translation + Health Information

ChatGPT OR Gemini OR Llama OR Claude + Translation + Health Information

Second, we searched PubMed (September 2025) using the terms 'Machine translation + Health'. Information'. The search was limited to the time period of 2014-2025. Third, we hand-searched reference lists of relevant articles.

 doi:10.20944/preprints202602.1740.v1

Study Selection

Studies evaluating the use of AI for language translation were selected for inclusion based on the following criteria: they employed one or more commercially-available AI systems; the content being translated was written health information for the public or standardized patient-facing materials; results were available for the AI system without any post-editing or human revision; and the studies provided empirical data on AI system performance (e.g., accuracy, errors, etc.). These criteria ensured the we focused on real-world applications of AI systems in health communication without the influence of human correction. Some studies included data on the performance of both the AI system alone and with human correction. In these cases, we only extracted results and conclusions for the AI system alone.

The outputs of the Google Scholar searches were screened independently by two authors (EV, LH). Discrepancies regarding inclusion were discussed and resolved between the two authors. Results of the PubMed search were screened by one author (SG) and classified as relevant, not relevant, or unsure. Two authors (SE, LH) reviewed the citations that were classified as unsure to determine relevance.

Data Extraction

Data were extracted on key study characteristics, including the first author, year of publication, study objectives, the type of resource translated, AI translation tools evaluated, any relevant non-AI comparators, languages involved, reported outcomes, results and the authors' conclusions/recommendations. Data were extracted by one author (EV or SG) into Excel spreadsheets. A second author (SE or LH) verified the extracted data.

Data Synthesis

Data are presented in tables and summarized descriptively.

Results

Ninety articles were identified in the Google searches, and an additional 399 from the PubMed search. After removing duplicates, 460 articles remained and were screened. An additional six references were identified through other sources (e.g., reference checking). A total of 19 articles were included (10-28). The characteristics of the included articles are shown in Tables 2 and 3.

Table 2. Characteristics and results of studies evaluating a single AI tool for language translation of patient-facing health information (n=11 studies).

First Author Year Country	Study Objectives	Resource translated	AI tools evaluated Non-AI comparison (where relevant) Languages translated	Outcomes Use of existing or study-specific outcomes (i.e., developed by study authors) Patient/public involvement in evaluation	Summary of results	Authors' conclusions

 doi:10.20944/preprints202602.1740.v1

Almahsees 2021 (10) Jordan	'The present research has chosen (Costa et al., 2015b) framework to assess the MT output of Google Translate in this language pair.'	Information from WHO, UNICEF, U.S. FDA, FCDO, and ECDC about the global pandemic COVID-19 Whole texts (n=628, from n=14 source texts)	Google Translate English to Arabic	Use of Costa et al. (2015) framework – taxonomy for authors to assess translation quality, based on analysis of the number of errors in the whole texts (n=628) including; orthographic, lexical, grammatical, and semantic errors. Existing outcomes No	Orthographic Errors (n=15, 1.46%) • There were no spelling errors • Arabic does not use the same punctuation system and GT made errors in this subcategory Lexical Errors (n=20, 3.90%) • Made omission errors (removal of words) in translation • Made fewer addition errors (adding extra words) Grammatical Errors (n=29, 4.26%) • Several mis selection errors (where it translates the wrong class or word i.e. a noun as a verb) • Contained mis ordering errors (words in the incorrect order) Semantic Errors (n=50, 8.53%) • Confusion of sense errors appeared several times • Several examples of wrong choice and idiom errors	'Google Translate committed a set of errors: semantic, grammatical, lexical, and punctuation. Such errors inhibit the intelligibility of the translated texts.' 'The study recommends that a review by a trained translator should post edit the output of MT systems to ensure the quality

 doi:10.20944/preprints202602.1740.v1

						of the output.'
Chen 2016 (13) _ United States	'The purpose of this study was to evaluate the accuracy of the Google Translate website when translating health information from English to Spanish and English to Chinese.'	Health education sheet for diabetes patients Sentences (n=6)	Google Translate Professional HT English to Spanish, Mandarin Chinese	Evaluation rubric, given to evaluators blinded by translation method, was adapted from Khana et al. (2011) that includes domains of fluency, adequacy, meaning, and severity on 5-point scale (1=low accuracy and 5=high accuracy) Existing outcomes No	GT performed better in the English to Spanish translations than the English to Chinese translations. HT performed better than GT translating English to Chinese, but not better translating English to Spanish. Fluency • Spanish: GT all at least good (scores ≥3), HT all excellent or perfect (scores 4.33-5) • Chinese: 3 sentences by GT had marginal or no fluency (score ≤2). All agreed S5 was not understandable. HT had all excellent or perfect fluency (scores 4.67-5). Adequacy • Spanish: GT and HT conveyed >75% original information (scores 4.33-5), except one sentence by HT conveyed 50% of original information (score 3)	'Google produced a more accurate translation from English to Spanish than English to Chinese.' 'The Chinese human translator provided much more accurate translation than Google did. The Spanish human translator, on the other hand, did not provide a significantly better translati

 doi:10.20944/preprints202602.1740.v1

					• Chinese: One sentence in GT conveyed <50% of original information (score 2.67). HT conveyed almost 100% of original information in all sentences (all scores of scored 5, except one 4.67) Meaning • Spanish: GT and HT both had almost same meaning as original information (scores 4.33-5), except one sentence from HT had partially same meaning as original sentence (score 3) • Chinese: In two sentences GT had less than partially same meaning as original information (score <3). All HT sentences had the same or almost same meaning (all scores of 5, except one 4.67). Severity • Spanish: One sentence by GT had unclear effect (score 4). Same	on compared to Google. Additionally, we identified some sentences translated by Google from English to Chinese that might lead to delayed patient care. Similarly, one sentence translated by the professional human translator from English to Spanish could also have a negative impact on

 doi:10.20944/preprints202602.1740.v1

					sentence by HT delayed necessary care (score 3). Otherwise, GT and HT scores were high (score 4.67-5). • Chinese: Two sentences by GT delayed necessary care for patients (score <3). All HT had no effect on patient care (all scores 5).	patients.'
Das 2019 (14) United States	'This study assesses the accuracy of a popular free machine translation service, in translating AAP anticipatory guidance safety guidelines for the top 20	AAP safety section for 12-month well-child anticipatory guidance handout Statements (n=9 original; translated into n=753 statements total across langua	Google Translate English to Spanish, Arabic, Bengali, Chinese, French, German, Greek, Haitian Creole, Hebrew, Hindi, Italian, Japanese, Korean, Polish, Portuguese, Punjabi, Russian, Tagalog, Urdu, Vietname	HCPs with native proficiency back-translated the MT statements to English, and then all back-translations were assessed by one reviewer. Accuracy was evaluated with a 5-point rubric from American Translators Association (1=minimal [frequent or serious errors] to 5=standard [fully and appropriately translated]) Existing outcomes No	Accuracy • All translations fell short of the professional standard except for Spanish (score 4.95). • Portuguese had a 'strong' translation range (score 4.33) • Arabic, Chinese, French, German, Greek, Haitian Creole, Hebrew, Italian, Japanese, Korean, Polish, Russian, Tagalog, Urdu had deficient to acceptable scores (scores of 2-3) • Bengali, Hindi, Punjabi and Vietnamese received the	'The results of this study suggest that Google Translate yields translations with mostly accurate and unchanged verbiage for only 2 of the 20 most commonly spoken languages in the United

