Viability of Machine Translation for Healthcare in Low-Resourced Languages

Hellina Hailu Nigatu[1], Nikita Mehandru[1], Negasi Haile Abadi[2], Blen Gebremeskel[1], Ahmed Alaa[1], Monojit Choudhury[3]

[1]UC Berkeley, [2] Lesan AI, [3] MBZUAI,

Correspondence: hellina_nigatu@berkeley.edu

Abstract

Machine Translation errors in high-stakes settings like healthcare pose unique risks that could lead to clinical harm. The challenges are even more pronounced for low-resourced languages where human translators are scarce and MT tools perform poorly. In this work, we provide a taxonomy of Machine Translation errors for the healthcare domain using a publicly available MT system. Preparing an evaluation dataset from pre-existing medical datasets, we conduct our study focusing on two low-resourced languages: Amharic and Tigrinya. Based on our error analysis and findings from prior work, we test two pre-translation interventions–namely, paraphrasing the source sentence and pivoting with a related language– for their effectiveness in reducing clinical risk. We find that MT errors for healthcare most commonly happen when the source sentence includes medical terminology and procedure descriptions, synonyms, figurative language, and word order differences. We find that pre-translation interventions are not effective in reducing clinical risk if the base translation model performs poorly. Based on our findings, we provide recommendations for improving MT for healthcare.[1]

1 Introduction

Language barriers exacerbate unequal access to healthcare (Al Shamsi et al., 2020; Slade and Sergent, 2025). Advances in Natural Language Processing (NLP) have the potential to ease these barriers with Machine Translation (MT) systems that facilitate communication across languages (Turner et al., 2013). MT systems are used in the medical sector to: 1) facilitate physician-patient communication when human interpreters are not available (Vieira et al., 2021; Mehandru et al., 2022), and (2) increase access to health information, for instance, by "translating public health informational materials." (Turner et al., 2013).

While MT systems hold great promise to facilitate communication across languages, commercially available MT tools may not always be designed with the healthcare domain in mind (Mehandru et al., 2022; Khoong and Rodriguez, 2022; Vieira et al., 2021). Additionally, there are limited avenues for verifying the outputs of the MT systems, especially in a critical setting like healthcare (Mehandru et al., 2023). These complications are further exacerbated for low-resourced languages as they are less likely to have human interpreters readily available (Mehandru et al., 2022). Moreover, training robust MT systems, especially for critical sectors like the medical sector, requires high-quality and quantity data, which is not available for low-resourced languages (Nekoto et al., 2020).

Errors in outputs from MT systems may pose a significant clinical risk (Khoong et al., 2019). MT tools could also misinform users who use them to access online health information (Saadany et al., 2024). For example, a machine translation output might incorrectly instruct an individual to "<u>move</u> the person completely," when the intended instruction was to "<u>immobilize</u> the person completely."[2] Such an error could have serious health consequences. Moreover, disparate performance of MT systems across languages could result in unequal access to care (Brewster et al., 2024). Given the consequences of errors in MT for healthcare, we pose the following research questions:

- **RQ1.1:** What is the error taxonomy for MT in healthcare for low-resourced languages?
- **RQ1.2:** How do MT errors differ between general health information and physician-patient communication in low-resourced languages?

[1] `https://github.com/hhnigatu/ViabilityofMTinHC/`

[2] Example taken from our set of experiments.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 10584–10598
November 4-9, 2025 ©2025 Association for Computational Linguistics

- **RQ2:** Without altering the underlying MT system, to what extent can pre-translation interventions, such as paraphrasing the source sentence or pivoting through a related language, reduce clinical risk?

In this paper, we answer these three research questions by focusing our investigation on two low-resourced languages: Tigrinya and Amharic. Our contributions can be summarized as follows:

- We contribute a dataset of human-annotated MT errors for general health information and physician-patient communication in two low-resourced languages (Sec. 3.1).
- In collaboration with a physician, we present a taxonomy of errors in MT for healthcare, with annotations and review by the physician, for our two target languages (Sec. 4).
- We provide insights into the benefits and drawbacks of two pre-translation interventions: paraphrasing the source sentence and pivoting with a related language (Sec. 5).
- Finally, we provide recommendations for design and research in improving MT for healthcare for low-resourced languages (Sec. 6).

Overall, we find that errors related to medical terminology and omission most frequently lead to high and life-threatening clinical risk. Further, we find that while pre-translation interventions reduce some errors, they preserve some errors and introduce new ones. Our research shows that there is still a long and challenging path for NLP to be useful in practice in the healthcare domain for low-resourced languages. We end our paper by highlighting the possibilities for building MT systems that can improve access to healthcare.

2 Related Work

Language barriers in medical settings are well-documented and present significant challenges for limited English proficiency (LEP) patients. LEP individuals are more likely to struggle with accessing care, experience worse health outcomes, and are at higher risk of diagnostic and medical errors, often exposed to preventable harm (Gangopadhyaya, 2021). These barriers disproportionately affect racially minoritized and uninsured patient populations (Kreienbrinck et al., 2024). Thus, there is an urgent need to improve communication with LEP patients to achieve equity in emergency care (Gutman et al., 2022).

A critical point in physician-patient communication is the discharge process, where patients are provided with written instructions containing crucial information about their diagnosis, treatment plan, and follow-up (Chen et al., 2016). These instructions are essential for the patient's understanding and post-care management. For low-resource language-speaking patients, interpreters are often unavailable.

A lack of diverse multilingual training data has jeopardized equitable applications of machine translation in medical settings. While Google Translate (GT) and ChatGPT have demonstrated comparable performance on discharge instructions in high-resource languages such as Spanish and Portuguese, both systems have performed significantly worse in terms of adequacy, fluency, and clinical risk on lower-resource languages, such as Haitian Creole. (Brewster et al., 2024; Robinson et al., 2023; Lankford and Way, 2024). These challenges highlight the need for more sophisticated models trained on medical text tailored to the specific needs of low-resource languages.

Acknowledging the importance of the severity of errors in Machine Translation, Sharou and Specia (2022) provide a general taxonomy of errors in Machine Translation and analyze *critical errors*, where mistranslations may lead to adverse consequences in contexts such as health, legal, and religious domains. To the best of our knowledge, there is no taxonomy of Machine Translation errors in healthcare for low-resourced languages.

3 Method

Here, we first describe how we prepared a dataset for evaluation in Sec. 3.1 and then detail our human evaluation scheme in Sec. 3.2.

3.1 Dataset

To answer our research questions, we first prepared an evaluation dataset relying on existing data in the healthcare domain. In this section, we describe the source datasets we used, detail our rationale for selecting an MT model, and describe our methods for selecting sentences.

3.1.1 Identifying Source Dataset

Our study focuses on two medical settings: general health information and patient-physician communication. As such, we built our evaluation dateset from two preexisting datasets:

AfriDOC-MT: For the context of general health information, we used the AfriDOC-MT(Alabi et al., 2025) dataset, which includes 334 health documents from the World Health Organization (WHO) website translated into 5 African languages, including Amharic.

