Eliciting the Translation Ability of Large Language Models via Multilingual Finetuning with Translation Instructions

Jiahuan Li*, **Hao Zhou**[1*], **Shujian Huang**[1†], **Shanbo Cheng**[2] and **Jiajun Chen**[1]
[1] National Key Laboratory for Novel Software Technology, Nanjing University, China
{lijh,zhouh}@smail.nju.edu.cn, {huangsj,chenjj}@nju.edu.cn
[2] Bytedance, China
chengshanbo@bytedance.com

Abstract

Large-scale pretrained language models (LLMs), such as ChatGPT and GPT4, have shown strong abilities in multilingual translation, without being explicitly trained on parallel corpora. It is intriguing how the LLMs obtain their ability to carry out translation instructions for different languages. In this paper, we present a detailed analysis by finetuning a multilingual pretrained language model, XGLM-7.5B, to perform multilingual translation following given instructions. Firstly, we show that multilingual LLMs have stronger translation abilities than previously demonstrated. For a certain language, the translation performance depends on its similarity to English and the amount of data used in the pretraining phase. Secondly, we find that LLMs' ability to carry out translation instructions relies on the understanding of translation instructions and the alignment among different languages. With multilingual finetuning with translation instructions, LLMs could learn to perform the translation task well even for those language pairs unseen during the instruction tuning phase.

1 Introduction

The emergence of large pretrained language models (LLMs) (Brown et al., 2020; OpenAI, 2023) has revolutionized the research of machine translation (Hendy et al., 2023; Garcia et al., 2023). These models have demonstrated remarkable multilingual translation capabilities, without requiring explicit training on parallel corpora. For instance, XGLM, a medium-sized multilingual language model, outperforms supervised models using only several examples as demonstrations (Lin et al., 2022); the cutting-edge LLM GPT4 has been shown to perform comparably to commercial translation systems on multiple language pairs (Jiao et al., 2023b).

Most existing research on LLMs for machine translation focuses on in-context learning (ICL), i.e., taking several parallel sentences as the demonstration to guide LLMs to perform translation (Vilar et al., 2023; Agrawal et al., 2023; Hendy et al., 2023; Zhu et al., 2023). However, these methods rely heavily on the in-context learning ability of LLMs. For smaller models, e.g., models with only 1B or 7B parameters, the relatively weak ICL ability may result in an underestimation of their potential translation ability.

Instead of relying on the ICL abilities, we propose to investigate the ability of LLMs by directly training them to follow translation instructions. Inspired by the recent success of instruction tuning (Wei et al., 2022; Chung et al., 2022), we organize multilingual translation tasks as different instances of the translation instruction, with each instance corresponding to a specific language pair. By training the LLMs to follow these instructions, i.e., with **m**ultilingual **F**inetuning with **T**ranslation **I**nstructions (mFTI), it is possible to better elicit translation ability inside LLMs.

Our results show that by training on a mixed dataset of 1,000 sentences per language pair, mFTI outperforms the 8-shot in-context learning by near 3 BLEU on average, showing a greater potential of LLMs' translation ability than previously demonstrated (Lin et al., 2022). In addition, we also discuss how mFTI improves the LLMs and which factors influence the performance.

To better understand why LLMs could follow these instructions, we design a mFTI setting where only a subset of the translation instructions, i.e., language pairs, are used for training. Thus LLMs need to generalize their instruction

*Equal contribution.

†Corresponding author.

Transactions of the Association for Computational Linguistics, vol. 12, pp. 576–592, 2024. https://doi.org/10.1162/tacl_a_00655
Action Editor: Barry Haddow. Submission batch: 7/2023; Revision batch: 7/2023; Published 4/2024.

following abilities for those language pairs unseen during mFTI. Surprisingly, mFTI elicits the translation ability not only for trained language pairs but also for those unseen during instruction training. With further experiments and analyses, we find that LLMs could learn translation behavior in general by being trained to translate even irrelevant language pairs. It is also interesting that with mFTI, LLMs learn to directly align languages through the use of pivot languages, which enhances the instruction-following ability for unseen language pairs.

2 Multilingual Finetuning with Translation Instructions

2.1 Overall Framework

Given a corpus of multilingual parallel sentences and their languages $\mathcal{M} = \{(l_s{}^i, l_t{}^i, \mathbf{x}^i, \mathbf{y}^i)\}$, where $l_s{}^i$ and $l_t{}^i$ are names of the source and target language of i-th parallel sentence $(\mathbf{x}^i, \mathbf{y}^i)$, respectively, mFTI leverages an instruction template $\mathcal{T}$ to organize the corpus $\mathcal{M}$ into a language modeling dataset $\mathcal{D}$. Each sentence d^i in $\mathcal{D}$ is an instantiation of the translation instruction with a specific sentence pair: $d^i = \mathcal{T}(l_s{}^i, l_t{}^i, \mathbf{x}^i, \mathbf{y}^i)$. The parameter of LLMs are then optimized using a standard next-token-prediction objective on $\mathcal{D}$:

$$\arg\max_{\theta} \sum_{i=1}^{|\mathcal{D}|} \sum_{j=1}^{|d^i|} \mathtt{log}\, p_\theta(d^i_j | d^i_{<j}), \qquad (1)$$

where θ are parameters of LLMs. The instruction template we adopt is

Translation: $[l_s]$: $\mathbf{x}$ $[l_t]$: $\mathbf{y}$,

where the prefix ''Translation:'' is used to indicate the translation task; the pattern ''[·]:'' is used to identify the name of the specific language.