 doi:10.20944/preprints202602.1740.v1

	foreign languages spoken in the United States.'	ges)	se		lowest scores (<2)	States: Spanish and Portuguese. For all other languages analyzed, Google Translate provided translations with obscured or changed meanings. Translations were least useful and weakest for Bengali, Hindi, Punjabi, and Vietnamese, with disruptive or inappropriate wording.'
Khanna	'We conduc	Instruction	Google Translate	Manual evaluation	Fluency • GT was	'We found

 doi:10.20944/preprints202602.1740.v1

2011 (18) United States	ted a pilot evaluation of an online translation tool as it relates to detailed, complex patient educational material.'	manual regarding warfarin use prepared by the AHRQ Sentences (n =45)	Professional HT English to Spanish	composed of five domains: Fluency, Adequacy, Meaning, Severity, and Preference, all measured on a Likert scale from 1 to 5. Evaluators were blinded to the method of translation. For fluency, evaluators did not have access to the original English sentences. Preferences were evaluated on 19 sentence pairs Automated Evaluation: METEOR[1] system, providing a score for each GT translated sentence using the professional translation as our reference. METEOR scores also compared to manual evaluations. Existing outcomes (severity a study-specific outcome) No	significantly worse than prof translations Adequacy • Not significantly different Meaning • Not significantly different Severity • GT had more errors of any severity than prof translations (39% vs 22%, p=0.05), but similar number of serious, clinically impactful errors (4% vs 2%, p=0.61). • One GT translation that was 25 words long and complex was erroneous and "dangerous to patient" (score 1). Evaluators all considered it nonsensical Preference • No overall preference for prof translation (score 3.2) • Greater complexity of sentence was significantly associated with preference for prof translations	that GT was inferior to the professional translation in grammatical fluency but generally preserved the content and sense of the original text. Out of 30 GT sentences assessed, there was one substantially erroneous translation that was considered potentially dangerous. Evaluators

 doi:10.20944/preprints202602.1740.v1

					METEOR • All correlations of the four domains (fluency, adequacy, meaning, severity) with METEOR scores were significant	preferred the professionally translated sentences for complex sentences, but when the English source sentence was simple—containing a single clause—this preference disappeared.'
Khoong 2019 (19) United States	'In this study, we assess the use of GT to translate emergency department (ED) dischar	Emergency department discharge instructions Sentences (n=647, from n=100 sets of instruc	Google Translate English to Spanish, Chinese (type not specified)	Bilingual translators back-translated the instructions to English for assessment. Accuracy was assessed for overall content accuracy (not word-for-word), and coded as a binary outcome. Potential for harm was	Accuracy: • Spanish: 92% of sentences accurately translated • Chinese: 81% of sentences accurately translated Potential harm • Spanish: 2% of sentences (28% of total inaccuracies) had potential for clinically	'Google Translate can be used to translate clinician-entered, patient-specific ED instructions for Spanish- and

 doi:10.20944/preprints202602.1740.v1

	ge instructions into Spanish and Chinese.'	tions)		assessed with an established rating system, and categorized in a binary way: clinically nonsignificant vs clinically significant/life-threatening potential harm. Readability scores were compared to accuracy and potential harm. Existing outcomes No	significant harm • Chinese: 8% of sentences (40% of total inaccuracies) had potential for clinically significant harm • Potential harm was significantly associated with Spanish translations that had a reading level higher than 8th grade.	Chinese-speaking patients. Potential for harm can be minimized by using clear communication practices. We recommend including English instructions and automated warnings regarding the use of machine translation.'
Lomas 2024 (21) United Kingdom	'This study sets out to pragmatically assess the ability of a	Section 1: Adapted patient information leaflet on epidur	ChatGPT-4 Section 1: NA Section 2: HT English to Mandarin	Section 1: Structured grading system, including a 5-point Likert scale (1 to 5) for three domains: fluency/readability (accuracy of language),	Section 1: Fluency • Portuguese scored the highest (score 4.7), followed by Hindi, Spanish, then Mandarin, and Bengali scored the lowest	'At present, whilst LLMs continue to improve, human translators

 doi:10.20944/preprints202602.1740.v1

	LLM to perform healthcare-related translation tasks.'	al analgesia from OAA Words (n=228) Section 2: Prompted ChatGPT to create a patient information sheet on epidurals, using a link to a website containing advice and information.	Chinese, Hindi, Spanish, Bengali, Portuguese	meaning (accuracy of content), and adequacy (overall assessment of translation), assessed by independent interpreters unaware of the source of translation. Section 2: Only fluency/readability was evaluated. Existing outcomes No	(score 2.7) Meaning • Mandarin scored the highest (score 4.7), followed by Hindi, then Spanish, and Bengali did the worst (score 4.0) Adequacy • Hindi and Spanish both got perfect scores (5.0), Portuguese had a high score (score 4.7), and Mandarin and Bengali performed worst (score 4.0 and 3.7 respectively) Section 2: Fluency • Portuguese had a perfect score (5.0), Hindi performed second best, then Mandarin and Bengali, and the worst score was Spanish (score 3.3)	should continue to be the first choice for interpretation services and offer benefits to communication beyond the fluency, accuracy and meaning assessments considered here. However, there are scenarios encountered within healthcare where these services are not available. For the languages

 doi:10.20944/preprints202602.1740.v1

						considered here, this study suggests that GPT-4 is capable of accurately conveying information and supports further research involving patients.'
Ray 2025 (23) United States	'To determine whether GPT-4o can generate high-quality Spanish translations of personalized patient instructions comparable to those	Pediatric patient instructions Words (approx. n=300 for each of 20 source texts)	ChatGPT 4.0 Professional HT English to Spanish	Professional medical translators used the MQM framework to assess the quality of GPT and HT (score 0-100), blinded to translation method. 8 error types deemed most relevant to clinical contexts were used: wrong medical term; mistranslation; addition; omission; grammar,	MQM Scores • Mean score GPT 98.3 vs. HT 96.7 • no significant difference in translation quality between the GPT and HT Translation Errors • Mean number of mistranslation errors of any severity was significantly higher for HT than GPT • No significant difference with other error types Preference • A mean of 52%	'GPT-4o, in its default, publicly accessible configuration, can translate Spanish translations of real-world personalized written patient instructions that

 doi:10.20944/preprints202602.1740.v1

	perfor med by profess ional human translat ors.'			spelling, or punctuation, language register; awkward style; culturally insensitive. Errors were also classified by severity, based on potential for clinical harm: neutral, minor, major, and critical. Preference for translation type was also recorded. Existing outcomes No	preferred or strongly preferred GPT • A mean of 20% preferred or strongly preferred HT • 28% neutral	are compara ble in quality to those generate d by professi onal human translato rs. Indepen dent professi onal medical translato rs preferre d the GPT-4o–generate d translati ons over the human translati ons, and error analysis revealed a higher rate of mistrans lation errors in the human translati ons. These