DischargeME: For the second context of physician-patient communication, we used sentences from the DischargeME(Xu, 2024) dataset, which includes 109k discharge summaries derived from the MIMIC-IV(Johnson et al., 2021) dataset accessed through PhysioNet(Goldberger et al., 2000). MIMIC-IV(Johnson et al., 2021) is a dataset of de-identified Emergency Room data from a hospital in Boston, USA.

3.1.2 Picking a Translation Model

Our main criterion for selecting an MT tool was that the MT system must be publicly accessible. Based on our pilot experiment results and findings from prior work that physicians may use publicly available tools like Google Translate to facilitate communication with their patients (Mehandru et al., 2022; Taira et al., 2021; Al-Jarf, 2024), we used Google Translate in our experiments[3].

3.1.3 Selecting Sentences for Evaluation

Our goal was to understand the errors in MT for healthcare in two low-resourced languages. Hence, we were interested in building an evaluation dataset that has a diverse representation of errors. We set two criteria for selecting sentences from the test splits of the two source datasets:

Pseudo-Fuzzy xScore Matching (Li and Specia, 2019): Prior work has shown that back-translation can help physicians detect critical MT errors (Mehandru et al., 2023). Relying on this finding, we used back-translation to select sentences whose back-translation differed significantly from the original sentence. We used Google Translate to translate the sentences into our respective target languages. We then translated the sentences back into English and used SentenceBERT (Reimers and Gurevych, 2019) to get sentence embeddings for the original and back-translated sentences. We then used cosine similarity to check the similarity between the original sentence and the back-translated sentence. Finally, we used stratified sampling, randomly selecting sentences in three pseudo-fuzzy score ranges: [0.0-0.35), [0.35-0.7), [0.7,1.0].

Medical Term Count: We first conducted a pilot study with about 100 medical sentences translated from English into the two languages. From our pilot study, we found that medical terminology was most frequently mistranslated. Hence, we used a medical Named Entity Recognition (NER) model, namely `blaze999/Medical-NER`[4] to extract medical terms from each sentence. We then set a threshold (n=7) for the number of medical terms per sentence by looking at the distribution of the count of medical terms.

Using the two selection criteria, we prepared an evaluation dataset with 500 parallel sentences in each language pair, with 250 sentences for general health information and 250 sentences for physician-patient communication per language pair, for a total of 1000 parallel sentences.[5]

3.2 Human Evaluation

For our human evaluation, we prepared an annotation guideline that four physicians independently verified. We adopted the annotation axis from Mehandru et al. (2023), where each sentence pair is evaluated for *adequacy*: whether the translation correctly preserves the meaning in the source sentence (Turian et al., 2006) and *clinical risk*: whether the translation had clinically insignificant, mild, moderate, high, or life threatening errors (Nápoles et al., 2015). Once we translated the sentences with Google Translate, the first and third authors, who are native speakers of each language, labeled the sentences for adequacy and clinical risk. Then, a physician who is the fourth author and speaks both languages looked through the annotations and verified the labels. Any disagreement between the physician and the annotators was resolved in frequent weekly meetings. We used a reflexive thematic analysis (Braun et al., 2006) approach to build our taxonomy: we started with the annotations for physician-patient communication and inductively identified themes in the

[3] We conducted experiments with NLLB but found that the performance was poor. We excluded the results from our main paper. However, we provide analysis of the pilot results in Appendix A

[4] https://huggingface.co/blaze999/Medical-NER

[5] The annotated dataset will be released following licensees of the source datasets. See Appendix C.

sentence pairs with errors. We then iteratively refined the themes with the general health information annotations, discussing the final set of themes among all authors.

4 Understanding Errors in MT for Healthcare

While MT systems are used within the healthcare sector, translation errors can pose clinical risks to patients (Sec. 1). Understanding the errors in MT for healthcare would (1) give insights to how we can improve MT systems and (2) help us identify where we should pay special attention to–and at times, avoid using–MT tools. To this end, we provide a taxonomy of errors in Machine Translation for healthcare in low-resourced language pairs (Sec. 4.1). We then provide an analysis of our evaluation dataset using our taxonomy (Sec. 4.2).

4.1 A Taxonomy of MT Errors in Healthcare

As discussed in Sec. 3, we used reflective thematic analysis to arrive at our taxonomy of MT errors. We focused on identifying themes in sentences that were mistranslated. Through our iterative qualitative analysis, we identified six major categories of errors:

- **Medical Terminology and Procedure Descriptions:** Medical terminology is (1) left untranslated, (2) transliterated, (3) omitted, or (4) mistranslated in the target language. For example, “antivenom” is translated to “anti-nutrients” when translating from English to Amharic.

- **Omission:** Words and phrases in the source sentence are sometimes omitted from the translation. The omitted words could be medical terms, negations, or full descriptive phrases. For instance, a sentence that had the phrase “strict non-weight bearing” was translated to “strict weight bearing”, omitting the negation and thereby resulting in a translation that is the opposite of the source sentence.

- **Synonyms:** Words that have synonyms in the source sentence were translated to their out-of-context version–for instance, “masses” was translated to mean “collection.”

- **Word Order and Tense Disagreement:** Some of the translations had tense disagreements or wrong word orders that altered the meaning of the translation. For instance, a sentence that said “you required blood cell transfusion” was translated to “you will require blood cell transfusion”, not conveying that the transfusion has already happened.

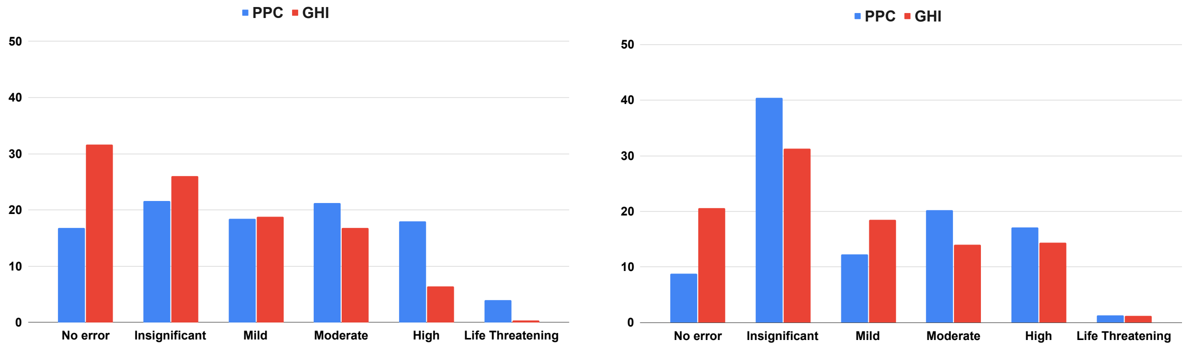
- **Figurative Language:** Common English phrases and idioms were translated literally, making the translated sentence hard to understand or culturally irrelevant. For instance, a sentence that had the phrase “aim to wean down this medication” was translated to “aim to cut off breasts” in Amharic, which poses a significant clinical risk.

