Machine Translation with Large Language Models: Prompting, Few-shot Learning, and Fine-tuning with QLoRA

Xuan Zhang
Johns Hopkins University
xuanzhang@jhu.edu

Navid Rajabi
George Mason University
nrajabi@gmu.edu

Kevin Duh
Johns Hopkins University
kevinduh@cs.jhu.edu

Philipp Koehn
Johns Hopkins University
phi@jhu.edu

Abstract

While large language models have made remarkable advancements in natural language generation, their potential in machine translation, especially when fine-tuned, remains underexplored. In our study, we conduct comprehensive experiments, evaluating 15 publicly available language models on machine translation tasks. We compare the performance across three methodologies: zero-shot prompting, few-shot learning, and fine-tuning. Central to our approach is the use of QLoRA, an efficient fine-tuning method. On French-English, QLoRA fine-tuning outperforms both few-shot learning and models trained from scratch. This superiority is highlighted in both sentence-level and document-level translations, with a significant BLEU score improvement of 28.93 over the prompting method. Impressively, with QLoRA, the enhanced performance is achieved by fine-tuning a mere 0.77% of the model's parameters.

1 Introduction

The rapid advancement of large language models (LLMs) is reshaping the field of natural language processing (NLP), marking a potential paradigm shift in future development (Zhao et al., 2023). Instead of crafting dedicated task-specific systems, a growing interest has been focusing on quickly adapting LLMs to specific tasks simply through prompting (Liu et al., 2023; Sanh et al., 2022). So far, studies have shown that prompting LLMs can match or even rival the performance of specialized systems on numerous NLP tasks (Radford et al.).

Among all the NLP tasks, the application of LLMs to machine translation (MT) is understudied. The optimal way to harness LLMs for MT remains an open question. While encoder-decoder-based LLMs (Xue et al., 2021; Liu et al., 2020; Costa-jussà et al., 2022) are inherently designed for the sequence-to-sequence demands of MT, the approach for leveraging decoder-only models is less straightforward.

Although there are initial attempts in this direction (Sia and Duh, 2022; Hendy et al., 2023; Moslem et al., 2023; Zhu et al., 2023), these studies mainly concentrate on prompting and few-shot learning, not exploiting the availability of bitext. Additionally, most work focus on exceptionally large LLMs like GPT3 (Brown et al., 2020) with its staggering 175 billion parameters, which are beyond the reach of non-commercial research groups for local training. This poses a significant hurdle for institutions with constrained computational resources, rendering the findings less applicable and relevant to many researchers.

In this paper, we aim to investigate the performance of LLMs on MT tasks, with a particular focus on decoder-based LLMs, a category less charted for MT applications. Our research focuses on a range of publicly available medium-sized LLMs. This includes models pretrained on English-centric datasets, such as GPT-Neo (Black et al., 2021), OPT (Zhang et al., 2022), LLaMA2 (Touvron et al., 2023), as well as those on multilingual datasets such as XGLM (Lin et al., 2021) and BLOOMZ (Muennighoff et al., 2022). We evaluate various versions of these models, with their parameter sizes spanning from 1.3 billion to 13 billion, totaling 15 models.

In our experiments, we explore zero-shot prompting, few-shot learning, and fine-tuning, where our emphasis on fine-tuning fills the gap in previous studies. For the fine-tuning process, we employ the QLoRA method (Dettmers et al., 2023), which enhances efficiency and minimizes memory usage by quantizing the model to 4-bit precision and limiting the number of trainable parameters. To the best of our knowledge, this is the first instance of QLoRA being applied to fine-tuning LLMs for MT tasks.

We also evaluate the performance of LLMs in document-level translation. Standard sequence-to-sequence MT models focus on translating one sentence at a time, overlooking discourse phenom-

ena and the broader context. Existing methods for document-level translation often pivot toward architectural modifications (Tu et al., 2018; Tan et al., 2019; Xu et al., 2021), leading to specialized models that need unique designs. Our objective is to evaluate the capability of LLMs in preserving long-term contextual coherence and to explore their potential in facilitating the development of a robust document-level translation system.

We demonstrate the effectiveness of fine-tuning on a French-English dataset – this language pair is selected due to its accessibility for LLMs, positioning it as an ideal starting point for research in this domain. Our experimental results, complemented by thorough analysis, reveal that:

- LLMs, when subjected to fine-tuning, are potent MT models. Through fine-tuning, they consistently outperform their zero-shot prompting counterparts, achieving an average improvement of 8 BLEU for sentence-level translation and 16.33 BLEU for document-level translation. Notably, the model *opt-13b* even sees a remarkable boost of 28.93 BLEU (from 4.56 to 33.49).

- There is a large variation in the performance across different LLMs. LLaMA 2 consistently outperforms others for both prompting and fine-tuning. BLOOMZ, initially lagging behind in prompting, ascends to top-tier models after fine-tuning. However, some models, despite benefiting from fine-tuning, either match or fall short of the performance of models trained from scratch. It is also noteworthy that larger models don't invariably outshine their smaller counterparts.

- When prompted, LLMs demonstrate superior performance in sentence-level translation. However, the application of fine-tuning yields more substantial enhancements in document translation, as reflected by both the BLEU and COMET scores. Notably, LLaMA 2 surpasses its performance in sentence-level translation when trained on documents.

- QLoRA accelerates the fine-tuning process without compromising model performance. To attain an equivalent BLEU score, it necessitates 21 times less training time and reduces the trainable parameters by 1370-fold compared to conventional fine-tuning.

