Multiway Parallel Corpus in Forced Migration Domain for Multilingual Machine Translation

Fatemeh Azadi, Samuel Larkin, Chi-kiu Lo 羅致翹
Digital Technologies Research Centre, National Research Council Canada
{fatemeh.azadi,samuel.larkin,chikiu.lo}@nrc-cnrc.gc.ca

Abstract

High-quality domain-specific parallel corpora play a significant role in improving the performance of machine translation (MT) and multilingual natural language processing (NLP) systems in a target domain. However, most existing multilingual parallel corpora focus on general-purpose data, and a majority of highly specialized domains such as forced migration are suffering from a lack of multilingual data. In this work, we present a new high-quality 4-way parallel corpus in the forced migration domain. The corpus consists of human-translated journal articles from Forced Migration Review in English, French, Spanish, and Arabic. Our corpus contains data aligned at both document and sentence level in four languages and provides a clean and reliable 4-way parallel resource for multilingual research in forced migration. Using this dataset, we benchmark several open-weight large language models (LLMs), an open-weight multilingual MT system, online closed MT systems, and a closed LLM across 12 translation directions. We further leverage our corpus to improve the MT quality of a top-performing multilingual foundation model with two common domain adaptation approaches, fine-tuning and few-shot prompting. Our results demonstrate the effectiveness of our corpus in improving the translation performance of current models in the forced migration domain.

Keywords: Multiway Parallel Corpus, Domain Adaptation, Multilingual Machine Translation

1. Introduction

Recent advances in machine translation (MT) and multilingual natural language processing (NLP) research have been largely driven by the availability of large-scale parallel corpora. However, the majority of existing multilingual parallel corpora cover only general domain data, often derived from web-crawled sources (such as CCMatrix (Schwenk et al., 2021b), ParaCrawl (Bañón et al., 2020), and FLORES (Goyal et al., 2022)), or cover only a limited number of domains and focus on English-centric translation directions. Thus, for the majority of non-English language pairs and domains, few or no parallel corpora are available.

Multilingual and domain-specific corpora play a crucial role not only for the training and domain adaptation of translation models but also for their evaluation, as they are essential for capturing specialized terminology and writing styles. Previous efforts have been made on the extraction of multilingual parallel data for some specific domains, such as scientific publications (Nakazawa et al., 2016; Soares et al., 2018; Roussis et al., 2022), patents (Soares et al., 2020), biomedical (Névéol et al., 2018; Neves et al., 2016; Soares and Krallinger, 2019), or legal texts (Steinberger et al., 2006).

The forced migration domain, which pertains to displacement and refugee issues, is a specialized domain with specific terminology. Table 1 shows an example of parallel sentences in this domain. The specialized terms present challenges for MT systems and other multilingual NLP tasks in this domain. Developing effective translation systems for this domain is important for multilingual communication and information exchange between researchers, policymakers, humanitarian organizations, and affected communities. However, to the best of our knowledge, there is no publicly available multilingual parallel corpus specifically designed for this domain, which limits the ability to train, adapt or evaluate translation systems or other NLP tasks for accurate and reliable use in this context.

In this paper, we present a new high-quality 4-way parallel corpus in the forced migration domain, covering the English, French, Spanish, and Arabic languages. The corpus is built from articles published in an open-access journal, the Forced Migration Review (FMR), which includes articles professionally translated across the four languages. The data is carefully processed, aligned across documents, and then sentence-aligned in each document to ensure high-quality alignments. The resulting corpus includes 979 documents and about 46K sentences aligned across all four languages, which provides a clean, high-quality resource for multilingual research in this domain, particularly for developing and evaluating domain-specific multilingual translation models.

With the remarkable multilingual translation capabilities that large language models (LLMs) have recently shown (Zhang et al., 2023; Zhu et al., 2024b), evaluating and enhancing their performance in specific domains has become an important research

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 4889–4901
11–16 May 2026.

language	sentence
Arabic	وفي حين أن هذه الاتفاقية لم تدخل قط قيد التنفيذ، تتوفر مبادئ دولية أخرى مثل التوجيهات الصادرة عن منظمة الهجرة الدولية واتفاقية تسهيل حركة الملاحة البحرية الدولية لعام ١٩٦٥ وصيغتها المعدلة لعام ٢٠١٨. . . .
English	It has never come into force but international principles do exist – such as the guidelines issued by the International Maritime Organization and the 1965 Convention on Facilitation of International Maritime Traffic (FAL Convention) and its 2018 amendment. . . .
French	Bien que cette convention ne soit jamais entrée en vigueur, certains principes internationaux, comme les lignes directrices publiées par l'Organisation maritime internationale ou la Convention visant à faciliter le trafic maritime international (Convention FAL) de 1965, et son amendement de 2018, . . .
Spanish	Nunca ha entrado en vigor, pero existen principios internacionales, como los lineamientos emitidos por la Organización Marítima Internacional y el Convenio para facilitar el tráfico marítimo internacional de 1965 (Convenio FAL) y su enmienda de 2018. . . .

Table 1: Example of Arabic, English, French, and Spanish sentences in forced migration domain.

direction. We carefully curated a test set from our corpus to reduce the risk of data contamination in existing LLMs, to provide a more reliable and fair comparison. Then, we benchmark a variety of open-weight LLMs (including Llama-3.1, Gemma-2, Aya-101 and Aya Expanse), an open-weight multilingual MT system (NLLB-200), online closed MT systems (including Google Translate, DeepL, and Microsoft Bing) and a closed LLM (GPT-4o) on this test set, across all twelve translation directions between the four languages in our corpus.