- **Measurement Units:** In some of the translations, ‘pound’ and ‘lbs’ were translated to ‘kgs’ without altering the number associated with the measurement. For instance, a sentence that told a patient to alert their doctor if their weight goes above ‘5 pounds’ was translated to ‘5 kilograms,’ which is inaccurate and could lead to clinical harm.

The lack of digitally available medical data in low-resourced languages and the low-resourced context of healthcare could explain the errors in Medical terminology and Synonyms. Synonyms may be more likely seen in MT datasets in their non-medical contexts–for instance, general MT data may lack sentences where “mass” refers to “a lump” but may have examples where it refers to “a collection.” Word order and tense disagreement could be the result of word-order differences in the language pairs: Amharic sentences follow a SOV or OSV (Gutman and Avanzati, 2013) and Tigrinya sentences follow a SOV (Appleyard, 2006) word order where as English has SVO (Assaiqeli et al., 2021) word order. While omission could happen in any MT setting, it poses a significant risk when it happens in the context of healthcare; for instance, by giving the patient a translated instruction with the opposite meaning of the source sentence. Our taxonomy also reveals culture-specific phrases, and differences in standard measurement units used by different countries could lead to clinically harmful errors.

4.2 Analysis

In this section, we use our taxonomy to ground our analysis of our evaluation dataset, giving concrete examples of how the errors described in our taxonomy lead to clinically harmful mistranslations.



(a) English-Amharic

(b) English-Tigrinya

Figure 1: Distribution of Clinical Risk for General Health Information and Physician-Patient Communication.

Target Lang.	Terminology	Mistranslation	Meaning
Tir	Gonorrhea Smallpox Antibiotic Diarrhea Fever	ሽኮርያ ፍንጣጣ ጸረ-ነፍሳት ተምላስ ሰዓል	Diabetes Syphilis Insecticide Vomiting Coughing
Amh	Stool Mass Seizure Blood- Count Incision	በርጩማ ጅምላ መናድ ደም ብዛት ቁርጭምጭሚት	Chair Collection Erosion Blood Pressure Ankle

Table 1: Frequently Mistranslated Medical Terms Across our Full Evaluation dataset.

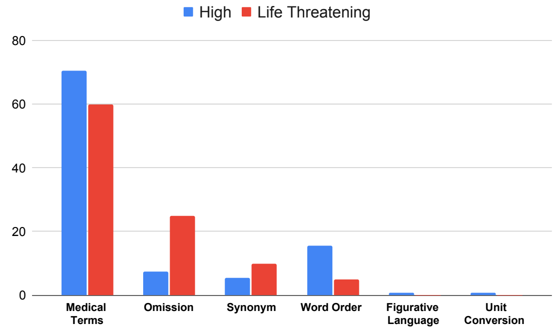


Figure 2: Percentage of High and Life-Threatening Errors (y-axis) wrt Error Taxonomy (x-axis) aggregated across the full evaluation dataset.

Figure 1 presents the distribution of the clinical risk level for English-Amharic and English-Tigrinya in general health information (GHI) and patient-physician communication (PPC) settings. We observe that for both languages, over 17% of the mistranslations could lead to a high clinical risk in patient-physician communication. Additionally, as Figure 1 shows, Physician-Patient communication translations have similar percentage of sentences in the two language pairs with 'Moderate', and 'High' clinical risks. Physician-Patient communication for English-Amharic language pair had the most 'Life Threatening' errors.

The different error types can appear in a single sentence or happen consistently. For instance, a sentence with the phrase "apply gauze dressing" was translated to "wear gasoline cloth" in Amharic. Here, the medical term "gauze" was mistranslated to "gasoline" and the word "dressing" which has a synonym, meaning "to put on cloth" was mistranslated to "cloth." We also find that the MT system would consistently output the same mistranslation for some medical terms, regardless of the source dataset. We present some examples in Table 1; for instance, in Tigrinya, the term "antibiotics" was consistently translated to "insecticide."

Further, we looked at which categories of errors were most frequently associated with high and life-threatening clinical risk. As Figure 2 shows, errors in translating Medical Terms and Procedures account for the majority of the mistranslations with life-threatening and high clinical risk. Further, Omission accounts for more of the life-threatening errors than high clinical risk. This is due sentences where the omission is (1) a negation, (2) a medical term such as name of medication, or (3) a full descriptive phrase that holds necessary information. In Table 2, we give additional examples of mistranslations from the most frequent categories.

We observed that sentences from the general health information data had fewer clinically harmful errors and fewer inadequate errors compared to

Category	Source Sentence	Translated Sentence	Target Lang.	Clinical Risk	Description
Patient-Physician Communication					
Medical Terminology and Procedures	This was handled by placing you on bowel regimen and[...]	ይህንን የተሳካነው እርስዎን ወደ አንጀት ስርዓት በማስቀመጥ እና[...]	Amh	High	"placing you on bowel regimen" is literally translated, which does not make sense in Amharic.
	You were given IV antibiotics which improved your infection, so we are sending you home with oral antibiotics	IV ጸረ-ነፍሳት ተዋሂቡካ እቲ ረኽሲ ድማ እናተመሓየሸ ዝኸይድ ዘሎ ይመስል፡ ስለዚ ሕጂ ብኣፍ ዝውሰድ ጸረ-ነፍሳት ናብ ገዛኻ ክትከይድ ትኽእል ኢኻ።	Tir	Life Threatening	"antibitoics" is translated to "insecticide," telling the patient they received IV insecticide and are being sent home with oral insecticide.
Omission	For [...]: Please take oral amoxicillin 500mg twice daily, for 7 days	እባክዎን በአፍ የሚወሰድ X 500mg በቀን ሁለት ጊዜ ለ7 ቀናት ይውሰዱ።	Amh	Life Threatening	The name of the medication, "amoxicillin" is omitted in the translation.
	Thankfully, your symptoms are improving so we moved you to acute rehabilitation[...]	ምስጋና ይግባእ።	Tir	High	Full phrase after "Thankfully" is omitted from the translation.
Synonyms	When you released from the hospital, [...] there was no bright red blood in the stool	ከወጣህ በኋላ [...] በርጩማ ውስጥ ምንም ደማቅ ቀይ ደም አልነበብህም.	Amh	High	"stool" is translated to the small chair you sit on.
	General Discharge Instructions: You received an abdominal operation [...]	ሓፈሻዊ መምርሒ ምፍሳስ ደም፡- ናይ ከብዲ መጥባሕቲ ጌርኪ ኣለኺ[...]	Tir	High	"Discharge" is translated to "something that flows out", in this case explicitly as "blood that flows out".
General Health Information					
Medical Terminology	Overview: Smallpox is an acute contagious disease caused by the variola virus, a member of the orthopoxvirus family.	ጠቃላይ እይታ፡ ፈንጣጣ የኦርቶዶክስ ቫይረስ ቤተሰብ አባል በሆነው በቫሪዮላ ቫይረስ የሚከሰት አጣዳፊ ተላላፊ በሽታ ነው።	Amh	High	"orthopoxvirus" is mistranslated to "orthodox" which can be misunderstood for the Orthodox religion.
	Three bacterial STIs [...] are usually curable with existing, effective single-dose or multiple-dose regimens of antibiotics.	ሰለስተ ብባክተርያዊ ጾታዊ ርክብ ዝፍጠሩ ሕማማት [...] መብዛሕትኡ ግዜ ብዝጸንሐ፡ ውጽኢታዊ ዝኾነ ሓደ ዶዝ ወይ ብዙሕ ዶዝ ስርዓታት ጸረ-ነፍሳት ይፍወሱ።	Tir	Life Threatening	"antibiotic" is translated to "insecticide"
Omission	Maternal syphilis, when untreated, treated late or not treated with penicillin, results in adverse birth outcomes (ABOs) [...]	የእናቶች ቂጥኝ ህክምና ካልተደረገለት ፣ ዘግይቶ በፔኒሲሊን ካልታከመ [...] የሚገመቱ X የወሊድ ውጤቶች[...] ያስከትላል ።	Amh	High	"adverse" is omitted from the translation, making the sentence seem like untreated syphilis could lead to births.
Synonyms	[...] show trends of high rates of quinolone resistance[...]	[...] ከፍተኛ የ quinolone ተከላካይነት አዝማሚያዎችን ያሳያሉ [...]	Amh	High	"resistance" is translated to "[the medication] being able to defend"
	Deficiencies in vitamin and mineral status, particularly of folate, iron, vitamin A, and zinc,[...].	ሕጽረት ኩነታት ቪታሚንን ማዕድናትን ብፍላይ ድማ ፎሌት፡ ሓጺን፡ ቪታሚን ኤን ዚንጎን፡ ኣብ መላእ ዓለም [...]	Tir	High	"iron" is translated literally to "steel"

