LLMs-in-the-loop Part-1:
Expert Small AI Models for Bio-Medical Text Translation

Bunyamin Keles, Murat Gunay[1], Serdar I.Caglar

AI Amplified (Aimped) Inc.,
1775 Tyson Blvd,
Floor 5 Tysons, VA 22102

{benjamin, joseph, forest} @aimped.ai

Abstract

Machine translation is indispensable in healthcare for enabling the global dissemination of medical knowledge across languages. However, complex medical terminology poses unique challenges to achieving adequate translation quality and accuracy. This study introduces a novel "LLMs-in-the-loop" approach to develop supervised neural machine translation models optimized specifically for medical texts. While large language models (LLMs) have demonstrated powerful capabilities, this research shows that small, specialized models trained on high-quality in-domain (mostly synthetic) data can outperform even vastly larger LLMs.

Custom parallel corpora in six languages were compiled from scientific articles, synthetically generated clinical documents, and medical texts. Our LLMs-in-the-loop methodology employs synthetic data generation, rigorous evaluation, and agent orchestration to enhance performance. We developed small medical translation models using the MarianMT base model. We introduce a new medical translation test dataset to standardize evaluation in this domain. Assessed using BLEU, METEOR, ROUGE, and BERT scores on this test set, our MarianMT-based models outperform Google Translate, DeepL, and GPT-4-Turbo.

Results demonstrate that our LLMs-in-the-loop approach, combined with fine-tuning high-quality, domain-specific data, enables specialized models to outperform general-purpose and some larger systems. This research, part of a broader series on expert small models, paves the way for future healthcare-related AI developments, including deidentification and bio-medical entity extraction models. Our study underscores the potential of tailored neural translation models and the LLMs-in-the-loop methodology to advance the field through improved data generation, evaluation, agent, and modeling techniques.

Keywords: Machine Translation, Medical Translation, Clinical Translation, Deep Learning, LLMs-in-the-loop, Expert Small Models

---

[1] Corresponding Author: joseph@aimped.ai

1. Introduction

The ability to think and communicate is significantly impacted by language, which has become increasingly important due to technological advancements and global internet access. English has emerged as a widely used lingua franca, with numerous individuals worldwide learning it to facilitate communication [1, 2]. In societies where multiple languages are spoken, effective communication is crucial for cross-border interactions and daily exchanges [3]. As a result, there is a growing demand for translation, which plays a vital role in academic language teaching and enables the global dissemination of information [3, 4]. However, translating poses challenges as it involves transferring meaning between different languages, and neglecting the diverse range of text types can hinder the accuracy and variety of translations [5].

Medical translation is crucial in bridging communication gaps in the healthcare field. Throughout history, languages such as Latin, Greek, Arabic, Hebrew, Syriac, and Persian have been translated for medical education [6, 7]. English emerged as the dominant language in medical education during the 19th century, and today, it is recognized as the universal language for scientific communication, with most medical research articles written in English. This is especially important considering the vast number of medical articles published annually, which amounts to approximately 10 million [8-10].

Medical translation plays a significant role in knowledge transfer across various language domains, encompassing drug prospectuses, medical books, patient notes, and articles [11-13]. With the widespread use of English in medical education and the preference among doctors in multilingual Arab countries, accurate and efficient translation of medical texts has become increasingly essential for interprofessional communication, prescription writing, and report generation [12, 14]. Particularly amid the global concern of the coronavirus pandemic, the demand for comprehensible medical information in English has grown exponentially, highlighting the importance of human translators in improving access to healthcare and enhancing telemedicine practices [15, 16].

Due to the high cost and time constraints associated with expert translation, there is a growing need to develop affordable, high-quality, and accessible machine translation (MT) solutions in the medical field [17, 18]. Effective medical translation is crucial in knowledge sharing and can significantly improve healthcare outcomes [19]. Although English is widely used in medicine, many medical studies are written in languages other than English, necessitating medical text translation between English and other languages.

In the medical field, professionals use various types of texts to communicate with each other, such as discharge summaries, medications, case studies, case notes, epicrises, academic studies, and imaging reports. These texts contain specific terms and require accurate and prompt translation to avoid errors.

The advancement of artificial intelligence algorithms has paved the way for developing machine translation (MT) systems that can effectively address medical translation requirements [20].

These systems leverage the analysis of paired text samples and can be trained to cater to specific medical domains, enabling the generation of accurate translations [20]. This study focuses on developing medical text translation models using the Neural Machine Translation (NMT) approach, which has gained attention for its ability to optimize translation performance [21]. NMT employs a sequence-to-sequence (seq2seq) architecture with Recurrent Neural Network (RNN) components, consisting of an encoder and a decoder, to convert source text into target text in different languages [22, 23]. To handle longer sequences, variants of RNN cells such as Long Short-Term Memory (LSTM) or Gated Recurrent Units (GRU) are utilized, ensuring relevant information transfer and enhancing translation accuracy over longer-term dependencies [24, 25].

2. Background

With the impact of globalization, interaction has increased in every field, such as medicine. In this context, accurate and effective translation of medical texts between different languages is vital for disseminating research findings, sharing clinical data, and overcoming language barriers so scientists can work in parallel with their colleagues in other nations.

Users may use a translation to get a general idea of a context. However, translation accuracy is crucial in some applications, such as medicine, where zero errors are required. A mistranslation between patient and physician can jeopardize patient safety. While Statistical Machine Translation (SMT) progress is slowing down, promising methods such as neural networks are needed [26]. Machine translation has also been applied to the

medical field due to advances in language technologies. For example, in the US, both regional and national translations of medical documents have been carried out [27]. While these translations were previously considered insufficient, developing technologies show they can soon provide translations of the required quality [19].

