An Evaluation of English to Spanish Medical Translation by Large Language Models: A Quantitative and Qualitative Analysis

Nicholas Riina* nicholas.riina@icahn.mssm.edu
Likhitha Patlolla* likhitha.patlolla@icahn.mssm.edu
Camilo Hernandez Joya camilo.hernandezjoya@icahn.mssm.edu
Roger Bautista roger.bautista@icahn.mssm.edu
Melissa Olivar-Villanueva melissa.olivar-villanueva@icahn.mssm.edu
Anish Kumar anish.kumar@icahn.mssm.edu
Icahn School of Medicine, Mount Sinai Hospital, New York, 10029, USA

*Both authors contributed equally

Abstract

Medical translation is a critical tool for overcoming the barriers of discordant cultural backgrounds and languages within the healthcare field. Large Language Models (LLMs) that advertise translation and multilingual capabilities, like ChatGPT, pose a newfound solution that could include unique abilities that a typical machine translation (MT) system does not exhibit (e.g. catering a translation for a specific patient, such as a child). This work compares the English to Spanish translation of three LLMs: ChatGPT3.5 Turbo, ChatGPT4o, and Aguila with the performance of Google Translate. Medical Translations were provided by MedlinePlus, a parallel dataset developed by the National Library of Medicine that consists of four categories of information for patients in English and Spanish: health topics, patient instructions, lab tests, and drug information. Each model translated 15,816 sentences which were scored by three automated metrics: BLEU, BERTscore, and METEOR. 100 sentences were also graded by three Spanish interpreters using metrics defined in this paper: Fluency (is the translation correct Spanish), Adequacy (does the translation convey the original meaning), and Patient-friendliness (is the translation written in language that a patient can easily understand). The human evaluated translations were then subject to qualitative analysis that examined frequent errors and word choice. Automated results indicated that Chat-GPT4o performed equivalently to Google Translate, with ChatGPT3.5 not far behind. Human rated scores found both Chat-GPT models to perform statistically similar to Google Translate in all three metrics. Aguila, the only model intended for primarily Spanish and Catalan use, surprisingly performed much worse than the other models. However, qualitative analysis of Aguila's translations reveal the use of terms that may reach a broader audience, rendering the Spanish used more accessible than the other models. It is important, as MT systems are applied to the medical field, that the translations provided by these models are not only factually correct and patient safe, but accessible by vulnerable populations. This work provides an evaluation of the most recent ChatGPT model's medical translations with a comparison to a well-researched system, Google Translate, using verified metrics. Our work also highlights small, yet important disparities between the Spanish use of LLMs with English as a primary language and other LLMs that are intended for Spanish use.

1 Introduction

It is well-understood that in today's increasingly diverse America, the healthcare field must overcome barriers of discordant races, ethnicities, cultures, and languages to deliver high-quality care to all patients. According to the 2020 census, approximately 8.3% of American residents speak English less than "very well" (US Census Bureau, 2020). Metropolitan areas are disproportionately home to a large number of immigrants (29% in New York City), many of them not proficient in English (Profile of the Foreign-Born Population in New York, New York, 2023). Urban settings experience amplified disparities in care for underserved populations, including immigrants, refugees and limited-English proficiency, or LEP, patients. LEP status has been linked to greater health disparities (e.g. via poorer preventative screening) and worse health outcomes (Cheng et al., 2007; Ponce et al., 2006; Shi et al., 2009). In such cities, medical education institutes and academic health centers are a crucial form of advocacy and social justice that address disparities through service-learning mechanisms (Rupert et al., 2022). For instance, several medical schools operate student-run free clinics (SRFC) for uninsured patients in their communities. However, language and medical literacy barriers in these patients present a challenge for trainees to ensure their patients, who often have many chronic conditions, understand their diagnosis, medication regime, necessary lifestyle changes, specialist referrals, etc. (Rupert et al., 2022). For SRFCs and other healthcare settings that deal with a large number of LEP patients, artificial-intelligence (AI) or machine translation (MT)-based solutions present a potential low-cost, convenient, and efficient tool to address language barriers in patients. However, maintaining accurate, patient-friendly translations without compromising medical accuracy is a limiting factor of such translation services. Thus, phone interpretation services such as CyraCom and Pacific Interpreters remain the standard practice for communicating with LEP patients. Unfortunately, phone interpretation itself is limited by potentially poor acoustics, lack of visual cues, and lack of context provided to the interpreter (Cho, 2023).

Large language models (LLMs) are deep-learning algorithms that are trained to accomplish various natural language processing tasks like text classification and text generation, among others. A chatbot, like ChatGPT, is a system which has been optimized for conversation with a user. ChatGPT, along with other LLMs, is reported to able to use multiple languages, and has anecdotally been reported as an effective translator between languages (Achiam et al., 2023). However, there has not been a formal study looking at the medical translation capabilities of LLM chatbots that were not formally trained for translation versus dedicated machine translation algorithms, like Google Translate (GT).

The goal of this research was to quantify the effectiveness of various chatbot LLMs for translating health information from English to Spanish in a patient-friendly manner. At a preliminary stage, we are evaluating the potential of four models with multilingual capabilities to provide translations of English take-home instructions and clinical information into Spanish.

2 Related Work

Automated medical translation has benefited from research in the fields of both machine translation and LLMs.

2.1 LLM Translation

Over the last 5 years, LLMs have been researched as translators and compared with other neural MTs like DeepL and GT (Jiao et al., 2023). Several popular LLMs have been used for research without any fine tuning. Yao et al. compares GPT-3.5-turbo-1106 with LLaMA2-7B alongside GT and NLLB on translating between English and four other languages. For English to Spanish translations, as in this paper, Yao et al. (2023) found GT to have a BLEU score of 42.9, GPT3.5 Turbo to have 47.9, LLaMA2 to have 44.6, and NLLB to show 48.8. Yao interestingly found GPT3.5 to out-perform GT. Hendy et al. (2023) found that GPT3.5, on zero-shot translation, performed slightly worse than Microsoft Translator for a variety of languages, with GPT3.5's BLEU scores ranging from 25.9 (ZH>EN) to 41.0 (RU>EN).

