On Creating an English-Thai Code-switched Machine Translation in Medical Domain

Parinthapat Pengpun[1,†], **Krittamate Tiankanon**[2,†], **Amrest Chinkamol**[2,3], **Jiramet Kinchagawat**[2], **Pitchaya Chairuengjitjaras**[2,4], **Pasit Supholkhan**[5], **Pubordee Aussavavirojekul**[2], **Chiraphat Boonnag**[7], **Kanyakorn Veerakanjana**[2,5,6], **Hirunkul Phimsiri**[4], **Boonthicha Sae-jia**[4], **Nattawach Sataudom**[2], **Piyalitt Ittichaiwong**[2,5,6,*], **Peerat Limkonchotiwat**[3,*]

[1]Bangkok Christian International School, [2]PreceptorAI team, CARIVA Thailand, [3]Vidyasirimedhi Institute of Science and Technology, [4]Chulalongkorn University, [5]Mahidol University, [6]King's College London, [7]Independent researcher

[†]Equal contribution, [*]Corresponding authors

piyalitt.itt@preceptorai.tech, peerat.l_s19@vistec.ac.th

Abstract

Machine translation (MT) in the medical domain plays a pivotal role in enhancing healthcare quality and disseminating medical knowledge. Despite advancements in English-Thai MT technology, common MT approaches often underperform in the medical field due to their inability to precisely translate medical terminologies. Our research prioritizes not merely improving translation accuracy but also maintaining medical terminology in English within the translated text through code-switched (CS) translation. We developed a method to produce CS medical translation data, fine-tuned a CS translation model with this data, and evaluated its performance against strong baselines, such as Google Neural Machine Translation (NMT) and GPT-3.5/GPT-4. Our model demonstrated competitive performance in automatic metrics and was highly favored in human preference evaluations. Our evaluation result also shows that medical professionals significantly prefer CS translations that maintain critical English terms accurately, even if it slightly compromises fluency. Our code and test set are publicly available https://github.com/preceptorai-org/NLLB_CS_EM_NLP2024.

1 Introduction

Medical-domain machine translation (MT) serves as a critical component in enhancing healthcare quality and disseminating medical knowledge. By providing accurate translations of medical research publications, MT enables local medical professionals without English proficiency to overcome the linguistic barrier and have access to more medical academic resources, which are predominantly written in English (Pecina et al., 2014; McLean et al., 2013). This accessibility is crucial for facilitating continuing medical education, which has been shown to be an effective strategy for healthcare professionals to enhance care quality and patient outcomes (Bloom, 2005; Randhawa et al., 2013).

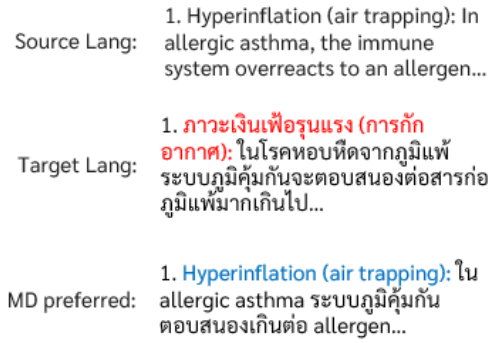


Figure 1: Example of how Google NMT alters the meaning of the sentence when translating from the source language (English) to the target language (Thai) and how it compares with the translations that Medical Doctors (MDs) prefer. "Hyperinflation" (abnormal increase of lung volume) is translated into Hyperinflation in economic context; "air trapping" (retention of air in the lungs distal to an obstruction) is translated into "air quarantine".

Despite the research in the English-Thai MT field, most of the common MT techniques are not yet suitable for the medical domain due to the need for precise translation of medical terminology. Translating medical terminology accurately is challenging due to the lack of equivalent Thai terms for some English medical keywords. Thus, it is understandable why common techniques of machine translation in the medical domain MT — such as Google NMT, No Language Left Behind (NLLB) (Team et al., 2022), GPT-4 (Achiam et al., 2023), or Gemini-Pro (Team et al., 2023) — cannot translate medical keywords precisely, as shown in Figure 1. Previous studies have aimed to improve translation accuracy through terminology integration (Nieminen, 2023; Semenov et al., 2023), yet their application to the Thai language and specifi-

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 6055–6073
November 12-16, 2024

cally in the medical field remains limited.

Rather than focusing on enhancing the accuracy of terminology translation, our objective is to preserve medical terms in their original English form within the translated output, thus removing the need to handle terminology translation. This strategy is characterized as **code-switching (CS)**, deviating from conventional monolingual translation practices. Apart from reducing the task's complexity, framing this problem as CS is also preferred by Medical Doctors (MDs). A previous study from the Thai COVID-19 administrative unit (Toomaneejinda et al., 2022) suggests that medical professionals prefer retaining medical and technical keywords in English, with the rest of the translation in Thai, to avoid potential translation inaccuracies. Other studies (Alqurashi, 2022; Nur'Aini and Fanani, 2019; Wood, 2018; Rodríguez Tembrás, 2016) also suggest that this phenomenon exists in other languages as well, including Arabic, Javanese, and Spanish. As shown in Figure 1, keeping medical terms in English preserves the original meaning and is preferred by MDs.

However, CS datasets are usually scarce, preventing researchers from developing a CS translator. Several initiatives have been made to address this issue. For example, the LinCE dataset (Aguilar et al., 2020) is one of the publicly available CS datasets focused on the general domain created to mitigate this problem. Additionally, various studies have focused on enhancing CS dataset efficiency through augmentation techniques (Gowda et al., 2022; Sugiyama and Yoshinaga, 2019; Menacer et al., 2019), pre-training techniques (Yang et al., 2020; Iyer et al., 2023), synthetic CS dataset generation techniques (Tarunesh et al., 2021; Appicharla et al., 2021; Xu and Yvon, 2021). However, these previous works did not focus on medical texts. Furthermore, the adaptation of such research to the Thai language context remains limited. This, in turn, leads to a significant scarcity of the English-Thai CS translation dataset, especially in the medical domain, as shown in Table 1.

In this work, we aim to achieve two objectives: (i) address the data scarcity in medical-domain English-Thai MT and (ii) validate our hypothesis that doctors prefer CS translations to monolingual ones in medical contexts. To achieve the first objective, we create a new English-Thai CS dataset for the medical domain. Our process begins by generating initial CS translations (pseudo-CS) of English medical texts using a widely available monolingual translator. During this translation process, we apply a keyword masking algorithm to preserve key medical terms. We then hire an annotator to post-process and clean a portion of the generated translations, as opposed to doing the whole translation process, to save both time and resources.

To achieve our second objective, we conduct comprehensive evaluations using both traditional MT metrics and MD evaluations to confirm our hypothesis. This involves evaluating 52 models, including an off-the-shelf translator, large language models (LLMs), and our fine-tuned CS models based on NLLB. Furthermore, we assess the translations for factual accuracy and MD preference by having MDs directly rate them and by distributing preference ranking questionnaires, respectively. Our findings reveal that our fine-tuned CS model based on NLLB is preferred by MDs due to its factual accuracy, even though it achieves a lower score in traditional metrics when compared to off-the-shelf translators like Google NMT. These results indicate that traditional MT metrics are inadequate for evaluating medical-domain translations and that MDs prefer CS translations over monolingual ones.

To summarize, our key contributions are:

- We propose the first benchmark dataset specifically designed for medical English-Thai CS translation.
- We develop the first open-source model tailored for medical English-Thai CS translation, which is preferred by MDs over Google NMT and GPT-3.5 systems.
- We present a comprehensive evaluation of various models on our benchmark, including 52 models, 8 metrics, and 27 MD evaluators. These results reveal a misalignment between traditional MT metrics and the judgments of medical professionals, and underscore the preference of MDs for CS translations.
- Our code, test set, and translation models are publicly available at `https://github.com/preceptorai-org/NLLB_CS_EM_NLP2024`.

2 Related Works

2.1 Neural Machine Translation (NMT)

NMT has gained prominence in both academic and commercial sectors, largely due to advancements in Transformer-based architectures (Vaswani et al., 2023). Various models designed for NMT, such as mT5 (Xue et al., 2021), mBART (Liu et al., 2020), OPUS (Tiedemann and Thottingal, 2020),

Table 1: Comparison of English-to-Thai translation datasets. Given #Samples is the number of samples, #Sentences is the number of sentences, and #English Tokens is the count of English tokens within all of the target translations. The Ratio En:All column reflects the proportion of English token usage compared to other languages within the target translations. The CS column calculation is based on the percentage of English tokens in target translations.

Dataset	#Samples	#Sentences	#English Tokens	Ratio En:All	Domain	CS?
FLORES-200	2,009	3,251	442	1.3%	Wikidata	✗
Thai US Embassy	615	9,303	11,176	4.7%	News	✗
SCB_MT_EN-TH_2020	1,001,752	1,084,328	8,124,662	1.4%	General	✗
Our Pseudo-CS	63,982	188,037	640,951	16.1%	Medical	✓

and NLLB (Team et al., 2022), have been developed. However, as previously mentioned, most of these NMT models cannot perform precise terminology translations, which disqualifies them from the medical domain.