Machine Translation (MT) is closely related to advanced rule-based systems or statistical phrase-based methods and dates back many years [28]. NMT, first introduced in 2013 [29], has become more popular recently [30]. In this context, a new neural network model called RNN Encoder-Decoder was proposed by researchers in 2014 [22]. In 2016, Google developed the Google Neural Machine Translation (GNMT) model using deep learning techniques [31, 32], which the German company DeepL developed as a CNN-based model in 2017. Today, open-source translation solutions trained by Hugging Face have paved the way for researchers to translate independently [30].

In this study, we present the results of medical translation models by fine-tuning MarianMT, which has recently been very popular in MT models, using medical texts from six different languages.

MarianMT is an efficient and self-contained NMT framework written entirely in C++ with minimal dependencies. It relies only on Boost, CUDA, or a BLAS library, allowing barrier-free optimization at all levels. MarianMT features its automatic differentiation engine based on dynamic computation graphs. This enables implementing new models, custom operators, and GPU kernels without changing the core library. Optimizations like meta-algorithms for multi-node training, efficient batched beam search, and compact model implementations can be done in performant C++ code. MarianMT demonstrates state-of-the-art results on the WMT2017 English-German news translation task. It also shows case studies for automatic post-editing and grammatical error correction research. Experiments exhibit MarianMT's high training and translation speeds, achieving 30x faster training than Nematus on 8 GPUs. MarianMT is a high-performance, flexible, self-contained NMT framework allowing C++-level optimization. Its efficiency makes it suitable for cutting-edge NMT research [33].

Recently developed Large Language Models (LLM) have been used in many fields. Moslem, Haque [34] showed that large language models such as GPT-3.5 can be used for adaptive MT. Their study indicates that GPT-3.5 can produce more consistent translations by imitating these examples when given sample translation sentences. It is also shown that GPT-3.5 can successfully perform tasks such as terminology extraction and terminology-constrained translation with zero training. The results show that GPT-3.5 can produce quality translations that compete with robust decoder-encoder-based MT systems. The work of Moslem, Haque [34] is essential because it shows that large language models have great potential in adaptive MT. There are services on the web that provide acceptable quality translations like Moslem, Haque [34] did. However, these translation services perform lower than their general performance in domain-specific translations such as medical areas.

It is known that in some cases larger models with more parameters, such as LLMs, do not produce better scores. In this context, in a study [35], it is stated that the Marian model [33] with 7.6 mn parameters produces better scores using the ClinSpEn-2022[2] benchmark dataset compared to much larger models such as Clinical-NLLB [36] (54 bn) and Clinical-WMT21fb (4.7 bn) [37].

Within the scope of this study, results are obtained with the AI Amplified MT model using the benchmark dataset used by Han, Gladkoff [35] and presented in the result section of this study. It is determined that the results we obtained are above the scores of all three studies given by Han, Gladkoff [35].

3. Methodology

Medical texts are generally characterized by the unique features of medical terminology and meaningful and long sentences in different fields. The primary objective of this study is to achieve zero-error translation of medical texts with grammatical features such as passive constructions and third person.

The literature survey shows that translation models have been developed by fine-tuning MarianMT in studies such as English-to-Malayalam [18, 38], English-Ukrainian [39] and improving the Norwegian Translation

---

[2] https://temu.bsc.es/clinspen/

model [40]. However, it can be stated that there are few medical translation models developed in different languages, and there is still a great need for medical text translation models.

Within the scope of this study, medical translation models have been developed in six languages (Table 1).

Table 1: Medical Translation Language Pairs

Languages "from" and "to" English
Deutsch
Turkish
French
Romanian
Spanish
Portuguese

For evaluation, the "Blue Score," "Rouge Score," "Meteor score," and "Bert score" were used to compare the results of the models. In many MT models, these scores compare the results obtained.

The Bilingual Under Evaluation (BLEU) score is a metric used to evaluate a generated sentence against a reference sentence, where a perfect match produces a score of one and an ideal mismatch produces a score of zero. [41]. The Metric for Evaluation of Translation with Explicit ORdering (METEOR) score is an automated metric for MT evaluation based on a generalized concept of unigram matching between machine- and human-generated reference translations [42].

Recall-Oriented Understudy for Gisting Evaluation (ROUGE) is a recall-oriented metric that compares summary quality by counting the number of overlapping words or n-grams between the summary and the ideal outline [43, 44]. BERT score is an automatic evaluation metric for text generation. Analogously to common metrics, it computes a similarity score for each token in the candidate sentence with each token in the reference [45].

3.1. Datasets

This study aims to translate medical texts into six different languages, as mentioned before. Custom datasets synthetically generated, and publicly available datasets were also used in the model development. Domain experts evaluate and curate high-quality language pairs derived from in-house labeled data and synthetically generated content. The sentence pairs and token numbers of the developed models are shown in Table 2.

Within the scope of this study, in addition to the data on the Opus.nlp [46] web page, scientific articles, and medical texts belonging to each language were scraped, and parallel corpora were prepared by us to be translated from English to the specified target languages and from these target languages to English. The scrapping process was based on the condition that the source language must have an English translation. Publicly available datasets published by language pairs were also used [47].