2 Related Work

2.1 LLM Applications

Leveraging LLMs across a spectrum of downstream natural language processing (NLP) tasks is now a prevailing approach. However, the optimal strategies for utilizing these models both effectively and efficiently remain an open question. Broadly speaking, there are three primary methods to build applications based on LLMs:

- **Zero-shot prompting.**[1] This involves querying LLMs with a prompt that hasn't been seen in the training data of the model. Such prompts typically provide specific task instructions along with the main query. Given the sensitivity of LLMs to the structure and content of prompts, careful prompt engineering is crucial to achieve optimal performance.

- **Few-shot learning.** Often referred to as in-context learning, few-shot learning is a technique where LLMs are provided with a handful of examples to guide their responses. Zero-shot prompting can be considered a subset of this, where no examples are given. In few-shot learning, these examples are integrated into the prompt template, serving as context to instruct the model on how to respond.

- **Fine-tuning.** The two methods above allow for task adaptation without the need for further training on the LLMs. In contrast, fine-tuning involves extending the training of the LLMs using additional, task-specific data. This is particularly beneficial when such tailored datasets are available.

Yang et al. (2023) survey the 'use cases' and 'no use cases' of LLMs for specific downstream tasks, considering the three aforementioned methods, and conclude that LLMs excel in most NLP tasks.

2.2 LLMs for MT

Recent literature has begun to explore the application of LLMs for MT, an area that remained relatively under-explored until now. Both Hendy et al. (2023) and Moslem et al. (2023) underscore the superiority of GPT3 (Brown et al., 2020), GPT3.5 and ChatGPT (Bawden and Yvon, 2023) in MT

[1]Throughout this paper, we refer to 'zero-shot prompting' simply as 'prompting'.

using prompting. However, the former also indicates that these models may not consistently outperform SOTA MT systems and commercial translators. In a comparative study, Zhu et al. (2023) experiment with various LLMs, including GLM-7.5B (Lin et al., 2021), OPT-175B (Zhang et al., 2022), BLOOMZ-7.1B (Muennighoff et al., 2022), and ChatGPT. Their findings suggest that while these decoder-only LLMs are competitive, they still lag behind when compared to the encoder-decoder-based multilingual language model NLLB (Costa-jussà et al., 2022). Briakou et al. (2023) studied the impact of LLM data on MT.

Prompting strategies for MT are studied by Vilar et al. (2023) for PaLM (Chowdhery et al., 2022) and Zhang et al. (2023) for GLM-130B (Zeng et al., 2022). They reveal several challenges associated with MT prompting, such as issues with copying, mistranslation of entities, and hallucination. These challenges are echoed by Bawden and Yvon (2023), which identify similar constraints with prompting on BLOOM (Scao et al., 2022). However, they show these limitations can be mitigated in a few-shot learning setting. Sia and Duh (2022) investigated a light-weight tuning method akin to prefix tuning (Li and Liang, 2021). Sia and Duh (2023) and Wang et al. (2023) expand the evaluation to document-level translation.

While prior studies have highlighted the potential of LLMs in MT, their focus has been primarily on in-context learning. A significant gap remains in the exploration of fine-tuning LLMs specifically for MT tasks. Additionally, there is an evident absence of research that provides a comprehensive comparison among prompting, few-shot learning, and fine-tuning methodologies. Recognizing this oversight, the primary objective of this paper is to address and bridge this research gap.

3 QLoRA

QLoRA (Dettmers et al., 2023) is an efficient fine-tuning approach that reduces the memory usage of training without compromising the 16-bit task performance. The approach involves quantizing a pre-trained model to 4-bit precision. Subsequently, a compact set of learnable Low-rank Adapter (LoRA, Hu et al. (2021)) weights are added, which can be tuned through backpropagation.

LoRA Motivated by the empirical findings of Li et al. (2018) and Aghajanyan et al. (2020), which suggest that LLMs possess a notably low intrinsic dimension for their parameters, LoRA hypothesizes a similar low intrinsic rank for weights during model adaptation. Thus, LoRA introduces a reparameterization aimed at reducing dimensions. Specifically, it employs a low-rank decomposition to represent the pretrained weights, resulting in newly-added adapter weight matrices, with the rank r anticipated to be considerably smaller than the original weight matrices' dimension. During fine-tuning, the pretrained weights are frozen, with only the newly incorporated adapter updated via back-propagation. A key observation is that as the rank r is reduced, there is a corresponding decrease in the number of adaptable parameters.

4 Experimental Setup

4.1 Datasets

In this study, we focus on the translation direction from French to English due to its significant demand for high-quality translation and the availability of substantial parallel data. Our fine-tuning set includes the commonly used Europarl (Koehn, 2005) and News Commentary dataset from WMT14[2]. The dev and test sets are the newstest2013 and newstest2014 datasets, respectively, from WMT14. These datasets are constructed from documents, thus enabling a natural evaluation of document-level translation. Table 1 summarizes the statistics of the datasets.

	#sents	#docs	avg.sents/doc
train	2,366,117	21,430	144
dev	3000	126	24
test	3003	169	18

Table 1: Dataset statistics.

4.2 Baseline

We compare the performance of systems built upon LLMs against an NMT model trained from scratch using the Amazon Sockeye framework (Hieber et al., 2022). The model architecture is a 12-layer transformer with a model size of 1024, 16 attention heads, and 4096 hidden units in the feed-forward layers. We employ byte pair encoding (BPE, Sennrich et al. (2016)) separately for each language, setting the number of BPE symbols to 30k for both languages. The model is trained with a batch size of 4096, an initial learning rate of 0.0002, and a

[2]https://www.statmt.org/wmt14/translation-task.html

Model	Release Time	Data	Size (B)
GPT-Neo (Black et al., 2021)	Mar, 2021	English-centric	1.3; 2.7
OPT (Zhang et al., 2022)	June, 2022	English-centric	1.3; 2.7; 6.7
LLaMA2 (Touvron et al., 2023)	July, 2023	English-centric	7; 13
XGLM (Lin et al., 2021)	Nov, 2022	Multilingual	1.7; 2.9; 4.5; 7.5
BLOOMZ (Muennighoff et al., 2022)	Nov, 2022	Multilingual	1.7; 3; 7.1

Table 2: Overview of evaluated LLMs.

plateau-reduce learning rate scheduler. Additionally, we apply a dropout and label smoothing of 0.1, use the Adam optimizer with a warm-up of 10k steps, and set the checkpoint interval to 4000. Training is halted if there is no improvement in performance on the dev set for 32 consecutive checkpoints. The model has 4 billion parameters and is trained on a single NVIDIA V100 with 32G GPU memory.