The emergence of Large Language Models (LLMs) has further changed the NMT landscape. LLMs, such as GPT-4, have demonstrated emergent abilities in machine translation, excelling in paragraph-level translations without the need for extensive fine-tuning on large parallel corpora (Wei et al., 2022; Kocmi et al., 2023). A few studies (Zhu et al., 2023; Robinson et al., 2023; Bawden and Yvon, 2023) have suggested that LLMs are not yet effective translators, especially in low-resource languages including Thai. Nevertheless, it has been shown that LLMs are proficient at generating CS data for many languages (Yong et al., 2023). To the best of our knowledge, no research has comprehensively investigated the performance of LLMs (both closed and open-source) in translating the Thai language, especially in the medical domain.

2.2 Evaluation Metrics for NMT

The assessment of Machine Translation (MT) quality is a continually evolving field of research. Several automated metrics have been proposed to measure MT quality through lexical analysis, including BLEU (Papineni et al., 2001), chrF (Popović, 2015), METEOR (Banerjee and Lavie, 2005), and Translation Edit Rate (Snover et al.). Furthermore, various neural-network-based metrics have been devised to enhance the measurement of MT quality using neural networks: COMET (Rei et al., 2020), Mask-Language-Modeling Score (Zheng et al., 2021), and BLEURT (Sellam et al., 2020). While these metrics provide effective means to assess translations, several studies have also shown their limitations, indicating that these metrics do not always correlate well with human evaluation (Mathur et al., 2020; Callison-Burch et al., 2006; Roy et al., 2021). It still remains unclear whether these metrics align well with the specific use cases of medical-domain MT, where the precise translation of terminology is more important than overall sentence fluency.

Human evaluation is also crucial, especially in a medical context where technically accurate and human-readable translations are necessary. Graham et al. (2013) attempted to better standardize crowd jurisdictions on with Likert-type continuous rating scales. After that, a band scale was proposed by (Menacer et al., 2019; Tarunesh et al., 2021) for more consistent qualitative evaluation among human judges. Bai et al. (2022); Askell et al. (2021) introduced the concept of the Elo Rating to benchmark multiple translation systems' performances. Elo Rating allows for a leaderboard-like relative comparison between these systems. All these works provided valuable perspectives on how to conduct human preference evaluations on NMT models. Using these studies as a basis of our human evaluation on translation models, we choose to use an improved Elo-based metric called the Glicko score, which was developed by Glickman to accounts for the uncertainty in Elo-based calculations (Glickman, 1995a,b, 1999).

3 Benchmark Data Collection

3.1 English Text collection

Our dataset of English medical texts was collected from an in-house LLM-based application designed to tackle intricate medical questions, with an emphasis on differential diagnosis and multiple-choice problems. The dataset consists of 10,000 medical excerpts, with an additional 250 excerpts reserved for testing purposes.

3.2 Pseudo-translation Masking and Generation

In the absence of an existing CS translator, we adopt a masking approach to create our CS translation dataset. Inspired by the Language Identification (LID) translation pipeline (Ramadurgam and Mundada), this method involves augmenting a standard monolingual translator with a keyword masking strategy. By identifying the important medical keywords and selectively translating the rest of the sentence, this method allows for the retaining of domain-specific terminology after translation. Using this, we establish a *pseudo-CS* translator, which forms the basis of our benchmark dataset.

The overview of the procedure for the Keyword masking algorithm is as follows:

1. Use GPT-4 to identify medical keywords in the original English sentence. We specifically chose GPT-4 for its capabilities in Named Entity Recognition (NER) of medical terms (see Appendix F for our evaluation) and its flexibility, which allows us to manually adjust the types of terms to include or exclude in order to mimic medical CS as closely as possible.
2. Replace each medical keyword with a unique placeholder. This results in an English text where medical terms are masked.
3. Process the masked English text through an MT model to obtain a masked Thai text. In this text, the non-medical parts are translated, while the placeholders remain untouched.
4. Substitute the unique placeholder tokens with their original English medical keywords to produce the final Thai-English pseudo-CS translation.

Expanding on Step 3, the masked sentences from the previous step are translated to generate pseudo-CS translations. All English excerpts and their corresponding masked versions are processed through the keyword-masked Google translation system, resulting in Thai pseudo-CS translations. To ensure proper alignment between the English and Thai+English (as in the target translations contain CS between Thai and English) content, both the original English excerpts and the CS translations are segmented into chunks of fewer than 256 tokens. We then re-validate that the English and Thai texts contain the same number of chunks.

Regarding the size of our dataset, our dataset size is competitive when compared to existing code-switched datasets. In terms of the number of samples, our dataset has 64K records, while a single language pair within the LinCE has 7k to 67k records. For the total token counts, our dataset has 640K tokens, while a single language pair within LinCE has 33k to 808k tokens. We split our dataset into 63,982 English-to-Thai CS translation pairs for training and 1,100 translation pairs for the test benchmark.

3.3 Test dataset Constitution

To ensure the quality of our test set, we employ human annotators to recheck its fluency with the instruction in Figure 4. After annotation, the dataset goes through an NLP pipeline to correct typos and adjust spacing. Subsequently, it then undergoes another round of validation by MDs to ensure its readability and factual accuracy. The MDs make further corrections to improve the accuracy of the translation chunks compared to their source text. This process ensures that every sentence and CS word is correct as verified by MDs; LLMs only serve to reduce the time spent here.

3.4 Training Data Procedure

As mentioned in the previous step, we utilize both human annotators and MDs to assess the quality of our test set. However, applying the same process to the training data would be 64 times more expensive than the test set. To mitigate this issue, we employ data augmentation and filtering techniques to improve the quality of our training dataset.

- **Data Augmentation**: Inspired by the back-translation augmentation method (Sugiyama and Yoshinaga, 2019), we prompt Gemini-Pro to rephrase the existing CS translation while retaining a roughly similar CS boundary. The rephrased CS sentences are then back-translated to generate corresponding English sentences, thereby constructing new translation pairs.
- **Filtering**: We filter the training CS translation dataset based on a rough measure of its quality. We use the COMET score metric (which assesses semantic similarity) to estimate the quality of the translation dataset and filter out samples that did not achieve a COMET score of at least 0.6.

4 Experimental Setup

4.1 Baseline Models

Off-the-Shelf Translator (1 model). We leverage Google NMT as our off-the-shelf translator, utiliz-

ing the version released on January 17, 2024.

Large Language Models (18 models). This set includes OpenThaiGPT 7B, OpenThaiGPT 13B, Typhoon 7B (Pipatanakul et al., 2023), SeaLLM 7B (Nguyen et al., 2023), Llama2 7B, Llama2 13B (Touvron et al., 2023), Google's Gemini-Pro, GPT 3.5, and GPT 4. Each large language model has two prompt variants: one prompted to generate monolingual translations (denoted as the "MN" variant) and another prompted to generate CS translations (denoted as the "CS" variant). All local LLMs (OpenThaiGPT, Typhoon, SeaLLM, Llama2) are evaluated using bfloat16 precision. The rest are accessed via API calls with default settings and a temperature of 0.1. The GPT models used are based on the 1106-preview version. The Gemini-Pro model is utilized as presented through the API on January 17th, 2024.

CS Baseline (6 models): We employ a state-of-the-art language translation model, NLLB. We utilized NLLB 3.3B as a base model and fine-tuned it on six variants of our training dataset (Section 3.4) as follows:

- NLLB-1: Initial 64k dataset (64k)
- NLLB-2: Augmentation of the 64k dataset (64k)
- NLLB-3: Initial dataset plus augmentation of the 64k dataset (128k)
- NLLB-4: Filtered 64k dataset (30k)
- NLLB-5: Filtered augmentation dataset (40k)
- NLLB-6: Filtered 64k dataset plus filtered augmentation dataset (70k)

It is important to note that our augmentation technique, which utilizes an LLM to rephrase translation pairs, likely results in an overall improvement in the COMET score of the augmented dataset. Setting a fixed COMET score threshold for dataset filtration results in the augmented filtered dataset containing more records than the initially filtered dataset. The exact training configurations are listed in Appendix C. In addition, the inference is performed using bfloat16 quantization.

4.2 Evaluation Metrics

We evaluate 52 translation systems—26 with the masking system and 26 without the masking system during the inference step (see Section 3.2)—using standard machine translation metrics and MD evaluators to further validate our results.

4.2.1 Machine Metric Evaluation

We evaluate all our translation models using our MD-annotated test set. The following metrics are employed for evaluation:

- Lexical score (BLEU (Papineni et al., 2001), chrF (Popović, 2015), METEOR (Banerjee and Lavie, 2005)).
- Translation Edit Rate, which includes Character Error Rate (CER) and Word Error Rate (WER).
- Semantic score (COMET (Rei et al., 2020; Guerreiro et al., 2023)).
- CS boundary F1 Score, inspired by (Sterner and Teufel, 2023). The CS boundary F1 Score is calculated using the common formula, i.e., the harmonic mean of *precision* and *recall*. *Precision* is defined as the proportion of correctly identified English words in the generated translation compared to those in the reference translation. *Recall* is the proportion of English words in the reference translation that are correctly identified in the generated translation.