To demonstrate the effectiveness and potential impact of our corpus for domain adaptation, we explore two common approaches to improve one of the best-performing open-weight models in our benchmark (Aya Expanse): fine-tuning and few-shot prompting. Our results show that using our multiway parallel, domain-specific corpus, both of these domain adaptation approaches improve the baseline model in almost all 12 translation directions.

In summary, our contributions are as follows:

- We present a new, high-quality 4-way parallel corpus in the forced migration domain, covering English, French, Spanish, and Arabic, which can be used as a valuable resource for any domain-specific multilingual NLP research, particularly machine translation.
- We conduct a benchmarking evaluation of the translation capability of several open-weight LLMs, open-weight multilingual MT system, online closed MT systems and closed LLM across all 12 translation directions in this specific domain.
- Using our corpus, we investigate fine-tuning and few-shot prompting for domain adaptation of a baseline model, and show that our corpus can be effectively used to improve the in-domain translation quality of open-weight LLMs, across nearly all language directions.

2. Related Work

Multiway parallel corpora have been shown to provide remarkable benefits in the field of machine translation (Freitag and Firat, 2020; Wu et al., 2024; Mu et al., 2024). Earlier efforts in creating multiway parallel corpora focused on the parliamentary (Koehn, 2005; Ziemski et al., 2016), wiki (Schwenk et al., 2021a), subtitle (Lison and Tiedemann, 2016; Qi et al., 2018; Reimers and Gurevych, 2020) or religious domains (Christodouloupoulos and Steedman, 2015; Agić and Vulić, 2019). There are works aiming to create multiway parallel corpora by automatically aligning multiple bilingual parallel corpora or web-crawled corpora from different publication sources (Dabre and Kurohashi, 2019; Ramesh et al., 2022; Xu et al., 2022). However, the quality of parallelism (i.e. translation and alignment quality) and content type selection bias in web crawled corpora are concerning (Thompson et al., 2024), which affects performance on NLP tasks, especially for low resource languages. Besides, these automatically created resources are mostly in the general domain.

At the same time, domain-specific parallel corpora have also been shown to play a significant role for the improvement of translation models, as they can help them adapt to the specialized terminology and style of those specific domains (Moslem et al., 2023; Wassie et al., 2025; Zhu et al., 2024a).

Several previous studies have addressed the extraction of multilingual parallel data for some particular domains, such as scientific publications (Nakazawa et al., 2016; Soares et al., 2018; Roussis et al., 2022), patents (Soares et al., 2020), med-

ical (Névéol et al., 2018; Neves et al., 2016; Soares and Krallinger, 2019; Anastasopoulos et al., 2020), legal (Steinberger et al., 2006), financial (Ghaddar and Langlais, 2020) or educational domains (Abdelali et al., 2014).

To the best of our knowledge, there is no publicly available multiway parallel corpus for the specialized domain of forced migration, which we aim to address in this paper.

3. Dataset

In this section, we present our 4-way parallel corpus in the forced migration domain. This corpus is created from articles published in four languages (English, Arabic, French, and Spanish) in the Forced Migration Review (FMR) journal.[1] FMR is an open-access journal that has been published since 1998 and covers articles related to forced migration, refugees, and displacement issues. Most issues of FMR are published in parallel in English, Arabic, French, and Spanish, making it a suitable resource for creating high-quality domain-specific 4-way parallel corpora that can be highly valuable for multilingual research. In the rest of this section, we first describe our corpus creation methodology and then provide some statistics and analysis of the resulting dataset.

3.1. Corpus Creation

In collaboration with the publisher of the Forced Migration Review (FMR) journal, we obtained access to a collection of multilingual issues published in this journal, which we used to create our four-way parallel corpus. This collection includes issues 34 to 75, where articles were provided as separate Word Document (`docx`) files.

We first perform manual document-level alignment, since article filenames were not always consistent across languages. As not all of these issues and articles were published in all languages, we only used the issues that were available in all four languages (i.e., English, Arabic, French, and Spanish) and excluded the articles missing in one or more languages from this data. After this filtering step, a total of 979 documents (articles) remained, which were then converted to plain text format for further processing.

Next, we apply sentence splitting on each document using the Portage Text Processing toolkit (Larkin et al., 2022) for English, French and Spanish, and CoreNLP[2] (Manning et al., 2014) for Arabic, to prepare the data for sentence-level alignment.

For sentence alignment, we use SentAlign[3] (Steingrimsson et al., 2023), a state-of-the-art sentence aligner based on the bilingual semantic similarity of sentence embeddings encoded by LaBSE (Feng et al., 2022). Since SentAlign is designed for aligning bitext, we align sentences within each document for the English-French, English-Spanish and English-Arabic language pairs separately. Using English as the pivot language, we merge these aligned bitext to form the aligned quadruplets across all four languages and obtain our 4-way parallel dataset. In cases where a sentence was missing or could not be reliably aligned in one language, the sentence and its aligned counterparts in other languages were discarded from the final 4-way aligned dataset.