Table 2: Examples of High and Life Threatening Clinical Risk Mistranslations in General Health Information and Physician-Patient Communication. Note that all sentences for the Physician-Patient Communication have been manually paraphrased to adhere to the source dataset license.

the patient-physician communication data: 46.6% of the patient-physician translations were inadequate, compared to 38.5% of the general health information translations that were inadequate, aggregated across both languages. Of the inadequate sentences in English-Amharic translation, there was a single sentence with life-threatening clinical risk for general health information, as compared to 7.75% life-threatening errors in the patient-physician translations. As Figure 1 shows, the patient-physician dataset had more errors that could cause high and life-threatening clinical risk in both languages. For Amharic, general health information translations had more clinically 'insignificant' errors. Since we only had the references available for the General Health Information dataset for Amharic, we calculated the BLEU score for the Google Translate output to complement our qualitative analysis. We find that the translation had a BLEU score of 47.82%. However, since we do not have the references for Tigrinya and for the physician-patient communication dataset in both languages, we could not do further analysis comparing metric score performance and qualitative analysis.

5 Intervening on Errors in MT for Healthcare

Using MT tools for healthcare in low-resourced language settings poses two challenges: (1) healthcare is a critical setting where errors in translation could lead to clinical harm and (2) MT tools generally perform poorly for low-resourced languages making the errors more likely to happen (Sec. 1). However, improving MT systems for low-resource languages is challenging mainly due to the lack of parallel data; a problem further complicated in the context of healthcare. Hence, we experimented with two pre-translation interventions to understand to what extent they can reduce clinically harmful errors without altering the underlying MT system. Below, we detail the pre-translation interventions and discuss our results.

5.1 Paraphrasing the Source Sentence

Based on our analysis in Sec. 4, we observed that ambiguity in the source sentence (for instance, due to word order or synonyms) or use of figurative language resulted in clinically harmful errors.

Hypothesis 1: Paraphrasing the source sentence will help reduce errors by removing ambiguity and simplifying the input to the MT system.

Prior work has experimented with post-editing translations with Large Language Models (Ki and Carpuat, 2024). However, we are working with low-resourced languages that are not well supported by the state-of-the-art language models (Ojo et al., 2025); hence, we paraphrased the source sentence instead. We used LLaMa-4 Maverick Instruct model (Meta) for paraphrasing. We designed our prompts by following the criteria from the medical translation guidelines (Txabarriaga, 2008; Edwards and Goodman, 2006; DPH, 2024). See Appendix B for more details.

Results Paraphrasing with an LLM reduced ambiguity in some sentences but also introduced new types of errors. The model would sometimes replace single medical terms with more elaborate details: for instance, 'adhesion', which was in the baseline translated to "ማጣበቂያዎችን" meaning "adhesives", was paraphrased to "tissue bands" and correctly transliterated in the final output. Paraphrasing also helped with unit conversion errors: while in the baseline translation "5 pounds" was translated to "5 kilograms", the paraphrasing converted the measurement units and resulted in "2.27 kgs." However, we noticed the model would sometimes add new information or details that were not present in the original sentence, leading to increased clinical risk. For instance, a sentence with the word "stump" was translated to "remaining part of your hand" even though the original sentence did not mention which limb the word "stump" was referring to.

5.2 Pivoting with a Related Language

Pivoting has been used in prior work to help improve machine translation performance (Kim et al., 2019). Pivoting is especially helpful when the data between the source language and the target language is small and when there is significant linguistic variation between the source and target languages (Paul et al., 2013).

Hypothesis 2: Pivoting through a related, higher-resourced language would reduce errors due to ambiguity.

In our pilot analysis (Appendix A), we found that only 6% of the MT output for Arabic, an Afro-Semitic language related to our two target languages, was inadequate, with two moderate clinical risks and one insignificant clinical risk. Hence, we selected Arabic as our pivot language. We use naive pivoting (Kim et al., 2019) where we translate from English to Arabic, then from Arabic to the target language. We used Google Translate for all pivoting experiments.

Results Pivoting improved translation for some ambiguous sentences that had synonyms; for instance, the sentence "[your disease]...has been progressing over the week" was mistranslated to mean "[your disease] has been improving over the week" in the baseline. However, after pivoting, the sentence was correctly translated as the Arabic sentence resolved the ambiguity caused by the word "progressing." However, pivoting does not solve all problems: when the error is due to medical term mistranslations, pivoting does not resolve the errors as the MT model does not have the medical terminology in the target language. Further, we find that some of the translations were exactly the same as the direct English-[target language] translation. Pivoting also increased the number of words that are wrongly transliterated, which could be due to the model trying to transliterate from an already transliterated term.

In summary, pre-translation interventions reduce some errors, preserve some errors, and introduce new ones. In Figure 6 and Figure 7, we show the distribution of clinical risk for the baseline and the two interventions for patient-physician communication for Amharic and Tigrinya, respectively. Both interventions increase the number of translations without error, with paraphrasing resulting in a larger increase. Further, we see that life-threatening errors are reduced with both interventions, although there are still life-threatening errors in some of the translations post-intervention. Additionally, in physician-patient communication, high clinical risk shows a 3.6 percentage point reduction for Amharic with paraphrasing and a 6.8 percentage point decrease with pivoting for Tigrinya. But while the interventions reduce some clinically harmful errors, as we described above, they introduce new errors (e.g. added details when paraphrasing with LLMs) or preserve old ones (e.g. errors due to the lack of medical terms in the target language).