This is a relatively standard NMT model, devoid of advanced techniques such as back translation, knowledge distillation, or ensembling, which could potentially elevate the model to state-of-the-art performance (Kocmi et al., 2022). However, the primary objective of this study is to compare the efficacy of using an off-the-shelf machine translation toolkit, which is widely accessible and requires minimal effort for machine translation practitioners, against building MT systems using LLMs. Importantly, both methods demand similar levels of effort in development, making this a fair comparison to ascertain the most efficient approach for practitioners and researchers alike.

4.3 Pretrained LLMs

We investigate a varied collection of pretrained LLMs accessible on HuggingFace (Wolf et al., 2020), all based on the transformer architecture. This collection comprises five distinct LLMs, each trained on either English-centric or multilingual data and available in multiple versions with varying parameter sizes. This results in a comprehensive assortment of 15 models, with parameter sizes ranging from 1.3 billion to 13 billion. Table 2 summarizes the models included in our study.

- **GPT-Neo** - a GPT-2 (Radford et al.) like causal language model trained on the Pile dataset (Gao et al., 2020), an 825 GiB English corpus.
- **OPT** - a suite of causal language models, where the largest one, OPT-175B, exhibits performance comparable to GPT-3 (Brown et al., 2020).
- **LLAMA 2** - pretrained on 2 trillion tokens of English-centric data. We used a fine-tuned version of the model, referred to as *LLAMA 2-CHAT*. This fine-tuned version demonstrates superior performance compared to open-source chat models across a wide range of benchmarks.
- **XGLM** - a multilingual language model trained on a balanced corpus covering 30 diverse languages with 500B tokens. The XGLM 7.5B outperforms GPT-3 on the FLORES-101 (Goyal et al., 2022) machine translation benchmark in few-shot learning scenarios.
- **BLOOMZ** - a multilingual BLOOM model (Scao et al., 2022) fine-tuned with the xP3 dataset (Muennighoff et al., 2022), which consists of multilingual datasets with English prompts, totaling 95 GiB of text.

The selection of these models enables us to assess the impact of various factors on translation performance, including the type of model (English-centric vs. multilingual) and model size. Additionally, the chosen sizes reflect the computational resources typically available to research institutes with limited GPU resources, such as university labs. This consideration ensures that our findings are applicable and accessible to a broad range of machine translation researchers and practitioners.

4.4 Prompted Tuning

We fine-tune LLMs using examples that include specifically formatted prompts (French: [fr sent] English:) and their corresponding responses ([en sent]). The dev set is also formatted in the same way. This approach customizes the model for the French-English machine translation task.

Sentence-level Prompts The inputs at the sentence level are formatted as follows:

French: [fr sent] English: [en sent] <eos>

We append the special token <eos> at the end of each sample to regulate the length of the text generated by the model. Without this, LLMs tend

to generate text continuously until they reach a predetermined length limit.

Document-level Prompts We use the given document boundaries to concatenate parallel sentences into document-level sequences. These parallel documents comprise an equal number of sentences in both languages. Our goal is to ensure that the models generate the same number of output sentences per document as the number of input sentences provided, facilitating sentence-level evaluation. We adopt the document mark-up used in Junczys-Dowmunt (2019), incorporating symbols for document start (<BEG>) and end (<END>), as well as sentence separators (<SEP>). In instances where documents exceed our sentence limit of 10, we substitute the <END> symbol with a break symbol (<BRK>) and commence the subsequent sequence with a continuation symbol (<CNT>) instead of <BEG> . Below is an example of a document input:

French: <BEG> [fr sent1] <SEP> [fr sent2] <SEP><END> English: <BEG> [en sent1] <SEP> [en sent2] <SEP><END>

4.5 Fine-tuning Setup

We configure the learning rate to 2e-4 and employ the Adam optimizer for the training process. A batch size of 32 is used, and the evaluation is performed every 1000 steps. The fine-tuning process is halted if there is no improvement in the model's performance over 16 consecutive checkpoints. For the LoRA configurations, the rank for the low-rank approximation is set to 64, and the scaling factor for the low-rank adaptation is set to 32. The trainable parameters are limited to the self-attention layers of the model. Additionally, a dropout rate of 0.05 is applied in the LoRA layer. The model weights are quantized to 4-bit precision to reduce memory requirements, and mixed-precision training is enabled, using a combination of float16 and float32 data types to accelerate the training process. Models with less than 3 billion parameters are trained on a single NVIDIA RTX GPU with 24GB of memory, while models with more than 3 billion but less than 7 billion parameters are trained on a single NVIDIA V100 GPU with 32GB of memory. For models with an even larger number of parameters, we employ multiple V100 GPUs and enable model parallelism by setting device_map="auto" . This is facilitated by the Accelerate library from Hugging Face, which automatically distributes the model across the available GPUs.

4.6 Evaluation Metrics

We use **BLEU** and **COMET** (Rei et al., 2020) as evaluation metrics to assess the performance of our models. For BLEU we use the SacreBLEU (Post, 2018) implementation, which standardizes tokenization and facilitates reproducibility.