 doi:10.20944/preprints202602.1740.v1

						findings demonstrate GPT-4o's ability to produce quality translations in Spanish—a language well represented in digital training data—without additional model optimization.'
Reategui-Rivera 2025 (24) United States	'This study evaluated the dual capability of GPT-4o to both summarize and translate English discharge notes into	Discharge summaries from the Medical Transcription Samples database (n=66)	ChatGPT 4.0 from OpenAI API English to Spanish	Each generated Spanish summary was assessed by a bilingual physician using a five-point Likert scale across four dimensions: Completeness, Correctness, Conciseness, and Writing Quality. Existing outcomes No	'For all metrics, responses were predominantly clustered at the higher end (ratings of 4 or 5).' 'Completeness received a rating of 5 in 48 responses (73%), a rating of 4 in 12 responses (18%), and fewer than 10% of responses were 3 or below' 'Correctness and Writing Quality were rated 5 in over 80% of cases' 'Conciseness had the lowest top-score concentration with 37 responses (56%)'	'These findings demonstrate that, under controlled conditions, GPT-4o can generate clinically accurate and linguistically fluent Spanish

 doi:10.20944/preprints202602.1740.v1

	Spanish'					summaries of discharge notes, offering a promising complementary tool for overcoming language barriers in healthcare.'
Taira 2021 (26) United States	'To perform a pragmatic assessment of GT for the written translation of commonly used ED discharge instructions in seven commonly spoken	Common emergency discharge instructions Phrases (n=20 per language, n=400 total)	Google Translate English to Spanish, Chinese (Mandarin and Cantonese), Vietnamese, Tagalog, Korean, Armenian, Farsi	MT scoring rubric, containing a 5-point Likert Scale for each domain: Fluency, Adequacy, Meaning and Severity. Evaluations were done by volunteer community members who speak the translated language, without healthcare or professional interpreter experience. Existing outcomes	'Mean scores for fluency, adequacy, meaning and severity were high running from 4.2-4.4' ... 'but varied by language' 'GT accurately conveyed meaning of 330/400, 82.5% instructions' Accuracy varied by language from 55% (Armenian) to 94% (Spanish) • 'Some of the translation errors reported by the volunteers made the GT translations nonsensical.' • Best to worst accuracy: Spanish>Tagalog >Korean>Chinese >Vietnamese>	'The important implication of our study is that, despite recent reports of improvement in accuracy and the suggestion that GT has a role for use in the clinical setting,

 doi:10.20944/preprints202602.1740.v1

	langua ges.'			Yes	Farsi>Armenian	we found that GT accuracy varies substant ially by languag e and is not yet a reliable tool in the clinical setting. Even for languag es in which the accuracy is high, there is still the potentia l for importa nt inaccura cies and the potentia l for patient harm.'
Turne r 2015 (27) Unite d States	'The purpos e of this study was to investi gate the	Health promot ion docum ents from differe	Google Translate Professio nal HT English to Tradition	Linguistic error analysis to identify MT errors, performed by native Chinese speakers with	'The most common machine translation errors were errors of word sense (40% of all errors) and word order (22% of all errors). MT errors observed and	'Additio nal work is needed to improve the

 doi:10.20944/preprints202602.1740.v1

	feasibility of using machine translation (MT) tools (e.g., Google Translate) followed by human postediting (PE) to produce quality Chinese translations of public health materials.' (post-editing results not of interest for this scan)	nt public health agencies in the United States (n=60)	al Chinese	formal training in linguistics. A categorization scheme for MT errors was developed, and all MTs were annotated based on this scheme by a native Chinese speaker with formal training in linguistics. Subsequently, aggregate error statistics were computed to gain insights into the most frequent error categories: word sense, word order, missing word, superfluous word, orthography/ punctuation, particle error, untranslated word, pragmatic error, and other grammar error. Study-specific (categorization scheme for errors) No	percentage of all errors: word sense (40%), word order (22%), missing word (16%), superfluous word (14%), other grammar error (3%), orthography/punctuation (3%), untranslated word and pragmatic errors (<0.5% each).	quality of MT from English to Chinese. Word sense and word order errors require the most attention for improvement.' 'Chinese machine translations have a different relative frequency of certain error types and lower quality overall. Compared to our previous studies on English-Spanish [8,13], we

 doi:10.20944/preprints202602.1740.v1

						found that the Chinese translations had high percentages of word order and word sense errors, which require more cognitive effort to correct [14–16].' 'Translation between English and Chinese presents a challenge due to very divergent syntactic structures (e.g., topic-comment structure in

 doi:10.20944/preprints202602.1740.v1

						Chinese vs subject-verb-object structure in English), frequent dropping of pronouns in Chinese, higher degree of morphology in English, and other linguistic differences.'
Ugas 2025 (28) Canada	'This study investigates the feasibility and utility of using machine translation (Google Transla	Patient education pamphlets related to radiation therapy written in plain language (n=5)	Google Translate Professional HT English to Vietnamese, Punjabi, Simplified Chinese, Portuguese, and Spanish	Translators assessed quality of MT vs. HT on domains of fluency, adequacy, meaning, and severity (of possible risk) on a 5-point Likert scale. Translators were blinded to the method of translation and were shown MT or HT versions in random order.	MT vs HT scoring 'Spanish and Vietnamese language pamphlets achieved the highest overall scores. There were significant differences between human and machine translation in favor of the former for all of the languages, although machine translation scored above 3/5 in 90% of the domains tested. There was no correlation between read ability scores and translation scores'	'Google Translate performs well in multiple translation domains despite its continued inferiority relative to

 doi:10.20944/preprints202602.1740.v1

	te) to translate patient education materials'			Three translators evaluated pamphlets in each language to minimize subjective factors related to the stringency or laxity of a given evaluator. Existing outcomes No		professional human translation. The high scoring of machine translated pamphlets, particularly in the most crucial domain of severity however, points to its potential adoption in a limited capacity in healthcare settings, with processes in place, like pre-screening for high-risk content

 doi:10.20944/preprints202602.1740.v1

						that may pose a threat to patient wellbeing.'

Abbreviations: AAP=American Academy of Pediatrics; AI=artificial intelligence; API= application programming interface; ECDC= European Centre for Disease Prevention and Control; FAQ=frequently asked questions; FCDO= Foreign, Commonwealth & Development Office; GPT= generative pre-trained transformer; HT=human translation; LKD=living kidney donation; LLM=large language model; METEOR= MT=machine translation; OPUS; MQM=Multidimensional Quality Metrics; MT=machine translation; NKF= National Kidney Foundation; NHS=National Health Service; OAA= Obstetric Anesthetists' Association; OPTN= Organ Procurement and Transplantation Network; UNICEF= United Nations Children's Emergency Fund; U.S. FDA=United States Food and Drug Administration; WHO=World Health Organization. [1]BLEU score: automated metric, range 0-1, with higher scores indicating better alignment with the reference translation (Block, 2025).[2]METEOR score: automated metric, range 0-1, with higher scores indicating better alignment with the reference translation (Block, 2025).[1]METEOR score: automated metric, range 0-1, with higher scores indicating better alignment with the reference translation (40).