6 Discussion

In this study, we investigated the viability of a general-purpose MT tool that is used in the healthcare setting for translating general health information and physician-patient communication data from English to two low-resourced languages. We tested two pre-translation interventions: paraphrasing the source sentence and pivoting with a related language, and found that, without altering the underlying MT model, pre-translation interventions are not efficacious in reducing clinically harmful errors (Sec. 5). Based on our findings, we provide the following concrete recommendations:

Low-data model improvements: Improving the underlying MT systems for the healthcare domain may pose a significant challenge due to the lack of digitally available parallel data in low-resourced languages (Sec. 2). However, future work can explore low-data interventions targeted towards the MT model: for instance, one potential avenue could be collecting translations for medical terminology and using the resulting dictionary as a data augmentation step.

Identify context of use within a domain: Within the healthcare domain, there is a difference between general health information and physician-patient communication in terms of the clinical risk and adequacy of MT outputs (Sec. 4). As such, MT tools that perform well for general health information may not necessarily have a similar level of performance for all clinical settings. Hence, it is imperative to distinguish between the specific contexts in which our MT system performs well.

Incorporating cultural-sensitivity in data collection: A shift in cultural context affects the effectiveness of MT outputs. Sentences with figurative language and measurement units were mistranslated in the target languages, resulting in clinically risky errors (Sec. 4). Further, some translations did not make sense in the target language cultural context (Sec 4). Data collection schemes for MT for health should therefore pay special attention to cultural context.

Accounting for intervention over-confidence: While pivoting and paraphrasing reduced ambiguities, they did not eradicate all errors. This could potentially be dangerous when the MT output unambiguously gives the wrong translation. Both interventions were not effective when the MT model did not have vocabulary for medical terms in the target language–for instance, regardless of the intervention "antibiotics" was translated to "insecticide" in Tigrinya. In such cases, the output of the MT system could unambiguously tell a patient to "take insecticide" which could lead to a life-threatening clinical harm.

Exploring other prompting strategies: While we tried paraphrasing as a possible intervention, we limited our experiment to one prompt, which incorporated medical translation guidelines. Future work could explore different prompting strategies, for instance, by incorporating additional context for the LLM or specifying unit conversion rules. Such approaches may draw on literature on interactive translation (Lyu et al., 2024; Santy et al., 2019; Knowles and Koehn, 2016). Future work could also explore incorporating our error taxonomy into prompts. However, we caution that using LLMs poses the risks discussed in Sec. 5, requiring careful consideration, especially for low-resourced languages where mitigation of mistranslations may not be easily accessible.

7 Conclusion

MT tools hold a promise to break the language barriers and are used in critical settings like healthcare to facilitate communication. However, our study shows there is still a long road ahead to make general-purpose MT tools useful in practice, particularly in critical settings like healthcare. Based on our findings, we provide a set of recommendations to improve MT for health in low-resource language contexts.

Limitations

Our study has several limitations. First, it focuses on only two low-resourced languages-Amharic and Tigrinya-which limits the generalizability of our findings to other linguistic communities. However, while we only study two language pairs, the languages are generally understudied within NLP research. As such, we provide insights that are informed by a medical professional for two understudied languages. Our evaluation scheme and dataset preparation can also be adopted to other low-resourced languages and communities. We only had the human resources to conduct a thorough analysis for the two low-resourced languages. Further, we conducted pilot studies with three additional languages that demonstrated the clinical harm was pronounced for the two low-resourced languages (Appendix A). Second, our evaluation is conducted at the sentence level, which does not account for discourse-level coherence or cumulative errors across longer clinical narratives. Yet, our results still demonstrate where MT models lead to clinically harmful errors. Third, our analysis for physician-patient communication relies on data from a single healthcare institution in the United States, specifically the MIMIC-IV-based DischargeME dataset. This dataset may not capture the variability in language use, medical practices, or patient communication styles present in other regions or healthcare systems, particularly those where the target languages are spoken. However, medical datasets are scarce, and these were the only datasets we had access to; future work could explore using other datasets from other contexts. Fourth, we use Google Translate as our MT model for our experiments, which limits the interventions we can test, as the underlying model is not available. However, as we have discussed in our Sec. 3, Google Translate is used in practice in the healthcare domain and our pilot results showed that the open-source model had worse performance. Future work could explore comparing performance across models.

Ethics Statement

In resource-limited medical settings, the evaluation and use of MT and LLMs require careful scrutiny beyond simple error rates. Our work demonstrates that even seemingly minor medical translation errors can have significant repercussions, particularly for vulnerable patient populations. A seemingly minor translation error or a misinterpreted query by a language model can inadvertently lead to misdiagnoses or treatment delays.

References

Reima Al-Jarf. 2024. Translation of medical terms by ai: A comparative linguistic study of microsoft copilot and google translate. In *International Conference on Artificial Intelligence and its Applications in the Age of Digital Transformation*, pages 220–235. Springer.

Hilal Al Shamsi, Abdullah G. Almutairi, Sulaiman Al Mashrafi, and Talib Al Kalbani. 2020. Implications of Language Barriers for Healthcare: A Systematic Review. *Oman Medical Journal*, 35(2):e122.

Jesujoba O. Alabi, Israel Abebe Azime, Miaoran Zhang, Cristina España-Bonet, Rachel

Bawden, Dawei Zhu, David Ifeoluwa Adelani, Clement Oyeleke Odoje, Idris Akinade, Iffat Maab, Davis David, Shamsuddeen Hassan Muhammad, Neo Putini, David O. Ademuyiwa, Andrew Caines, and Dietrich Klakow. 2025. AFRIDOC-MT: Document-level MT Corpus for African Languages. ArXiv:2501.06374 [cs].

D. Appleyard. 2006. Tigrinya. In Keith Brown, editor, *Encyclopedia of Language & Linguistics (Second Edition)*, pages 715–717. Elsevier, Oxford.

Aladdin Assaiqeli, Mahendran Maniam, and Mohammed Farrah. 2021. Inversion and word order in English: A functional perspective. *Studies in English Language and Education*, 8(2):523–545. Number: 2.

Virginia Braun, , and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101.

Ryan C. L. Brewster, Priscilla Gonzalez, Rohan Khazanchi, Alex Butler, Raquel Selcer, Derrick Chu, Barbara Pontes Aires, Marcella Luercio, and Jonathan D. Hron. 2024. Performance of ChatGPT and Google Translate for Pediatric Discharge Instruction Translation. *Pediatrics*, 154(1):e2023065573.

Xuewei Chen, Sandra Acosta, Adam Etheridge Barry, et al. 2016. Evaluating the accuracy of google translate for diabetes education material. *JMIR diabetes*, 1(1):e5848.