2.2 Experiment Setup

Backbone Language Model We consider XGLM-7.5B (Lin et al., 2022) as our backbone language model. XGLM-7.5B is a massive multilingual auto-regressive language model, which is trained on a massive corpus of 500 billion tokens comprising 30 diverse languages. Low-resource languages have been up-sampled during training, making it an ideal backbone model for multilingual translation research.

Languages Following Lin et al. (2022), our evaluation involves 13 languages that are covered in the pretraining corpus of XGLM, i.e., English (En), German (De), French (Fr), Catalan (Ca), Finnish (Fi), Russian (Ru), Bulgarian (Bg), Chinese (Zh), Korean (Ko), Arabic (Ar), Swahili (Sw), Hindi (Hi), and Tamil (Ta). Among these languages, En, De, Fr, Ru, and Zh are high-resource languages (with ratios in the XGLM pretraining data greater 4%); Ko, Fi, Ar, and Bg are medium-resource languages (with ratios between 0.5%–4%); and Ca, Hi, Ta, and Sw are low-resource languages (with ratios under 0.5%).

Evaluation Datasets Following previous work (Lin et al., 2022), we evaluate translation models on the FLORES-101 dataset (Goyal et al., 2022), which provides manual translations of 1012 sentences in 101 languages.

Finetuning Datasets Our finetuning dataset primarily comes from WikiMatrix (Schwenk et al., 2021). WikiMatrix provides a parallel corpus for 1620 different language pairs, including many non-English language pairs, which enables a systematic investigation for the translation of languages other than English. We also leverage the MultiCCAligned (El-Kishky et al., 2020) corpus for language pairs that are not contained in WikiMatrix, including Hi-Sw, Ko-Sw, Ta-Sw, Sw-Hi, Sw-Ko, and Sw-Ta.

Optimization Details We finetune all models using the Adam (Kingma and Ba, 2014) optimizer with the learning rate fixed as $5e-6$. We use a fixed batch size of 80 sentences and finetune models for 1 epoch or 2000 steps (depending on the size of the training corpus) for all experiments.

3 Understanding the Potential Translation Ability of LLMs

In this section, we first assess the overall translation performance of mFTI by comparing it to few-shot in-context learning.[1] We then present a detailed analysis of how the corpus for mFTI influences the translation quality.

[1] We randomly select 8 examples from the FLORES-101 dev split as the demonstration for ICL. Random selection strategy has been shown to be good enough in much previous work (Vilar et al., 2023; Zhu et al., 2023). The template we use for ICL is <src_text> = <tgt_text>, which shows good performance according to Zhu et al. (2023).

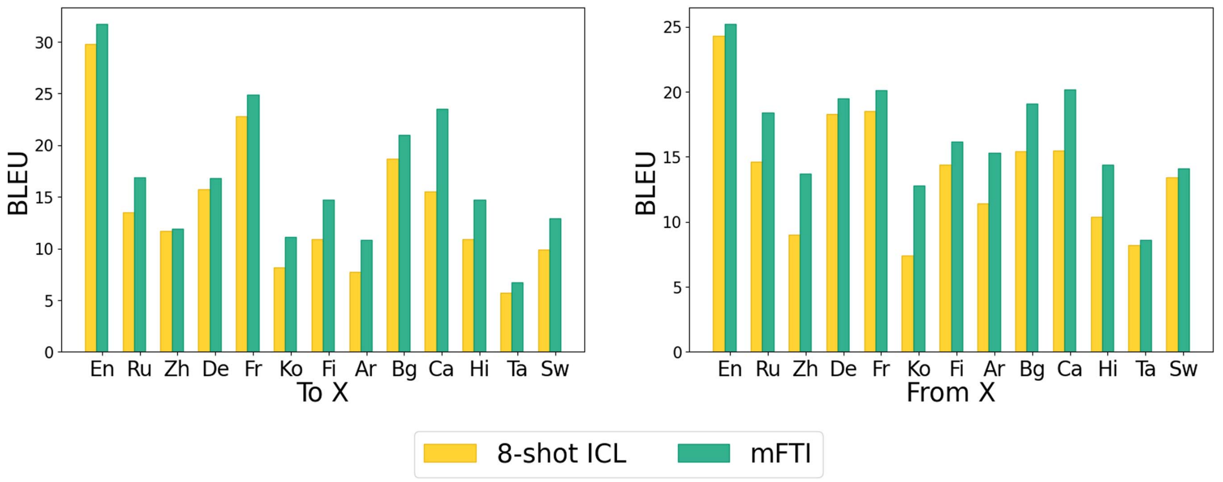

Figure 1: Translation performance of 8-shot ICL and mFTI using 1000 sentences per language pair. Languages are ordered by the data amount in the pretraining corpus.