On the other hand, unlike BLEU, which depends on the n-gram overlap between the machine-generated translation and the reference translation, COMET models are trained on a comprehensive dataset comprising human translations and human quality assessments. This dataset is used to predict translation quality while also taking the source side into account. This approach enables COMET to provide a more holistic evaluation that includes fluency, adequacy, and preservation of meaning. We employ the latest model, *Unbabel/wmt22-comet-da*, for our evaluation. This model scales the scores between 0 and 1, where a score approaching 1 indicates a high-quality translation.

By employing both BLEU and COMET, we can ensure that our evaluation is robust and comprehensive, accounting for not only the lexical similarity between the translations and the references but also the overall quality and preservation of meaning in the translations. Moreover, COMET may serve as a superior evaluation metric when assessing the zero-shot performance of LLMs compared to BLEU. As we demonstrated in Section 7, the outputs from LLMs often excel in preserving meanings but might receive a low score if evaluated solely based on n-gram matching.

5 Sentence-level Translation

In this section, we assess the sentence-level translation performance of pretrained LLMs using prompting versus fine-tuned LLMs (Section 5.1). We investigate the effects of incorporating or not incorporating QLoRA during the fine-tuning process (Section 5.2). Additionally, we analyze the impact of varying QLoRA hyperparameters (Section 5.3), including the rank of the low-rank approximation (Section 5.3.1), and the trainable parameters (Section 5.3.2). We also conduct experiments with different sizes of fine-tuning data and compare the results of fine-tuned LLMs with the baseline NMT model (Section 5.4). Lastly, we explore few-shot learning with varying numbers of shots and diverse prompts (Section 5.5).

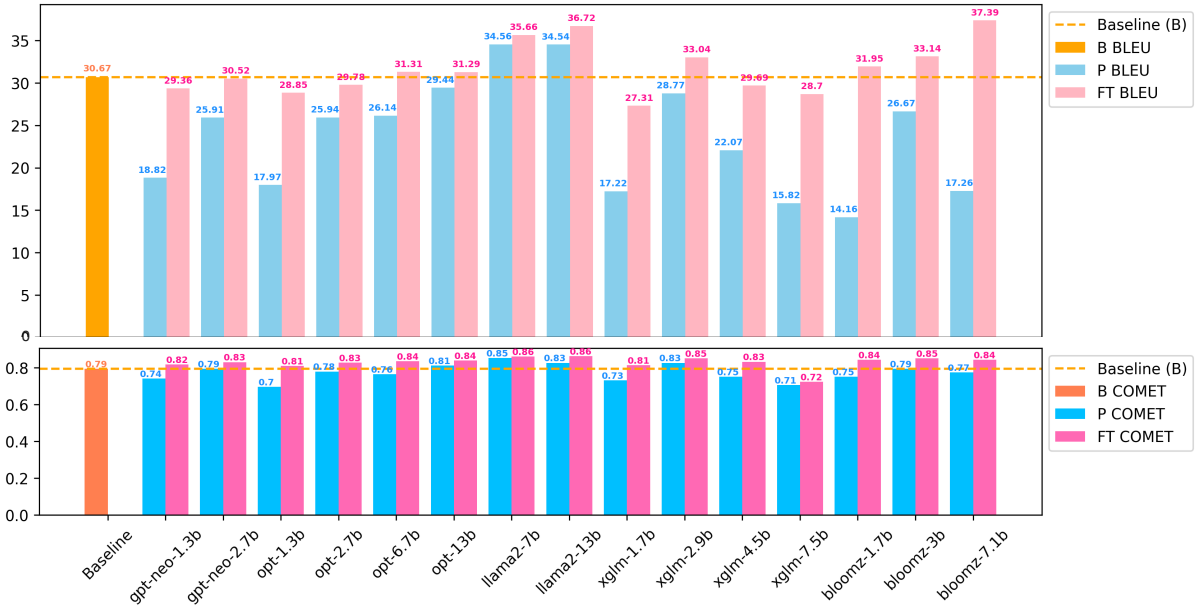

Figure 1: Prompting (*P*) vs. QLoRA fine-tuning (*FT*) on sentence-level translation using various pretrained LLMs. *Baseline* is the NMT system described in Section 4.2. Rank r for QLoRA is set to 64.[3]

5.1 Main Results

We present the results of prompting and QLoRA fine-tuning in Figure 1. Key observations are:

- While there is a significant disparity in BLEU scores, the same is not observed in COMET. All models exhibit comparable COMET scores. The top-performing fine-tuned model, *llama2-13b*, outperforms the *baseline* from 0.837 to 0.862. This indicates that while all models produce semantically coherent translations, their lexical choices, which affect BLEU scores, might differ.
- In terms of BLEU, the *baseline* model surpasses most prompted LLMs, with the exception of *LLAMA 2*. Specifically, *llama2-7b* achieves the highest performance at 34.56 BLEU, marking a 3.89 BLEU improvement over the *baseline*.
- 8 out of the 15 fine-tuned LLMs exceed the *baseline*. This includes both English-centric and multilingual models. The standout model is *bloomz-7.1b* achieving a BLEU score of 37.39, a 6.72 BLEU enhancement compared to the *baseline*.
- Fine-tuning invariably boosts LLM performance on average by 8 BLEU points, with *bloomz-7.1b* witnessing the most substantial leap of 20.13 BLEU.
- No clear advantage is discerned when contrasting prompted multilingual models with English-centric ones. For instance, the multilingual *bloomz-1.7b* scores the lowest at 14.16 BLEU. Yet, when evaluating the fine-tuning gains over prompting, multilingual models average an 11.32 BLEU improvement, surpassing the 5.02 BLEU of their counterparts.
- Bigger models do not consistently outshine their smaller counterparts. For instance, after fine-tuning, *bloomz-1.7b* trumps the larger *opt-13b* (31.95 vs. 31.29 BLEU). Within the same architecture, models with more parameters typically fare better, but there are exceptions, like with *XGLM*, where the *4.5b* and *7.5b* versions lag behind the *2.9b* variant.