This literature shows that LLMs are competitive translators without any fine tuning or training examples. Brown et al. (2020) also found that GPT model architectures improve in performance with exposure to correct examples

from zero-shot, one-shot, and multi-shot learning.

2.2 Machine Medical Translation

The medical field is especially challenging for translation due to an abundance of medical jargon. Thus, a MT system is tasked with either translating the medical jargon into medical jargon in the target language or explaining the medical term in the target language's common terms. Skianis et al. (2020) shows that BLEU and METEOR scores both improve dramatically for English to French translations by statistical MT and neural MT systems when finetuned with medical terminology datasets. In their study, the medical terminology datasets were constructed from 5 datasets of English and French medical jargon. Pretrained Neural MTs (a pretrained Convolutional Neural Network from fairseq) had an improvement of BLEU score from 42.93 to 53.40 after pretraining with the medical terminology. However, this is unhelpful for low-health literacy patients.

Electronic health records of patients are commonly studied with MT systems as they are a rich source of clinical data (Johnsi Rani et al., 2019; Liu & Cai, 2015; Weng et al., 2019; Zeng-Treitler et al., 2010). Again, linguistics properties of health records are often vastly different from those of conversations between clinicians and patients, which are often the use case for medical translation. Other studies therefore have focused on MT translation of public health education texts (Almahasees et al., 2021; Chen et al., 2017; Das et al., 2019; Dew et al., 2015; Khanna et al., 2011; Kirchhoff et al., 2011; Turner et al., 2015), patient instructions (Lester et al., 2021; Miller et al., 2018; Taira et al., 2021) and general patient-provider communication (Kapoor et al., 2022; Patil & Davies, 2014; Turner et al., 2019). Automatic evaluation was used less often than human evaluation. Results from these studies demonstrate that MTs like GT are somewhat successful at medical translation, though some errors, especially with longer sentences, may relay dangerously inaccurate information.

3 Novel Contributions

To our knowledge, this is the first work that examines the most recent ChatGPT model, GPT4o, on medical translation from English to Spanish. This work also contributes to the literature by comparing neural MTs with translation by LLMs using commonly used automated scoring metrics, and newly applies these metrics to evaluate patient-provider communication. Finally, our study looks at LLMs developed with primarily English usage compared with one LLM that is intended for Spanish chat. Our qualitative analysis finds small yet important differences in the Spanish word choice among different models and highlights areas where medical MTs fall short.

4 Methods

In this section we will discuss the selection and cleaning of the dataset, methods applied for automated scoring and human evaluation, and the models tested and corresponding prompts.

4.1 Dataset

The MedlinePlus English-Spanish corpus encompasses 7,033 articles with information in four categories--health topics (e.g. strokes, diabetes, etc.), patient instructions, lab tests, and drug information--provided by the US National Library of Medicine. The dataset contains free health information for patients in both English and Spanish written in a patient-friendly manner. This corpus is representative of the types of conversations that a clinical Spanish interpreter may encounter. The Spanish articles are exact translations of the English articles and used as reference translations for human and automatic evaluation of all LLM translations.

4.2 Data Preparation

The articles were chosen from each category at random, to

Category	Sentences Translated
Patient instructions	3,014
Health topics	3,289
Lab tests	3,259
Drug information	6,254
Total	15,816

Table 1: Sentences translated for each category of information in the MedlinePlus dataset.

translate a minimum of 3000 lines per category. The total amount of lines translated was proportional to the size of each category. After a file was selected for translation, each sentence was separated and paired with its Spanish counterpart. Files that did not have the same number of sentences between English and Spanish were not used. After the file was parsed into sentences, formatting symbols and speaker designations were stripped.

4.3 Models

Three LLMs models were used. GPT3.5 turbo version gpt-3.5-turbo-0125 and GPT4o version gpt-4o-2024-05-13 (Achiam et al. 2023) These GPT models were selected since they did not require a paid OpenAI subscription and were more accessible to patients and providers. GPT4o is also reported to have translation capabilities. These models were both accessed through the OpenAI API. The prompt used mirrored that in He (2024) and is shown below:

messages= [

{"role": "system", "content": "You are a medical translator. Translate the following into Spanish while preserving the file format"},

{"role": "user", "content": SENTENCE TO TRANSLATE'}]

The third LLM is Aguila, an LLM finetuned with 26 billion tokens of Spanish and Catalan data that was designed for chat in Spanish and Catalan. Aguila was developed by the Barcelona Supercomputing center by finetuning Falcon-7B with a dataset that was approximately 40% Spanish, 40% Catalan, and 20% English (mapama247 2023). 455 million words in the dataset were medical terms. The prompt used for translations is shown below:

'The sentence "SENTENCE TO TRANSLATE'" translated into Spanish is'

The final MT used was GT, a neural MT system based on a transformer architecture. GT is a common benchmark for translation tasks and has been shown to be effective with translating medical Spanish (Khoong et al. 2019). GT was accessed through the Google Translate API and no prompt was used (Googletrans).

4.4 Scoring metrics

Three automated scoring metrics and three human evaluation metrics were used for this paper. The automated scoring metrics used in this paper include BLEU, METEOR, and BERTscore (Papineni et al., 2002; Banerjee & Lavie, 2005; Zhang et al; 2020). The METEOR metric is an n-gram based metric that is proven to correlate better than BLEU with human judgements on sentence-level translations, as it also better accounts for synonyms and morphological variants. BERTscore, a more recently developed metric, uses a pretrained BERT model to assess the cosine similarity between model embeddings of the translation and the reference, better accounting for paraphrases and distant clause dependencies. These three-scoring metrics have been used often when evaluating MTs as evidenced by the metareviews by Zappatore & Ruggeiri (2024) and Dew et. al. (2018).