Table 3. Characteristics and results of studies evaluating multiple AI tools for language translation of patient-facing health information (n=8 studies).

First Author Year Country	Study Objectives	Resource translated	AI tools evaluated Non-AI comparison (where relevant)	Languages translated	Outcomes Use of existing or study-specific outcomes (i.e., developed by study authors) Patient/public involvement in evaluation	Summary of results	Authors' conclusions
Andalib 2025	'This study used	American Acade	GPT4, Google Transla	English to Spanis	BLEU[2] scores (accuracy), precision,	Accuracy • Google Translate	'Our findings indicate

 doi:10.20944/preprints202602.1740.v1

(11) United States	the bilingual evaluation understudy (BLEU) method to assess translation quality and investigated the ability of AI to simplify patient education materials (PEMs) in Spanish.'	my of Orthopedic Surgery patient education materials (n=78)	te Professional HT	h	readability and feature analysis evaluated by study authors. Existing outcomes No	significantly better on BLEU scores compared with GPT4; GPT4 showed moderate success with improvement using prompt optimization Precision • Google Translate significantly better on precision scores compared with GPT4 Readability • Readability scores significantly improved after simplification of text with GPT4 compared to original translations, but did not achieve	that while Google Translate provides superior accuracy in translating medical texts, LLMs, such as ChatGPT, demonstrate moderate success and offer significant benefits in simplifying complex medical information into more comprehensible formats.' 'Our recommended dual approach—leveraging Google Translate for accuracy and ChatGPT for simplification—presents a

 doi:10.20944/preprints202602.1740.v1

						targeted grade level that was specified in the prompts Feature analysis • Syllable complexity of original English text is important predictor of translation accuracy for both AI tools (features examined were number of words, average number of words per sentence, total number of syllables, average number of syllables per sentence)	practical solution for enhancing patient education and engagement.'
Brewster 2024 (12)	'We aimed to assess	Pediatric discharge	Google Translate, ChatGP	English to Spanish,	Translation evaluation rubric, given to clinician	Adequacy • Spanish: GT and GPT	'Machine translation platforms have

 doi:10.20944/preprints202602.1740.v1

United States	the performance of Google Translate and ChatGPT for multilingual translation of pediatric discharge instructions'	instructions (n=20)	T-4.0 Professional HT	Brazilian Portuguese, Haitian Creole	evaluators blinded to method of translation, that includes 4 accepted domains of translation accuracy and quality along a 5-point Likert scale: adequacy (preserved information), fluency (grammatical correctness), meaning (preserved connotation), and severity (clinical harm). Readability was compared to domain scores and method of translation. Preferences for which translation, if any, the evaluators preferred was also captured. Existing outcomes No	significantly higher scores than HT. • Portuguese: GPT significantly higher scores than HT; no difference between GT and prof translation. • Creole: HT significantly better than GT and GPT. Fluency • Spanish: GT and GPT significantly higher scores than HT. • Portuguese: no differences. • Creole: HT significantly better than GT and GPT. Meaning • Spanish: GT and GPT	comparable performance to professional translations for Spanish and Portuguese but shortcomings in quality, accuracy, and preference persist for Haitian Creole.'

 doi:10.20944/preprints202602.1740.v1

						significantly higher scores than HT. • Portuguese: no differences. • Creole: HT significantly better than GT and GPT. Severity (clinically significant harm or delay of care) • Spanish: HT significantly lower scores (better) than GPT, but similar scores to GT. • Portuguese: no differences. • Creole: HT significantly better than GT and GPT (both AI tools had significant number of clinically meaningful errors).	

 doi:10.20944/preprints202602.1740.v1

						Readability • Not significantly associated with nearly all domain scores, nor with any translation source preferences. Overall preference • Spanish: GT > GPT > HT • Portuguese: HT > GPT > GT • Creole: HT > GPT = GT • Preference for HT significantly greater than GT and GPT for Portuguese and Creole, but not Spanish.	
Garcia Valen	'This study evaluat	Kidney transplant	ChatGPT 3.5, ChatGP	English to Spanis	Linguistic accuracy and cultural	Linguistic accuracy • Mean	'These high linguistic and

 doi:10.20944/preprints202602.1740.v1

cia 2024 (15) United States	ed ChatGPT's capabilities in translating 54 English kidney transplant frequently asked questions (FAQs) into Spanish using two versions of the AI model, GPT-3.5 and GPT-4.0.'	frequently asked questions sourced from Donate Life America's website (19 FAQs sourced from OPTN, 15 from NHS, and 20 from NKF) Questions (n=54)	T 4.0	h	sensitivity were evaluated by nephrologists with a detailed rubric scale ranging from 1 to 5 (1=lower/poor performance and 5=higher/excellent performance) Existing outcomes No	scores across all questions were: GPT 3.5: 4.89 (0.31) vs. GPT 4.0: 4.94 (0.23), no significant difference in mean scores, regardless of question or source • Comparisons within each GPT version by source were also not significantly different Cultural sensitivity • No significant difference in	cultural sensitivity scores demonstrate Chat GPT effectively translated the English FAQs into Spanish across systems. The findings suggest Chat GPT's potential to promote health equity by improving Spanish access to essential kidney transplant information.'

 doi:10.20944/preprints202602.1740.v1

						mean scores for GPT 3.5 vs. GPT 4.0, regardless of question or source • Mean scores across all questions were: 4.96 (0.19) for both GPT versions Comparisons within each GPT version by source were also not significantly different	
Garcia Valencia 2025 (16) United States	'This study utilized ChatGPT versions 3.5 and 4.0 to translate 27 frequently asked questions	Kidney transplant FAQs sourced from Donate Life America's website Questions (n=27)	ChatGPT 3.5, ChatGPT 4.0	English to Spanish	Linguistic accuracy and cultural sensitivity were evaluated by nephrologists with a detailed rubric scale ranging from 1 to 5 (1=lower/poor performance and 5=higher/excellent	Linguistic accuracy • Mean scores across all questions were: GPT 3.5: 4.89 (0.32) vs. GPT 4.0: 5.00 (0.00), no significa	'ChatGPT 4.0 demonstrates strong potential to enhance health equity by improving Spanish-speaking Hispanic patients' access to LKD informatio

 doi:10.20944/preprints202602.1740.v1

	(FAQs) from English to Spanish, sourced from Donate Life America's website.'				performance) Existing outcomes No	nt differences between versions • Excellent ratings were given for 89% of translations for GPT 3.5 vs. 100% for GPT 4.0, no significant differences between versions Cultural sensitivity • Mean scores across all questions were: 4.89 (0.32) for both GPT versions 89% of translations were of excellent quality for both GPT versions	n through accurate and culturally sensitive translations.'
Haq 2024	'This study's	Neurology	ChatGPT-4	English to	Four metrics were used:	Claude had best performance	'There is room for

 doi:10.20944/preprints202602.1740.v1

(17) United States	purpose is to compare the ability to translate instructional and educational medical documents across multiple languages.'	patient education material Descriptions of neurological conditions (n=5)	Omni (o), Gemini 1.5 Pro, Claude 3.5 Sonnet	Spanish, Urdu, Arabic	accuracy, clarity, comprehensiveness, and readability (6th grade level), with each graded on a Likert scale from 1 to 5. Language- and domain-level performance scores were presented as percentages. Existing outcomes No	(total score 737/900). Gemini was second best (694/900), and lowest score was ChatGPT (669/900). Claude performed best overall for Spanish and Urdu translations, but Gemini had a higher score for Arabic. All LLMS performed better in Spanish versus Urdu and Arabic. Accuracy • Best to worst scores: Claude> Gemini> ChatGPT (all over 75%) • Claude had highest accuracy for Spanish and Arabic. ChatGPT was	improvement in the LLMs capability to translate patient education material into Urdu and Arabic, while the existing capability for translating into Spanish is much more satisfactory .'