3.1 Translation Ability of LLMs

We finetune XGLM on 156 language pairs spanning all 13 languages. Since our goal is to elicit the translation ability of LLMs using a small number of examples, we limit the number of parallel sentences to 1000 per language.

mFTI Better Elicits Translation Ability than Few-shot ICL. Figure 1 shows the average BLEU for translation to and from language X, respectively. Full results on each langauge direction can be found in Appendix A. It is clear that mFTI leads to better translation performance than 8-shot ICL for all language pairs (3 BLEU on average). For some languages, the gap is up to 8 BLEU (e.g., translating into Catalan). This demonstrates the effectiveness of mFTI in eliciting LLM's translation ability. It also shows that LLMs have a greater potential for multilingual translation than we saw with ICL (Lin et al., 2022).

Even for translating to and from English, mFTI still outperforms 8-shot ICL, but with a much smaller gap. This indicates that LLMs with ICL are better at performing tasks that involve English rather than other languages, but they still have the potential to perform even better.

	To X	**From X**
Data Amount in Pretraining	0.39	0.36
Similarity To English		
Geography	0.93	0.87
Syntax	0.85	0.80
Phylogeny	0.71	0.75
Phonology	0.50	0.49
Inventory	0.51	0.41

Table 1: Spearman correlation between average translation performance (in BLEU) and possible influence factors (data amount in pretraining, similarity to English). The performance of translating to and from language X is calculated separately.

XGLM is Still an English-centric Model. The translation performance for each language varies greatly. Considering that the number of sentences used in mFTI is the same for each language, one may suspect that the translation performance of each language largely depends on the amount of its pretraining data. For this reason, the languages in Figure 1 are listed in descending order of their data amount in the XGLM pretraining. However, there are clear fluctuations. For example, Russian and Chinese are the two languages with the largest portion of pretraining data other than English, but their translation performance is much worse than some other languages, such as French.

We calculate the Spearman correlation between the translation performance and possible influence factors, namely, data amount in pretraining and similarity to English. For data amount, we use the size of the pretraining corpus reported in Lin et al. (2022). For similarity to English, we adopt lang2vec,[2] which is a toolkit for querying

[2] `https://github.com/antonisa/lang2vec`.

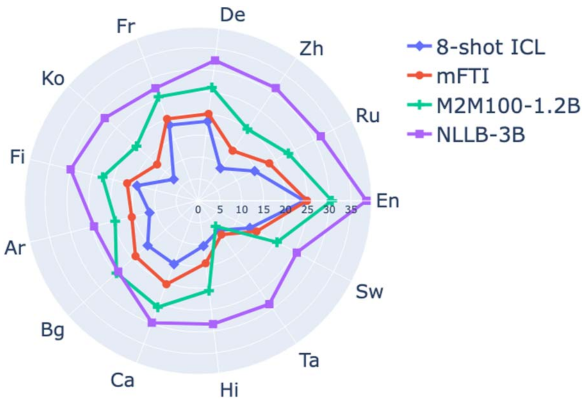

Figure 2: Comparison of mFTI with conventional supervised machine translation models. Performances are evaluated in BLEU.

the URIEL typological database, to get each language's feature vector of different perspectives including *geography*, *syntax*, *phylogeny*, *phonology*, and *inventory*.[3]

As shown in Table 1, the translation performance indeed has a positive correlation with data amount in pretraining (0.39/0.36). However, the similarity between a specific language and English plays a more important role in determining the final translation performance. All considered features demonstrate a higher correlation coefficient than the data amount in pretraining. This indicates that XGLM is still a predominantly English-centric model. Based on these observations, we suggest taking the relation between different languages into consideration when collecting and sampling data for pretraining multilingual LLMs.

It is not Trivial for LLM-based Models to Outperform Conventional Supervised MT Models. To better posit the performance of mFTI, we compare it with two conventional supervised MT models, i.e., M2M-1.2B (Fan et al., 2020) and NLLB-3B (Costa-jussà et al., 2022), in Figure 2.[4] We can see that despite mFTI significantly improving over 8-shot ICL and sometimes achieving comparable performance to M2M-615M, it still lags behind the stronger NLLB-3B by a large margin, rendering the challenge to adopt a medium-sized LLM to outperform large-scale supervised MT models.