In conclusion, while directly prompted LLMs do not universally outperform train-from-scratch MT models, certain LLMs, such as LLAMA 2, defy this trend. Moreover, fine-tuning consistently proves beneficial, with the potential to elevate even underperforming LLMs, like *bloomz-7.1b*, to top-tier performance.

	params(%)	**#GPUs**	**time(hrs)**
No QLoRA	27.40	4	52
QLoRA	0.02	1	10

Table 3: Fine-tuning *xglm-2.9b* with and without QLoRA to achieve the BLEU score of 30.05.[4]Only the self-attention layers are tuned. The rank r for QLoRA approximation is set to 2.

[3]We also report TER in Appendix A.

[4]We train the model without QLoRA for 96 hours in total, and 30.05 is the BLEU score obtained at the best checkpoint.

r	2	4	8	16	32	64	128	256	512
train params(%)	0.02	0.05	0.09	0.19	0.39	0.77	1.53	3.01	5.85
BLEU	31.69	31.72	32.28	32.52	32.80	**33.04**	30.60	30.09	30.31
COMET	0.845	0.846	0.847	0.848	0.849	**0.850**	0.837	0.835	0.836

Table 4: QLoRA fine-tuning results on XGLM 2.9B with various rank r choices. All the weights except for self-attentions are frozen.

5.2 QLoRA vs. No QLoRA

To assess QLoRA's efficacy, we contrast it with the original approach, a more resource-intensive choice: fine-tuning without QLoRA, which excludes both quantization and low-rank adaptation. We train the *xglm-2.9b* model using its native 32-bit precision, necessitating the use of 4 NVIDIA v100s. This is compared against a model fine-tuned with QLoRA set at $r = 2$. For consistency, only the self-attention layers are unfrozen in both models. The comparative results are presented in Table 3.

Achieving a BLEU score of 30.05, the model fine-tuned without QLoRA requires 52 hours across 4 GPUs, totaling 208 GPU hours. In contrast, the QLoRA-enhanced model completes in just 10 hours, marking a 21-fold acceleration and utilizing 1370 times fewer trainable parameters (0.02% compared to 27.4%).

5.3 QLoRA Hyperparameters

We investigate the impact of selecting different ranks for LoRA and the unfrozen parameters for fine-tuning. We present the results for XGLM 2.9B.

5.3.1 Rank r

The rank r of the decomposition matrices influences the number of trainable parameters, with a larger r resulting in more trainable parameters. We assess the performance associated with different choices of r, ranging from 2 to 512, in Table 4, while only unfreezing the self-attention layers.

With $r = 64$, the model attains its optimal performance. However, either reducing or increasing the number of trainable parameters adversely affects the model's performance. Interestingly, when $r = 512$, the performance deteriorates even more than when $r = 2$, despite the fact that the latter converges more quickly due to a smaller number of trainable parameters.

5.3.2 Trainable Parameters

Next, we aim to determine which part of the model should be fine-tuned. To do this, we unfreeze the parameters in different layers of the XGLM 2.9B model. As illustrated in Table 5, we experiment with unfreezing parameters from various layers, including the self-attention layers, embedding layers, fully-connected feed-forward layers, and the LM head layers. The results indicate that fine-tuning only the self-attention layer is sufficient to yield the best performance.

Params	a	a+e	a+e+f	a+e+f+l
BLEU	**31.69**	30.09	30.30	28.39
COMET	**0.845**	0.837	0.834	0.826

Table 5: QLoRA fine-tuning results on XGLM 2.9B with different trainable parameters. *a*: self-attentions; *e*: embeddings; *f*: fully-connected feed-forward layers; *l*: lm head. Rank r is set to 2.

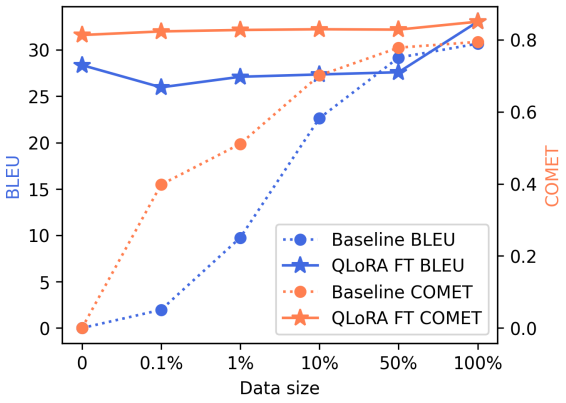

Figure 2: The performance of the *baseline* system and fine-tuned XGLM 2.9B trained with different amounts of data.

5.4 Data Curves

The performance of a traditional MT model is closely tied to the volume of its training data, as highlighted by (Koehn and Knowles, 2017). However, for LLMs, which have already benefited from vast training datasets, does this correlation still hold? To investigate, we compare the responses of both MT model types to varying training data sizes. We incrementally adjust the dataset size from 0.1% (2,366 examples) to its entirety and then train the

Prompt 1	{ **French:** [fr sent] **English:** [en sent] } x **K** **French:** [fr sent] **English:**
Prompt 2	{ **Translate French to English: French:** [fr sent] **English:** [en sent] } x **K** **Translate French to English: French:** [fr sent] **English:**
Prompt 3	**Translate French to English:** { **French:** [fr sent] **English:** [en sent] } x **K** **Translate French to English: French:** [fr sent] **English:**
Prompt 4	**Translate French to English:** **French:** { [fr sent] } x **K** **English:** { [en sent] } x **K** **Translate French to English: French:** [fr sent] **English:**
Prompt 5	{ **French:** [fr sent] **Translate to English:** [en sent] } x **K** **French:** [fr sent] **Translate to English:**

Table 6: Prompts used in **K**-shot learning. The substrings within {} are repeated **K** times.