Human evaluation metrics are still considered best practice despite being subjective and labor-intensive, as it allows application of cultural and contextual knowledge that reference-based methods lack. The human rated metrics we used were adapted from metrics used in the Workshop on Machine Translation. These metrics include Adequacy and Fluency scoring (WMT06-07), relative ranking (WMT07-16), and average score and z score (WMT17). Adequacy and Fluency are scores of translation accuracy and language accuracy, respectively. A high Adequacy score reflects a translation that contains all the semantic meaning of the reference text. A high Fluency score reflects a translation that is grammatically correct. Average and z score are the Fluency and Adequacy rankings after being normalized within each scorer (Harison, 2023).

These metrics were all ordinal and scored on a scale of 1-5. The definitions provided to human scorers are below:

Fluency score: Is it fluent Spanish? 5 is completely fluent, 1 is not fluent at all.

Adequacy score: Does it convey the original meaning? 5 is conveys original meaning perfectly, 1 is doesn't convey original meaning at all.

Patient-friendliness score: Is it written in language that a patient can easily understand? 5 is completely patient-friendly, 1 is not patient-friendly at all.

All the human scorers are interpreters at a student run free clinic associated with the Icahn School of Medicine. Scorers were provided with all translations from one model at a time. After completing all the evaluations, the scorers reported patterns and observations of frequent errors and model differences, which are discussed in the qualitative analysis.

5 Results

The results of MT translation will be presented first as automated metrics, human evaluation metrics, and qualitative analysis respectively. Following will be an analysis of human evaluation quality. Patient-friendliness was a new metric defined in this paper to capture how understandable a medical translation is for the general patient population. This metric is especially important for medical translation where medical terminology provides a unique challenge and patient understanding is especially critical.

5.1 Automated Evaluation Scores

The number of sentences translated per category of information is presented in Table 1. Score distributions from all three metrics were tested for normalcy and equal variance with the Shapiro Wilk Test and Levene's Test respectively. All the data

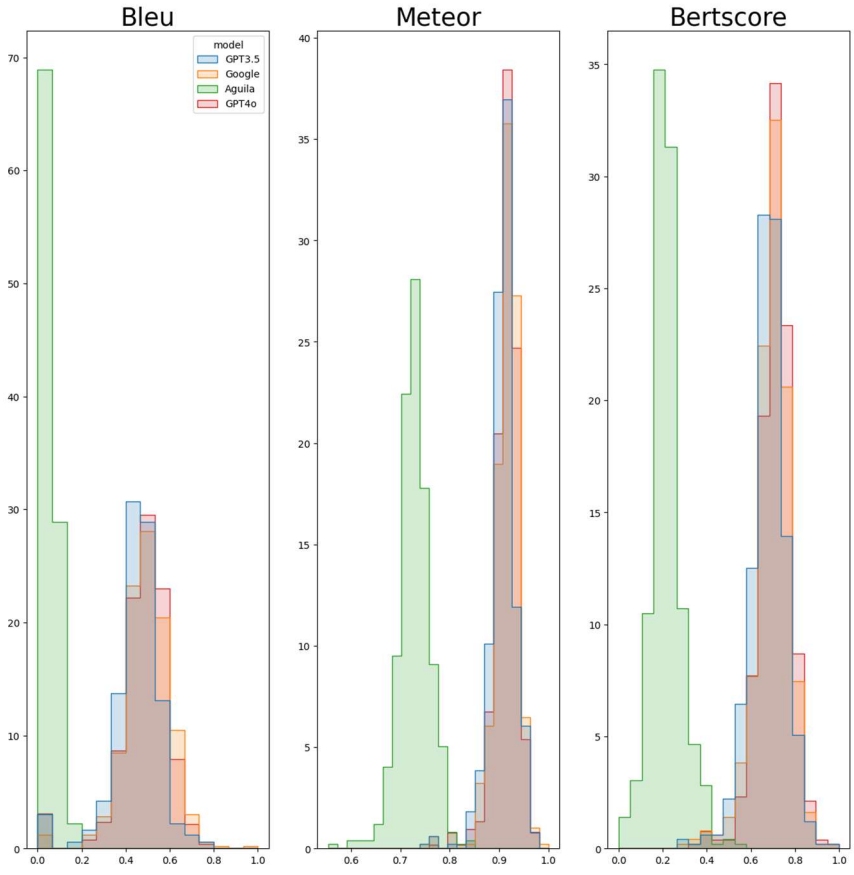


Figure 1: Distribution of Automated Scores by model. Scores for Aguila (green) are significantly lower while the other three models are almost identical. Google Translate is orange, GPT3.5 is blue and GPT4o is red.

Automated Scores		BLEU			METEOR			BERTscore		
Model A	Model B	Mean of Model A	Mean of Model B	P value	Mean of Model A	Mean of Model B	P value	Mean of Model A	Mean of Model B	P value
Aguila	Google Translate	0.0564 +/- 0.029	0.493 +/- 0.111	3.82 e-14	0.215 +/- 0.069	0.697 +/- 0.0828	0.00	0.729 +/- 0.031	0.915 +/- 0.0249	0.00
Aguila	ChatGPT3.5	0.0564 +/- 0.029	0.449 +/- 0.118	9.13 e-14	0.215 +/- 0.069	0.678 +/- 0.084	0.00	0.729 +/- 0.031	0.908 +/- 0.026	3.28 e-13
Aguila	GPT4o	0.0564 +/- 0.029	0.482 +/- 0.120	1.69 e-13	0.215 +/- 0.069	0.709 +/- 0.076	2.18 e-13	0.729 +/- 0.031	0.914 +/- 0.024	5.62 e-13
Google Translate	ChatGPT3.5	0.493 +/- 0.111	0.449 +/- 0.118	9.58 e-09	0.697 +/- 0.0828	0.678 +/- 0.084	1.67 e-03	0.915 +/- 0.0249	0.908 +/- 0.026	2.91 e-05
Google Translate	GPT4o	0.493 +/- 0.111	0.482 +/- 0.120	4.06 e-01	0.697 +/- 0.0828	0.709 +/- 0.076	5.86 e-02	0.915 +/- 0.0249	0.914 +/- 0.024	9.24 e-01
ChatGPT3.5	GPT4o	0.449 +/- 0.118	0.482 +/- 0.120	6.85 e-05	0.678 +/- 0.084	0.709 +/- 0.076	2.39 e-09	0.908 +/- 0.026	0.914 +/- 0.024	2.30 e-04