3.2 mFTI Brings Consistent Improvements across Different Metrics, LLMs, and Finetuning Strategies

In order to understand the universal effectiveness of mFTI, we present experiments on more LLMs, i.e., BLOOM-7b1 (Scao et al., 2022) and LLaMA (Touvron et al., 2023), and parameter-efficient finetuning strategy LoRA (Hu et al., 2022). We report the performance averaged on 156 translation directions evaluated by both sacreBLEU (Post, 2018) and COMET (Rei et al., 2022)[5] in Table 2.[6]

Firstly, we can see that methods based on XGLM-7.5B performs significantly better than BLOOM-7B and LLaMA-7B. This is because many low-resource languages are ill-represented in BLOOM and LLaMA. Secondly, mFTI consistently outperforms 8-shot ICL in terms of BLEU and COMET on all three studied LLMs, regardless of the finetuning strategy, which demonstrates the universal effectiveness in different scenarios. Contrary to previous findings (Jiao et al., 2023a), we did not find LoRA to perform better than full finetuning. We hypothesize that learning translation on 156 pairs simultaneously is more challenging and requires more model capacity, making full finetuning a better choice than LoRA in this scenario.

3.3 mFTI Enhances Direct Language Alignment

A distinct difference between ICL and mFTI is that mFTI could learn from more parallel sentences and update the model if needed. It is interesting to see what changes after the update. Many previous studies (Zhang et al., 2023; Jiao et al., 2023b) have shown that translating by pivoting through English significantly improves ICL's translation performance. To this end, we compare performance gains of pivot translation using ICL and mFTI, respectively.

Figure 3 presents the result. Each value in the grid is the BLEU difference before and after pivoting through English. We can first observe that pivoting through English indeed improves translation performance for ICL, up to 10 BLEU in some language pairs. However, after mFTI, the gap has been significantly reduced. Considering the fact the mFTI achieves an average 3 BLEU

[3] We refer readers to Littell et al. (2017) for details on how the feature vector is obtained.

[4] We also include the performance evaluated by COMET in Appendix C.

[5] We use the wmt22-comet-da version.

[6] Detailed hyperparameters are in Appendix B.

	BLOOM-7B		LLaMA-7B		XGLM-7.5B	
	BLEU	**COMET**	**BLEU**	**COMET**	**BLEU**	**COMET**
8-shot ICL	8.4	60.9	9.0	61.0	13.9	73.4
mFTI (LoRA)	9.0	64.3	9.5	63.9	16.7	77.0
mFTI (Full Finetuning)	**10.2**	**65.4**	**9.8**	**66.0**	**16.9**	**77.7**

Table 2: Averaged translation performance on all 156 language pairs of 8-shot ICL and mFTI using different LLMs and finetuning strategies.

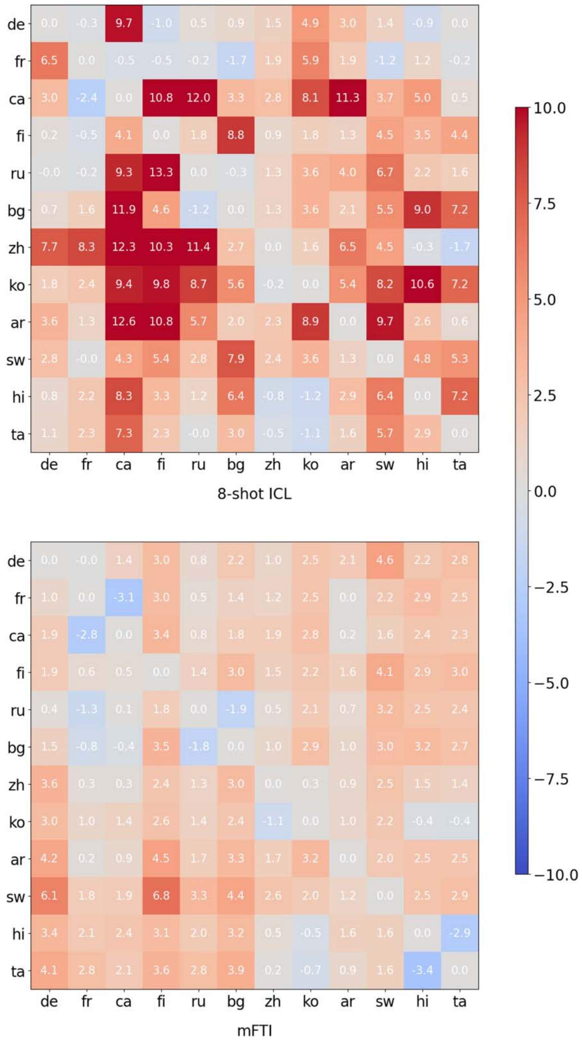

Figure 3: Changes of BLEU score after pivoting through English for 8-shot ICL and mFTI.

higher than ICL, the reduction of benefits from pivoting through English compared to direct translation may indicate a better direct alignment between languages.