As the journal is published online and openly accessible, there is a potential risk that some of its content may have been seen by existing LLMs. Thus, using this data for evaluating existing models on translation for the forced migration domain and across the four languages could lead to data contamination, and thus unfair comparisons and overestimated performance. We reserve the most recent issue published in May 2025 as our held-out data for validation and evaluation to ensure uncontaminated and reliable evaluation of models' translation capability in the following benchmarking and domain adaptation experiments. To further reduce the risk of contamination, we remove any sentences in the held-out dataset that have 13-gram overlaps (following Brown et al. (2020)) with sentences in the CCMatrix corpus.

Through this methodology, we produced a clean, high-quality 4-way parallel corpus in the domain of forced migration issues, aligned at both the document and sentence levels. This corpus supports the development and evaluation of domain-specific multilingual machine translation, as well as other multilingual research.

3.2. Data Statistics

Table 2 presents detailed statistics on the multiway parallel corpus that we created, including the number of documents and sentences in each set. As it shows, our corpus consists of 979 4-way aligned documents and 46,301 aligned sentences in each language. The training set contains 953 documents (about 44k sentence quadruplets). The held-out data described in the previous subsection (around 1.8k sentence quadruplets) is further divided into the development set and the testing set. The test set contains the first 1000 sentence quadruplets of the held-out set, whereas the development set contains the remaining ~800 sentence quadruplets.

[1] https://www.fmreview.org/

[2] https://stanfordnlp.github.io/CoreNLP

[3] https://github.com/steinst/SentAlign

	# documents	# sentences	# words			
			en	fr	es	ar
Train	953	44490	1.13M	1.41M	1.37M	1.10M
Dev	12*	811	21463	26183	26432	20939
Test	15*	1000	24316	30743	30358	25262
Total	979	46301	1.18M	1.47M	1.43M	1.14M

Table 2: Number of documents, sentences and words per language for each train, validation and test datasets. * indicates that part of a document is in **Dev** and the rest is in **Test** making the total one less document.

	Overlap Ratio
English	60.25%
French	54.44%
Spanish	52%
Arabic	40.13%

Table 3: The overlap ratio of sentences in our corpus with CCMatrix.

3.3. Data Analysis

We perform a manual evaluation to validate the accuracy of the sentence alignment. We sample 500 4-way aligned sentences from the collection, and alignments were examined separately in English-French, English-Spanish and English-Arabic language pairs. For each language pair, the alignment error rate was calculated as the percentage of incorrect alignments over the size of the evaluation set. Our evaluation shows that the error rate is less than 1% for all three language pairs, which indicates that our multiway parallel corpus is aligned with high accuracy.

To demonstrate the novelty of our corpus, we examined its coverage in the CCMatrix corpus, a large-scale multilingual dataset constructed from Common Crawl data. Assessing the level of this overlap provides useful insight into the distinctiveness and value of our dataset, beyond its advantages of being domain-specific and multiway parallel. Similar to the formation of the held-out dataset, we follow Brown et al. (2020) and considered a sentence in our dataset to be overlapping if it shares at least one 13-gram with a sentence in CCMatrix, or if it is found entirely in CCMatrix when it is shorter than 13-gram. We use the English-French, English-Spanish and English-Arabic sections of CCMatrix to compare against our corpus. Table 3 shows the overlap ratio for each language between our corpus and CCMatrix is around 40-60%. This shows that our corpus contains a significant amount of potentially unseen data to further train publicly available MT systems and pretrained LLMs.

In order to show the quality of our corpus, we compare the similarity scores for aligned sentence pairs in our corpus with those obtained from CCMatrix. For this purpose, we randomly select a set of 46301 sentence pairs from CCMatrix, yielding a corpus of the same size as our corpus, for English-Arabic, English-French, and English-Spanish language pairs. We then compute the cosine similarity between the aligned sentence pairs based on the sentence embeddings obtained from LaBSE.[4]

Figure 1 shows the distribution of LaBSE similarity scores for sentence pairs from our corpus and CCMatrix. For all three language pairs, our corpus shows a greater proportion of sentence pairs with higher similarity scores. More precisely, for English-Arabic, about 84% of the sentence pairs in our corpus have similarity scores of 0.8 or above, compared to around 74% in CCMatrix. Similarly, for English-French and English-Spanish, these ratios are about 92% vs. 80% and 96% vs. 86%, indicating that our corpus contains more similar sentence pairs, and thus translations and alignments in our corpus are cleaner and of higher quality.

4. Benchmarking Experiments

As the baselines for our experiments, we first benchmark various open-weight LLMs, an open-weight multilingual MT system, online closed MT systems and a closed LLM across all 12 translation directions between English, Arabic, French, and Spanish in the forced migration domain.