Table 2: Medical Translation Base Model, Sentence Pairs, and Token Numbers

Models	Base model (Helsinki-nlp/opus)	Sentence Pairs and Token Numbers					
		Sentence Pairs			Token Numbers		
		Train	Val.	Test	Train	Val.	Test
en-es-en	mt-en-es/mt-es-en	85,439	2,004	2,005	124,533,333	294,835	319,815
en-de-en	mt-en-de/mt-de-en	1,450,808	2,913	2,914	232,339,417	469,100	468,107
en-fr-en	mt-en-fr/mt-fr-en	1,905,850	2867	2868	25,521,977	377,787	381,422
en-tr-en	tatoeba-en-tr/mt-tr-en	852,402	1,712	1,712	55,426,674	109,203	112,049
en-ro-en	mt-tc-big-en-pt/mt-roa-en	1,438,810	3,615	3,616	57,599,188	29,852,963	29,852,093
en-pt-en	mt-tc-big-en-pt/mt-tc-big-pt-en	1920276	2889	2890	158,078,074	82,144,779	82,144,380

OPUS is a continuously updated collection of translated texts from the web. In the OPUS project, data is made available to users in various formats to help transform and align free online datasets, add linguistic annotation, and make it easier to work with the data. The main goal of Opus is to support data-driven NLP, mainly statistical MT. For this reason, it provides parallel datasets in various formats that can be used to train standard MT models. However, OPUS provides the data "as is" without guarantees or warranties. All preprocessing and

alignment in OPUS datasets are done automatically, and no manual corrections are made [46]. In addition to our in-house datasets, SciElo [48], Mespen [49], EMEA [50], and ELRC [51] datasets were also used in the language translation model studied in the context of the developed medical translation models.

3.2. Experimental Setup and Metrics

We evaluated the performance of our developed medical translation models using proprietary test datasets from AI Amplified. The evaluation metrics employed in this study include BLEU Score, BERT Score, METEOR Score, and ROUGE Score. We compared our results against those obtained from Google Translate, DeepL, and OPUS translation models across all language pairs.

For the English-to-German medical translation model specifically, we extended our comparison to include Claude-3. The Claude-3-Opus model achieved scores of 0.511 for BLEU, 0.745 for ROUGE, and 0.741 for METEOR on our test set. Due to practical constraints such as time and cost, this broader comparison including large language models was limited to this language pair (Table 2).

To further assess the quality of our translations, we innovatively employed ChatGPT and Claude AI as impartial judges. These LLM-based judges compared the outputs of Google Translate against our developed MT models (Tables 7 and 8). This approach was chosen over human translation to mitigate time and cost constraints while still providing valuable insights into translation quality.

For this comparative evaluation, we randomly selected 100 sentences from our test set. The prompts used to instruct the LLM judges in this MT evaluation process are detailed below.

3.2.1 Claude-3-Opus prompt:

<<Human: You are a professional translator and interpreter specializing in healthcare and biomedical. You will compare and evaluate the two translations.

I want you to help me in assessing translations from English to Turkish. The translated texts are from the healthcare and biomedical domain. Your evaluation should focus on adherence to medical terminology, simplicity, clearness, and accuracy, and you are expected to assign a score ranging from 0 to 100 to each translation. You MUST think, evaluate yourself, and do not generate text. Your ultimate output MUST only consist of the evaluation scores for each translation only in a JSON format as in <JSON></JSON> XML tags.

The output format:
<JSON> {"Model-1": score, "Model-2": score} </JSON>
Do you understand your task?

Assistant: Certainly, I understood. You will provide two different translations from English to Turkish, and I will evaluate as a medical translator expert and assign scores to each one by adherence to medical terminology, simplicity, consistency, and accuracy. The format will be a JSON file containing only scores. Ok, let us start.

Human: English text: I have a headache.
Model-1: Başım ağrıyor.
Model-2: Başımda ağrı var.>>

3.2.2 GPT-4-Turbo prompt:

<<System: You are a professional translator and interpreter specializing in healthcare and biomedical. You will compare and evaluate between two translations.

User: I want you to help me assess translations from English to Turkish. The translated texts are from the healthcare and biomedical domain. Your evaluation should focus on adherence to medical terminology, simplicity, clearness, and accuracy, and you are expected to assign a score ranging from 0 to 100 to each translation. You MUST think, evaluate yourself, and do not generate text. Your ultimate output MUST only consist of the evaluation scores for each translation in JSON format. Do you understand your task?

Assistant: Certainly, I understood. You will provide two different translations from English to Turkish, and I will evaluate as a medical translator expert and assign scores to each one by adherence to medical terminology, simplicity, consistency and accuracy. The format will be a JSON file containing only scores. Ok, let us start.

User: English text: I have a headache.
Model-1: Baş ağrım var.
Model-2: Başımda ağrı var.
Assistant: {"Model-1": 90, "Model-2": 80}>>

The potential of Large Language Models (LLMs) in translation studies was evaluated as part of this research. A notable challenge when using LLMs for translation is their inability to provide fuzzy-match suggestions. To address this, we randomly selected 10,000 sentences from the medical domain and generated their embeddings using the "multilingual-e5-large" model [52]. These embeddings were stored using FAISS libraries. We then selected 725 sentences from our existing test set for prediction. During the prediction process, we retrieved the two sentences most similar to the input sentence from the pre-generated set of 10,000, based on cosine similarity scores, to create a fuzzy match. The performance scores of these LLM-based predictions are presented in Table 3-6.

For the development of our Machine Translation (MT) models, we prepared datasets in TMX, Moses, and Dublin Core formats. Data preparation involved several steps: removing HTML tags, decoding Unicode characters, and calculating character count ratios between sentence pairs. To mitigate anomalies, we selected sentence pairs with character count ratios between 0.8 and 1.25, considering the specific characteristics of each language. This process resulted in the creation of train, test, and development datasets, with the number of sentence pairs and tokens for each detailed in Table 2.