	BLEU				**COMET**			
	0-shot	**1-shot**	**5-shot**	**10-shot**	**0-shot**	**1-shot**	**5-shot**	**10-shot**
Prompt 1	27.08	29.15	29.72	29.62	0.814	0.828	0.833	0.834
Prompt 2	28.36	29.46	29.86	29.95	0.813	0.830	0.836	0.835
Prompt 3	28.36	29.33	29.86	29.74	0.813	0.831	0.835	0.834
Prompt 4	28.36	29.46	28.66	27.83	0.813	0.830	0.829	0.825
Prompt 5	11.82	28.76	29.80	29.70	0.631	0.827	0.834	0.834

Table 7: Few-shot learning results on XGLM 2.9B.

baseline model and fine-tune the LLMs. The outcomes of this experiment are depicted in Figure 2.

The *baseline* curve validates the assumption that performance improves with increased data availability. In contrast, LLMs make a robust debut; even without additional training data, they achieve a BLEU score comparable to the *baseline* trained on half the dataset. Yet, their performance does not consistently improve with more data. In fact, fine-tuning with less than 50% (1.2 million examples) of the data seems counterproductive, diminishing performance until the full dataset comes into play.

5.5 Few-shot Learning

In this section, we evaluate the few-shot learning performance of LLMs. Few-shot learning is also denoted as **K**-shot, with **K** representing the number of examples provided before the query, where in our case, examples are randomly sampled from the training set. We also compare the impact of 5 slightly varied prompts, detailed in Table 6. The results of the experiments are presented in Table 7.

When **K** $>=$ 1, the model consistently outperforms the 0-shot scenario. For *prompt 5*, 1-shot dramatically enhances the model's capability, elevating the BLEU score from 11.82[5] to 28.76. However, the performance does not exhibit a linear growth with increasing **K**; it plateaus. In the case of *prompt 4*, augmenting **K** even diminishes the performance.

In our experiments, the choice of prompt is particularly impactful for 0-shot performance, especially when comparing *prompt 5* to the others. However, this impact seems to lessen when examples are presented before the query.

6 Document-level Translation

In this section, we delve into the proficiency of LLMs in document-level translation. Our primary observations, contrasting the prompted and fine-tuned LLMs, are detailed in Section 6.1. Additionally, we explore the influence of document length, measured by the number of sentences per document, in Section 6.2.

6.1 Main Results

Figure 3 presents the results for document-level translations. Key takeaways include:

[5]We observed many empty generations when prompting with **Prompt 5**. One hypothesis is that the prompt is ambiguous and the model is confused about what to translate.

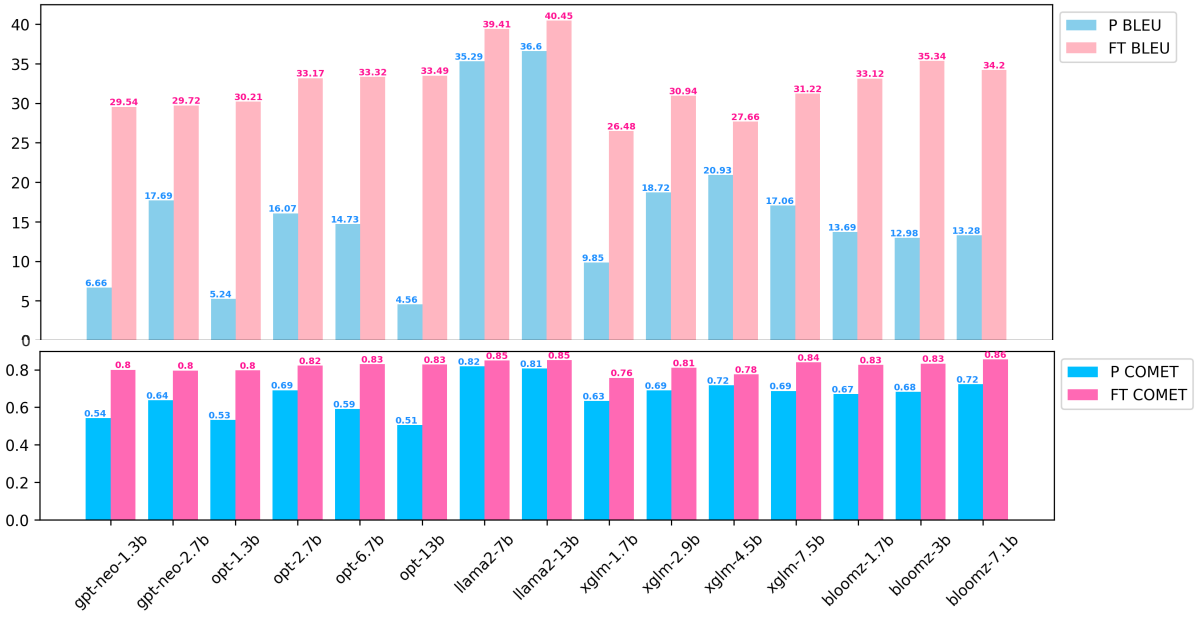

Figure 3: Prompting (*P*) vs. QLoRA fine-tuning (*FT*) on document-level translation using various pretrained LLMs. Rank r for QLoRA is set to 64.