Details on the implementation of these metrics are provided in Appendix D.

4.2.2 Human Evaluation

Anticipating a lower number of human respondents, we only selected the MD-preferred models for human evaluation. To ensure that each model is compared against each other at least 30 times within a comprehensive evaluation of 52 models, it would require at least 39,000 data points, or approximately 390 respondents, to achieve a statistically significant result. By selecting only 8–11 models, we can reduce this number to 2,000 data points or 20 respondents. Our methodology is as follows.

Before human evaluation We assess the *factual* accuracy of translations produced by each model by soliciting evaluations from four medical professionals. These professionals assess each translation's factual correctness using our specific rubric. The evaluation process is outlined as follows:

- 10 English texts are randomly selected from our test set and translated using 52 different translation systems.
- Medical professionals are instructed to individually rate each translation for factual correctness on a scale from 1 to 7, according to a detailed rubric provided in Table 4. Each medical professional is unaware of the translations' source models, and the sequence of translations they evaluate is randomized to prevent bias.
- The score for each model, as rated by an evaluator, is determined by calculating the median of the scores assigned to its translations.
- An arithmetic mean of the median scores across

all evaluators is then calculated to assign each model its preliminary final score.

Subsequent human evaluations are conducted only on model categories (differentiated by base translation model and usage of keyword masking) that achieved ratings higher than the Google NMT.

Human Preference Evaluation We perform a human preference evaluation to determine which models are preferred by crowd-sourced medical practitioners, assessing their preference for translations as well as their performance on our human dataset. Note that, in this step, we only ask medical professionals on our chatbot platform to assist in evaluating translations.

- We evaluate 10 translation models and the human label based on the previous step. This involves selecting one model from each category identified in the last step.
- We design a self-administered, web-based survey using a ranking format to enhance participants' experiences (Revilla and Höhne, 2020). Given that ranking five items requires approximately 40 seconds (Sauro et al., 2023) and our items consisted of a few sentences, we include ten ranking questions, each estimated to take approximately 1.5 minutes to complete.
- For each participant, we randomly sample 10 English texts from our benchmark test set. For each text, we present five versions of the translations, each randomly selected from the list of "comprehensible" models along with the human-annotated translation.
- Participants are asked to rank each translation sample based on the factual accuracy of the sentence and their preference (as shown in Figure 2). We specifically instruct them to disregard the proportion of English text retained in the translation (as shown in Figure 8).
- Subsequently, we use the preference data from the human evaluation to calculate the *Glicko* Rating, measuring the comparative preference of each model against the others and the human annotator. The initial Glicko rating is set according to the standard, with $r = 1500$ and $RD = 350$.

Moreover, we implement a simple filter to monitor each participant's response time to the survey. Participants who completed the survey in less than 5 minutes were flagged as potentially invalid, and their choice ordering was re-examined to confirm the validity of their responses. A row is flagged as an invalid record if the choice order remains nearly identical across questions despite variations in the translation model.

Figure 2: Example Questionnaire User Interface

5 Main Experimental Results

The full evaluation results for all 52 models on our dataset are presented in Table 2. We categorize the results into two groups: (i) traditional machine translation (MT) metrics and (ii) human preference.

Traditional MT metrics. We present our two best models: NLLB-1 (initial 64k dataset) and NLLB-4 (filtered 64k dataset) with Mask. These models demonstrate remarkable results among open-source models and achieve competitive results against closed-source models. NLLB-1 (without masking) achieved the top CS F1 score in its category, showing remarkable performance compared to off-the-shelf models and LLMs. NLLB-4 with Mask also obtained a competitive CS F1 score compared to those models equipped with masks, rivaling GPT-4 + Mask. In conclusion, these results underscore the importance of training models on source data rather than relying on off-the-shelf models. The NLLB results show that we achieve comparable outcomes to those of Google-NMT and Gemini-Pro on machine translation metrics, namely BLEU and chrF. On the other hand, the code-switch metric (CS F1) indicates that NLLB models retain medical

Table 2: Full Evaluation Result using our dataset. The "MN" suffix indicates that the LLM employs a Monolingual translation prompt, whereas the "CS" suffix denotes the use of a CS translation prompt. The term "Mask" indicates the system's use of a keyword masking algorithm, as described in Section 3.2. Each NLLB variant is labeled according to the dataset used for training the NLLB, as detailed in Section 4.1. "Fact." indicates factual accuracy score as described in Section 4.2.2

Model Variant	CS F1	BLEU	chrF	CER	WER	COMET	METEOR	Fact.
Gemini-Pro-CS	0.132	0.353	0.595	0.599	0.686	0.849	0.622	5.750
Gemini-Pro-MN	0.110	0.352	0.599	0.426	0.526	**0.854**	0.630	5.375
Google-NMT	0.119	**0.385**	**0.617**	**0.392**	**0.480**	0.815	**0.650**	3.125
GPT-3.5-CS	0.141	0.208	0.504	0.833	0.987	0.671	0.491	3.625
GPT-3.5-MN	0.114	0.205	0.504	0.599	0.775	0.687	0.494	3.875
GPT-4-CS	0.340	0.314	0.601	0.636	0.757	0.850	0.593	**6.250**
GPT-4-MN	0.132	0.282	0.581	0.511	0.660	0.847	0.597	5.125
Llama2-13B-CS	0.058	0.012	0.189	5.226	6.104	0.153	0.129	1.125
Llama2-13B-MN	0.082	0.022	0.224	3.326	4.139	0.163	0.174	1.000
Llama2-7B-CS	0.074	0.012	0.171	5.361	6.141	0.159	0.117	0.500
Llama2-7B-MN	0.086	0.015	0.197	3.852	4.761	0.162	0.143	0.500
OpenThaiGPT-13B-CS	0.039	0.094	0.446	2.343	2.538	0.394	0.388	2.125
OpenThaiGPT-13B-MN	0.036	0.094	0.465	1.978	2.208	0.425	0.396	2.375
OpenThaiGPT-7B-CS	0.045	0.046	0.308	12.620	13.224	0.310	0.237	2.875
OpenThaiGPT-7B-MN	0.027	0.068	0.344	9.954	10.223	0.369	0.282	2.750
SeaLLM-7B-CS	0.035	0.017	0.242	11.678	11.165	0.235	0.188	2.000
SeaLLM-7B-MN	0.076	0.032	0.329	8.340	8.259	0.321	0.259	1.705
Typhoon-7B-CS	0.021	0.012	0.220	18.946	18.434	0.186	0.168	1.875
Typhoon-7B-MN	0.023	0.013	0.239	18.111	19.020	0.174	0.176	1.875
NLLB	0.107	0.140	0.432	0.610	0.906	0.530	0.405	2.500
NLLB-1	**0.475**	0.253	0.487	0.491	0.593	0.678	0.502	4.375
NLLB-2	0.230	0.262	0.548	0.448	0.612	0.720	0.546	3.375
NLLB-3	0.380	0.257	0.520	0.472	0.604	0.702	0.521	4.000
NLLB-4	0.452	0.272	0.520	0.461	0.577	0.710	0.532	3.875
NLLB-5	0.193	0.255	0.544	0.458	0.627	0.715	0.546	3.250
NLLB-6	0.286	0.264	0.539	0.456	0.606	0.711	0.541	4.000
Gemini-Pro-CS + Mask	0.628	0.301	0.512	0.668	0.716	0.704	0.543	5.500
Gemini-Pro-MN + Mask	0.644	0.314	0.529	0.461	0.517	**0.726**	0.562	5.750
Google-NMT + Mask	**0.647**	**0.327**	**0.531**	**0.458**	**0.509**	0.656	**0.564**	5.000
GPT-3.5-CS + Mask	0.574	0.212	0.463	0.839	0.953	0.631	0.468	5.250
GPT-3.5-MN + Mask	0.536	0.215	0.474	0.662	0.755	0.623	0.478	5.000
GPT-4-CS + Mask	0.612	0.265	0.500	0.682	0.758	0.724	0.515	**6.000**
GPT-4-MN + Mask	0.619	0.275	0.517	0.556	0.634	0.705	0.535	4.750
Llama2-13B-CS + Mask	0.052	0.011	0.164	6.050	7.205	0.142	0.110	1.000
Llama2-13B-MN + Mask	0.100	0.023	0.199	4.201	5.363	0.156	0.148	0.750
Llama2-7B-CS + Mask	0.013	0.005	0.127	6.091	7.175	0.144	0.079	0.500
Llama2-7B-MN + Mask	0.024	0.008	0.150	4.188	5.712	0.161	0.101	0.750
OpenThaiGPT-13B-CS + Mask	0.052	0.072	0.369	1.831	2.215	0.275	0.313	2.250
OpenThaiGPT-13B-MN + Mask	0.078	0.066	0.384	2.119	2.715	0.293	0.309	1.375
OpenThaiGPT-7B-CS + Mask	0.043	0.038	0.266	11.545	12.430	0.226	0.202	1.250
OpenThaiGPT-7B-MN + Mask	0.063	0.062	0.307	6.760	7.068	0.271	0.258	2.125
SeaLLM-7B-CS + Mask	0.048	0.016	0.223	10.167	9.953	0.204	0.166	1.375
SeaLLM-7B-MN + Mask	0.163	0.033	0.306	8.119	8.009	0.259	0.240	1.625
Typhoon-7B-CS + Mask	0.080	0.011	0.199	18.283	18.291	0.170	0.147	1.875
Typhoon-7B-MN + Mask	0.113	0.010	0.218	17.891	18.786	0.172	0.150	1.750
NLLB + Mask	0.523	0.183	0.423	0.556	0.719	0.533	0.424	4.125
NLLB-1 + Mask	0.578	0.237	0.457	0.515	0.605	0.645	0.479	4.625
NLLB-2 + Mask	0.637	0.240	0.475	0.506	0.612	0.644	0.489	4.750
NLLB-3 + Mask	0.605	0.237	0.464	0.511	0.608	0.648	0.481	5.125
NLLB-4 + Mask	0.599	0.250	0.472	0.502	0.596	0.651	0.493	4.875
NLLB-5 + Mask	0.642	0.242	0.478	0.504	0.609	0.645	0.493	3.625
NLLB-6 + Mask	0.628	0.241	0.473	0.505	0.605	0.646	0.489	4.750

keywords more effectively than off-the-shelf MT models.