Following data cleaning, sentence alignment, and corpus preparation based on predetermined ratios, we employed sentence transformers to assess the semantic relationships between sentence pairs. The semantic closeness of each pair was quantified using cosine similarity. We retained only those sentence pairs with a cosine similarity score of 0.90 or higher, thus finalizing our datasets through these filtration and transformation processes. Table 2 also presents the embeddings utilized in the development of our medical translation models.

The translation models were optimized using the following hyperparameters: a batch size of 16 for both training and evaluation, epoch sizes ranging from 20 to 50, and a learning rate of 2e-5. Model development involved training sessions lasting between 40 and 130 hours, depending on the specific language pair and dataset size.

4. Result

Extensive research in language translation reveals a continued demand for high-end translation services, particularly in the medical field. To meet this need, we have introduced a medical translation model covering six languages meticulously designed for use by healthcare professionals and various stakeholders.

This model performed remarkably well when evaluated against our in-house test datasets. The medical translation field currently needs more shared open-source benchmark test data. In conclusion, the findings presented by AI Amplified originate from our in-house test datasets, which have been precisely crafted in line with AI best practices and algorithmic principles.

The performance of our translation models has been rigorously evaluated using well-established benchmark scores such as BERT, BLEU, METEOR, and ROUGE Scores, which have been widely used in numerous studies (Tables 3, 4, 5, 6).

Table 3: Comparison of Results of Trained Language Models (BERT Scores)

Model/ Languages	BERT Score				
	Aimped	DeepL	Google	Base-Helsinki	GPT4-Turbo
de-to-en	**0.969**	0.935	0.937	0.906	0.955
en-to-de	**0.928**	0.920	0.916	0.884	0.886
en-to-es	**0.957**	0.930	0.931	0.925	0.926
es-to-en	**0.981**	0.935	0.939	0.930	0.967
en-to-fr	**0.950**	0.940	0.948	0.932	0.937
fr-to-en	**0.976**	0.946	0.950	0.934	0.969
en-to-pt	**0.954**	0.945	0.947	0.940	0.938
pt-to-en	**0.979**	0.947	0.951	0.972	0.972
en-to-ro	**0.964**	0.921	0.924	0.931	0.851
ro-to-en	**0.993**	0.953	0.955	0.950	0.973
en-to-tr	0.886	**0.892**	0.882	0.826	-
tr-to-en	**0.959**	0.928	0.926	0.876	-

Analysis of the BERT and BLEU scores demonstrates that our models achieve highly satisfactory and statistically significant results. The strategy of fine-tuning with high-quality, domain-specific datasets has yielded exceptionally positive outcomes in specialized medical translation tasks. These results underscore the effectiveness of our approach in capturing the nuances of medical terminology and context.

Table 4: Comparison of Results of Trained Language Models (BLEU Scores)

Model/ Languages	BLEU Score				
	Aimped	DeepL	Google	Base-Helsinki	GPT4- Turbo
de-to-en	**65.44**	60.2	60.43	45.83	54.13
en-to-de	**56.49**	52.59	49.77	36.93	44.3
en-to-es	**66.63**	51.8	51.27	48.7	48.99
es-to-en	**69.6**	55.36	56.19	52.17	52.12
en-to-fr	**62.15**	56.8	60.52	51.62	56.49
fr-to-en	**66.3**	62.02	63.07	52.54	58.4
en-to-pt	**62.03**	55.73	56.98	51.85	51.52
pt-to-en	**66.89**	58.16	61.05	58.2	54.59
en-to-ro	**75.04**	48.1	50.19	54.49	38.06
ro-to-en	**89.71**	61.5	62.58	59.37	58.06
en-to-tr	42.99	**45.06**	42.38	28.83	-
tr-to-en	48.86	**50.62**	49.37	29.2	-

Despite the relatively modest size of our datasets, our models achieve notable performance due to the high quality of the curated sentence pairs. This underscores the importance of data quality over quantity in developing effective translation models. Moreover, the efficiency of our approach is highlighted by the short fine-tuning times required, which not only accelerates model development but also suggests potential for rapid adaptation to new medical subdomains or emerging terminology.

Table 5: Comparison of Results of Trained Language Models (METEOR)

Model/ Languages	METEOR Score				
	Aimped	DeepL	Google	Base-Helsinki	GPT4- Turbo
de-to-en	**0.835**	0.812	0.815	0.716	0.794
en-to-de	**0.792**	0.758	0.748	0.677	0.690
en-to-es	**0.853**	0.756	0.757	0.739	0.739
es-to-en	**0.882**	0.800	0.811	0.782	0.793
en-to-fr	**0.827**	0.785	0.816	0.764	0.790
fr-to-en	**0.867**	0.824	0.853	0.800	0.831
en-to-pt	**0.829**	0.792	0.797	0.769	0.771
pt-to-en	**0.872**	0.831	0.849	0.833	0.829
en-to-ro	**0.869**	0.722	0.732	0.759	0.575
ro-to-en	**0.956**	0.834	0.848	0.830	0.827
en-to-tr	**0.686**	0.699	0.675	0.553	-
tr-to-en	0.769	0.775	**0.779**	0.600	-

Our models demonstrate consistently high performance across multiple evaluation metrics. In addition to the strong BERT and BLEU scores previously mentioned, the METEOR and ROUGE scores also indicate exceptional translation quality. These results collectively underscore the robustness of our approach.