Table 2: Automated score means, standard deviations, and P values from the Games Howel Post Hoc Significance Test. BERTscore reported as F1 score. The only non-significant difference (p=0.05) is between Google Translate and ChatGPT4o and highlighted in green. The maximum scores are highlighted yellow.

was found to be significantly non-normal and to have significantly non-equal variance with p = 0.05. The score distribution is shown in a set of histograms in Figure 1. The means of each score and significant difference are reported in Table 2. All models were significantly different from each other except GPT4o and GT, the two top performing models. Interestingly, GPT3.5 Turbo and GPT4o are significantly different. Aguila performed much worse than the other models in all scoring metrics.

5.2 Human Evaluation Scores

Due to a strong right-skew in the human scored data (Appendix B), analysis assumed non-normal distributions. The Kruskal-Wallis Test, a non-parametric test for significance between multiple, non-normally distributed distributions of ordinal data, was used. The test was performed for the Fluency, Adequacy, and Patient-friendliness scores to assess differences between the models. These are all less than the alpha (p = 0.05) indicating that there are significant differences between models, which was individually assessed with a Games-Howell post hoc test (Table 3). GPT3.5 Turbo, GPT4o, and GT all scored similarly, with GPT4o scoring slightly better than the other two. Aguila again scored the worst in all metrics.

Human Evaluation Scores		Fluency			Patient-friendliness			Adequacy		
Model A	Model B	Mean of Model A	Mean of Model B	P value	Mean of Model A	Mean of Model B	P value	Mean of Model A	Mean of Model B	P value
Aguila	Google Translate	3.38 +/- 1.43	4.89 +/- 0.37	0.0	2.91 +/- 1.40	4.90 +/- 0.37	8.25 e-14	3.64 +/- 1.45	4.72 +/- 0.59	0.0
Aguila	ChatGPT3.5	3.38 +/- 1.43	4.81 +/- 0.52	4.39 e-14	2.91 +/- 1.40	4.92 +/- 0.31	1.57 e-13	3.64 +/- 1.45	4.76 +/- 0.54	2.02 e-14
Aguila	ChatGPT4o	3.38 +/- 1.43	4.95 +/- 0.25	9.33 e-15	2.91 +/- 1.40	4.97 +/- 0.21	7.92 e-14	3.64 +/- 1.45	4.79 +/- 0.47	0.0
Google Translate	ChatGPT3.5	4.89 +/- 0.37	4.81 +/- 0.52	1.41 e-01	4.90 +/- 0.37	4.92 +/- 0.31	8.13 e-01	4.72 +/- 0.59	4.76 +/- 0.54	8.88 e-01
Google Translate	ChatGPT4o	4.89 +/- 0.37	4.95 +/- 0.25	9.31 e-02	4.90 +/- 0.37	4.97 +/- 0.21	4.66 e-01	4.72 +/- 0.59	4.79 +/- 0.47	2.60 e-02
ChatGPT3.5	ChatGPT4o	4.81 +/- 0.52	4.95 +/- 0.25	2.09 e-04	4.92 +/- 0.31	4.97 +/- 0.21	9.48 e-01	4.76 +/- 0.54	4.79 +/- 0.47	1.02 e-01

Table 3. Human evaluated score means, standard deviation, and P values from the Games Howel Post Hoc Significance Test. The only non-significant differences (p = 0.05) are between Google Translate and both ChatGPT4o and ChatGPT3.5 and is highlighted in green. The maximum scores are highlighted in yellow.

5.3 Human Evaluation Qualitative Analysis

Qualitative feedback from scorers reported that GT, GPT3.5, and GPT4o produce very similar translations, and both GPTs capture and translate meaning as well as GT. All three were also very good at providing patient-friendly translations, provided that the input itself is patient-friendly. One scorer noted that any drop in Patient-friendliness score would be due to the input itself containing some jargon (this may be due to the random selection of individual sentences without their surrounding context). Another scorer noted that the only consistent error made by all three of these models is the dropping of articles in front of certain words, i.e. *glucosa en sangre o azúcar en sangre instead of la glucosa en sangre o el azúcar en sangre.* Aguila would make errors quite frequently, including adding inaccurate information, conjugating incorrectly, including Catalan words, and altering crucial semantic relationships within sentences. Table 3 provides a list of common errors with examples. However, two scorers noted that amongst its few successful translations, Aguila's word choice was more accessible and patient-friendly compared to the other models. For instance, the GPTs and GT used *revestimiento del estómago* in contrast to Aguila's usage of *mucosa estomacal* to translate stomach