 doi:10.20944/preprints202602.1740.v1

						better for Urdu. Clarity • Best to worst scores: Claude> Gemini> ChatGPT (all over 70%) • Claude had the highest clarity for Spanish and Urdu. Gemini was better for Arabic. Comprehensiveness • Best to worst scores: Claude> Gemini> ChatGPT (all over 75%) • Claude had the highest comprehensiveness	

 doi:10.20944/preprints202602.1740.v1

						score for Spanish and Urdu, and tied with Gemini for Arabic. Readability • Best to worst scores: Claude> Gemini> ChatGPT (all over 70%) • Claude had the highest readability for Spanish, tied with ChatGPT for Urdu, and Gemini was best for Arabic.	
Kong 2025 (20) Unite	'We evaluated the translati	Emergency department	ChatGPT-4, Google Transla	English into Spanish,	Professional translators back-translated the	Accuracy • GPT significa	'At the sentence level, both GPT and

 doi:10.20944/preprints202602.1740.v1

d States	on accuracy and potential for harm of ChatGPT-4 and Google Translate in translating from English to Spanish, Chinese and Russian.'	discharge instructions Sentences (n=316, from n=50 sets of instructions)	te	Chinese (type not specified), Russian	instructions to English for assessment. Physicians assessed accuracy (binary variable), and potential for harm (clinically nonsignificant vs significant/potentially life threatening). Performance of GT vs their earlier publication [Khoong et al 2019] was also assessed. Readability scores were compared to accuracy and potential harm. Existing outcomes N	ntly more accurate than GT for Chinese and Russian, but similar in Spanish • Compared to previous study, GT sentence accuracy rates improved significantly for Spanish and Chinese • At instruction set level 16% of Spanish, 24% of Chinese, and 56% of Russian	Google Translate were ≥90% accurate in translating English to Spanish and English to Chinese, but less accurate in translating English to Russian. GPT was more accurate than Google Translate in translating across the three languages, with significant differences for Chinese and Russian.' 'These results support the potential of machine translation tools to mitigate

 doi:10.20944/preprints202602.1740.v1

						GPT translations had at least one inaccuracy • At the instruction set level 24% of Spanish, 56% of Chinese and 66% of Russian GT translations had at least one inaccuracy • Russian low readability scores (>8th grade) were associated with sentence-level GPT	gaps in translation services for low stakes written communication from English to Spanish, while also strengthening the case for caution and for professional oversight in non-low-risk communication.'

 doi:10.20944/preprints202602.1740.v1

						inaccuracies • Inaccurate Chinese sentence translations by GT were significantly associated with low readability Potential harm • At sentence level,≤1% had potentially harmful mistranslations for all languages, and there was no significant difference between GPT and GT	

 doi:10.20944/preprints202602.1740.v1

						• Spanish: 0% of GPT and 6% GT instruction sets had a mistranslation that could cause harm. • Chinese: 4% of GPT and 6% for GT instruction sets had a potentially harmful mistranslation • Russian: 6% of GPT and 6% GT instruction sets had a potentially harmful	

 doi:10.20944/preprints202602.1740.v1

						mistranslation	
Rao 2024 (22) United States	'This study's purpose is to compare ChatGPT *vs.* GT in its ability to accurately translate instructional and educational medical documents across multiple languages.'	Post-operative discharge instructions for two pediatric urology topics Sentences (n=132)	Google Translate, ChatGPT 3.5	English to Spanish, Vietnamese, Russian	Reviewers from the Language Services department, who were blinded to which documents were ChatGPT vs GT, recorded the percentage of incorrect translations, as well as total errors across all sentences. Standardized error categories were also used, including: Flow, Form, Language, Meaning, Style guide, and Terminology, and were categorized as minor, major, or critical. Existing outcomes No	Incorrect Translation • Spanish: GPT 3.8% vs. GT 18.1% • Russian: GPT 35.6% vs. GT 41.6% • Vietnamese: GPT 24.2% vs. GT 10.6% Total Errors • GT 93 errors vs. GPT 84 errors • Most errors came from the Russian Translations (GT 47 vs. GPT 55) • GPT performed better than GT in Spanish (5 vs. 24), but had	'ChatGPT excelled considerably compared to GT for Spanish translation. However, it was outperformed by GT for the Vietnamese translations, and both ChatGPT and GT produced low quality translations in Russian. ChatGPT has an unacceptably high rate of translation error in Vietnamese and Russian and should not be used alone to translate medical

 doi:10.20944/preprints202602.1740.v1

						more errors in Vietnamese (32 vs. 14) Error Types Distribution • GPT: Meaning 46%, Language 26%, Flow 20%, Form 5%, Style Guide 3% • GT: Flow 48%, Meaning 28%, Language 19%, Style Guide 4%, Terminology 1% • GPT had critical error in Vietnamese and Russian where it completely replaced and made up	documents from English to these languages. While it shows promise in the translation of Spanish materials, its utility for additional languages requires further study and it remains unreliable for use without human oversight.'

 doi:10.20944/preprints202602.1740.v1

						information compared to the source. Did not occur with GT.	
Riina 2024 (25) United States	'This work compares the English to Spanish, sourced from Donate Life America's website.'	Information for patients: health topics, patient instructions, lab tests, drug information Sentences (n=15,816)	ChatGPT 3.5 Turbo, ChatGPT 4.0, Aguila, Google Translate	English to Spanish	METEOR[1], BLEU[2], and BERTscore[3] automatic scoring methods used to create an average score for all translated sentences, and three human evaluation metrics (for 100 sentences) were rated by Spanish interpreters on a scale from 1 to 5: fluency (correct translation), adequacy (convey original meaning), patient-friendliness (patient can easily understand) scores. Qualitative feedback was also reported	Automated evaluation • Overall, GPT and GT were the top two performing models, all other models were significantly different from each other, and Aguila performed the worst METEOR Scores • GPT4o= GT (not significant differences between versions BLEU Scores • GT=GPT	'The automated evaluation results demonstrate that GPT3.5 and GPT4o perform similarly to GT for medical translation accuracy across all scoring metrics: BLEU, METEOR, and BERTscore (Table 2). Analysis with the Games-Howell nonparametric post hoc test highlights that all three automated

 doi:10.20944/preprints202602.1740.v1

					from the interpreters. Existing outcomes No	4o (not significant difference) > GPT3.5 > Aguila BERTscore Scores • GT=GPT 4o (not significant difference) > GPT3.5 > Aguila Human Evaluation • GPT3.5 Turbo,GPT4o, and GTscored similarly, GPT4o performing slightly better • Augila scored significantly worse in all metrics Fluency	scoring metrics were not significantly different between GT and GPT4o (P = 0.05). GPT3.5 scored slightly, but significantly lower on all three metrics. Aguila performed worse than the other models for all scoring metrics.'