Human preference. As shown in Table 3, human preference evaluation received responses from 23 medical doctors (MDs). The Glicko rating calculation results show that both NLLB models are preferred over Google NMT and LLMs like GPT-3.5. Both models are also almost equally preferred when compared to translations from Gemini-Pro models. Thus, we can summarize that machine translation metrics might not fully satisfy medical doctors' preferences. The results from the MT metrics contradict human preferences, which we will discuss further in Section 6.1. Confirming our hypothesis, we also found that MDs preferred CS translation over translating all words into the target language, as indicated by the CS F1 metric. MD preferences are discussed in more detail in the following section, Section 6.2.

6 Discussion

6.1 Automated Metrics Versus Factual Accuracy

The evaluation results reveal an unexpected outcome: Google NMT consistently achieves top scores across nearly all machine metrics among the 52 models, despite the lack of medical terminology preservation. Similarly, Google NMT with Mask dominates in almost every automated metric among

Table 3: Models sorted by their Glicko ratings with 95% confidence interval. Our fine-tuned NLLB models' scores are highlighted below

Model	Glicko MD
Human Annotated	1638.57 ± 49.39
Gemini-CS	1500.00 ± 50.31
Google NMT	1398.61 ± 50.07
GPT-3.5-MN	1316.40 ± 48.52
GPT-4-CS	1578.93 ± 49.84
NLLB-1	**1549.55** ± 52.05
Gemini-MN + Mask	1555.71 ± 55.18
Google NMT + Mask	1480.98 ± 53.12
GPT-3.5-CS + Mask	1394.69 ± 51.03
GPT-4-CS + Mask	1564.60 ± 48.10
NLLB-3 + Mask	**1532.28** ± 50.79

the masked models (a better rank breakdown can be seen in Table 7). Nevertheless, a closer examination of individual samples still reveals that Google NMT frequently translates medical terminology imprecisely (as shown in Figure 3). We hypothesize that Google NMT's superior performance in automated metrics is due to its fluency in translating non-essential parts of the medical text, which constitutes the bulk of our dataset. Conversely, the accuracy of medical-domain translations rather depends on the precise translation of critical medical terms, an area where Google NMT falls short. This is further supported by the minimal correlation between most automated metric scores and factual accuracy, especially among models that are rated higher than 3 in factual accuracy (see Figure 5).

In fact, the CS F1 metric addresses this issue by focusing on the preservation of key medical terms, demonstrating a stronger positive correlation with factual accuracy ratings. However, it is still not a comprehensive metric, as it only assesses the retention of English keywords without considering the quality of the Thai translation. A trade-off consideration between the retention of precise medical terms and the fluency of the overall translation may be necessary to develop a more suitable automated metric for medical translation tasks.

6.2 MD Evaluation

Our human evaluation within the MD population further supports our hypothesis that traditional automated metrics are not well-suited for medical-domain MT. This is shown by the significant correlation observed between Glicko ratings and both factual accuracy scores ($r = 0.698$) and CS F1 scores ($r = 0.516$), as opposed to the weak correlation (less than 0.3) between traditional automated metrics and Glicko ratings (seen in Figure 7).

Moreover, an in-depth analysis of questionnaire responses (shown in Figure 3) also presents a consistent picture. Google NMT provides a fluent Thai translation of the English text, but medical terminologies are still imprecisely translated. On the other hand, our NLLB model, despite exhibiting less fluency, successfully retains critical medical terminology in English. This also aligns with our hypothesis that traditional automated metrics tend to measure the fluency of the translation but not the precision of medical terminology translation. Therefore, in medical-domain translations, traditional automated metrics might not be adequate for

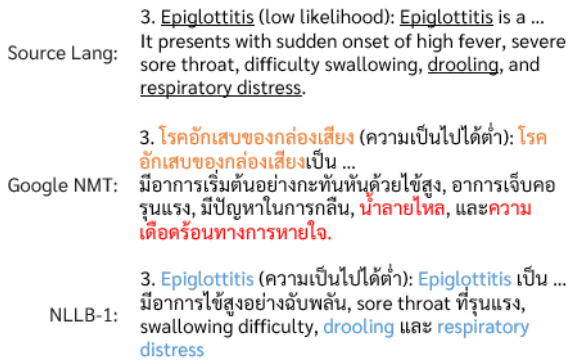


Figure 3: Real samples where our internal MDs and external MDs both report a preference for NLLB-1 CS translation over Google NMT. Red sections indicate medical keywords that Google NMT does not translate precisely. Orange sections indicate medical keywords that Google NMT translates precisely, but retaining them in English is still preferred. Blue sections indicate medical keywords that are retained in English and convey their meaning precisely.

7 Conclusion

This paper presents an approach for performing MT in the medical domain using a CS translation to generate translations preferred by medical professionals. We developed a method for generating CS translation data, trained a CS translation model leveraging this data, and evaluated its performance against multiple strong baselines. The experimental results demonstrate that although most automated metrics might be suitable for measuring translation fluency, they are inadequate for assessing factual accuracy or medical doctors' preference in the translations. While current MT technologies may offer monolingual translations with high fluency, medical professionals exhibit a clear preference for CS medical translations that accurately preserve crucial terms in English, even at the expense of fluency.

Limitations

There are inherent risks associated with machine translation (MT), particularly the potential for misinterpreting medical terminology and technical terms. While our models have shown promising results, there is still a possibility of inaccuracies in translation that could affect daily practice.

Moreover, there is still some potential for further improvement. We have not conducted extensive human preference evaluations on all 52 models because doing so would require more than 390 MD respondents, whom we cannot find or hire. Also, we have not optimized prompts for LLMs to produce the best CS translations yet. Our MDs deemed the translations generated by these prompts acceptable internally, so we selected them. Lastly, we have not conducted an extensive hyperparameter search for NLLB training. To limit the cost of the fine-tuning process, we selected the standard learning rate and learning rate scheduler that is used throughout the field and fixed it for the entire fine-tuning process of NLLB.

Ethical Considerations

Our human annotators were undergraduate students majoring in linguistics at a university in Thailand. We ensured that they received monthly monetary compensation at an industry-competitive salary. We compensated our annotators by first measuring their annotation speed in terms of the number of words processed per hour. After that, we established a monthly target for the annotators to achieve, and we paid our annotators a fixed salary.

Other human evaluators who respond to our questionnaire participate voluntarily. The participants were promised free usage of our upcoming product as compensation by randomly selecting five participants. In this regard, we have to collect their names and emails to prevent spamming attempts. For the remaining participants, we informed them that we would compensate for their work by releasing a questionnaire dataset without the respondents' information to the public domain, which we will release under a CC-BY-NC 4.0 license.

Regarding the licensing of models, we strictly adhere to the intended uses outlined by their respective licenses. The NLLB weight checkpoints we use as our pretrained weights are licensed under CC-BY-NC 4.0, which allows us to distribute our newly fine-tuned NLLB weights to the public for non-commercial use. We have also adhered to the Llama2 and SeaLLM Licenses by not using their outputs to enhance any language model and by restricting their use to research benchmarking purposes only. Additionally, we followed OpenAI and Google Gemini's Terms and Conditions strictly: we did not compete with OpenAI and Gemini's models but rather used them fairly for research purposes.

All local LLM inferences, NLLB fine-tuning, and NLLB inferences for translation were performed on a single A100 GPU, also with the maximum amount of batching possible. We used a total of 60 GPU hours for fine-tuning NLLB, 24 GPU hours for performing local LLM inferences, and 3.5 GPU hours for performing 24 variants of NLLB inferences.

Regarding a potential leak of personal information, our source English texts inherently contain no personal information, as they are outputs from our own LLM product with no personal information in the prompt. We conduct an initial screening of the test benchmark dataset regardless, which confirms the absence of personal information in any of the English texts. We also instructed our internal annotator to remove any identifying information in case any is found within the annotation process. Another potential concern arises when we collect names and email addresses from MD evaluators. However, we use this information solely for spam tracking purposes and do not disclose or utilize this personal data for any other reason, except to contact individuals later regarding compensations for free usage.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Gustavo Aguilar, Sudipta Kar, and Thamar Solorio. 2020. LinCE: A Centralized Benchmark for Linguistic Code-switching Evaluation. ArXiv:2005.04322 [cs] version: 1.