Type of Error	Example
Added additional information not present in original sentence	funciona cambiando el nivel de ciertos neurotransmisores en el cerebro", **que se traduce a "cambiando el nivel de neurotransmisores en el cerebro"; es decir no es literal, sino más bien metafórico, ya que el cerebro es una red neuronal y no una "cantidad" de sustancias sino de conexiones y neurotransmisores.**
Added irrelevant commentary in English	"LASIK is unable to permanently change the shape of the cornea. **The translator was not perfect, but his translation is very good**."
Inaccurate translation	""tu puedes prevenir la **gastroenteritis** bebiendo liquido" *should have been* "tu puedes prevenir la **enfermedad por calor** bebiendo liquido"
Conjugation errors	"si se los **ingiera**" *should have been* "si se los **ingiere**"
Formality errors	"**Toma** moxifloxacino" *should have been* "**Tome** moxifloxacino"
Incorrect use of articles	"**la** chance" *should have been* "**el** chance"
Impaired semantic relationships	"**No** se absorbe bien en el estómago vacío y lleno" *should have been* "se absorbe bien en el estómago vacío y lleno"

Table 3. Frequent errors made by Aguila.

lining. *Mucosa* is more descriptive and can be understood even if a person does not know what the stomach lining is, while understanding *revestimiento* is dependent on whether the patient knows this less frequently used term.

GPT3.5 and GPT4o also sometimes used the more patient-friendly term with Aguila, whereas GT consistently used less accessible, more formal terms. For instance, Aguila and GPT3.5 used *la parte inferior de la espalda* and *la parte baja de la espalda*, respectively, instead of *zona lumbar*, which GT used, to translate lower back. GT's word choice is dependent on understanding the names of the zones of the back, which many patients likely do not. Finally, while GT and GPT3.5 use the word *afección* to translate condition, Aguila and GPT4o use *condición*. While *afección* can be used, it has another meaning that means attachment, so the use of this word can be slightly confusing. A more widely understood translation, and the actual direct translation of the word condition, is *condición.*

5.4 Human Evaluation Metrics Validation

To gain insight into the consistency of each scoring metric across judges to judge each metric's validity, we evaluated each scoring metric across judges with intraclass correlation (Appendix A) and visually (Appendix B).

In Appendix A, the ICC was calculated for Random Fixed rates and was reported as a single ICC where each rater is evaluated compared to their own mean, and an average where each rater is evaluated compared to the group mean. The highest ICCs were found with the Adequacy score and with the set containing all the scores. The ICC was recalculated after normalizing each scorer's responses with z-score normalization and all the ICCs increased. The final ICCs were all above 0.7 and were significant with p value = 0.05. Patient-friendliness had the lowest ICC.

6 Discussion

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 non-parametric post hoc test highlights that all three automated 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.

Human evaluation also corroborated the pattern discerned by automated metrics. GPT4o was the top performing model for all categories. Notably, GPT4o scored significantly higher than GT for Adequacy and significantly higher than GPT3.5 in Fluency (Table 3). Otherwise, there were no significant differences in the scores for GT, GPT4o, and GPT3.5. Once again, Aguila performed worse than the other models in all categories. However, out of all metrics, it scored best in patient-friendliness.

Aguila was notably inconsistent with its translation accuracy. Despite some successful translations, the qualitative analysis found that the Spanish model made grammatical errors as well as translation errors. For instance, Aguila often added new information to the sentence and often incorrectly translated semantic relationships (e.g. *This medication can be taken vs This medication cannot be taken*) (Table 3). Both types of errors pose dangers to patients if the information transmitted to the patient is distorted. However, two scorers reported that Aguila occasionally utilized the most patient-friendly lexicon of the three models (e.g. *mucosa estomacal* instead of *revestimiento del estómago*). The lexicon of Aguila in these instances were described as 'more conversational language' and words that are suited for a larger audience. We hypothesize this may result from Aguila's development as a LLM fine-tuned with mostly Spanish/Catalan as opposed to an LLM used predominantly in English that is able to translate into other secondary languages like ChatGPT. More research is required to identify why the existences between the word choice of Aguila and the other two models differed. These results highlight the need for medical MT systems to be evaluated for the accessibility in terms of word choice in addition to the quality of their translations. GPT4o also used more patient-friendly and conversational terms, such as *condición* instead of *afección*, when compared to GT and GPT3.5.

Overall, despite some miniscule grammatical errors GT, GPT3.5, and GPT4o translated effectively without dangerously changing the original meaning of the sentence. One limitation of this study is that the translations were not graded by patients or bilingual physicians, but by medical students who interpret for the free clinic associated with the Icahn School of Medicine. Clinical research with patients and/or physicians is needed to determine if the ChatGPTs and GT are effective medical translators.

The human evaluation metrics were verified using ICC scores. High ICC scores above 0.7 for Fluency, Adequacy, Patient-friendliness demonstrate a strong similarity between scorers for each metric. While Fluency and Adequacy were human evaluation metrics adapted from the Workshop on Machine Translation in 2006, the Patient-friendliness metric was created in this study for the purpose of discerning differences in word choice. However, Patient-friendliness was not significantly higher for the ChatGPTs compared to GT as hypothesized.

Our results were limited by using only one prompt for each model without an exhaustive search for the optimal prompt. Additionally, as human scorers evaluated translations from one model at a time, they could have developed a bias for a certain score for each model. Still, this method of scoring was chosen so that scorers could discern patterns in the translations of each model. Finally, noting that Patient-friendliness had the lowest ICC score, it is possible that a clearer description of this measurement could better standardize evaluator interpretations, a suggestion that was also reflected by testimonies from human scorers. One scorer interpreted Patient-friendliness as primarily accounting for word choice, while another scorer mentioned they gave higher Patient-friendliness

scores when a model explained a medical term instead of just translating to the corresponding Spanish medical term. These two views differ yet both can be interpreted as patient-friendliness.