Massachusetts Department of Public Health DPH. 2024. Language Access Plan.

Martin B. Edwards and Neville W. Goodman. 2006. Use of the passive voice. In *Medical Writing: A Prescription for Clarity*, 3 edition, pages 139–144. Cambridge University Press, Cambridge.

Anuj Gangopadhyaya. 2021. Black Patients are More Likely Than White Patients to be in Hospitals with Worse Patient Safety Conditions.

A Goldberger, L Amaral, L Glass, J Hausdorff, P Ivanov, R Mark, J.E Mietus, G.B Moody, C.K Peng, and H.E Stanley. 2000. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals.

Alejandro Gutman and Beatriz Avanzati. 2013. Amharic.

Colleen K. Gutman, K. Casey Lion, Carla L. Fisher, Paul L. Aronson, Mary Patterson, and Rosemarie Fernandez. 2022. Breaking through barriers: the need for effective research to promote language-concordant communication as a facilitator of equitable emergency care. *Journal of the American College of Emergency Physicians Open*, 3(1):e12639.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2021. MIMIC-IV.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Elaine C Khoong and Jorge A Rodriguez. 2022. A research agenda for using machine translation in clinical medicine. *Journal of General Internal Medicine*, 37(5):1275–1277.

Elaine C. Khoong, Eric Steinbrook, Cortlyn Brown, and Alicia Fernandez. 2019. Assessing the Use of Google Translate for Spanish and Chinese Translations of Emergency Department Discharge Instructions. *JAMA Internal Medicine*, 179(4):580–582.

Dayeon Ki and Marine Carpuat. 2024. Guiding Large Language Models to Post-Edit Machine Translation with Error Annotations. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4253–4273, Mexico City, Mexico. Association for Computational Linguistics.

Yunsu Kim, Petre Petrov, Pavel Petrushkov, Shahram Khadivi, and Hermann Ney. 2019. Pivot-based transfer learning for neural machine translation between non-English languages. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 866–876, Hong Kong, China. Association for Computational Linguistics.

Rebecca Knowles and Philipp Koehn. 2016. Neural interactive translation prediction. In *Conferences of the Association for Machine Translation in the Americas: MT Researchers' Track*, pages 107–120, Austin, TX, USA. The Association for Machine Translation in the Americas.

Annika Kreienbrinck, Saskia Hanft-Robert, and Mike Mösko. 2024. Usability of technological tools to overcome language barriers in health care: a scoping review protocol. *BMJ open*, 14(3):e079814.

Séamus Lankford and Andy Way. 2024. Leveraging llms for mt in crisis scenarios: a blueprint for low-resource languages. *arXiv preprint arXiv:2410.23890*.

Zhenhao Li and Lucia Specia. 2019. Improving Neural Machine Translation Robustness via Data Augmentation: Beyond Back Translation. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 328–336. ArXiv:1910.03009 [cs].

Chenyang Lyu, Zefeng Du, Jitao Xu, Yitao Duan, Minghao Wu, Teresa Lynn, Alham Fikri Aji, Derek F. Wong, and Longyue Wang. 2024. A paradigm shift: The future of machine translation lies with large language models. In *Proceedings of* *the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1339–1352, Torino, Italia. ELRA and ICCL.

Nikita Mehandru, Sweta Agrawal, Yimin Xiao, Ge Gao, Elaine Khoong, Marine Carpuat, and Niloufar Salehi. 2023. Physician Detection of Clinical Harm in Machine Translation: Quality Estimation Aids in Reliance and Backtranslation Identifies Critical Errors. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11633–11647, Singapore. Association for Computational Linguistics.

Nikita Mehandru, Samantha Robertson, and Niloufar Salehi. 2022. Reliable and Safe Use of Machine Translation in Medical Settings. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 2016–2025, New York, NY, USA. Association for Computing Machinery.

Meta. The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. 2020. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online. Association for Computational Linguistics.

Team NLLB, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation. ArXiv:2207.04672 [cs].

Anna M. Nápoles, Jasmine Santoyo-Olsson, Leah S. Karliner, Steven E. Gregorich, and Eliseo J. Pérez-Stable. 2015. Inaccurate Language Interpretation and Its Clinical Significance in the Medical Encounters of Spanish-speaking Latinos. *Medical Care*, 53(11):940–947.

Jessica Ojo, Odunayo Ogundepo, Akintunde Oladipo, Kelechi Ogueji, Jimmy Lin, Pontus Stenetorp, and David Ifeoluwa Adelani. 2025. AfroBench: How Good are Large Language Models on African Languages? ArXiv:2311.07978 [cs].

Michael Paul, Andrew Finch, and Eiichrio Sumita. 2013. How to Choose the Best Pivot Language for Automatic Translation of Low-Resource Languages. *ACM Transactions on Asian Language Information Processing*, 12(4):14:1–14:17.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Nathaniel R Robinson, Perez Ogayo, David R Mortensen, and Graham Neubig. 2023. Chatgpt mt: Competitive for high-(but not low-) resource languages. *arXiv preprint arXiv:2309.07423*.

Hadeel Saadany, Ashraf Tantawy, and Constantin Orasan. 2024. Cyber Risks of Machine Translation Critical Errors : Arabic Mental Health Tweets as a Case Study. ArXiv:2405.11668 [cs].

Sebastin Santy, Sandipan Dandapat, Monojit Choudhury, and Kalika Bali. 2019. INMT: Interactive neural machine translation prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 103–108, Hong Kong, China. Association for Computational Linguistics.

Khetam Al Sharou and Lucia Specia. 2022. A taxonomy and study of critical errors in machine translation. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 171–180, Ghent, Belgium. European Association for Machine Translation.

Sam Slade and Shane R. Sergent. 2025. Language Barrier. In *StatPearls*. StatPearls Publishing, Treasure Island (FL).

Breena R Taira, Vanessa Kreger, Aristides Orue, and Lisa C Diamond. 2021. A pragmatic assessment of google translate for emergency department instructions. *Journal of General Internal Medicine*, 36(11):3361–3365.

Joseph P. Turian, Luke Shea, and I. D. Melamed. 2006. Evaluation of Machine Translation and its Evaluation:. Technical report, Defense Technical Information Center, Fort Belvoir, VA.

Anne M. Turner, Hannah Mandel, and Daniel Capurro. 2013. Local Health Department Translation Processes: Potential of Machine Translation Technologies to Help Meet Needs. *AMIA Annual Symposium Proceedings*, 2013:1378–1385.

Rocío Txabarriaga. 2008. IMIA Guide on Medical Translation.

Lucas Nunes Vieira, OHagan , Minako, , and Carol OSullivan. 2021. Understanding the societal impacts of machine translation: a critical review of the literature on medical and legal use cases. *Information, Communication & Society*, 24(11):1515–1532. Publisher: Routledge _eprint: https://doi.org/10.1080/1369118X.2020.1776370.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Justin Xu. 2024. Discharge Me: BioNLP ACL'24 Shared Task on Streamlining Discharge Documentation.