3.4 Influencing Factors of mFTI

The Quality of the Finetuning Corpus is Crucial. Recent work on instruction tuning demonstrates that the quality of instruction data is crucial for achieving good performances (Zhou et al., 2023). We observe a similar trend when performing mFTI. Specifically, we construct high and low-quality finetuning corpora by selecting parallel sentences according to their attached LASER[7] similarity score from the full set of parallel sentences. According to the results in Table 3, finetuning with high-quality parallel sentences can improve the BLEU score by around 2 points compared to finetuning with low-quality parallel sentences, emphasizing the importance of corpus quality, validating the importance of the quality of finetuning corpus.

	BLEU
Low quality	15.0
High quality	16.9

Table 3: The translation performance of finetuned XGLM as the quality of finetuning corpus varies.

The Effectiveness of mFTI Scales with Model Size and Training Examples. Figure 4 shows translation performance when varying the number of training examples per language pair (1k, 2k, 4k, 8k, 16k, 32k) and the number of model parameters (564M, 1.7B, 2.9B, 4.5B, 7.5B). As we can see, it follows a standard log-linear scaling law in terms of both the number of training examples and model size, which is consistent with findings in the previous work (Kaplan et al., 2020).

4 Understanding the Ability of Carrying Out Translation Instructions

In this section, we present a comprehensive analysis on how mFTI improves the model's ability to carry out translation instructions.

[7] https://github.com/facebookresearch/LASER.

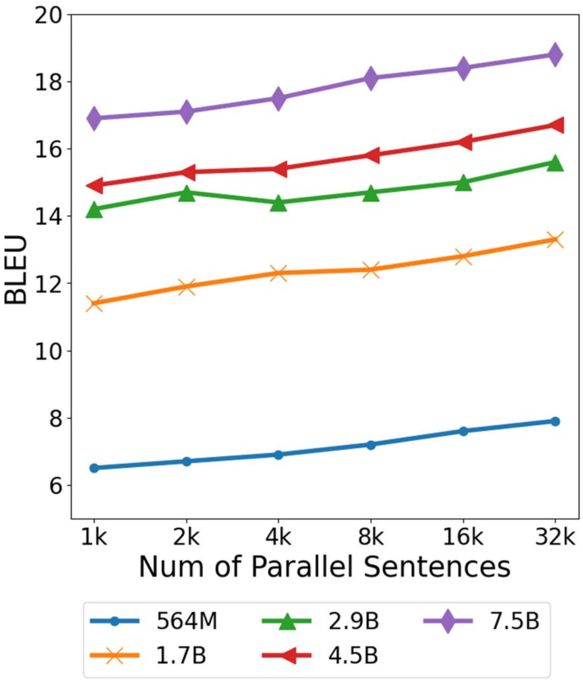

Figure 4: The translation performance of finetuned XGLM as the number of model parameters and training examples scales.

We begin by presenting an overarching experiment where we intentionally withhold certain language pairs during the mFTI process, which allows us to study models' ability to carry out translation instructions under different conditions.

Furthermore, we delve deeper into our analysis by exploring how mFTI enhances LLMs' ability to carry out translation instructions from following perspectives: better understanding of translation instructions (Section 4.3 and Section 4.4) and better alignment between languages to execute translation instructions (Section 4.5).

4.1 Manipulating Conditions

In Section 3, we have presented results in a fully supervised setting, where all testing language pairs are seen during instruction tuning. To provide further insights into LLMs' generalization ability across language pairs, we simulate a more realistic scenario where there may be a lack of source and/or target language sentences during the instruction tuning process.

More specifically, from the 13 selected languages, we hold out 6 languages as unseen languages. We further partition the remaining 7 languages into three groups: Only-Source (languages only appear on the source side), Only-Target (languages only appear on the target side), and Source-Target (languages appear on both the source and target side). We then form language pairs from these partitions following the requirement of partitions. This allows us to assess mFTI's performance under the following conditions:

- **Seen Both Sides** Both the source side and target side language appear in the finetuning corpus. This can be further divided to:
  - **Same Direction**. The same translation direction is trained during mFTI.
  - **Reversed Direction**. The same translation direction does not appear when training, but the reversed direction does.
  - **Unseen Direction**. The translation pair (neither the same nor the reverse) does not appear when training.
- **Unseen Src**. Only the target language sentences appear when training.
- **Unseen Tgt**. Only the source language sentences appear when training.
- **Unseen Both Sides**. Neither source language nor target language sentences appear in the finetuning corpus.

4.2 mFTI Learns to Follow Translation Instruction across Conditions

We finetune XGLM on the corpus described in the previous section. Since there are 16 language directions in the training corpus, we denote the finetuned model as mFTI-16. The model finetuned on all language pairs is denoted as mFTI-all. Table 4 shows the results.

mFTI-16 Brings Improvements on Most Settings, Yet Much Less Than mFTI-all. Firstly we can see that mFTI-16 brings improvements on most settings except Reversed Direction, demonstrating the effectiveness of mFTI-16. However, the improvements are less when compared mFTI-all, even for the Same Direction partition. This can be attributed to fewer language pairs when finetuning, which we will discuss in Section 4.3.