7 Conclusion

This study sought to quantify the translation capabilities of LLM chatbots like GPT3.5, GPT4o, and Aguila for use in healthcare contexts. These models are not specifically designed for translation, but have capabilities that typical MTs lack, such being tasked with translating for a particular target audience. To our knowledge, this is the first study to employ automated evaluation metrics to translate a large test set representing clinical patient-provider communication. This is also the first to evaluate the newest ChatGPT model, GPT4o, in this manner and context. This work's findings confirm that the widely accessible LLM chatbots GPT3.5 Turbo and GPT4o indeed have medical MT capabilities on par with GT to translate clinical communication from English to Spanish. They hold promise for use in a variety of healthcare settings, from creating public health education texts to explaining physical examinations and inquiring about symptoms to providing take-home patient instructions. The Spanish chatbot Aguila was less successful at translating from English to Spanish, although when successful, its Spanish lexicon was much more conversational and accessible than the other models. Further studies should seek to evaluate LLM chatbots' performances at completing various clinical translation tasks in a real clinical setting, as well as explore more translation prompts.

8 Acknowledgements

This work was guided by Dr. Eric Oermann and Xujin Chris Liu with the OLAB at New York University. The work was also guided by Dr. Yasmin Meah and the East Harlem Health Outreach Partnership, the free clinic associated with the Icahn School of Medicine at Mount Sinai.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). Gpt-4 technical report. arXiv preprint arXiv:2303.08774.

Almahasees, Z., Meqdadi, S., & Albudairi, Y. (2021). Evaluation of google translate in rendering English COVID-19 texts into Arabic. 17(4), 2065–2080. https://doi.org/10.3316/informit.228360028176817

Banerjee, S., & Lavie, A. (2005). METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. *Meeting of the Association for Computational Linguistics*, 65–72.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.

Chen, X., Acosta, S., & Barry, A. E. (2017). Machine or Human? Evaluating the Quality of a Language Translation Mobile App for Diabetes Education Material. *JMIR Diabetes*, *2*(1), e7446. https://doi.org/10.2196/diabetes.7446

Cheng, E. M., Chen, A., & Cunningham, W. (2007). Primary Language and Receipt of Recommended Health Care Among Hispanics in the United States. *Journal of General Internal Medicine*, *22*(2), 283–288. https://doi.org/10.1007/s11606-007-0346-6

Cho, J. (2023). Interpreters as Translation Machines: Telephone Interpreting Challenges as Awareness Problems. *Qualitative Health Research*, *33*(12), 1037–1048. https://doi.org/10.1177/10497323231191712

Das, P., Kuznetsova, A., Zhu, M., & Milanaik, R. (2019). Dangers of Machine Translation: The Need for Professionally Translated Anticipatory Guidance Resources for Limited English

Proficiency Caregivers. *Clinical Pediatrics*, *58*(2), 247–249. https://doi.org/10.1177/0009922818809494

Dew, K., Turner, A. M., Desai, L., Martin, N., Laurenzi, A., & Kirchhoff, K. (2015). PHAST: A Collaborative Machine Translation and Post-Editing Tool for Public Health. *AMIA Annual Symposium Proceedings*, *2015*, 492–501. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4765627/

Dew, K. N., Turner, A. M., Choi, Y. K., Bosold, A., & Kirchhoff, K. (2018). Development of machine translation technology for assisting health communication: A systematic review. *Journal of Biomedical Informatics*, *85*, 56–67. https://doi.org/10.1016/j.jbi.2018.07.018

Googletrans. Free and unlimited google translate API for python¶. (n.d.-a). https://py-googletrans.readthedocs.io/en/latest/

Google Cloud. (2024, June 6). *Evaluating models | AutoML Translation Documentation*. Google Cloud. https://cloud.google.com/translate/automl/docs/evaluate

Hakami, H., & Bollegala, D. (2017). A classification approach for detecting cross-lingual biomedical term translations. *Natural Language Engineering*, *23*(1), 31–51. https://doi.org/10.1017/S1351324915000431

Harison, T. (2023, October). *Human evaluation metrics*. Machine Translate. https://machinetranslate.org/human-evaluation-metrics

He, S. (2024). Prompting ChatGPT for Translation: A Comparative Analysis of Translation Brief and Persona Prompts. arXiv preprint arXiv:2403.00127.

Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., ... & Awadalla, H. H. (2023). How good are gpt models at machine translation? a comprehensive evaluation. arXiv preprint arXiv:2302.09210.

Jiao, W., Wang, W., Huang, J. T., Wang, X., & Tu, Z. (2023). Is ChatGPT a good translator? A preliminary study. arXiv preprint arXiv:2301.08745, 1(10).

Jimeno Yepes, A., Prieur-Gaston, É., & Névéol, A. (2013). Combining MEDLINE and publisher data to create parallel corpora for the automatic translation of biomedical text. *BMC Bioinformatics*, *14*, 146. https://doi.org/10.1186/1471-2105-14-146

Johnsi Rani, J., Gladis, D., & Mammen, J. (2019). Regional Language Support for Patient-inclusive Decision Making in Breast Cancer Pathology Domain. *International Journal of Recent Technology and Engineering (IJRTE)*, *8*, 8392–8399. https://doi.org/10.35940/ijrte.C6518.098319

Kapoor, R., Corrales, G., Flores, M. P., Feng, L., & Cata, J. P. (2022). Use of Neural Machine Translation Software for Patients With Limited English Proficiency to Assess Postoperative Pain and Nausea. *JAMA Network Open*, *5*(3), e221485. https://doi.org/10.1001/jamanetworkopen.2022.1485

Khanna, R. R., Karliner, L. S., Eck, M., Vittinghoff, E., Koenig, C. J., & Fang, M. C. (2011). Performance of an online translation tool when applied to patient educational material. *Journal of Hospital Medicine*, *6*(9), 519–525. https://doi.org/10.1002/jhm.898

Khoong, E. C., Steinbrook, E., Brown, C., & Fernandez, A. (2019). Assessing the use of Google Translate for Spanish and Chinese translations of emergency department discharge instructions. JAMA internal medicine, 179(4), 580-582.