H.S. Alqurashi. 2022. Investigating the code switching phenomenon in private medical workplaces: A case study of some private hospitals in saudi arabia. *Journal of Language and Linguistic Studies*, 18(4):344–361.

Ramakrishna Appicharla, Kamal Kumar Gupta, Asif Ekbal, and Pushpak Bhattacharyya. 2021. IITP-MT

at CALCS2021: English to Hinglish neural machine translation using unsupervised synthetic code-mixed parallel corpus. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 31–35, Online. Association for Computational Linguistics.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, and Jared Kaplan. 2021. A General Language Assistant as a Laboratory for Alignment. ArXiv:2112.00861 [cs].

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. ArXiv:2204.05862 [cs].

Satanjeev Banerjee and Alon Lavie. 2005. 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*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Rachel Bawden and François Yvon. 2023. Investigating the translation performance of a large multilingual language model: the case of BLOOM. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 157–170, Tampere, Finland. European Association for Machine Translation.

Bernard S Bloom. 2005. Effects of continuing medical education on improving physician clinical care and patient health: a review of systematic reviews. *International journal of technology assessment in health care*, 21(3):380–385.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of Bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256, Trento, Italy. Association for Computational Linguistics.

Mark E Glickman. 1995a. A comprehensive guide to chess ratings. *American Chess Journal*, 3(1):59–102.

Mark E Glickman. 1995b. The glicko system.

Mark E. Glickman. 1999. Parameter Estimation in Large Dynamic Paired Comparison Experiments. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 48(3):377–394.

Thamme Gowda, Mozhdeh Gheini, and Jonathan May. 2022. Checks and Strategies for Enabling Code-Switched Machine Translation. ArXiv:2210.05096 [cs].

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous Measurement Scales in Human Evaluation of Machine Translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2023. xcomet: Transparent machine translation evaluation through fine-grained error detection.

Vivek Iyer, Edoardo Barba, Alexandra Birch, Jeff Pan, and Roberto Navigli. 2023. Code-switching with word senses for pretraining in neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12889–12901, Singapore. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, and Mariya Shmatova. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual Denoising Pre-training for Neural Machine Translation. ArXiv:2001.08210 [cs].

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics.

Michelle McLean, Deborah Murdoch-Eaton, and Sami Shaban. 2013. Poor english language proficiency hinders generic skills development: a qualitative study of the perspectives of first-year medical students. *Journal of Further and Higher Education*, 37(4):462–481.

Mohamed Menacer, David Langlois, Denis Jouvet, Dominique Fohr, Odile Mella, and Kamel Smaïli. 2019. Machine Translation on a parallel Code-Switched Corpus. In *Canadian AI 2019 - 32nd Conference on Canadian Artificial Intelligence*, Lecture Notes in Artificial Intelligence, Ontario, Canada.

Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, Hang Zhang, and Lidong Bing. 2023. SeaLLMs – Large Language Models for Southeast Asia. ArXiv:2312.00738 [cs].

Tommi Nieminen. 2023. OPUS-CAT Terminology Systems for the WMT23 Terminology Shared Task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 912–918, Singapore. Association for Computational Linguistics.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, Renqian Luo, Scott Mayer McKinney, Robert Osazuwa Ness, Hoifung Poon, Tao Qin, Naoto Usuyama, Chris White, and Eric Horvitz. 2023. Can generalist foundation models outcompete special-purpose tuning? case study in medicine.

Indah Nur'Aini and Qori Fanani. 2019. Code-switching between nurse-patient communication: Bilingual interaction society. *British (Jurnal Bahasa dan Sastra Inggris)*, 8:79.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, page 311, Philadelphia, Pennsylvania. Association for Computational Linguistics.

Pavel Pecina, Ondřej Dušek, Lorraine Goeuriot, Jan Hajič, Jaroslava Hlaváčová, Gareth JF Jones, Liadh Kelly, Johannes Leveling, David Mareček, Michal Novák, et al. 2014. Adaptation of machine translation for multilingual information retrieval in the medical domain. *Artificial intelligence in medicine*, 61(3):165–185.

Wannaphong Phatthiyaphaibun, Korakot Chaovavanich, Charin Polpanumas, Arthit Suriyawongkul, Lalita Lowphansirikul, Pattarawat Chormai, Peerat Limkonchotiwat, Thanathip Suntorntip, and Can Udomcharoenchaikit. 2023. Pythainlp: Thai natural language processing in python.

Kunat Pipatanakul, Phatrasek Jirabovonvisut, Potsawee Manakul, Sittipong Sripaisarnmongkol, Ruangsak Patomwong, Pathomporn Chokchainant, and Kasima Tharnpipitchai. 2023. Typhoon: Thai Large Language Models. ArXiv:2312.13951 [cs].

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Kalyani Ramadurgam and Surabhi Mundada. Translating Code-Switched Texts From Bilingual Speakers.

Gurdeeshpal Randhawa, Mariella Ferreyra, Rukhsana Ahmed, Omar Ezzat, and Kevin Pottie. 2013. Using machine translation in clinical practice. *Canadian family physician Medecin de famille canadien*, 59(4):382–383.

Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020. COMET: A Neural Framework for MT Evaluation. ArXiv:2009.09025 [cs].

Melanie Revilla and Jan Karem Höhne. 2020. How long do respondents think online surveys should be? new evidence from two online panels in germany. *International Journal of Market Research*, 62(5):538–545.

Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. ChatGPT MT: Competitive for High- (but Not Low-) Resource Languages. In *Proceedings of the Eighth Conference on Machine Translation*, pages 392–418, Singapore. Association for Computational Linguistics.

Vanesa Rodríguez Tembrás. 2016. Code-switching as a facework strategy within doctor-patient communication in a galician clinic. *Textos en Proceso*, 2:94–121.

Devjeet Roy, Sarah Fakhoury, and Venera Arnaoudova. 2021. Reassessing automatic evaluation metrics for code summarization tasks. *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*.

Jeff Sauro, Dylan Atkins, David Du, and Jim Lewis. 2023. How hard is it to rank items in surveys?

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Kirill Semenov, Vilém Zouhar, Tom Kocmi, Dongdong Zhang, Wangchunshu Zhou, and Yuchen Eleanor Jiang. 2023. Findings of the WMT 2023 Shared Task on Machine Translation with Terminologies. In *Proceedings of the Eighth Conference on Machine Translation*, pages 663–671, Singapore. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A Study of Translation Edit Rate with Targeted Human Annotation.

Igor Sterner and Simone Teufel. 2023. TongueSwitcher: Fine-Grained Identification of German-English Code-Switching. In *Proceedings of the 6th Workshop on Computational Approaches to Linguistic Code-Switching*, pages 1–13, Singapore. Association for Computational Linguistics.

Amane Sugiyama and Naoki Yoshinaga. 2019. Data augmentation using back-translation for context-aware neural machine translation. In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 35–44, Hong Kong, China. Association for Computational Linguistics.

Ishan Tarunesh, Syamantak Kumar, and Preethi Jyothi. 2021. From Machine Translation to Code-Switching: Generating High-Quality Code-Switched Text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3154–3169, Online. Association for Computational Linguistics.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

NLLB Team, 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].

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Anuchit Toomaneejinda, Amnat Paksasuk, and Pornthip Chertchinnapa. 2022. Identifying patterns of english code-mixing: A look into thai communication of the center for covid-19 situation administration. *Journal of Liberal Arts, Thammasat University*, 22(3):492–510.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. ArXiv:2307.09288 [cs].

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention Is All You Need. ArXiv:1706.03762 [cs].

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models.

N. Wood. 2018. Departing from doctor-speak: a perspective on code-switching in the medical setting. *Journal of General Internal Medicine*, 34:464–466.

Jitao Xu and François Yvon. 2021. Can you traducir this? machine translation for code-switched input. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 84–94, Online. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. ArXiv:2010.11934 [cs].

Zhen Yang, Bojie Hu, Ambyera Han, Shen Huang, and Qi Ju. 2020. Csp: Code-switching pre-training for neural machine translation. pages 2624–2636.

Zheng Xin Yong, Ruochen Zhang, Jessica Forde, Skyler Wang, Arjun Subramonian, Holy Lovenia, Samuel Cahyawijaya, Genta Winata, Lintang Sutawika, Jan Christian Blaise Cruz, Yin Lin Tan, Long Phan, Long Phan, Rowena Garcia, Thamar Solorio, and Alham Aji. 2023. Prompting multilingual large language models to generate code-mixed texts: The case of south East Asian languages. In *Proceedings of the 6th Workshop on Computational Approaches to Linguistic Code-Switching*, pages 43–63, Singapore. Association for Computational Linguistics.

Yuanhang Zheng, Zhixing Tan, Meng Zhang, Mieradilijiang Maimaiti, Huanbo Luan, Maosong Sun, Qun Liu, and Yang Liu. 2021. Self-Supervised Quality Estimation for Machine Translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3322–3334, Online

and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Multilingual machine translation with large language models: Empirical results and analysis.