- In contrast to sentence-level translation, prompted LLMs face challenges with document-level translation. 4 out of the 15 LLMs register BLEU scores below 10. However, consistent with sentence-level findings, *LLAMA 2* continues to stand out in zero-shot performance, with the *7b* and *13b* versions achieving impressive BLEU scores of 35.29 and 36.6, respectively.
- Fine-tuning demonstrates significant promise for document-level translations, enhancing the BLEU scores of their prompted counterparts by an average of 16.33. The most notable improvement is seen in *opt-13b*, which witnesses a BLEU increment of 28.93 (from 4.56 to 33.49).
- Unlike sentence-level translation, where COMET scores remain consistent across all models, document-level translation displays a more pronounced variance. This variability is particularly evident in prompted models but diminishes in fine-tuned ones.
- Trends observed in sentence-level translation (Section 5.1) persist in the document-level context: (1) Both English-centric and multilingual models deliver comparable performance. (2) Larger models do not consistently surpass their smaller counterparts. (3) Top-performing models in sentence-level translation, such as *LLAMA 2* and *BLOOMZ*, maintain their dominance in document-level translation.

6.2 Input Lengths

We construct the examples for fine-tuning by merging sentences from an original document, as described in Section 4.4. By varying the number of sentences per document – specifically, 5, 10, and 15 – we present our findings in Table 8. Notably, fine-tuning with sets of 10 consecutive sentences yields the optimal performance, registering a 30.94 BLEU and a 0.811 COMET score.

#sents	5	10	15
BLEU	29.07	**30.94**	28.32
COMET	0.781	**0.811**	0.742

Table 8: QLoRA fine-tuning on documents composed of different numbers of sentences[6]. XGLM 2.9B systems are fine-tuned with $r = 64$, and self-attentions and embeddings as trainable parameters.

7 Qualitative Analysis

Figure 4 shows translations of two French sentences using various LLMs. In the first example, when prompted, *bloomz-7.1b* replicates the source sentence verbatim, neglecting to translate. While this does not occur for every test set sample, as shown in the second example, similar behavior is noted in other prompted LLMs. On the other hand,

[6]We match the length of text documents to that of train documents.

French	L'ONU donne un bilan même plus élevé avec 979 morts et 1 902 blessés.
English	The UN has reported even higher numbers with 979 dead and 1,902 injured.
bloomz-7.1b P	L'ONU donne un bilan même plus élevé avec 979 morts et 1 902 blessés.
bloomz-7.1b FT	The UN gives a higher figure with 979 dead and 1 902 wounded.<eos>.<eos>.<eos>.
llama2-13b P	979 deaths and 1,902 injuries, according to the UN's latest tally.
llama2-13b FT	The UN gives an even higher death toll of 979 and 1 902 injured.<eos>The UN gives an even higher death toll of 979 and 1 902 injured.<eos>The UN gives an even higher death toll of 979 and 1 902 injured.<eos>The
French	L'affaire NSA souligne l'absence totale de débat sur le renseignement
English	NSA Affair Emphasizes Complete Lack of Debate on Intelligence
bloomz-7.1b P	French: The NSA case highlights the complete absence of debate on intelligence.
bloomz-7.1b FT	The NSA Case Sheds Light on the Absence of Debate on Intelligence<eos>.<eos>.<eos>.
llama2-13b P	The NSA case highlights the complete lack of debate on intelligence gathering.
llama2-13b FT	The NSA Scandal Highlights the Lack of Intelligence Debate<eos>eos>eos>

Figure 4: Translations from prompted (**P**) and fine-tuned (**FT**) LLMs.

the translation using *llama2-13b P*, though not mirroring the reference verbatim, retains the original sentence's meaning. Both fine-tuned LLMs produce proper translations with the initial segment of the generated sequences. *Bloomz-7.1b* appends a <eos> token post-translation, while *llama2b-13b* reiterates its translation multiple times. Both outputs necessitate post-processing, specifically truncating the output at the first occurrence of the <eos> token.

In the second example, the LLM-generated translations retain the meaning of the reference translation, showcasing LLMs' potential in the translation tasks.

8 Conclusions

In this study, we investigate the capabilities of LLMs in performing machine translation tasks. Through comprehensive experiments, we assess the effectiveness of prompting, few-shot learning, and fine-tuning using QLoRA for French-English translation. Our key findings are:

1. The proficiency of LLMs in machine translation varies. While **LLAMA 2** consistently outperforms its counterparts, other models, when relying solely on few-shot learning, often lag behind models trained from scratch.

2. Fine-tuning invariably enhances performance, particularly for models that struggle with few-shot learning and for translating documents. It can transform a seemingly inadequate model into a top-tier translation model, as seen with *bloomz-7.1b*.

3. QLoRA, due to its efficiency, can be a superior alternative to original fine-tuning methods.

4. Fine-tuning LLMs with QLoRA can be a promising and new paradigm for machine translation practice.

In the future, we are interested in exploring two primary avenues. (1) While our current study demonstrates the promise of LLMs trained on English-centric data for French-to-English translations, it raises intriguing questions: Would these results hold true for other language pairs, especially for low-resource languages? And would there be a noticeable difference in performance between English-centric and multilingual LLMs in such scenarios? (2) Our experiments are confined to decoder-based LLMs. Moving forward, we are also interested in comparing these models against their encoder-decoder counterparts, such as mT5(Xue et al., 2021), mBART (Liu et al., 2020), NLLB (Costa-jussà et al., 2022).

Limitations

Single dataset and language pair Our experiments are confined to a single dataset and the French-English language pair. It remains unclear if our findings are generalizable to other datasets and language pairs.

Medium-sized LLMs We have only experimented with medium-sized LLMs due to computational resource constraints. The necessity of fine-tuning for significantly larger LLMs remains an open question.

Acknowledgements

This work is supported in part by an Amazon Initiative for Artificial Intelligence (AI2AI) Faculty Research Award.