 doi:10.20944/preprints202602.1740.v1

						• GPT4o = GT =GPT3.5 > Aguila Adequacy • GPT4o = GPT3.5 =GT > Aguila Patient-friendliness • GPT4o = GPT3.5 =GT > Aguila Qualitative feedback from scorers • GT, GPT3.5, and GPT4o produce very similar translations, and both GPTs capture and translate meaning as well as GT. • Aguila made several	

 doi:10.20944/preprints202602.1740.v1

						different types of errors quite frequent ly	

Abbreviations: AAP=American Academy of Pediatrics; AI=artificial intelligence; API= application programming interface; BLEU= bilingual evaluation understudy; ECDC= European Centre for Disease Prevention and Control; FAQ=frequently asked questions; FCDO= Foreign, Commonwealth & Development Office; GPT= generative pre-trained transformer; HT=human translation; LKD=living kidney donation; LLM=large language model; METEOR= MT=machine translation; OPUS; MQM=Multidimensional Quality Metrics; MT=machine translation; NKF= National Kidney Foundation; NHS=National Health Service; OAA= Obstetric Anesthetists' Association; OPTN= Organ Procurement and Transplantation Network; UNICEF= United Nations Children's Emergency Fund; U.S. FDA=United States Food and Drug Administration; WHO=World Health Organization. [1]BLEU score: automated metric, range 0-1, with higher scores indicating better alignment with the reference translation (Block, 2025).[2]METEOR score: automated metric, range 0-1, with higher scores indicating better alignment with the reference translation (Block, 2025). [3]BERT score: automated metric, range 0-1, with higher scores indicating higher accuracy (Riina, 2025).

Characteristics of included studies

Studies were published between 2011 and 2025, with the median year of publication being 2024 and the majority (n=12, 63%)(11, 12, 15-17, 20-25, 28) published in 2024 or 2025. Most studies were conducted in the United States (n=16, 84%)(11-20, 22-27) with single studies from Jordan (10), the United Kingdom (21), and Canada (28). Eleven studies examined a single AI system (58%)(10, 13, 14, 18, 19, 21, 23, 24, 26-28), while eight studies compared two or more systems (42%)(11, 12, 15-17, 20, 22, 25). The most common AI system evaluated was Google Translate (n=13, 68%)(10-14, 18-20, 22, 25-28) followed by ChatGPT (n=11, 58%)(11, 12, 15-17, 20-25). Studies using ChatGPT evaluated different models including 3.5 and 4.0. Single studies evaluated Gemini, Claude, and Aguila (17, 25). The resources used for translation across studies covered a wide range of clinical topics, e.g., orthopedics, obstetrics, neurology, radiation, COVID-19, diabetes, etc. The resources in four studies pertained to pediatrics (12, 14, 22, 23), and in three studies the focus was emergency department instructions (19, 20, 26). In all studies, resources were translated from English to other languages. Nine studies (47%)(10, 11, 15, 16, 18, 23-25, 27) evaluated translation into one other language, while 10 studies (53%)(12-14, 17, 19-22, 26, 28) evaluated translations into several languages and up to 20 different languages. The most common languages were Spanish and Chinese, evaluated in 17 (89%)(11-26, 28) and eight (42%)(13, 14, 19-21, 26-28) studies, respectively.

Evaluations of AI systems for language translation of patient-facing health information

A wide range of outcomes were evaluated across studies; further, similar outcomes or concepts were evaluated in different ways. The most common outcome (or concept) was errors (n=5) or accuracy (n=6). Different types of errors were assessed, e.g., orthographic, lexical, grammatical, and semantic. Three studies used automated scores for accuracy including BLEU, METEOR, and BERT. The concepts of adequacy (n=7), fluency (n=7), meaning (n=6), and severity/harm (n=8) were commonly evaluated and many studies referenced the same source for these concepts. Three studies evaluated "preferences", with evaluators being professional translators in two studies and clinicians (native speakers in the language of interest) in one study. One study evaluated "patient-friendliness" based on interpreter scoring, and two studies (by the same author group) evaluated "cultural

 doi:10.20944/preprints202602.1740.v1

sensitivity" by clinicians who were fluent in the language. Only one study involved patients or members of the public in their evaluations. A wide range of outcomes were evaluated across studies; further, similar outcomes or concepts were evaluated in different ways. The most common outcome (or concept) was errors (n=5)(10, 22-24, 27) or accuracy (n=6)(14-17, 19, 20). Different types of errors were assessed, e.g., orthographic, lexical, grammatical, and semantic. Three studies used automated scores for accuracy including BLEU, METEOR, and BERTscore (11, 18, 25). The concepts of adequacy (n=7)(12, 13, 18, 21, 25, 26, 28), fluency (n=7)(12, 13, 18, 21, 25, 26, 28), meaning (n=6)(12, 13, 18, 21, 26, 28), and severity/harm (n=8)(12, 13, 18-20, 23, 26, 28) were commonly evaluated and many studies referenced the same source for these concepts (18). Three studies evaluated "preferences", with evaluators being professional translators in two studies (18, 23) and clinicians (native speakers) in one study (12). One study evaluated "patient-friendliness" based on interpreter scoring (25), and two studies (by the same author group) evaluated "cultural sensitivity" by clinicians who were fluent in the language (15, 16). Only one study involved the public in evaluating Google Translate (26). No studies involved patients in assessments.

Eleven studies evaluated a single AI system: Google Translate (n=8)(10, 13, 14, 18, 19, 26-28) and ChatGPT 4.0 (n=3)(21, 23, 24). These studies evaluated translations for a variety of languages. Five studies that evaluated Google Translate did not recommend it's use (13, 14, 18, 26, 27); these studies were published in 2021 or earlier (median year 2016). Three studies evaluating Google Translate (published in 2019, 2021 and 2025) recommended it's use with either post-editing by human translators (10, 28) or other supports (i.e., patients receive verbal instructions while reading translated instructions)(19). Three studies evaluating ChatGPT 4.0 (published in 2024 and 2025) recommended it's use with caveats, i.e., under controlled conditions, with human translator review, or in cases where human translators are not available (21, 23, 24). Two of these studies evaluated translation from English to Spanish only (23, 24).