However, it is crucial to note a significant limitation in our evaluation process. Due to the scarcity of standardized, open-source test sets specifically designed for benchmarking medical translation models, all reported scores are derived from test sets based on our proprietary in-house datasets. While this allows for a consistent evaluation across our models, it may limit direct comparability with other studies or systems using different datasets.

This limitation underscores the need for the development and adoption of standardized, publicly available test sets in the field of medical translation. Such resources would greatly enhance the ability to compare different approaches and models across the research community[3].

Table 6: Comparison of Results of Trained Language Models (ROUGE Scores)

Model/ Languages	ROUGE Score				
	Aimped	DeepL	Google	Base-Helsinki	GPT4- Turbo
de-to-en	**0.841**	0.817	0.822	0.723	0.769
en-to-de	**0.798**	0.765	0.758	0.673	0.668
en-to-es	**0.866**	0.775	0.776	0.763	0.732
es-to-en	**0.886**	0.812	0.819	0.790	0.768
en-to-fr	**0.847**	0.809	0.842	0.796	0.790
fr-to-en	**0.871**	0.830	0.858	0.807	0.810
en-to-pt	**0.850**	0.814	0.825	0.779	0.774
pt-to-en	**0.877**	0.834	0.855	0.809	0.804
en-to-ro	**0.897**	0.804	0.809	0.810	0.773
ro-to-en	**0.961**	0.849	0.855	0.841	0.803
en-to-tr	0.742	**0.752**	0.733	0.626	-
tr-to-en	0.787	**0.790**	**0.790**	0.634	-

AI-Amplified's translation models demonstrate exceptional performance across multiple evaluation metrics. The scores obtained for BERT, BLEU, METEOR, and ROUGE consistently rank among the highest in our comparative analysis, with only a few minor exceptions. This comprehensive evaluation underscores the robustness and accuracy of our models in medical translation tasks.

In this work, we also compared the results of Google Translate and our MT model [53] in six languages using two LLM models, GPT4-Turbo (Figure 1 and Table 7) and Claude-3-Opus (Figure 2 and Table 8), as referees.
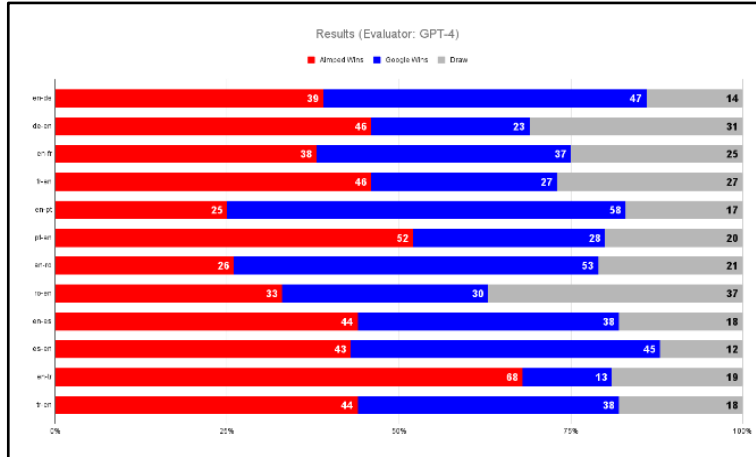

Figure 1: Refereed by GPT-4-Turbo Results

In the MT comparison for GPT4-Turbo with a total of 1200 instances, AI-Amplified outperformed Google Translate in 504 instances, while Google Translate performed better in 437 instances, and both models achieved the same score in 259 instances.

---

[3] https://github.com/ai-amplified/models/tree/main/medical_translation/test_data

Table 7: AI Amplified MT & Google Translate Translation Comparison Results
(Refereed by GPT4-Turbo)

MT Models & Languages	outperformed		Draw
	Aimped	Google Translate	
en-de	39	47	14
de-en	46	23	31
en-fr	38	37	25
fr-en	46	27	27
en-pt	25	58	17
pt-en	52	28	20
en-ro	26	53	21
ro-en	33	30	37
en-es	44	38	18
es-en	43	45	12
en-tr	68	13	19
tr-en	44	38	18
Total	**504**	**437**	**259**

In the context of language samples, using GPT4-Turbo as a referee, our models produce better scores than Google Translate models in most translations.
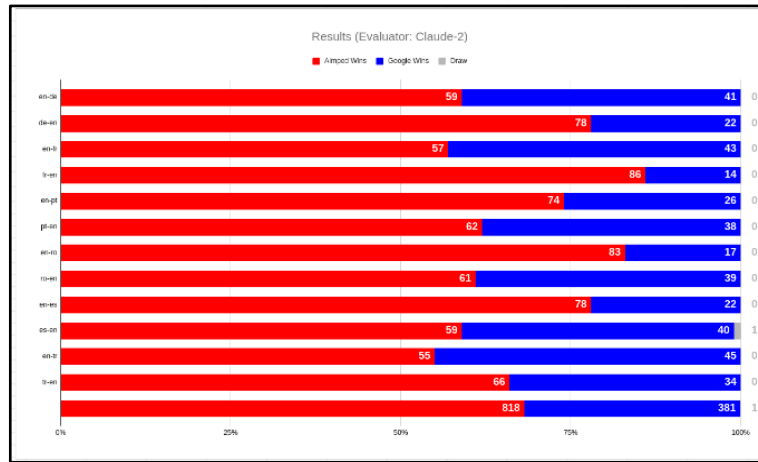

Figure 2: Refereed by Claude-3-Opus Results

In the MT comparison for Claude-3-Opus with a total of 1200 samples, AI-Amplified outperformed Google Translate in 818 samples, while Google Translate outperformed in 381 samples, with both models achieving the same score in only one sample.