References

Armen Aghajanyan, Luke Zettlemoyer, and Sonal Gupta. 2020. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. *arXiv preprint arXiv:2012.13255*.

Rachel Bawden and François Yvon. 2023. Investigating the translation performance of a large multilingual language model: the case of bloom. *arXiv preprint arXiv:2303.01911*.

Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow. If you use this software, please cite it using these metadata.

Eleftheria Briakou, Colin Cherry, and George Foster. 2023. Searching for needles in a haystack: On the role of incidental bilingualism in PaLM's translation capability. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9432–9452, Toronto, Canada. 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, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

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.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Felix Hieber, Michael Denkowski, Tobias Domhan, Barbara Darques Barros, Celina Dong Ye, Xing Niu, Cuong Hoang, Ke Tran, Benjamin Hsu, Maria Nadejde, et al. 2022. Sockeye 3: Fast neural machine translation with pytorch. *arXiv preprint arXiv:2207.05851*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Marcin Junczys-Dowmunt. 2019. Microsoft translator at wmt 2019: Towards large-scale document-level neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 225–233.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39.

Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. 2018. Measuring the intrinsic dimension of objective landscapes. *arXiv preprint arXiv:1804.08838*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. 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 4582–4597, Online. Association for Computational Linguistics.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, et al. 2021. Few-shot learning with multilingual language models. *arXiv preprint arXiv:2112.10668*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

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. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Yasmin Moslem, Rejwanul Haque, and Andy Way. 2023. Adaptive machine translation with large language models. *arXiv preprint arXiv:2301.13294*.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

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.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2022. Multitask prompted training enables zero-shot task generalization. In *ICLR 2022-Tenth International Conference on Learning Representations*.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *54th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725. Association for Computational Linguistics (ACL).

Suzanna Sia and Kevin Duh. 2022. Prefix embeddings for in-context machine translation. In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 45–57, Orlando, USA. Association for Machine Translation in the Americas.

Suzanna Sia and Kevin Duh. 2023. In-context learning as maintaining coherency: A study of on-the-fly machine translation using large language models. In *Proceedings of Machine Translation Summit XIV (Volume 1: Research Track)*.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231.

Xin Tan, Longyin Zhang, Deyi Xiong, and Guodong Zhou. 2019. Hierarchical modeling of global context for document-level neural machine translation. 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 1576–1585.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Zhaopeng Tu, Yang Liu, Shuming Shi, and Tong Zhang. 2018. Learning to remember translation history with a continuous cache. *Transactions of the Association for Computational Linguistics*, 6:407–420.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. Prompting PaLM for translation: Assessing strategies and performance. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15406–15427, Toronto, Canada. Association for Computational Linguistics.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models. *arXiv preprint arXiv:2304.02210*.

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

Hongfei Xu, Deyi Xiong, Josef Van Genabith, and Qiuhui Liu. 2021. Efficient context-aware neural machine translation with layer-wise weighting and input-aware gating. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3933–3940.

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. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Bing Yin, and Xia Hu. 2023. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *arXiv preprint arXiv:2304.13712*.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study. *arXiv preprint arXiv:2301.07069*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Lingpeng Kong, Jiajun Chen, Lei Li, and Shujian Huang. 2023. Multilingual machine translation with large language models: Empirical results and analysis. *arXiv preprint arXiv:2304.04675*.

A TER on Sentence-level Translations

The Translation Edit Rate (TER) is a metric introduced by Snover et al. (2006) to quantify the amount of human editing required to align a system's output with a reference translation. Specifically, TER is calculated as the ratio of the total edits made to the length of the reference translation. Such edits encompass insertions, deletions, single-word substitutions, and shifts in word sequence. A lower TER indicates better alignment with the reference. As illustrated in Figure 5, when evaluated using TER, LLMs do not exhibit a noticeable improvement over the baseline model.

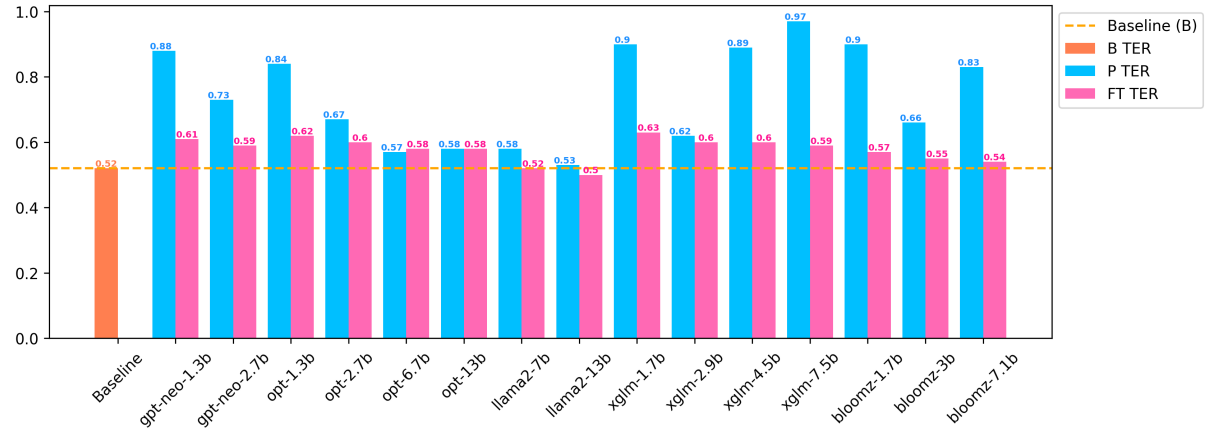

Figure 5: Prompting (*P*) vs. QLoRA fine-tuning (*FT*) on sentence-level translation using various pretrained LLMs. *Baseline* is the NMT system described in Section 4.2. Rank r for QLoRA is set to 64.