4.1. Setup

In our experiments, we examined a variety of open-weight LLMs, an open-weight multilingual MT system, online closed MT systems and a closed LLM. To avoid data contamination, all models and systems, except the closed MT systems, were selected such that all were released before the publication date of our held-out test dataset. Open-weight LLMs include Llama-3.1 8B,[5] Gemma-2 9B,[6] Aya-

[4] https://huggingface.co/sentence-transformers/LaBSE

[5] https://huggingface.co/meta-llama/Meta-Llama-3.1-8B-Instruct

[6] https://huggingface.co/google/gemma-2-9b-it

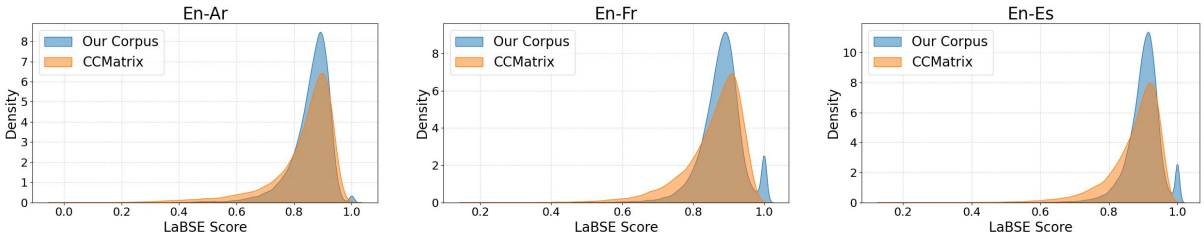

Figure 1: Distribution of LaBSE similarity scores for our corpus against CCMatrix

101,[7] and Aya Expanse 8B.[8] The open-weight multilingual MT system is NLLB-200 3.3B.[9] The closed LLM is GPT-4o.(OpenAI et al., 2024) The online closed MT systems include Google Translate,[10] DeepL,[11] and Microsoft Bing Translator.[12]

For the open-weight LLMs, the translation inference is performed using Hugging Face's pipeline with default parameters. We use the following prompt for both inference and fine-tuning the LLMs in our zero-shot prompting experiments, where `<src_lang>` and `<tgt_lang>` are the source and target languages and `<test_sentence>` is the test input to be translated:

```
Translate the following text from
<src_lang> into <tgt_lang>.
<test_sentence>
```

4.2. Evaluation

We use our test corpus, which contains 1000 4-way aligned sentences, to evaluate the models for translation across all 12 possible translation directions. We report translation quality based on two widely used automatic MT evaluation metrics, spBLEU (Goyal et al., 2022) and COMET (Rei et al., 2022). We compute spBLEU using the implementation in sacreBLEU[13] (Post, 2018) and for COMET, we use the `wmt22-comet-da` model.

4.3. Results

Table 4 and 5 present the results of the benchmarking evaluation on the selected LLMs and MT systems across all 12 translation directions in the forced migration domain, in terms of spBLEU and COMET, respectively.

Our results show that online MT systems generally outperform LLMs and open-weight multilingual MT systems in all translation directions. However, their performances on translating to or from Arabic are considerably weaker than translating between the other three languages.

One peculiar aspect of GPT-4o's performance is that it produced translations that consistently achieve better spBLEU scores but much lower COMET scores than those produced by the open-weight models.

Among open-weight models, Aya Expanse and NLLB appear as the two best performing models, consistently outperforming the other three models in all translation directions. While Aya Expanse achieves the best performance in most directions (except Arabic→English) in terms of COMET, NLLB achieves, on average, higher spBLEU scores in half of the translation directions. Since COMET is better at reflecting translation adequacy while spBLEU is reflecting subword matching with the reference translation, that means Aya Expanse might be producing more adequate translations than NLLB.

5. Fine-tuning Experiments

We now examine the effectiveness of our corpus for domain adaptation and improving the translation performance of existing models. We consider Aya Expanse, one of the overall best performing open-weight models, as the baseline, and we fine-tune it using our in-domain 4-way parallel corpus.

5.1. Setup

The advantage of multilingual LLMs or MT systems is the simplicity of encapsulating multiple translation directions in one model. Therefore, we decided to fine-tune Aya Expanse jointly in all 12 translation directions at the same time using our in-domain 4-way parallel corpus. As full fine-tuning is very compute-intensive and requires too much GPU memory for our available hardware, we leverage LoRA (Hu et al., 2022) for efficient fine-tuning. For this purpose, we use the LLaMA-Factory (Zheng

[7] https://huggingface.co/CohereLabs/aya-101

[8] https://huggingface.co/CohereLabs/aya-expanse-8b

[9] https://huggingface.co/facebook/nllb-200-3.3B

[10] https://translate.google.com/

[11] https://www.deepl.com/en/translator

[12] https://www.bing.com/translator

[13] nrefs:1|case:mixed|eff:no|tok:flores200|smooth:exp|version:2.5.1

Model	en→fr	en→es	en→ar	fr→en	fr→es	fr→ar	es→en	es→fr	es→ar	ar→en	ar→fr	ar→es	Avg.
Open-Weight LLMs													
Llama-3.1 8B	36.41	41.61	22.74	34.69	29.47	12.00	42.11	27.93	12.89	26.12	19.39	21.77	27.26
Gemma-2 9B	40.24	45.36	26.76	37.61	33.12	21.50	45.99	33.87	23.93	30.73	25.62	26.24	32.58
Aya-101	31.54	36.24	20.31	31.35	27.62	17.53	38.31	28.73	18.25	26.90	22.34	23.78	26.91
Aya Expanse 8B	**40.40**	45.59	**33.52**	38.24	**33.87**	**27.20**	46.44	**34.90**	**29.45**	33.02	27.85	29.05	34.96
Open-Weight Multilingual MT System													
NLLB-200 3.3B	40.33	**45.72**	33.24	**40.67**	33.48	26.10	**49.26**	33.57	27.85	**35.13**	**28.50**	**29.45**	**35.28**
Closed LLM													
GPT-4o	42.76	47.57	36.21	38.34	34.60	29.34	47.28	35.01	31.22	32.49	28.91	29.81	36.13
Online Closed MT Systems													
Google Translate	**47.81**	48.52	**38.50**	41.09	35.73	**31.02**	50.86	**41.42**	**33.12**	**39.74**	**36.51**	**34.65**	**39.91**
DeepL	43.33	**49.74**	38.08	**42.05**	**35.98**	30.47	**52.74**	39.04	32.71	39.09	32.29	32.05	38.96
Microsoft Bing	45.56	48.69	35.59	41.07	34.93	28.47	50.23	38.13	30.87	36.20	31.38	32.50	37.80

Table 4: Models' performance as measured with spBLEU↑ for all translation directions between English (en), Arabic (ar), French (fr), and Spanish (es). The best score for each direction is shown in **bold**.