Table 8: AI Amplified MT & Google Translate Translation Comparison Results
(Refereed by Claude-3-Opus)

MT Models & Languages	outperformed		Draw
	Aimped	Google Translate	
en-de	59	41	0
de-en	78	22	0
en-fr	57	43	0
fr-en	86	14	0
en-pt	74	26	0
pt-en	62	38	0
en-ro	83	17	0
ro-en	61	39	0
en-es	78	22	0
es-en	59	40	1
en-tr	55	45	0
tr-en	66	34	0
Total	**818**	**381**	**1**

In the context of language samples, using Claude-3-Opus, as a referee, our models produce better scores than Google Translate models in all translations. Sentence pairs of our MT models developed by AI-Amplified are available on our GitHub page[4].

To investigate transfer learning methods of clinical texts with multilingual pre-trained language models (MPLMs) in a study evaluating the performance of medical text translations between English and Spanish [35], Clinical-Marian [33], Clnical-NLLB [36] and Clinical-WMT21fb [37] models were compared. The ClinSpEn-2022[5] competition dataset was used for the evaluation. BLUE, METEOR, ROUGE, and COMET scores were compared. The ClinSpEn-2022 dataset used in this study was evaluated as a benchmark and the scores we obtained as AI Amplified were compared with those obtained by Han, Gladkoff [35] (Table 9).

Table 9: Comparison of Translation Model Performance (ClinEspen 22 benchmark)

Task–I: Clinical Cases (CC) EN→ES					
MT-fine-tuning Models	SACREBLEU	METEOR	COMET	BLEU	ROUGE-L-F1
Clinical-Marian	38.18	0.6338	0.4237	0.365	0.6271
Clnical-NLLB	37.74	0.6273	0.4081	0.3601	0.6193
Clinical-WMT21fb	34.3	0.5868	0.3448	0.3266	0.5927
Aimped	**38.43**	**0.6444**	-	**0.366**	**0.6288**

Task–II: Clinical Terms (CT) EN←ES					
MT-fine-tuning	SACREBLEU	METEOR	COMET	BLEU	ROUGE-L-F1
Clinical-Marian	26.87	0.5885	0.9791	0.2667	0.672
Clinical-NLLB	28.57	0.5873	1.029	0.2844	0.671
Clinical-WMT21fb	24.39	0.584	0.8584	0.2431	0.6699
Aimped	**37.85**	**0.6141**	-	**0.3777**	**0.7146**

Task–III: Ontology Concept (OC) EN→ES					
MT-fine-tuning	SACREBLEU	METEOR	COMET	BLEU	ROUGE-L-F1
Clinical-Marian	39.1	0.6262	0.9495	0.3675	0.7688
Clinical-NLLB	41.63	0.6072	0.918	0.3932	0.7477
Clinical-WMT21fb	40.71	0.5686	0.9908	0.3859	0.7199
Aimped	**44.14**	0.579	-	**0.4302**	0.7283

The results show that the MT model we developed as AI-Amplified scores above the scores obtained by Han, Gladkoff [35] with the relevant benchmark dataset in most cases. Han, Gladkoff [35] state that the Marian model is much smaller than the other two models. However, the Marian model surpassed these two models in terms of the score it obtained. As a result, it is revealed that models with many more parameters do not always produce good scores. Data quality and fine-tuning may be more important than model size alone. Therefore, it is shown that LLMs may not necessarily be better and that the quality of the data set and training is also essential.

5. Conclusion

Machine Translation (MT) has become increasingly crucial in artificial intelligence due to globalization and technological advancements. The need for accurate translations, particularly in sharing scientific research and medical terminology across languages, has grown exponentially. This study demonstrates that AI-based MT, specifically tailored for the medical domain, plays a critical role in meeting this demand.

Our research introduced a novel "LLMs-in-the-loop" approach to develop specialized neural machine translation models for medical texts. We focused on six language pairs: English to and from Spanish, German, French, Romanian, Turkish, and Portuguese. The models employ encoder-decoder architectures with LSTM units, trained on custom parallel corpora compiled from scientific articles and medical texts.

---

[4] https://github.com/ai-amplified

[5] https://temu.bsc.es/clinspen/

Key findings of our study include:

1. Small, specialized models trained on high-quality in-domain data can outperform larger, general-purpose models. For instance, our English-German model achieved BERT and BLEU Scores of 0.969 and 65.44, surpassing GPT-4 Turbo's scores of 0.955 and 54.13 respectively.

2. The LLMs-in-the-loop methodology, incorporating synthetic data generation, rigorous model evaluation, and agent orchestration, significantly enhanced model performance.

3. When evaluated by LLM models like GPT-4 and Claude-3 acting as expert judges, our models were preferred over Google Translate in 68-86% of sample evaluations across all languages.

4. Balancing domain-specific data with out-of-domain data (10-20% ratio) during fine-tuning helps maintain translation quality for general expressions while excelling in medical terminology.

5. The importance of starting with a strong general translation model before fine-tuning with domain-specific data was highlighted.

Our approach of combining medical domain-specific datasets with a proportion of out-of-domain data yielded successful results. However, we noted that fine-tuning a general language model with purely domain-specific data can decrease translation quality for out-of-domain texts. Our method of merging domain-specific and general datasets mitigated this issue.

While our results are promising, it's important to note that they were obtained using our proprietary datasets due to limited availability of open-source medical translation test sets. This underscores the need for more publicly available resources in this field.