A Pilot Study

We conducted a pilot study with 3 additional target languages: Bengali, Hindi, and Arabic. For our pilot study, we randomly selected 50 sentences from the DischargeME dataset (Xu, 2024). We selected sentences from the Phase I test set; we used the Phase II test set for our main experiments. Using the annotation scheme described in Sec. 3, 2 native speakers of each language independently labeled the data, and disagreements were resolved in joint meetings through discussions between the annotators. We selected the languages for our pilot by relying on the language taxonomy provided by Joshi et al. (2020). We selected one language from each class from Class 1- Class 6 in the Joshi et al. (2020) paper, where languages are classified by the amount of data they have available.

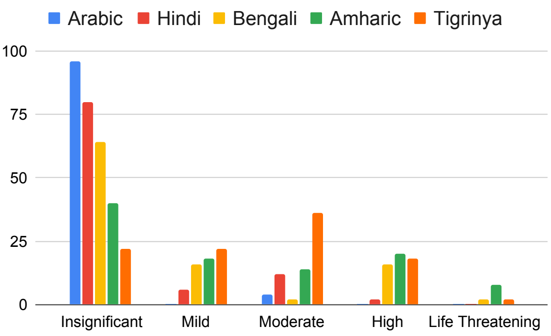


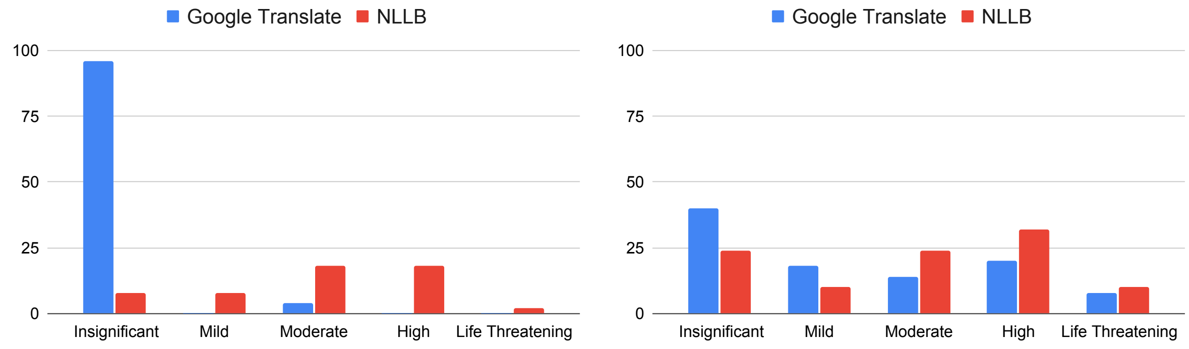
Figure 3: Comparison of Clinical Risk in Translation Errors across Pilot Study Languages.

Comparison of Errors Across Languages: We translated the 50 sentences into the 5 target languages using Google Translate. As Figure 3 shows, there is a linear relation between the amount of data available for a language and the amount of insignificant errors in the translation. Arabic, which is the highest resourced from all five target languages, had 96% of translations with insignificant errors, followed by Hindi with 80% of translations having clinically insignificant errors. Further, the three lower-resourced languages–Bengali, Amharic, and Tigrinya–all had some translations with life-threatening errors in translation.

Comparison of MT models: As discussed in Sec. 3, our criterion for selecting MT model was that the model had to be publicly available. In our pilot, we experimented with NLLB(NLLB et al., 2022), namely the NLLB-600M model. We selected Arabic and Amharic from our target languages and translated the sentences using the NLLB-600M model. As Figure 4, we find that NLLB output greatly increased high and life-threatening errors for both languages. Qualitatively, we find that the NLLB model would leave medical terminology untranslated for Amharic and omit details such as medication dosage in Arabic. We give qualitative examples in Table 3. Based on these observations and evidence from prior work that Google Translate has been used in clinical settings (Sec. 1), we selected Google Translate as our translation model for our experiments.

B Paraphrasing Prompt Design

To design out prompt for the paraphrasing experiment described in Sec 5, we relied on existing guidelines for translation in the medical setting



(a) English-Arabic

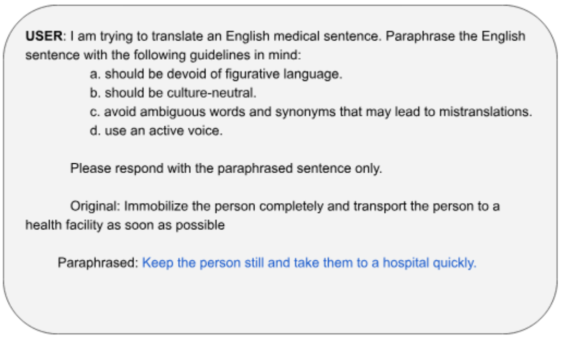
(b) English-Amharic

Figure 4: Comparison of MT models–Google Translate and NLLB–for Clinical Risk in Translation

Source Sentence	**Translation**	**Error Description**
MEDICATIONS: 1) Take Amoxicillin every 7 hours around the clock	መድሃኒቶች: 1) Amoxicillin በየ 7 ሰዓቱ በየሰዓቱ ይውሰዱ	The name of the medication was left untranslated. the phrase "around the clock" was mistranslated to "every hour."
If it is draining, apply a gauze dressing secured with paper tape	ማፍሰስ ካለበት በወረቀት ቴፕ የታሸገ የጋዝ ማሰሪያ ሊተገበር ይችላል ።	"gauze" is mistranslated to "gasoline"
3) Please START Advil 8	3) እባክዎን START Advil 8	all words except "please" are left untranslated.

Table 3: Examples of pilot translations with NLLB for the Amharic DischargeME dataset. Sentences have been paraphrased manually for display to keep with the dataset guidelines.

from Txabarriaga (2008), Edwards and Goodman (2006) and DPH (2024). In discussing formatting for source documents, the guidelines in DPH (2024) and Txabarriaga (2008) state that source text should be culturally neutral and that source text should not include figurative language. Further, DPH (2024) and Edwards and Goodman (2006) state that source text should be written in active voice. Incorporating this with our findings from Sec. 4, we designed our prompt as shown in Figure 5. We prompted the model using the HuggingFace API (Wolf et al., 2020) at a temperature of 0.5, with 2048 maximum tokens and nucleus sampling with p=0.7.



USER: I am trying to translate an English medical sentence. Paraphrase the English sentence with the following guidelines in mind:

a. should be devoid of figurative language.
b. should be culture-neutral.
c. avoid ambiguous words and synonyms that may lead to mistranslations.
d. use an active voice.

Please respond with the paraphrased sentence only.

Original: Immobilize the person completely and transport the person to a health facility as soon as possible

Paraphrased: Keep the person still and take them to a hospital quickly.

Figure 5: Example prompt and model completion (in text color blue) for paraphrasing.