A Fully listed Model Categories

- Google-translate-based model with keyword masking
- NLLB-based model with keyword masking
- OpenThaiGPT-7B-based model with keyword masking
- OpenThaiGPT-13B-based model with keyword masking
- Typhoon-7B-based model with keyword masking
- SeaLLM-7B-based model with keyword masking
- LLama2-7B-based model with keyword masking
- LLama2-13B-based model with keyword masking
- Gemini-Pro-based model with keyword masking
- GPT-3.5-based model with keyword masking
- GPT-4-based model with keyword masking
- Google-translate-based model without keyword masking
- NLLB-based model without keyword masking
- OpenThaiGPT-7B-based model without keyword masking
- OpenThaiGPT-13B-based model without keyword masking
- Typhoon-7B-based model without keyword masking
- SeaLLM-7B-based model without keyword masking
- LLama2-7B-based model without keyword masking
- LLama2-13B-based model without keyword masking
- Gemini-Pro-based model without keyword masking
- GPT-3.5-based model without keyword masking
- GPT-4-based model without keyword masking

B Prompts for Large Language Model Translation

CS Translation Prompt

```
You are a linguist with expertise in medicine
and had your training in Thailand.
You are well acquainted to how's Thai MD
usually code switched between Thai Language and
English when they're communicating medical-related
information among each other.
For instance, you never translate the following
English medical terms and jargons, symptoms,
technical terms, and pharmaceutical terms into Thai.

Hence, task is to examine the medical-related
information text input and translated them
into Thai with the previously given constraint
and information.
```

Monolingual Translation Prompt

```
Translate the following text input into Thai
in Medical Context
```

GPT4 Medical NER Prompt

```
Annotate the medical report with HTML-like tags.
The output should start with <annotated> and end
with </annotated>.
Use the following tags to annotate the
respective terms:
- <patho> for pathological and medical symptoms
terms
- <pharm> for pharmaceutical terms and drugs' names
- <taxo> for scientific names and taxonomical-
like names
- <anato> for anatomical terms
- <chem> for chemical names
- <med> for medical practices and jargons
FYI:
- Drug names sometimes start with a single charactor
followed by full stop then full name.
For example: A. Parafivir, B. Paracetamol.
- Anatomical terms must include limbs, organs, cells,
and organelle.
```

C NLLB Training Configuration

```
LoraConfig:
    r = 16,
    lora_alpha = 16,
    target_modules = ["q_proj", "v_proj"],
    lora_dropout = 0.1,
    bias = "none",


TrainingArguments:
    num_train_epochs = 10,
    evaluation_strategy = "steps",
    logging_strategy ="steps",
    save_strategy ="steps",
    eval_steps=5000,
    logging_steps=500,
    save_steps=5000,

    bf16=True,

    seed=42,
    data_seed=42,
```

```
warmup_ratio = 0.1,
learning_rate=10e-5,

per_device_train_batch_size= 3,
per_device_eval_batch_size= 4,

load_best_model_at_end=True,
metric_for_best_model="loss",
```

D Evaluation Metric Implementation details

The evaluation environment was established using Python 3.11, incorporating the following key libraries and their respective versions:

- **PyTorch 2.2.0:** Used for neural network-based computations and model loading, supporting the latest deep learning model features and optimizations.
- **NLTK 3.8.1:** Provided tools for text processing and evaluation metrics, including BLEU, METEOR, and CHRF scores.
- **PyThaiNLP 4.0.2:** Essential for processing the Thai language, used specifically for tokenizing Thai text and for the implementation of the NewMM tokenizer (Phatthiyaphaibun et al., 2023).
- **JiWER 3.0.2:** Employed for calculating Word Error Rate (WER) and Character Error Rate (CER), key in assessing model performance in speech recognition tasks.
- **Unbabel Comet 2.2.1:** Employed for calculating the COMET score using the XCOMET-XL (Guerreiro et al., 2023) model.

We implemented a Python script on our own to calculate the Glicko rating based on (Glickman, 1999). The RD/Glicko evaluation was established using an initial rating of 1500 and an RD of 350. All the ratings are calculated at once, eliminating the need for nondeterminism. The 95% confidence interval is reported using 2 times the RD.

E Disclaimer for Participants

Notice to Participants

- This study focuses exclusively on medical questions.
- Our system leverages LLM technology currently under development. Do not use the output as medical facts.
- Participant's inputs, system outputs and feedback will be reviewed and used to improve the system capability.
- To comply with Thai PDPA law, do not disclose real patient information or any patient identifiable information in general. Use hypothetical clinical cases only.

F GPT-4's Medical NER Performance

Although our confidence in GPT-4's capabilities in medical keyword extraction was already substantial, based on its performance in various analyses (Nori et al., 2023), we have conducted an experiment to determine the medical NER performance of multiple systems. Results are shown in Table 5

Table 5: Performance of multiple LLMs and tools used to perform medical NER. CS F1 refers to the F1 score in identifying medical keywords denoted as "CS F1".

Model	CS F1	Recall	Precision
GPT-4	0.30	0.49	0.25
GPT-3.5	0.29	0.47	0.23
Gemini-Pro	0.28	0.40	0.25
BioMedNER	0.03	0.05	0.03

The complete evaluation of GPT-4 reveals scores of 0.488 and 0.253 for average CS recall and average CS precision, respectively, making it the top system for low-resource languages. Among the options considered, GPT-4 stands out as a competent system, particularly in its ability to detect keywords akin to how medical doctors code-switch in Thai. Although the initial mask detection is not perfect, we asked MDs to review our test dataset. If a mask is missing or incorrect, we request the MD to examine and correct it for us. Consequently, the test dataset is accurately masked.

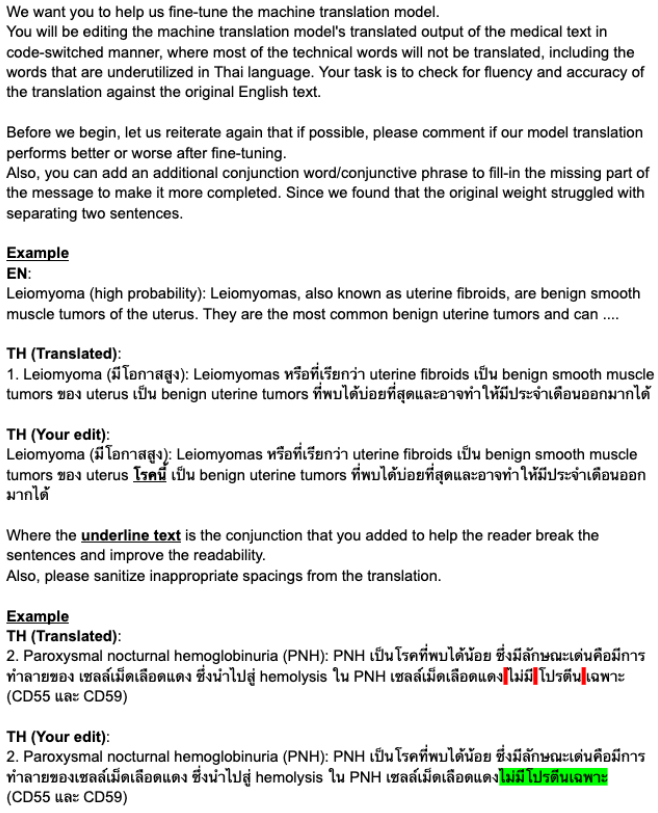
G Medical Doctor Annotator Instruction

Instruction

- Please review the annotations from human annotators in the Google Spreadsheet provided.
- Look for any serious errors in the labels.
- Pay attention to whether any technical words have been lost during the cleaning process.
- Leave a comment in the spreadsheet for any errors that you may find.

Table 4: Internal Evaluation metric

Band Score	**Factual Correctness**	
7	Fully contain all information, no addition or loses of information.	
6	Fully contain all information but might add some information that does not improve (increased in clarity) of the source text.	Fully contain all information and also correctly adding information that enhanced the source text.
5	Fully contain all information but might have some hallucination added in the translation but does not distort the information.	Losses of information that can safely disregarded.
4	Fully contain all information but have some hallucination added and minorly distort information.	Losses of information in such a way that might distort the information if the reader does not pay attention.
3	Fully contain all information but have a noticeable amount of hallucination that majorly distort the information.	Losses of information that majorly distort the information in such a way that misled the reader.
2	The reader can barely gain information from the translation.	Hallucination heavily distorted the information that led to misunderstanding by the reader.
1	The reader cannot gain the information from the translation.	The translation not relevant to the source text.



We want you to help us fine-tune the machine translation model.
You will be editing the machine translation model's translated output of the medical text in code-switched manner, where most of the technical words will not be translated, including the words that are underutilized in Thai language. Your task is to check for fluency and accuracy of the translation against the original English text.

Before we begin, let us reiterate again that if possible, please comment if our model translation performs better or worse after fine-tuning.
Also, you can add an additional conjunction word/conjunctive phrase to fill-in the missing part of the message to make it more completed. Since we found that the original weight struggled with separating two sentences.