Eight studies evaluated multiple AI systems or versions: Google Translate and ChatGPT 4.0 (n=3)(11, 12, 20); Google Translate and ChatGPT 3.5 (n=1)(22); Google Translate, ChatGPT 4.0, ChatGPT 3.5, Aguila (n=1)(25); ChatGPT 4.0 and ChatGPT 3.5 (n=2)(15, 16); ChatGPT 4.0, Gemini and Claude (n=1)(17). These studies were all published in 2024 or 2025. One study (assessing Spanish translations) found Google Translate more accurate than ChatGPT and recommended a dual approach with Google Translate for translation and ChatGPT for simplification of text (11). A second study (assessing Spanish, Brazilian, Portuguese, Haitian, Creole translations) of Google Translate and ChatGPT found comparable performance of both systems to professional translation; however, performance varied by language and authors recommended further validation even in languages with better performance before use in clinical practice (12). A third study (assessing Spanish, Chinese, Russian translations) of Google Translate and ChatGPT showed both systems had good accuracy for Spanish and Chinese but lower accuracy for Russian (20). The authors suggested potential for use in low stakes written communication from English to Spanish but professional oversight in other situations (non-low-risk communication) and languages. A study evaluating Google Translate, ChatGPT 3.5, ChatGPT 4.0, and Aguila (assessing Spanish translations) found that Google Translate, ChatGPT 3.5 and ChatGPT 4.0 were effective although ChatGPT4.0 was the highest performing; Aguila performed worse than the other systems (25). Authors concluded that Google Translate, ChatGPT 3.5, and ChatGPT 4.0 hold promise but more evaluation with real-world clinical tasks is needed. A study (assessing Spanish, Russian, Vietnamese translations) comparing Google Translate and ChatGPT 3.5 showed variability by language (22). Overall, the authors concluded that neither system performed adequately; however, ChatGPT 3.5 could be used for Spanish with human oversight. Two studies (assessing Spanish translations) by the same authors evaluated ChatGPT 3.5 and 4.0 and found high accuracy and cultural sensitivity (these two studies may report on subsets of the same data, email to corresponding author was sent for confirmation); authors concluded that both systems have significant potential and require ongoing development and evaluation across different medical contexts and languages (15, 16). Finally, a study (Spanish, Urdu, Arabic) comparing ChatGPT 4.0, Gemini 1.5 Pro and Claude 3.5 Sonnet found that Claude performed best overall but there was

 doi:10.20944/preprints202602.1740.v1

variability across systems and languages, with systems performing better for Spanish than Urdu and Arabic; authors concluded that all systems have potential (17).

Discussion

This is the first study, to our knowledge, that identifies and synthesizes studies evaluating AI systems for language translation of patient-facing health information. These technologies hold significant promise for improving equitable access to health information, particularly for communities that have long faced barriers in health care due to language inequities

We identified 19 studies evaluating commercially-available AI systems, most commonly Google Translate and ChatGPT. The studies were diverse in terms of clinical contexts and languages with the most common being Spanish and Chinese. Overall, studies generally found AI systems to be potentially useful for language translation in clinical contexts. However, most authors concluded that further validation is needed before widespread clinical use, or AI systems needed to be used with human oversight. Further, results varied across languages underscoring the need for language-specific evaluations. A preferred AI system (e.g., Google Translate vs. ChatGPT) was not clearly identified as performance varied slightly based on outcomes assessed and languages of interest.

Among the included studies, there was a wide range of outcomes evaluated. The most common concepts were related to errors or accuracy. In addition, many studies assessed a combination of fluency, adequacy, meaning, and severity, and referenced a common source for these outcomes (18). Few studies evaluated preference, patient friendliness, and cultural sensitivity. Moreover, only one study assessed these outcomes using the intended users of the health information, that is, patients or members of the public (26), and no studies included other intended users of the health information. While errors were commonly evaluated, the potential implications of errors need to be considered. For instance, some errors (e.g., orthographic) that do not introduce clinical inaccuracies may be acceptable to patients in the interests of delivering information in their native language. Some studies assessed severity in terms of the potential for errors or inaccuracies to result in clinical harm. The included studies offer a suite of outcomes for other researchers to consider in future evaluations.

In addition to a variety of outcomes, studies employed different methods for assessing the outcomes. For example, studies involved clinicians who were fluent in the language of interest, interpreters, and professional translators. In some studies, back-translation of the AI-translated text was done and the back-translations were assessed by individuals who may not have been fluent in the language of interest. It is unknown how variations in the different methods might influence results. This area of research would benefit from standardized approaches to ensure valid results and findings that are comparable across studies.

While this study provides an important foundation for future evaluations of AI-assisted language translation of patient-facing health information, we acknowledge some limitations. This is not a comprehensive identification of all AI systems; our focus was on those that are commercially available. We only searched for studies using Google Scholar and PubMed, and through reference screening; therefore, relevant studies may have been missed. However, we did appear to reach saturation with different combinations of key words. A potential concern for the literature in this area overall is reporting biases including publication bias, i.e., increased likelihood of publication based on positive results. Finally, the field of AI and its application to health is rapidly evolving and newer systems may perform differently than the results reported here. For example, while writing this report, GPT5 was released.

Conclusions

Numerous studies have evaluated the use of AI systems for language translation of patient-facing health information. Results are promising in terms of the *potential* for AI systems to facilitate translation; yet most authors recommend use of AI with additional supports such as human verification. Across studies, there was no clear preference between the two most commonly evaluated

 doi:10.20944/preprints202602.1740.v1

systems – Google Translate and ChatGPT. Variation in performance of AI systems based on language was observed and needs to be considered in future evaluations and recommendations for use. Consistency in terms of outcomes and methods of evaluation will assist with comprehensive understanding of the strengths and limitations of AI systems for language translation of health information. We only identified one study involving a public sample for evaluations, and none including patients. Therefore, future studies should involve these end users in both their design and outcome assessments.

Acknowledgments: Dr. Lisa Hartling is supported by a Canada Research Chair in Knowledge Synthesis and Translation. All Authors have no conflicts of interest to declare.