C Data Access and Release

As discussed in Sec. 3.1, we prepared our evaluation dataset from the AfriDOC-MT(Alabi et al., 2025) and the DischrageME dataset(Xu, 2024). AfriDOC-MT (Alabi et al., 2025) is a publicly available dataset that includes translations from the WHO website. Hence, we will release the 500 sentences from AfriDOC-MT publicly, with a Creative Commons license CC BY-NC-SA 3.0 identical to the original dataset[6].

On the other hand, the DischargeME (Xu, 2024) dataset contains de-identified data about patients from an emergency room in a hospital in the US (Johnson et al., 2021). It is accessible through PhysioNet (Goldberger et al., 2000) under restricted data agreements. Hence, we will release our annotations for the 500 sentences taken from the DischrageME (Xu, 2024) data through PhysioNet (Goldberger et al., 2000) following the same

[6] Link to source dataset license information: https://github.com/masakhane-io/afridoc-mt

Source Sentence	Translation	Error Description
Maternal syphilis, when untreated, treated late or not treated with penicillin, results in adverse birth outcomes (ABOs) estimated in 50-80% of cases, depending on the stage of syphilis.	እናታዊ ሲፊሊስ ሳንተክነይ እንደሌለው፣ ወይም በመዘግየት ወይም በፔኒሲሊን አንደተሳነ በሚድን እንደሌለው፣ በአንደኛው ደረጃው ሲፊሊስ እንደምትገኝ ከመቶ 50 እስከ 80 ያህል የውል ፍጹም ችግሮች (ABOs) ይመነገራሉ።	Translation included words that have no meaning in the language (in red font) and mistranslation of the term "adverse birth outcomes" which was translated to "a contract's absolute problems," unrelated to the medical context (in orange font).
This makes the preparation of correct antivenoms an ongoing problem.	ይህ የትክክለኛ የመርዛማ መድኃኒቶች እንቅስቃሴ ማዘጋጀት ቀጣይ ችግር ያደርገዋል።	"antivenoms" was translated to "poisonous medicines"
Ticks also transmit Borreliosis (Lyme disease), which is a bacterial infection.	እትም እንደሆነ ባክቴርያላዊ በሽታ የሆነውን ቦሬሊዮሲስ (ላይም በሽታ) ደግሞ ይላክታል።	"Ticks" was translated to "(publication) edition"; the last word indicated with red font has no meaning in the language.

Table 4: Examples of pilot translations with GPT-4o for the Amharic AfriDOC-MT dataset.

data restrictions as the original dataset.

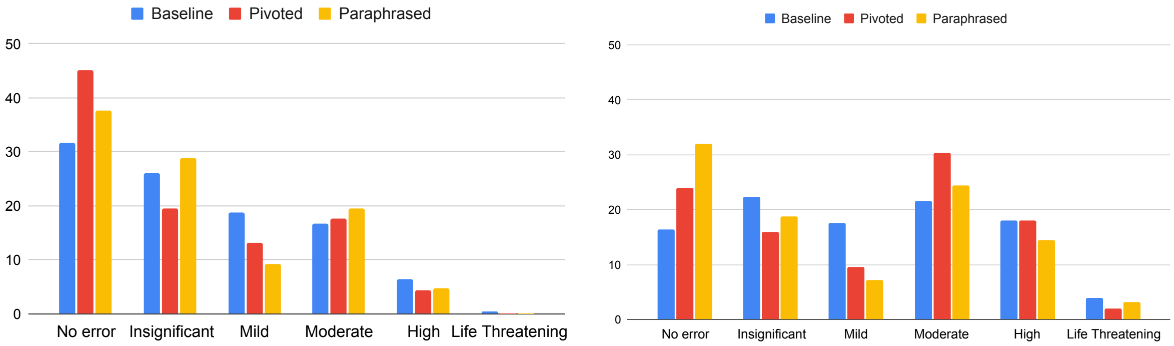
Steps taken to present restricted data: While it was important to provide examples in this paper, the DischrageME (Xu, 2024) dataset is restricted under data agreements. Hence, we manually paraphrased all examples we give in this paper for data taken from the DischrageME (Xu, 2024) dataset. We replaced names of medications with other medication names and changed dosage and measurement values: for instance, in Table 2, we replaced the original medication in the first example for Omission with "amoxicillin".

D Additional Results

In this section, we provide additional results for our main experiments.

Clinical Risk Distribution As discussed in Sec. 5, paraphrasing the source sentence and using a related language for pivoting did not effectively reduce clinically harmful errors overall. In Figure 6, we present the distribution of clinical risk for Amharic in the baseline setting and in the two pre-translation interventions. We observe that Pivoting shifted translations to no error for Amharic in the General Health Information setting. Both pivoting and paraphrasing reduced the life-threatening error for the General Health Information setting to 0. However, in the Physician-Patient Communication setting, both interventions still have life-threatening errors. For Tigrinya, we again see that pivoting shifted translations to no error for General Health Information in Figure 7. However, both settings still have life-threatening errors in the Physician-Patient Communication setting.

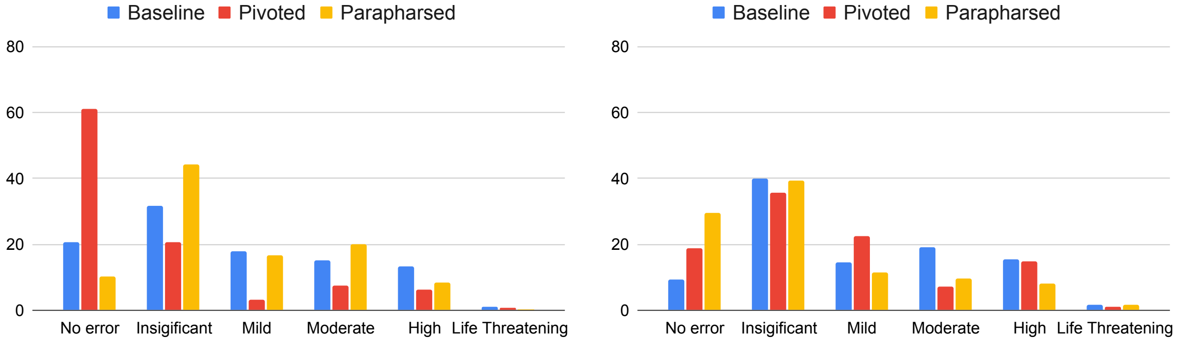
Translation with LLM In our pilot runs, we translated a few sentences using GPT-4o to see if we can use it as a viable MT model for our experiments. However, we found that the translation quality was poor for our target languages. In Table 4, we give a few examples describing the observed errors. We find that the translations included words that do not exist in the language and when there were mistranslations of medical terms, the mistranslations usually were not in the medical context.



(a) General Health Information

(b) Physician-Patient Communication

Figure 6: Clinical Risk Distribution For English-Amharic in General Health Information (a) and Physician-Patient Communication (b) across baseline, pivoting and paraphrased settings.



(a) General Health Information

(b) Physician-Patient Communication

Figure 7: Clinical Risk Distribution For English-Tigrinya in General Health Information (a) and Physician-Patient Communication (b) across baseline, pivoting and paraphrased settings.