Model	en→fr	en→es	en→ar	fr→en	fr→es	fr→ar	es→en	es→fr	es→ar	ar→en	ar→fr	ar→es	Avg.
Open-Weight LLMs													
Llama-3.1 8B	85.08	86.73	78.22	86.97	84.54	65.47	88.19	81.42	62.37	82.99	77.12	80.58	79.97
Gemma-2 9B	86.98	88.04	83.04	87.80	86.31	81.03	89.09	85.67	81.85	85.74	82.89	83.94	85.20
Aya-101	85.16	86.84	83.16	86.43	85.20	81.65	87.52	84.09	81.84	84.82	81.50	83.40	84.30
Aya Expanse 8B	**87.37**	**88.39**	**86.22**	**87.89**	**86.50**	**84.77**	**89.11**	**86.04**	**85.15**	86.39	**83.85**	**84.85**	**86.38**
Open-Weight Multilingual MT System													
NLLB-200 3.3B	87.01	88.06	85.32	87.84	86.15	83.98	88.79	84.92	84.28	**86.47**	83.42	84.27	85.88
Closed LLM													
GPT-4o	82.36	84.30	82.13	83.80	82.17	80.86	85.25	80.79	81.27	82.07	79.36	80.94	82.11
Online Closed MT Systems													
Google Translate	**88.31**	89.03	**87.68**	88.25	86.71	**85.64**	89.83	86.82	86.35	87.70	85.20	**85.99**	87.29
DeepL	87.97	**89.36**	87.38	**88.75**	**87.09**	85.51	**90.18**	**87.10**	**86.57**	**87.76**	**85.23**	85.95	**87.40**
Microsoft Bing	87.92	88.70	85.88	88.05	86.29	83.77	89.51	86.17	84.59	86.83	83.75	85.02	86.37

Table 5: Models' performance as measured with COMET↑ for all translation directions between English (en), Arabic (ar), French (fr), and Spanish (es). The best score for each direction is shown in **bold**.

et al., 2024) platform, with a learning rate of $2e-6$, a training batch size of 4, a warm-up ratio of 0.01, a weight-decay of 0.01, and a LoRA rank of 16. We perform fine-tuning for 5 epochs using 4 NVIDIA A100 GPUs. The translation inference and evaluation setup is the same as described in section 4.1 and 4.2.

5.2. Results

Table 6 shows translation quality of all 12 directions after fine-tuning, in terms of both spBLEU and COMET. Our results show that the fine-tuned models have statistically significant ($p \leq .05$)[14] improvements in 11 out of the 12 translation directions. Although the increase in COMET scores is relatively small indicating that fine-tuning had a limited impact on the translation adequacy, the improvements according to spBLEU are more substantial in most directions. This suggests that fine-tuning with our in-domain corpus improves the lexical overlap with references, which could mean an improvement in the choice of terminology for the targeted domain.

Table 7 shows an example of the translation output from the baseline model and the fine-tuning experiment for the French→English part of the test set. As we can see, for the French term "subi une agression", the baseline model has translated into "having experienced an attack" which is literal but does not fit the context and domain. After fine-tuning, the model output becomes "having been assaulted", which is a more accurate translation of the term in the forced migration domain.

Figure 2 shows the distribution of sentence-level COMET differences (ΔCOMET) between our fine-tuned model and our baseline model for the Arabic→English translation direction. Approximately 67.30% of the translations were of identical (ΔCOMET $= 0$) or better quality (ΔCOMET > 0). Note that some translations have an improvement greater than 0.1 whereas we do not observe any translations that degraded more than 0.1. Only 32.70% of translations had a slight degradation in COMET. All other translation directions exhibit similar distribution of COMET differences between

[14] Statistical significance tests performed using sacreBLEU and comet-compare.

Translation Direction	0-shot Prompting				5-shot Prompting			
	Baseline Model		Fine-tuned Model		Baseline Model		Fine-tuned Model	
ar→en	33.02	86.39	36.71†	**87.08**†	34.64†	86.75†	**37.29**†	**87.08**†
ar→es	29.05	84.85	30.77†	85.04†	29.83†	**85.12**†	**30.95**†	85.10†
ar→fr	27.85	83.85	**32.44**†	**84.21**†	29.67†	84.10†	31.30†	84.11†
en→ar	33.52	86.22	**34.05**†	**86.50**†	34.03†	86.39	33.95†	86.49
en→es	45.59	88.39	45.85	88.49	**46.14**†	**88.68**†	45.92	88.45
en→fr	40.40	87.37	43.77†	87.61†	**44.17**†	**87.62**†	44.06†	87.47
es→ar	29.45	85.15	29.73	**85.46**†	29.72	85.37	**29.75**	85.29
es→en	46.44	89.11	49.78†	**89.46**†	47.65†	89.27†	**49.80**†	89.36†
es→fr	34.90	86.04	**38.64**†	86.21†	38.31†	**86.35**†	38.50†	86.12
fr→ar	27.20	84.77	27.85†	**84.92**	27.72†	84.95	**27.86**†	84.79
fr→en	38.24	87.89	41.00†	88.13†	39.05†	88.06†	**41.77**†	**88.27**†
fr→es	33.87	86.50	34.67†	86.66†	34.37†	**86.77**†	**34.83**†	86.71†
Avg.	34.96	86.38	37.11	**86.65**	36.28	86.62	**37.17**	86.60