Kirchhoff, K., Turner, A. M., Axelrod, A., & Saavedra, F. (2011). Application of statistical machine translation to public health information:

A feasibility study. *Journal of the American Medical Informatics Association: JAMIA*, *18*(4), 473–478. https://doi.org/10.1136/amiajnl-2011-000176

Lankford, S., Afli, H., Ní Loinsigh, Ó., & Way, A. (2022). gaHealth: An English–Irish Bilingual Corpus of Health Data. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 6753–6758). European Language Resources Association. https://aclanthology.org/2022.lrec-1.727

Lester, C. A., Ding, Y., Li, J., Jiang, Y., Rowell, B., & Vydiswaran, V. G. V. (2021). Human versus machine editing of electronic prescription directions. *Journal of the American Pharmacists Association*, *61*(4), 484-491.e1. https://doi.org/10.1016/j.japh.2021.02.006

Liu, B., & Huang, L. (2021). ParaMed: A parallel corpus for English–Chinese translation in the biomedical domain. *BMC Medical Informatics and Decision Making*, *21*(1), 258. https://doi.org/10.1186/s12911-021-01621-8

Liu, W., & Cai, S. (2015). Translating Electronic Health Record Notes from English to Spanish: A Preliminary Study. In K. B. Cohen, D. Demner-Fushman, S. Ananiadou, & J. Tsujii (Eds.), *Proceedings of BioNLP 15* (pp. 134–140). Association for Computational Linguistics. https://doi.org/10.18653/v1/W15-3816

mapama247. (2023, July 19). Introducing Ăguila, a new open-source LLM for Spanish and Catalan. Medium. https://medium.com/@mpamies247/introducing-a%CC%8Cguila-a-new-open-source-llm-for-spanish-and-catalan-ee1ebc70bc79

Manzini, E., Garrido-Aguirre, J., Fonollosa, J., & Perera-Lluna, A. (2022). Mapping layperson medical terminology into the Human Phenotype Ontology using neural machine translation models. *Expert Systems with Applications*, *204*, 117446. https://doi.org/10.1016/j.eswa.2022.117446

Mauser, A., Hasan, S., & Ney, H. (n.d.). *Automatic evaluation measures for statistical machine translation—System optimization*.

Miller, J. M., Harvey, E. M., Bedrick, S., Mohan, P., & Calhoun, E. (2018). Simple Patient Care Instructions Translate Best: Safety Guidelines for Physician Use of Google Translate. *Journal of Clinical Outcomes Medicine, 25*(1), 18–27.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. (2002). BLEU: A method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. https://doi.org/10.3115/1073083.1073135

Patil, S., & Davies, P. (2014). Use of Google Translate in medical communication: Evaluation of accuracy. *BMJ*, *349*, g7392. https://doi.org/10.1136/bmj.g7392

Ponce, N. A., Hays, R. D., & Cunningham, W. E. (2006). Linguistic Disparities in Health Care Access and Health Status Among Older Adults. *Journal of General Internal Medicine*, *21*(7), 786–791. https://doi.org/10.1111/j.1525-1497.2006.00491.x

Profile of the foreign-born population in New York, New York. (2023). Vera Institute of Justice. https://www.vera.org/downloads/publications/profile-of-foreign-born-population-new-york-city.pdf

Renato, A., Castaño, J., Ávila, P., Berinsky, H., Gambarte, L., Park, H., Pérez, D., Otero, C., & Luna, D. (2024). *A Machine Translation Approach for Medical Terms*. 369–378. https://www.scitepress.org/Link.aspx?doi=10.5220/0006555003690378

Rupert, D. D., Alvarez, G. V., Burdge, E. J., Nahvi, R. J., Schell, S. M., & Faustino, F. L. (2022). Student-Run Free Clinics Stand at a Critical Junction Between Undergraduate Medical Education, Clinical Care, and Advocacy. *Academic Medicine : Journal of the Association of American Medical Colleges*, *97*(6), 824–831. https://doi.org/10.1097/ACM.0000000000004542

Shi, L., Lebrun, L. A., & Tsai, J. (2009). The influence of English proficiency on access to care. *Ethnicity & Health*, *14*(6), 625–642. https://doi.org/10.1080/13557850903248639

Skianis, K., Briand, Y., & Desgrippes, F. (2020). Evaluation of Machine Translation Methods applied to Medical Terminologies. In E. Holderness, A. Jimeno Yepes, A. Lavelli, A.-L. Minard, J. Pustejovsky, & F. Rinaldi (Eds.), *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis* (pp. 59–69). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.louhi-1.7

Taira, B. R., Kreger, V., Orue, A., & Diamond, L. C. (2021). A Pragmatic Assessment of Google Translate for Emergency Department Instructions. *Journal of General Internal Medicine*, *36*(11), 3361–3365. https://doi.org/10.1007/s11606-021-06666-z

Turner, A. M., Choi, Y. K., Dew, K., Tsai, M.-T., Bosold, A. L., Wu, S., Smith, D., & Meischke, H. (2019). Evaluating the Usefulness of Translation Technologies for Emergency Response Communication: A Scenario-Based Study. *JMIR Public Health and Surveillance*, *5*(1), e11171. https://doi.org/10.2196/11171

Turner, A. M., Dew, K. N., Desai, L., Martin, N., & Kirchhoff, K. (2015). Machine Translation of Public Health Materials From English to Chinese: A Feasibility Study. *JMIR Public Health and Surveillance*, *1*(2), e4779. https://doi.org/10.2196/publichealth.4779

US Census Bureau. (2020, April). *People That Speak English Less Than "Very Well" in the United States*. Census.Gov. https://www.census.gov/library/visualizations/interactive/people-that-speak-english-less-than-very-well.html

Weng, W.-H., Chung, Y.-A., & Szolovits, P. (2019). Unsupervised Clinical Language Translation. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 3121–3131. https://doi.org/10.1145/3292500.3330710

Wu, C., Xia, F., Deleger, L., & Solti, I. (2011). Statistical Machine Translation for Biomedical Text: Are We There Yet? *AMIA Annual Symposium Proceedings*, *2011*, 1290–1299. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3243244/

Yao, B., Jiang, M., Yang, D., & Hu, J. (2023). Benchmarking llm-based machine translation on cultural awareness. arXiv preprint arXiv:2305.14328.