Language Position Shift Between Training and Testing Has Negative Effects on Translation Performance. The translation performance of mFTI-16 on Reversed Direction degrades by 0.8

	Seen Both Sides					
	Same Direction	Reversed Direction	Unseen Direction	Unseen Src	Unseen Tgt	Unseen Both Sides
8-shot ICL	14.5	14.5	11.2	13.5	13.6	14.6
mFTI-16	15.7(+1.2)	13.7(−0.8)	12.6(+1.4)	14.9(+1.4)	14.5(+0.9)	15.3(+0.7)
mFTI-all	16.7	16.8	14.6	17.6	17.0	18.4

Table 4: Translation performance under different data conditions. mFTI-16: XGLM multilingual finetuned with translation instructions on a mixture of 16 language pairs described in Section 4.1.

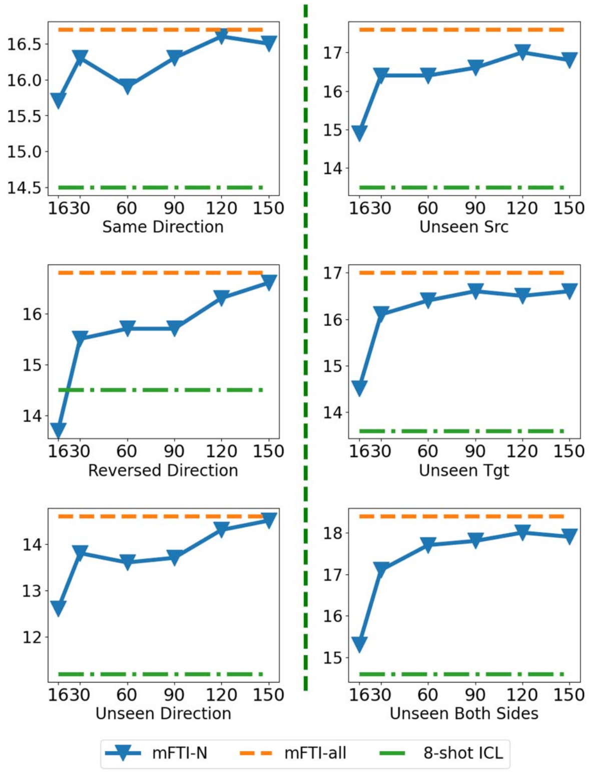

Figure 5: Translation performance on different partitions as the number of language pairs grows. *Left*: partitions where sentences of both source and target language are seen when training. *Right*: partitions where source and/or target language sentences are unseen when training.

BLEU compared to 8-shot ICL. By inspecting the translation results, we find that mFTI-16 suffers from severe off-target problems, i.e., generating translations in wrong target languages. We hypothesize that this could be attributed to the shift in the relative positions of the source and target languages during training.

Seeing Target Languages When Finetuning is Better Than Source Languages. When there are unseen languages in the language direction, the improvement on Unseen Src is much larger compared to Unseen Tgt, indicating the understanding of the specified target language may be more important than the source language.

Unseen Both Sides Also Benefit From mFTI Training. The most surprising phenomenon is that language pairs from Unseen Both Sides partition also benefit from mFTI, with an improvement of 0.7 BLEU compared to 8-shot ICL. Since mFTI-16 does not see any sentences of the source and target languages, the improvements indicate a better understanding of the translation instruction, which we will discuss in Section 4.4.

4.3 Instruction Tuning with More Language Pairs Leads to Better Translation Performance

Previous instruction-tuning works show that scaling the number of tasks significantly benefits the unseen tasks (Chung et al., 2022). Observing the performance gap of Same Direction between mFTI-16 and mFTI-all, we gradually add more language pairs to mFTI-16, and plot the translation performance on each partition in Figure 5. In order to isolate possible effects of additional monolingual sentences, we only add language pairs that exclude the studied 13 languages.[8]

It can be seen that as the number of language pairs grows, the translation performance of all partitions generally increase, validating the importance of more language pairs. Notably, the performance of the Reversed Direction partition is significantly boosted, outperforming 8-shot ICL by a large margin when increasing the number of language pairs from 16 to 30.

Surprisingly, the performance of the Unseen Both Sides partition improves the most. Since no data of language pairs in Unseen Both Sides are added, this indicates the ability of instruction-following on these language pairs has been significantly enhanced, which we will discuss in the next section.

[8] Detailed language pairs are in Appendix D.

4.4 mFTI Generalizes the Understanding of Translation Instruction to Unseen Directions

In this section, we aim to understand how mFTI facilitates the understanding of instructions from a more fine-grained view, i.e., specific *language directions* and *instruction-following errors*.