Table 6: Results of using our 4-way parallel corpus for fine-tuning the baseline model (Aya Expanse 8B) and 5-shot prompting. Scores are reported as spBLEU/COMET. The best score for each direction is shown in **bold**. Scores marked with a † are statistically significant ($p \leq .05$) different than the 0-shot prompting baseline.

Source	En 2024, 42 % des patients de MSF ont déclaré avoir **subi une agression**, contre 17 % l'année précédente.
Reference	In 2024, 42% of MSF patients reported **experiencing assault**, compared with 17% the previous year.
Baseline Model Output	In 2024, 42% of MSF patients reported **having experienced an attack**, up from 17% the previous year.
Fine-tuned Model Output	In 2024, 42% of MSF patients reported **having been assaulted**, up from 17% the previous year.
5-shot Prompting Output (Baseline Model)	In 2024, 42% of MSF patients reported **having experienced an assault**, up from 17% the previous year.

Table 7: A French→English translation example, showing the baseline, fine-tuned model and 5-shot prompting outputs.

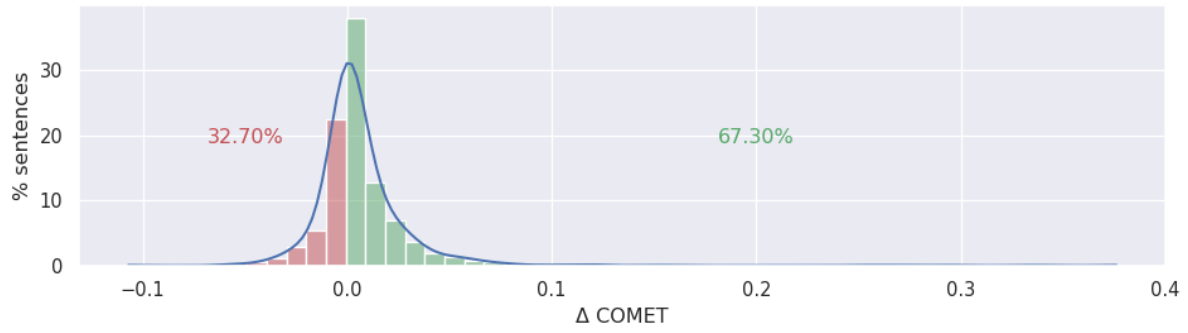

Figure 2: Distribution of sentence-level COMET differences (ΔCOMET) between our fine-tuned model and our baseline model for the Arabic→English translation direction. The blue line represents the univariate kernel density estimate (KDE) over the COMET differences. Positive differences indicate an improvement in translation quality in terms of COMET scores.

the baseline model and the fine-tuned model with roughly the same ratio of sentences with improved translation quality.

6. Few-shot Prompting Experiments

Since supervised fine-tuning is more compute-intensive, we investigate the potential of using our corpus for few-shot prompting, which is a less tech-

nically demanding approach to improve the translation quality of the model.

6.1. Setup

We retrieve the k most similar training examples for each test sentence and add them to the prompt with their target language translations to provide in-domain context to the model. We experiment with different values of $k \in \{0, 1, 3, 5, 7\}$. We use the Sentence-Transformers (Reimers and Gurevych, 2019) model `all-MiniLM-L6-v2`[15] to compute sentence embeddings and their similarities. Using the k most similar training examples to the test input, we construct our few-shot prompt as follows:

```
Translate the following texts from
<src_lang> into <tgt_lang>.
<src_lang>:  "<source_example_1>"
<tgt_lang>:  "<target_example_1>"
     ⋮              ⋮
<src_lang>:  "<source_example_k>"
<tgt_lang>:  "<target_example_k>"
<src_lang>:  "<test_sentence>"
<tgt_lang>:
```

The evaluation setup is the same as described in sections 4.2.

6.2. Results

Figure 3 presents the spBLEU results for different values of $k \in \{0, 1, 3, 5, 7\}$. As shown, even adding a single example to the prompt ($k = 1$) can remarkably improve the quality of translation in most translation directions. The best performance is achieved with $k = 5$ for most language pairs, while using more examples ($k = 7$) seems to give little to no additional improvement, or even degradation in some directions.

Table 6 shows the results of 5-shot versus 0-shot prompting for both the baseline and fine-tuned models. Although some score differences with respect to their 0-shot prompting baseline might seem small, those marked with † are nonetheless statistically significant ($p \leq .05$).[16] For the baseline model, using 5-shot prompting significantly improves the performance scores in 9 out of 12 translation directions, showing that even without fine-tuning, giving a few in-domain examples similar to the input can help the model produce better domain-specific translations. However, for most translation directions, the improvements are smaller than those achieved with fine-tuning. This shows that few-shot prompting is less desirable for domain-specific MT systems with long-term frequent usage.