Zappatore, M., & Ruggieri, G. (2024). Adopting machine translation in the healthcare sector: A methodological multi-criteria review. *Computer Speech & Language*, *84*, 101582. https://doi.org/10.1016/j.csl.2023.101582

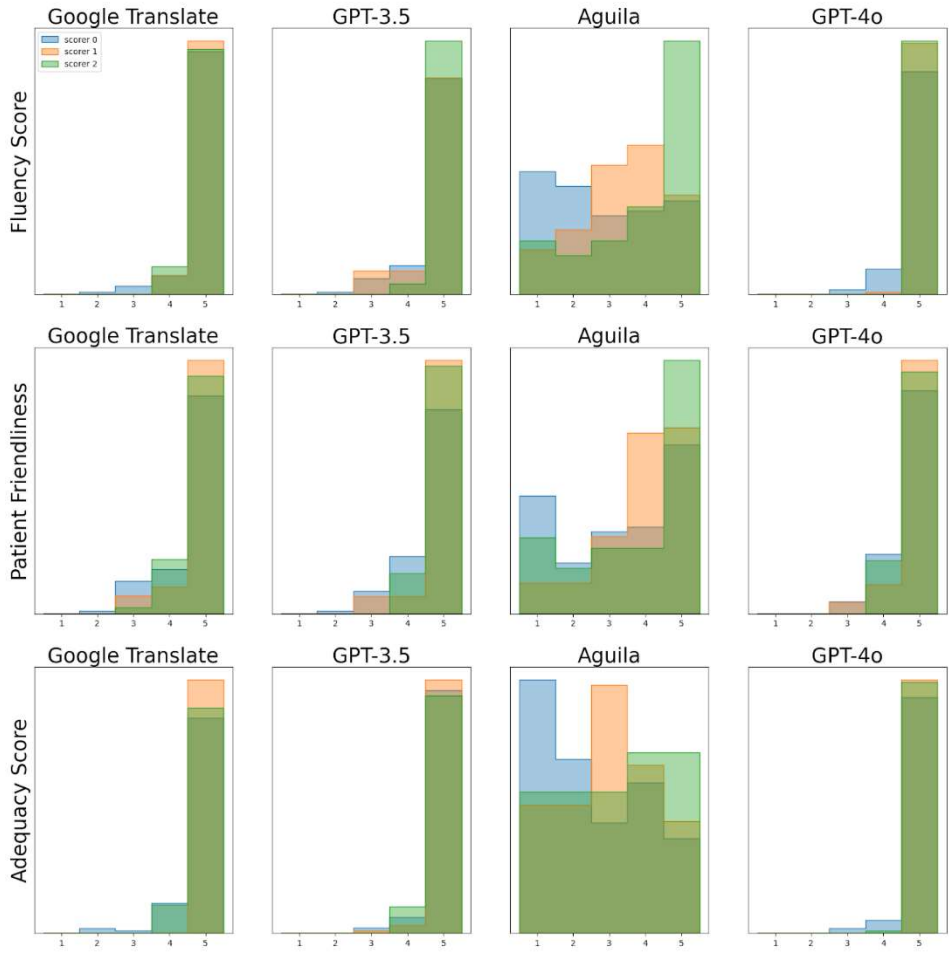
Zeng-Treitler, Q., Kim, H., Rosemblat, G., & Keselman, A. (2010). Can Multilingual Machine Translation Help Make Medical Record Content More Comprehensible to Patients? *MEDINFO 2010* (pp. 73–77). IOS Press. https://doi.org/10.3233/978-1-60750-588-4-73

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). BERTScore: Evaluating Text Generation with BERT. *ICLR 2020*. https://arxiv.org/pdf/1904.09675.pdf

Appendix

ICC Scores [95% CL]	Patient-friendliness	p value	Adequacy	p value	Fluency	p value	Total	p value
Single Random, Fixed Raters	0.409[0.27-0.54]	1.370 e-13	0.497[0.36-0.62]	1.641 e-19	0.397[0.21-0.55]	1.400 e-15	0.49[0.30-0.64]	7.69 e-22
Average Random, Fixed Raters	0.675[0.53-0.78]	1.370 e-13	0.748[0.63-0.83]	1.641 e-19	0.664[0.45-0.79]	1.4 e-15	0.742[0.57-0.84]	7.693 e-22
Rater-Normalized Single Raters	0.443[0.32-0.56]	2.02 e-13	0.549[0.44-0.65]	3.272 e-20	0.491[0.37-0.60]	3.201 e-16	0.579[0.47-0.68]	1.188 e-22
Rater-Normalized Average Raters	0.705[0.59-0.79]	2.02 e-13	0.785[0.70-0.85]	3.273 e-20	0.743[0.64-0.82]	3.201 e-16	0.804[0.73-0.86]	1.189 e-22

Appendix A. Intra-class correlation (ICC) scores for each score and for entire model. ICCs all increased when scores were normalized with z-score normalization within each judge group. The maximum score occurred in normalized average raters and was 0.785, indicating strong coherence across evaluators.



Appendix B. Score Variance for Human Evaluations for each model and Human Evaluator. Human evaluated scores shown for each human evaluator and model. Notice the strong right sided skew for each model, which is slightly more evenly distributed for Aguila. The strong skew of the results shows that the GPTs and GT performed much more consistently well than Aguila.