For the language directions, we select Ru→Fr (high-resource), Bg→Ar (medium-resource), and Ca→Ta (low-resource) from the Unseen-Both Sides partition to study mFTI's effectiveness under different resource settings.

For instruction errors, we identify the following four major problems in translations:

- *Source Copy* (**SC**): This error occurs when the model simply copies the source sentence as the translation without making any meaningful changes. We identify this error by calculating the sentence-level BLEU score between the translations and the source sentences. If the BLEU score is above 80, it indicates that the translation is nearly identical to the source.
- *Off-target translation* (**OT**): In this case, the model fails to produce sentences in the target language. We detect this error by using a language identification tool, such as *fasttext*, to determine the language of the generated translations.
- *Over/under translation* (**OU**): This error refers to situations where the model produces translations that are significantly longer or shorter than references. We consider translations with a length ratio above 2 or below 0.5 as over- or under-translations, respectively.
- *Oscillatory hallucination* (**OH**): This error occurs when the model gets stuck in a specific translation state and generates repeated n-grams until reaching the maximum length. We define translations with n-grams that consecutively repeat at least three times as oscillatory hallucinations.

4.4.1 Adding Irrelevant Language Pairs Reduces SC, OT and OU Ratios

In Section 4.3, we show that additional language pairs in mFTI lead to improved BLEU scores even for the Unseen Both Sides partition. We provide an in-depth analysis here from the aforementioned fine-grained views. We plot the trends of translation and instruction-following performance, and the ratios of 4 specific instruction-following errors as the number of additional language pairs grows. The results are in Figure 6.

More Language Pairs Reduce Instruction-Following Errors and Improve Translation Performance. Firstly, we can see that as more language pairs are added to the training corpus, instruction-following errors on Unseen-both language pairs are gradually reduced, leading to improvements in BLEU scores. Comparing different language pairs, we can see that high- and medium-resource language pairs generally perform better than low-resource language pairs on all four types of errors. Since all these language directions are unseen when instruction finetuning, it highlights the importance of language skills acquired during the pretraining phase.

SC: Solved. It can be observed that after adding about 30-60 language pairs, the model learns to avoid the SC problem, indicating this is a relatively easy problem to solve.

OU: Decreased to the Level of mFTI-all. We can further see that adding more language pairs is also effective for reducing OU errors, as the error ratios significantly decrease as the number of language pairs grows. Notably, after scaling the number of language pairs to 150, the OU ratios of three unseen language pairs are comparable to supervised full finetuning. This demonstrates the effectiveness of mFTI.

OT: Decreased, but not to a Satisfactory Level. Turning to the OT ratio, we observe that it also decreases as the number of language pairs grows. However, even after scaling the number of language pairs to 150, the OT ratio still cannot be decreased to the level of mFTI-all.

OH: No Effect. Finally, we can see that with the increment in the number of language pairs, the OH ratio does not show a clear decreasing trend, which we will further discuss in the next section.

4.4.2 Joint Training with Monolingual Generation Instructions Helps Reduce OH and OT Problems More Efficiently

In the previous section, we find that the off-target (OT) and oscillatory hallucination (OH) on some