Example
EN:
Leiomyoma (high probability): Leiomyomas, also known as uterine fibroids, are benign smooth muscle tumors of the uterus. They are the most common benign uterine tumors and can ....

TH (Translated):
1. Leiomyoma (มีโอกาสสูง): Leiomyomas หรือที่เรียกว่า uterine fibroids เป็น benign smooth muscle tumors ของ uterus เป็น benign uterine tumors ที่พบได้บ่อยที่สุดและอาจทำให้มีประจำเดือนออกมากได้

TH (Your edit):
Leiomyoma (มีโอกาสสูง): Leiomyomas หรือที่เรียกว่า uterine fibroids เป็น benign smooth muscle tumors ของ uterus **โรคนี้** เป็น benign uterine tumors ที่พบได้บ่อยที่สุดและอาจทำให้มีประจำเดือนออกมากได้

Where the **underline text** is the conjunction that you added to help the reader break the sentences and improve the readability.
Also, please sanitize inappropriate spacings from the translation.

Example
TH (Translated):
2. Paroxysmal nocturnal hemoglobinuria (PNH): PNH เป็นโรคที่พบได้น้อย ซึ่งมีลักษณะเด่นคือมีการทำลายของ เซลล์เม็ดเลือดแดง ซึ่งนำไปสู่ hemolysis ใน PNH เซลล์เม็ดเลือดแดง[ไม่มี]โปรตีน[เฉพาะ (CD55 และ CD59)

TH (Your edit):
2. Paroxysmal nocturnal hemoglobinuria (PNH): PNH เป็นโรคที่พบได้น้อย ซึ่งมีลักษณะเด่นคือมีการทำลายของเซลล์เม็ดเลือดแดง ซึ่งนำไปสู่ hemolysis ใน PNH เซลล์เม็ดเลือดแดงไม่มีโปรตีนเฉพาะ (CD55 และ CD59)

Figure 4: Instruction text for human annotators

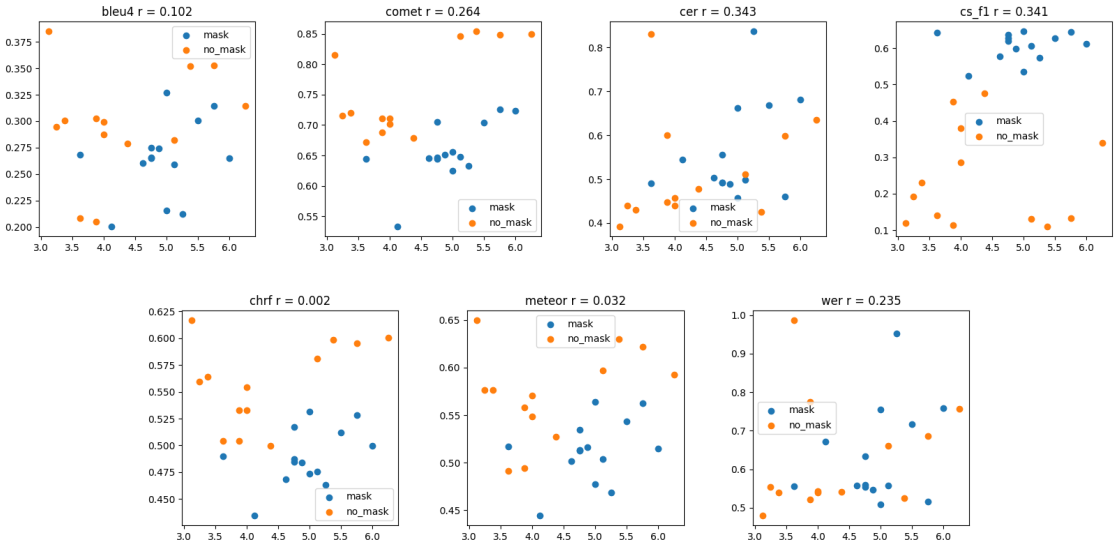


Figure 5: Plots of factual score of each model that pass 3 factual accuracy score against machine evaluation metric.Masked model are labeled in blue and models without masked are labeled in orange.

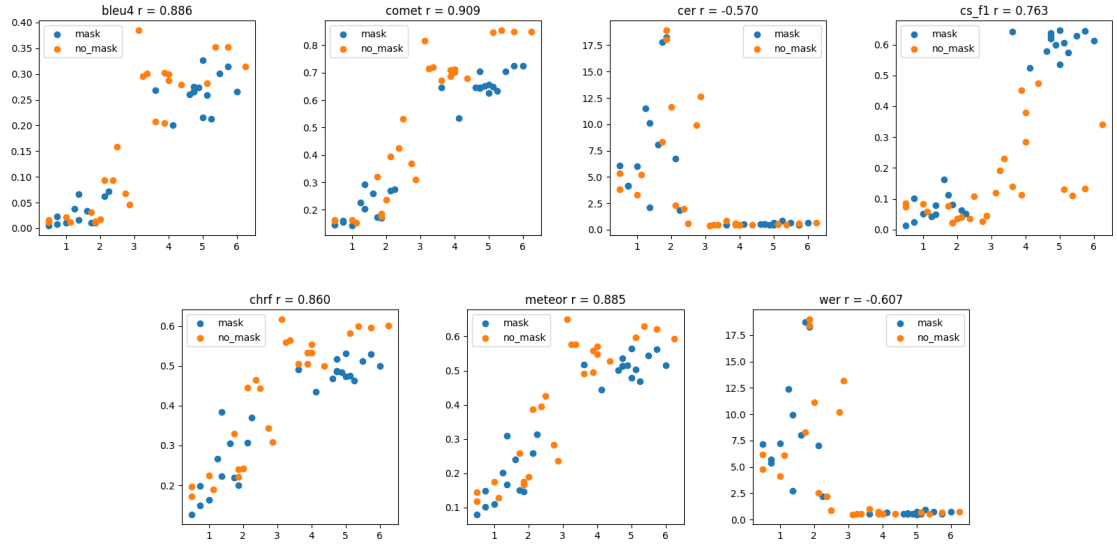


Figure 6: Plots of a factual score of all models against machine evaluation metric. The model below 3 factual accuracy score gives an extremely high value of CER and WER, so we decided to exclude them to not skew the Pearson r coefficient in the 5.

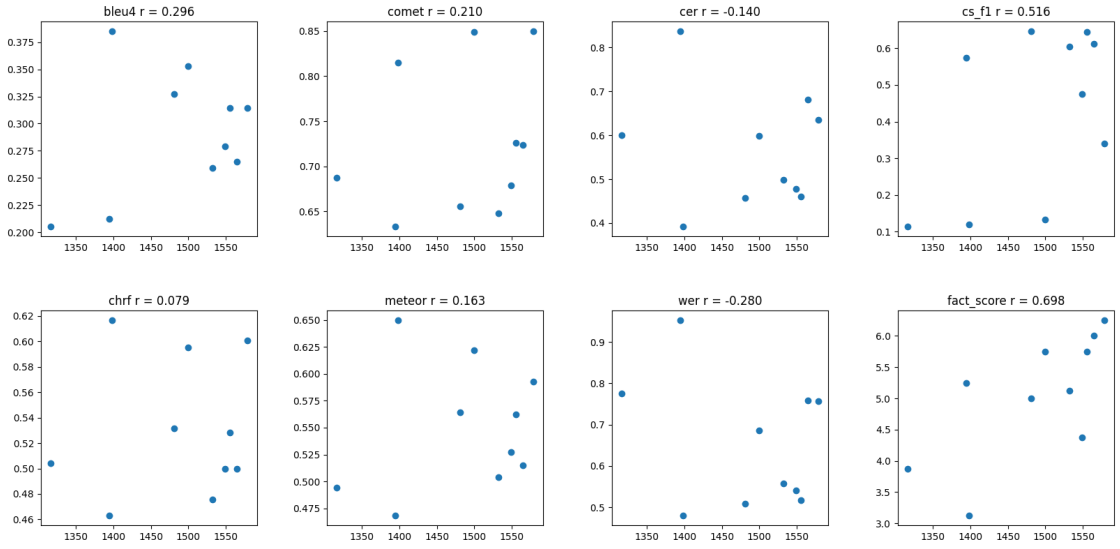


Figure 7: Plots of Glicko score of all human evaluated models against machine evaluation metric and Factual accuracy score.