Looking ahead, the development of AI models working with more languages and larger datasets holds great potential for advancing medical translation. As AI Amplified, we are committed to continually improving our datasets and models to achieve higher performance scores.

This study not only demonstrates the effectiveness of our specialized MT models in the medical field but also lays the groundwork for future developments in healthcare-related natural language processing, including planned models for deidentification and healthcare entity extraction.

Our medical translation models are accessible on our website [53][6], where demo translations can be tested. We believe these tools will significantly contribute to the global health community by facilitating knowledge sharing and accelerating scientific communication across language barriers.

Limitations

Our translation model has been meticulously designed and extensively trained to address the complex requirements of the healthcare and biomedical domains. While it excels within these highly specialized fields, it is crucial to acknowledge certain limitations:

1. Domain Specificity: The model's performance may not maintain the same high standards when applied to domains outside healthcare and biomedicine. Users should carefully consider this when contemplating applications in other fields.

2. Data Dependency: The model's proficiency is closely tied to the medical datasets used in training. While comprehensive, these datasets may not cover all possible medical scenarios or emerging terminologies.

3. Language Pair Constraints: The model's effectiveness has been demonstrated for specific language pairs. Its performance may vary for language combinations not included in this study.

4. Evaluation Metrics: While we used a range of metrics (BLEU, BERT, METEOR, ROUGE) and LLM-based evaluations, these may not capture all aspects of translation quality, particularly nuanced medical contexts.

---

[6] https://aimped.ai/models?page=1&domain=Healthcare&task=Translation

5. Continuous Updates: Medical knowledge and terminology evolve rapidly. Regular updates to the training data will be necessary to maintain the model's relevance and accuracy.

Future Improvements:

1. Expanding the diversity and volume of medical datasets could potentially yield higher performance scores and broader applicability.
2. Incorporating more language pairs and dialectal variations could enhance the model's global utility.
3. Developing domain-adaptive techniques could improve the model's versatility across different medical sub-specialties.

We encourage users to be mindful of these limitations when deploying the model and to contribute to its ongoing refinement through feedback and real-world application insights.

References

1.   Al-Sobhi, B.M.S. and A.S. Preece, *Teaching English speaking skills to the Arab students in the Saudi school in Kuala Lumpur: Problems and solutions.* International Journal of Education and Literacy Studies, 2018. 6(1): p. 1-11.

2.   Jenkins, S.R., *The Narrative Arc of TATs: Introduction to the JPA Special Section on Thematic Apperceptive Techniques.* J Pers Assess, 2017. 99(3): p. 225-237.

3.   Anderson, L., L. Gavioli, and F. Zanettin, *Translation and Interpreting for Language Learners. An introduction (TAIL).* 2018.

4.   Kim, M., *Research on translator and interpreter education*, in *The Routledge Handbook of Translation Studies.* 2013.

5.   Al-Otaibi, G.M., *An Integrated Approach of TBLT and GBP to Medical Translation Instruction: A Suggested Model.* International Journal of Linguistics, 2022. 14(2).

6.   Fischbach, H., *Translation and Medicine.* American Translators Association Scholarly Monograph Series. 1998, Amsterdam: John Benjamins Publishing Company.

7.   McMorrow, L., *Breaking the Greco-Roman mold in medical writing*, in *Translation and Medicine.* 1998, John Benjamins Publishing Company. p. 13-13.

8.   Kaplan, R.B., *English–the accidental language of science.* The dominance of English as a language of science: Effects on other languages and language communities, 2001. 84: p. 3-26.

9.   Montgomery, S.L., *English and Science: realities and issues for translation in the age of an expanding lingua franca.* The Journal of Specialised Translation, 2009. 11(1): p. 6-16.

10.   Salager-Meyer, F., *Origin and development of English for Medical Purposes. Part I: Research on written medical discourse.* Medical Writing, 2014. 23(1): p. 49-51.

11.   Al-Jarf, R.S., *Multiple Arabic Equivalents to English Medical Terms.* International Linguistics Research, 2018. 1(1).

12.   Alasbahy, W. and M. Shamsi, *Translating Medical Terminologies.* Journal of English Studies in Arabia Felix, 2023. 2(1): p. 1-9.

13.   Haddad, R., *Translation of Medical Terminology: An Analytical Study.* 1997, Yarmouk University.

14.   medlinkstudents.com. *Study Medicine in Europe in English.* 2023 [cited 2023 05.10.2023]; Available from: https://www.medlinkstudents.com/study-medicine-in-europe/.

15.   Dušek, O., et al., *Machine translation of medical texts in the khresmoi project.* In Proceedings of the Ninth Workshop on Statistical Machine Translation, 2014: p. 221-228.

16.   Yu, P., Y. Zhu, and K. Wong, *Model and Verification of Medical English Machine Translation Based on Optimized Generalized Likelihood Ratio Algorithm.* Journal of Sensors, 2021. 2021: p. 1-7.

17. Karliner, L.S., et al., *Do professional interpreters improve clinical care for patients with limited English proficiency? A systematic review of the literature.* Health services research, 2007. 42(2): p. 727-754.

18. Randhawa, G., et al., *Using machine translation in clinical practice.* Can Fam Physician, 2013. 59(4): p. 382-3.

19. Wolk, K. and K. Marasek, *Neural-based machine translation for medical text domain. Based on European Medicines Agency leaflet texts*, in *Conference On Enterprise Information Systems/International Conference On Project Management/Conference On Health And Social Care Information Systems And Technologies, Centeris/Projman / Hcist 2015*. 2015. p. 2-9.