Another interesting result is that combining 5-shot prompting with the fine-tuned model shows very small or no further gains in translation performance. This shows that the domain knowledge learned by these two adaptation approaches are largely overlapping and thus using the same in-domain training data, there is no need to apply these two adaptation approaches on top of each other.

Table 7 shows an example of the translation output from the baseline model with 0-shot prompting and 5-shot prompting for the French→English part of the test set. Similar to the previous section, the baseline model with 5-shot prompting is also able to adapt to the forced migration domain and more accurately in translating the French term “subi une agression” to “having experienced an assault” instead of “having experienced an attack”.

7. Conclusion

In this paper, we introduced a new high-quality 4-way parallel corpus in the domain of forced migration. Using this corpus, we benchmarked several open-weight LLMs, an open-weight multilingual MT system, online closed MT systems and a closed LLM in this specific domain, across all 12 translation directions between the four languages in our dataset (i.e., English, French, Spanish and Arabic).

We also explored two approaches for leveraging our corpus to improve translation performances of current open-weight models: fine-tuning the model and few-shot prompting. The results show that both of these approaches can improve the translation performance of the baseline model for this specific domain, demonstrating the effectiveness of our corpus for domain adaptation.

This domain-specific, multiway parallel corpus does not only support further research on multilingual machine translation and domain adaptation in future, but it also supports other topics of multilingual natural language processing research such as extracting specialized terminology, and cross-lingual information retrieval. In the future, we aim to further expand our corpus, incorporating new publications from the FMR journal, or other multilingual resources in this domain.

8. Limitations

Forced Migration Review is published in a mixture of different language variants in the four languages depending on the translation services employed. The Arabic is standard Arabic, possibly mixed with a variety of different Arabic variations because the

[15] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

[16] Statistical significance tests performed using sacreBLEU and comet-compare.

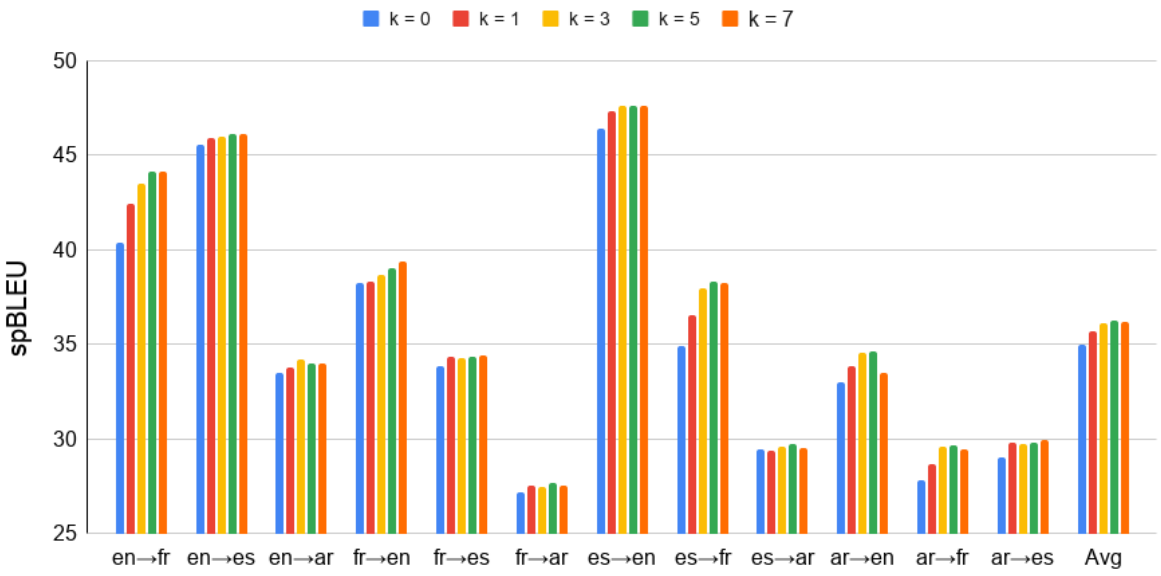

Figure 3: spBLEU results of few-shot prompting for different number of given examples (k).

translation was done by a diverse group of translators from different countries. The English is British English. The French is standard French. The Spanish up until Issue 70 is predominantly European Spanish and it is predominantly Latin American Spanish after Issue 70. We acknowledge that our corpus does not cover all the variations of the four languages and may result in potential biases in the language variations of the text generated by systems developed using our corpus. We recommend developers thoroughly evaluate and mitigate these potential biases before deploying systems trained on our corpus.

9. Acknowledgments

We are grateful for the publisher of the Forced Migration Review providing access to the collection of the journal issues.

10. Bibliographical References

Ahmed Abdelali, Francisco Guzman, Hassan Sajjad, and Stephan Vogel. 2014. The AMARA corpus: Building parallel language resources for the educational domain. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1856–1862, Reykjavik, Iceland. European Language Resources Association (ELRA).

Željko Agić and Ivan Vulić. 2019. JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Florence, Italy. Association for Computational Linguistics.