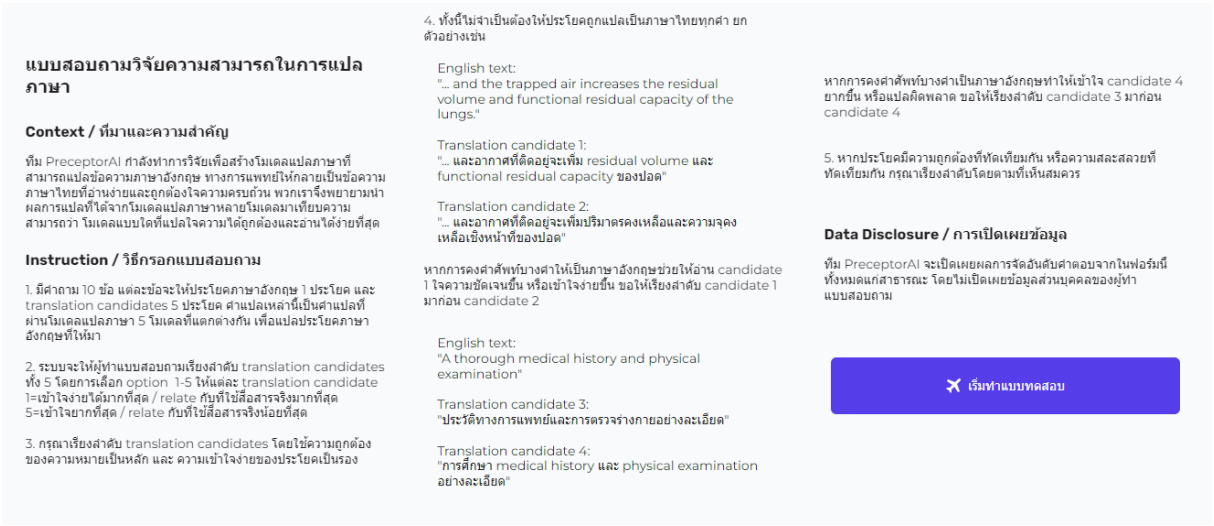


Figure 8: Instruction page for respondents to respond to our questionnaire.

Table 6: Factual Evaluation Result per MD

Model Variant	A*	B*	C*	D*	Fact. Score
Gemini-Pro-CS	4.5	6.5	5	7	5.75
Gemini-Pro-MN	6.5	6.5	3.5	5	5.375
Google-NMT	3	4	3.5	2	3.125
GPT-3.5-CS	3	5	3	3.5	3.625
GPT-3.5-MN	3	5	3.5	4	3.875
GPT-4-CS	5	7	6	7	6.25
GPT-4-MN	4.5	5.5	5	5.5	5.125
Llama2-13B-CS	1	1.5	0	2	1.125
Llama2-13B-MN	1	1	0	2	1
Llama2-7B-CS	1	0	0	1	0.5
Llama2-7B-MN	1	0	0	1	0.5
OpenThaiGPT-13B-CS	2	1	3	2.5	2.125
OpenThaiGPT-13B-MN	2.5	3	1.5	2.5	2.375
OpenThaiGPT-7B-CS	2	4	2	3.5	2.875
OpenThaiGPT-7B-MN	2.5	4.5	1	3	2.75
SeaLLM-7B-CS	2	3	1	2	2
SeaLLM-7B-MN	2	2.5	0	2.5	1.75
Typhoon-7B-CS	1.5	2	2	2	1.875
Typhoon-7B-MN	1.5	1	2	3	1.875
NLLB	1.5	3.5	1.5	3.5	2.5
NLLB-1	3.5	5	4.5	4.5	4.375
NLLB-2	3	4	2.5	4	3.375
NLLB-3	3.5	5.5	3.5	3.5	4
NLLB-4	2.5	5	3.5	4.5	3.875
NLLB-5	3	4.5	2.5	3	3.25
NLLB-6	3.5	5.5	3.5	3.5	4
Gemini-Pro-CS + Mask	5	6.5	5	5.5	5.5
Gemini-Pro-MN + Mask	5	7	4.5	6.5	5.75
Google-NMT + Mask	4	6	4.5	5.5	5
GPT-3.5-CS + Mask	5	7	4	5	5.25
GPT-3.5-MN + Mask	5.5	5.5	4.5	4.5	5
GPT-4-CS + Mask	5	6.5	6	6.5	6
GPT-4-MN + Mask	3.5	6.5	3.5	5.5	4.75
Llama2-13B-CS + Mask	3	3	1	2	1
Llama2-13B-MN + Mask	1	0	0	2	0.75
Llama2-7B-CS + Mask	1	0	0	1	0.5
Llama2-7B-MN + Mask	1	1	0	1	0.75
OpenThaiGPT-13B-CS + Mask	1	1	0	2	2.25
OpenThaiGPT-13B-MN + Mask	2	1.5	0	2	1.375
OpenThaiGPT-7B-CS + Mask	1.5	1	1	1.5	1.25
OpenThaiGPT-7B-MN + Mask	1	4	0.5	3	2.125
SeaLLM-7B-CS + Mask	1.5	2	0	2	1.375
SeaLLM-7B-MN + Mask	1.5	2.5	0	2.5	1.625
Typhoon-7B-CS + Mask	1.5	1	2	3	1.875
Typhoon-7B-MN + Mask	1	1	2	3	1.75
NLLB + Mask	3.5	6	2	5	4.125
NLLB-1 + Mask	5	6	4.5	3	4.625
NLLB-2 + Mask	3.5	6	4	5.5	4.75
NLLB-3 + Mask	4	6.5	4	6	5.125
NLLB-4 + Mask	3.5	6.5	4	5.5	4.875
NLLB-5 + Mask	2	5	3	4.5	3.625
NLLB-6 + Mask	4	6	3.5	5.5	4.75

Table 7: Model's rank on each score type

Rank of each score	CSF1	BLEU	chrF	CER	WER	COMET	METEOR	Fact.
Gemini-Pro-CS	24.5	2	4	20.5	19	3	3	3.5
Gemini-Pro-MN	29	3	3	2	4	1	2	6
Google-NMT	26	1	1	1	1	5	1	27
GPT-3.5-CS	23	25	15.5	27	28	17	20	23.5
GPT-3.5-MN	27	26	15.5	20.5	25	15	17	21.5
GPT-4-CS	18	5.5	2	23	23	2	5	1
GPT-4-MN	24.5	8	5	15.5	18	4	4	8.5
Llama2-13B-CS	39	46	48	37	37	50	48	45
Llama2-13B-MN	33	40	41	33	33	45	41	46.5
Llama2-7B-CS	37	46	49	38	38	48	49	51
Llama2-7B-MN	32	43	47	34	34	46	47	51
OpenThaiGPT-13B-CS	45	29.5	28	32	31	30	30	33.5
OpenThaiGPT-13B-MN	46	29.5	24	30	29	29	29	31
OpenThaiGPT-7B-CS	43	35	35	48	48	33	37	28
OpenThaiGPT-7B-MN	48	32	33	44	45	31	33	29
SeaLLM-7B-CS	47	41	39	47	46	38	39	35
SeaLLM-7B-MN	36	38	34	43	43	32	34	39.5
Typhoon-7B-CS	51	46	43	52	50	41	42	37
Typhoon-7B-MN	50	44	40	50	52	42	40	37
NLLB	30	28	29	22	26	28	28	30
NLLB-1	15	16	18	10	6	16	16	17
NLLB-2	20	13	6	3	14.5	8	8.5	25
NLLB-3	17	14	11.5	9	8	14	14	19.5
NLLB-4	16	10	11.5	7.5	5	11	13	21.5
NLLB-5	21	15	7	5.5	16	9	8.5	26
NLLB-6	19	12	8	4	11	10	11	19.5
Gemini-Pro-CS + Mask	5.5	7	14	25	20	13	10	5
Gemini-Pro-MN + Mask	2	5.5	10	7.5	3	6	7	3.5
Google-NMT + Mask	1	4	9	5.5	2	18	6	10.5
GPT-3.5-CS + Mask	12	24	26	28	27	25	26	7
GPT-3.5-MN + Mask	13	23	21	24	22	26	25	10.5
GPT-4-CS + Mask	8	11	17	26	24	7	15	2
GPT-4-MN + Mask	7	9	13	18.5	17	12	12	14
Llama2-13B-CS + Mask	40.5	48.5	50	39	41	52	50	46.5
Llama2-13B-MN + Mask	31	39	45.5	36	35	49	45	48.5
Llama2-7B-CS + Mask	52	52	52	40	40	51	52	51
Llama2-7B-MN + Mask	49	51	51	35	36	47	51	48.5
OpenThaiGPT-13B-CS + Mask	40.5	31	32	29	30	35	31	32
OpenThaiGPT-13B-MN + Mask	35	33	31	31	32	34	32	42.5
OpenThaiGPT-7B-CS + Mask	44	36	38	46	47	39	38	44
OpenThaiGPT-7B-MN + Mask	38	34	36	41	39	36	35	33.5
SeaLLM-7B-CS + Mask	42	42	42	45	44	40	43	42.5
SeaLLM-7B-MN + Mask	22	37	37	42	42	37	36	41
Typhoon-7B-CS + Mask	34	48.5	45.5	51	49	44	46	37
Typhoon-7B-MN + Mask	28	50	44	49	51	43	44	39.5
NLLB + Mask	14	27	30	18.5	21	27	27	18
NLLB-1 + Mask	11	21.5	27	17	9.5	22.5	24	16
NLLB-2 + Mask	4	20	20	14	14.5	24	21.5	14
NLLB-3 + Mask	9	21.5	25	15.5	12	20	23	8.5
NLLB-4 + Mask	10	17	23	11	7	19	18.5	12
NLLB-5 + Mask	3	18	19	12	13	22.5	18.5	23.5
NLLB-6 + Mask	5.5	19	22	13	9.5	21	21.5	14