References

1. Siddiqua S. Revolutionizing communication: Overcoming language barriers with AI and NLP. *Journal of Nonlinear Analysis and Optimization*. 2023;14(2).
2. PJ MM. Assessing the Effectiveness of Mobile Translation Applications in Cross-Lingual Communication: A Content Analysis Study. *Journal of Media & Management*. 2025;7(11):1–7.
3. Och F. Statistical Machine Translation Live. 2006. Available from: https://research.google/blog/statistical-machine-translation-live/ [Accessed Dec 15 2025].
4. Faes F. Linguee's Founder Launches DeepL in Attempt to Challenge Google Translate. *Slatore.* 2017. Available from: https://slator.com/linguees-founder-launches-deepl-attempt-challenge-google-translate/ [Accessed Dec 15 2025].
5. Stryker C. What are large language models (LLMs)? *IBM.* 2025. Available from: https://www.ibm.com/think/topics/large-language-models#:~:text=How%20Large%20Language%20Models%20Work,representations%20from%20layer%20to%20layer. [Accessed Dec 22 2025].
6. Heath M, Hvass AMF, Wejse CM. Interpreter services and effect on healthcare - a systematic review of the impact of different types of interpreters on patient outcome. *J Migr Health.* 2023;7:100162.
7. Khalid N, Qayyum A, Bilal M, Al-Fuqaha A, Qadir J. Privacy-preserving artificial intelligence in healthcare: Techniques and applications. *Comput Biol Med.* 2023;158:106848.
8. Al Shamsi H, Almutairi AG, Al Mashrafi S, Al Kalbani T. Implications of Language Barriers for Healthcare: A Systematic Review. *Oman Med J.* 2020;35(2):e122.
9. Genovese A, Borna S, Gomez-Cabello CA, et al. Artificial intelligence in clinical settings: a systematic review of its role in language translation and interpretation. *Ann Transl Med.* 2024;12(6):117.
10. Almahasees Z MS, Albudairi Y. Evaluation of google translate in rendering English COVID-19 texts into Arabic. *Journal of Language and Linguistic Studies.* 2021;17.
11. Andalib S, Spina A, Picton B, Solomon SS, Scolaro JA, Nelson AM. Using AI to Translate and Simplify Spanish Orthopedic Medical Text: Instrument Validation Study. *JMIR AI.* 2025;4:e70222.
12. Brewster RCL, Gonzalez P, Khazanchi R, et al. Performance of ChatGPT and Google Translate for Pediatric Discharge Instruction Translation. *Pediatrics.* 2024;154(1).
13. Chen X, Acosta S, Barry AE. Evaluating the Accuracy of Google Translate for Diabetes Education Material. *JMIR Diabetes.* 2016;1(1):e3.
14. Das P, Kuznetsova A, Zhu M, Milanaik R. Dangers of Machine Translation: The Need for Professionally Translated Anticipatory Guidance Resources for Limited English Proficiency Caregivers. *Clinical Pediatrics (Philadelphia).* 2019;58(2):247–9.
15. Garcia Valencia OA, Thongprayoon C, Jadlowiec CC, et al. AI-driven translations for kidney transplant equity in Hispanic populations. *Scientific Reports.* 2024;14(1):8511.
16. Garcia Valencia OA, Thongprayoon C, Jadlowiec CC, et al. Advancing health equity: evaluating AI translations of kidney donor information for Spanish speakers. *Frontiers in Public Health.* 2025;13.
17. Haq M, Ur Rehman MM, Derhab M, Saeed R, Kalia J. Bridging Language Gaps in Neurology Patient Education Through Large Language Models: a Comparative Analysis of ChatGPT, Gemini, and Claude. *medRxiv.* 2024:2024.09.23.24314229.

 doi:10.20944/preprints202602.1740.v1

18. Khanna RR, Karliner LS, Eck M, Vittinghoff E, Koenig CJ, Fang MC. Performance of an online translation tool when applied to patient educational material. *Journal of Hospital Medicine.* 2011;6(9):519–25.
19. Khoong EC, Steinbrook E, Brown C, Fernandez A. Assessing the Use of Google Translate for Spanish and Chinese Translations of Emergency Department Discharge Instructions. *JAMA Internal Medicine*. 2019;179(4):580–2.
20. Kong M, Fernandez A, Bains J, et al. Evaluation of the accuracy and safety of machine translation of patient-specific discharge instructions: a comparative analysis. *BMJ Quality and Safety*. 2025.
21. Lomas A, Broom MA. Large language models for overcoming language barriers in obstetric anaesthesia: a structured assessment. *International Journal of Obstetric Anesthesia.* 2024;60:104249.
22. Rao P, McGee LM, Seideman CA. A Comparative assessment of ChatGPT vs. Google Translate for the translation of patient instructions. *Journal of Medical Artificial Intelligence.* 2024;7.
23. Ray M, Kats DJ, Moorkens J, et al. Evaluating a Large Language Model in Translating Patient Instructions to Spanish Using a Standardized Framework. *JAMA Pediatrics.* 2025;179(9):1026–33.
24. Reategui-Rivera CM, Finkelstein J. Leveraging Generative AI to Overcome Language Barriers in Healthcare. *Studies in Health Technology and Informatics.* 2025;328:86–90.
25. Riina N, Patlolla L, Hernandez Joya C, Bautista R, Olivar-Villanueva M, Kumar A. An Evaluation of English to Spanish Medical Translation by Large Language Models. *Association for Machine Translation in the Americas Conference Proceedings.* 2024.
26. Taira BR, Kreger V, Orue A, Diamond LC. A Pragmatic Assessment of Google Translate for Emergency Department Instructions. *Journal of General Internal Medicine*. 2021;36(11):3361–5.
27. Turner AM, Dew KN, Desai L, Martin N, Kirchhoff K. Machine Translation of Public Health Materials From English to Chinese: A Feasibility Study. *JMIR Public Health Surveillance*. 2015;1(2):e17.
28. Ugas M, Calamia MA, Tan J, et al. Evaluating the feasibility and utility of machine translation for patient education materials written in plain language to increase accessibility for populations with limited english proficiency. *Patient Education and Counselling*. 2025;131:108560.
29. Wikipedia contributors. Google Translate. *Wikipedia.* 2025. Available from: https://en.wikipedia.org/wiki/Google_Translate. [Accessed Dec 22 2025].
30. Costa-Jussà MR, Cross J, Çelebi O, et al. No language left behind: Scaling human-centered machine translation. *arXiv Preprint.* 2022: 220704672.
31. BotPress. List of languages supported by ChatGPT. *BotPress.* 2023. Available from: https://botpress.com/blog/list-of-languages-supported-by-chatgpt.[Accessed Aug 11 2025].
32. Claude AI Hub. Claude AI's Global Reach: Multilingual Capabilities and Supported Languages of Claude 3. *Claude AI Hub.* 2024. Available from: https://claudeaihub.com/claude-ai-languages/#gsc.tab=0. [Accessed Aug 11 2025].
33. Wiik L. Google's Gemini Pro — How Multilingual is it? Medium2024 [Available from: https://medium.com/@lars.chr.wiik/googles-gemini-pro-how-multilingual-is-it-c88ed07d0857.
34. MetaAI. A family of AI research models that enable more natural and authentic communication across languages. *Meta*. 2025. Available from: https://ai.meta.com/research/seamless-communication/. [Accessed Aug 11 2025].
35. DeepL Help Center. DeepL Translator languages. *DeepL.* 2025. Available from: https://support.deepl.com/hc/en-us/articles/360019925219-DeepL-Translator-languages. [Accessed Aug 11 2025].
36. Üstün A, Aryabumi V, Yong Z, et al. Aya model: An instruction finetuned open-access multilingual language model. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*. 2024;1.
37. SmartCat Languages. Supported Languages. *SmartCat.* 2025. Available from: https://smartcat.com/Home/Languages. [Accessed Aug 11 2025].
38. Github Gist. Lokalise Languages. *GitHub.* 2022. Available from: https://gist.github.com/bodrovis/383df4c6cc16834a36ccf5fecf212ac3/stargazers. [Accessed Aug 11 2025].

 doi:10.20944/preprints202602.1740.v1

39. Alen C. Available languages in TextUnited. *Text United*. 2023. Available from: https://textunitedhelp.zendesk.com/hc/en-us/articles/14462106281239-Available-languages-in-TextUnited. [Accessed Aug 11 2025].
40. Block P, Schaefer J, Maurer F, Storf H. Quality of Machine Translations in Medical Texts: An Analysis Based on Standardised Evaluation Metrics. *Studies in Health Technology and Information*. 2025;331:63–72.