Antonios Anastasopoulos, Alessandro Cattelan, Zi-Yi Dou, Marcello Federico, Christian Federmann, Dmitriy Genzel, Franscisco Guzmán, Junjie Hu, Macduff Hughes, Philipp Koehn, Rosie Lazar, Will Lewis, Graham Neubig, Mengmeng Niu, Alp Öktem, Eric Paquin, Grace Tang, and Sylwia Tur. 2020. TICO-19: the translation initiative for COvid-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jef-

frey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Christos Christodouloupoulos and Mark Steedman. 2015. A massively parallel corpus: the bible in 100 languages. *Language resources and evaluation*, 49(2):375–395.

Raj Dabre and Sadao Kurohashi. 2019. Mmcr4nlp: Multilingual multiway corpora repository for natural language processing.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Markus Freitag and Orhan Firat. 2020. Complete multilingual neural machine translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 550–560, Online. Association for Computational Linguistics.

Abbas Ghaddar and Phillippe Langlais. 2020. SEDAR: a large scale French-English financial domain parallel corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3595–3602, Marseille, France. European Language Resources Association.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Samuel Larkin, Eric Joanis, Darlene Stewart, Michel Simard, George Foster, Nicola Ueffing, and Aaron Tikuisis. 2022. Portage Text Processing.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland. Association for Computational Linguistics.

Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023. Adaptive machine translation with large language models. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 227–237, Tampere, Finland. European Association for Machine Translation.

Yongyu Mu, Peinan Feng, Zhiquan Cao, Yuzhang Wu, Bei Li, Chenglong Wang, Tong Xiao, Kai Song, Tongran Liu, Chunliang Zhang, and JingBo Zhu. 2024. Revealing the parallel multilingual learning within large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6976–6997, Miami, Florida, USA. Association for Computational Linguistics.

Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. 2016. ASPEC: Asian scientific paper excerpt corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2204–2208, Portorož, Slovenia. European Language Resources Association (ELRA).

Aurélie Névéol, Antonio Jimeno Yepes, Mariana Neves, and Karin Verspoor. 2018. Parallel corpora for the biomedical domain. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Mariana Neves, Antonio Jimeno Yepes, and Aurélie Névéol. 2016. The scielo corpus: a parallel corpus of scientific publications for biomedicine. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*

(LREC'16), pages 2942–2948, Portorož, Slovenia. European Language Resources Association (ELRA).

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan

Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. 2024. Gpt-4o system card.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. 2018. When and why are pre-trained word embeddings useful for neural machine translation? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 529–535, New Orleans, Louisiana. Association for Computational Linguistics.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2022. Samanantar: The largest publicly available parallel corpora collection for 11 Indic languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Dimitrios Roussis, Vassilis Papavassiliou, Prokopis Prokopidis, Stelios Piperidis, and Vassilis Katsouros. 2022. SciPar: A collection of parallel corpora from scientific abstracts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2652–2657, Marseille, France. European Language Resources Association.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021a. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021b. CCMatrix: Mining billions of high-quality parallel sentences on the web. 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 6490–6500, Online. Association for Computational Linguistics.

Felipe Soares and Martin Krallinger. 2019. Bvs corpus: A multilingual parallel corpus of biomedical scientific texts.

Felipe Soares, Viviane Moreira, and Karin Becker. 2018. A large parallel corpus of full-text scientific articles. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Felipe Soares, Mark Stevenson, Diego Bartolome, and Anna Zaretskaya. 2020. ParaPat: The multi-million sentences parallel corpus of patents abstracts. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3769–3774, Marseille, France. European Language Resources Association.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş, and Dániel Varga. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Steinthor Steingrimsson, Hrafn Loftsson, and Andy Way. 2023. SentAlign: Accurate and scalable sentence alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 256–263, Singapore. Association for Computational Linguistics.

Brian Thompson, Mehak Dhaliwal, Peter Frisch, Tobias Domhan, and Marcello Federico. 2024. A shocking amount of the web is machine translated: Insights from multi-way parallelism. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1763–1775, Bangkok, Thailand. Association for Computational Linguistics.

Aman Kassahun Wassie, Mahdi Molaei, and Yasmin Moslem. 2025. Domain-specific translation with open-source large language models: Resource-oriented analysis.

Di Wu, Shaomu Tan, Yan Meng, David Stap, and Christof Monz. 2024. How far can 100 samples go? unlocking zero-shot translation with tiny multi-parallel data. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15092–15108, Bangkok, Thailand. Association for Computational Linguistics.

Yulin Xu, Zhen Yang, Fandong Meng, and Jie Zhou. 2022. EAG: Extract and generate multi-way aligned corpus for complete multi-lingual neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8141–8153, Dublin, Ireland. Association for Computational Linguistics.

Xuan Zhang, Navid Rajabi, Kevin Duh, and Philipp Koehn. 2023. Machine translation with large language models: Prompting, few-shot learning, and fine-tuning with QLoRA. In *Proceedings of the Eighth Conference on Machine Translation*, pages 468–481, Singapore. Association for Computational Linguistics.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. 2024. LlamaFactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 400–410, Bangkok, Thailand. Association for Computational Linguistics.

Shaolin Zhu, Menglong Cui, and Deyi Xiong. 2024a. Towards robust in-context learning for machine translation with large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16619–16629, Torino, Italia. ELRA and ICCL.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024b. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534, Portorož, Slovenia. European Language Resources Association (ELRA).