20. Koehn, P., et al., *Open source toolkit for statistical machine translation: Factored translation models and confusion network decoding.* callison-burch.github.io, 2007.

21. Bahdanau, D., K. Cho, and Y. Bengio, *Neural Machine Translation by Jointly Learning to Align and Translate.* 2015.

22. Cho, K., et al., *Learning phrase representations using RNN encoder-decoder for statistical machine translation.* arXiv preprint arXiv:1406.1078, 2014.

23. Sutskever, I., O. Vinyals, and Q.V. Le, *Sequence to sequence learning with neural networks.* Advances in neural information processing systems, 2014. 27.

24. Chung, J., et al., *Empirical evaluation of gated recurrent neural networks on sequence modeling.* arXiv preprint arXiv:1412.3555, 2014.

25. Hochreiter, S. and J. Schmidhuber, *Long short-term memory.* Neural Comput, 1997. 9(8): p. 1735-80.

26. Costa-Jussà, M.R., M. Farrús, and J. Serrano Pons, *Machine translation in medicine: A quality analysis of statistical machine translation in the medical domain.* 2012.

27. Kirchhoff, K., et al., *Application of statistical machine translation to public health information: a feasibility study.* J Am Med Inform Assoc, 2011. 18(4): p. 473-8.

28. Stein, D., *Machine translation: Past, present and future.* Language technologies for a multilingual Europe, 2018. 4(5).

29. Kalchbrenner, N. and P. Blunsom. *Recurrent continuous translation models.* in *Proceedings of the 2013 conference on empirical methods in natural language processing.* 2013.

30. Pilch, A., R. Zygała, and W. Gryncewicz. *Quality Assessment of Translators Using Deep Neural Networks for Polish-English and English-Polish Translation.* in *2022 12th International Conference on Advanced Computer Information Technologies (ACIT).* 2022. IEEE.

31. Coldewey, D. and F. Lardinois, *DeepL schools other online translators with clever machine learning.* Techcrunch. com, 2017.

32. Wu, Y., et al., *Google's neural machine translation system: Bridging the gap between human and machine translation.* arXiv preprint arXiv:1609.08144, 2016.

33. Junczys-Dowmunt, M., et al., *Marian: Fast neural machine translation in C++.* arXiv preprint arXiv:1804.00344, 2018.

34. Moslem, Y., R. Haque, and A. Way, *Adaptive machine translation with large language models.* arXiv preprint arXiv:2301.13294, 2023.

35. Han, L., et al., *Neural Machine Translation of Clinical Text: An Empirical Investigation into Multilingual Pre-Trained Language Models and Transfer-Learning.* arXiv preprint arXiv:2312.07250, 2023.

36. Costa-jussà, M.R., et al., *No language left behind: Scaling human-centered machine translation.* arXiv preprint arXiv:2207.04672, 2022.

37. Tran, C., et al., *Facebook ai wmt21 news translation task submission.* arXiv preprint arXiv:2108.03265, 2021.

38.   Saji, J., et al. *English-to-Malayalam Machine Translation Framework using Transformers*. in *2022 IEEE 19th India Council International Conference (INDICON)*. 2022. IEEE.

39.   Maksymenko, D., et al., *Controllability for English-Ukrainian Machine Translation by Using Style Transfer Techniques*. 2023.

40.   Hellebust, D. and I.A. Lawal, *Improving Norwegian Translation of Bicycle Terminology Using Custom Named-Entity Recognition and Neural Machine Translation*. Electronics, 2023. 12(10): p. 2334.

41.   Papineni, K., et al. *Bleu: a method for automatic evaluation of machine translation*. in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 2002.

42.   Banerjee, S. and A. Lavie. *METEOR: An automatic metric for MT evaluation with improved correlation with human judgments*. in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*. 2005.

43.   Ganesan, K., *Rouge 2.0: Updated and improved measures for evaluation of summarization tasks*. arXiv preprint arXiv:1803.01937, 2018.

44.   Lin, C. *Rouge: A package for automatic evaluation of summaries*. in *Association for Computational Linguistics.Text summarization branches out*. 2004.

45.   Zhang, T., et al., *Bertscore: Evaluating text generation with bert*. arXiv preprint arXiv:1904.09675, 2019.

46.   Opus.nlp. *The open parallel corpus*. 2023 [cited 2023 05.11.2023]; Available from: https://opus.nlpl.eu/.

47.   OPUS. *OPUS Dataset*. 2023 [cited 2023 11.10.2023]; Available from: http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf.

48.   SciElo. *SciElo Dataset*. 2023 [cited 2023 12.10.2023]; Available from: https://figshare.com/articles/dataset/A_Large_Parallel_Corpus_of_Full-Text_Scientific_Articles/5382757.

49.   Mespen. *Mespen Dataset*. 2023 [cited 2023 12.10.2023]; Available from: https://zenodo.org/record/3562536.

50.   EMEA. *EMEA Dataset*. 2023 [cited 2023 11.10.2023]; Available from: https://joint-research-centre.ec.europa.eu/system/files/2016-05/2014_08_LRE-Journal_JRC-Linguistic-Resources_Manuscript.pdf.

51.   ELRC. ELRC Dataset. 2023 [cited 2023; Available from: https://shorturl.at/J8Xye.

52.   huggingface.co. multilingual-e5-large. 2023 [cited 2023 11.10.2023].

53.   AI Amplified. AI Amplified, Home of Exclusive AI. 2023 [cited 2023 18.10.2023]; Available from: https://aimped.ai/models?page=1&domain=Healthcare.