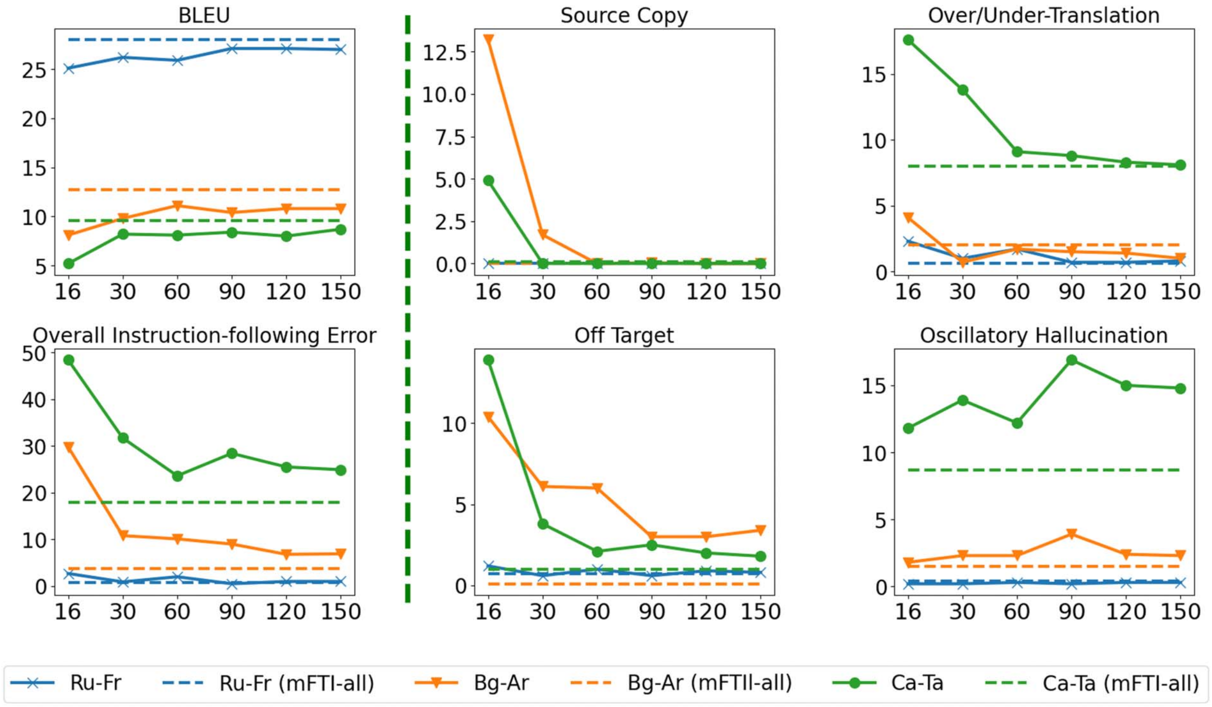

Figure 6: Trends of translation and instruction-following performance on 3 Unseen-both language pairs when scaling up the number of language pairs during mFTI. The left 2 figures show the BLEU score and overall instruction-following error ratios, respectively. The remaining 4 figures show the ratios of 4 specific error types, respectively, i.e., source copy, off-target, over/under translation, and oscillatory hallucination. The x-axis denotes the number of training language pairs. The y-axis denotes the percentage of translations with specific error types.

	Ru→Fr			**Bg→Ar**			**Ca→Ta**		
	OT⇓	**OH⇓**	**BLEU⇑**	**OT⇓**	**OH⇓**	**BLEU⇑**	**OT⇓**	**OH⇓**	**BLEU⇑**
mFTI-16	1.2	0.2	25.1	10.4	1.8	8.1	13.9	11.8	5.2
+ *unseen-mono*	0.9	0.1	25.4	2.6	0.9	10.4	4.4	6.3	6.3
mFTI-150	0.8	0.3	27.0	3.4	2.3	10.8	1.8	14.8	8.7
+ *unseen-mono*	0.7	0.2	27.4	0.5	1.5	12.0	1.2	5.1	9.3
mFTI-all	0.7	0.2	28.0	0.1	1.5	12.7	1.0	8.7	9.6

Table 5: BLEU score, off-target ratio and oscillatory hallucination ratio before and after adding monolingual sentences to the finetuning corpus. Scores where adding monolingual sentences leads to improved quality are with green background.

language pairs cannot be fully solved to the level of mFTI-all by adding more irrelevant language pairs. We note that both problems are only related to the target language: The OT problem can be attributed to models' inability to relate target language names to the corresponding scripts of the language, and the OH problem might be caused by the poor modeling of the target languages. We hypothesize that finetuning models on instructions of monolingual generation, i.e., given a language name, generates fluent sentences from that language, and should help ease these problems.

To this end, we organize the monolingual sentences of the held-out languages into monolingual generation instructions. The template we adopt is "$[l_i]$: $\mathbf{y}$". We then finetune XGLM on the dataset composed of translation instructions and these monolingual generation instructions.

We report the BLEU score, OT ratio, and OH ratio, in Table 5. Firstly we can see that adding monolingual generation instructions for the three Unseen Both Side language pairs can help mitigate the OT and OH problem in most scenarios, leading to better translation performance. Notably, by combining more irrelevant language pairs and monolingual sentences, the gap between mFTI-150 with monolingual sentences and mFTI-all has significantly diminished, despite that the model has never seen parallel sentences of the tested language before.

4.5 mFTI Improves Language Alignment via Pivot Languages

Besides the understanding of translation instruction, another crucial knowledge that models must grasp to carry out the instruction is the alignment between source and target languages. However, in scenarios where direct parallel sentences are not available, models have limited access to alignment information. This situation resembles the zero-shot setting commonly studied in multilingual translation research (Gu et al., 2019; Zhang et al., 2020; Arivazhagan et al., 2019; Liu et al., 2021). In this section, we aim to investigate the ability of mFTI to establish meaningful alignments through pivot languages in this scenario.

Specifically, for the three Unseen Both Sides language pairs X→Y studied in the previous section, i.e., Ru→Fr, Bg→Ar and Ca→Ta, we start from the mFTI-150 setting, and add parallel sentences of X→En and En→Y to the training corpus. We then perform mFTI using these augmented corpora and evaluate the model's performance on test sentences that do not contain instruction-following errors. As knowledge of language alignments is the last requirement for carrying out translation instructions once the model has learned to execute translation instructions correctly, the performance on these sentences serves as a reliable indicator of the model's proficiency in language alignment.

The result is in Table 6. First, we can see that mFTI-150 and 8-shot ICL perform comparably, both significantly worse than mFTI-all. Since the tested three language pairs are unseen in mFTI-150, this indicates that similar to mFTI-150, the main role of ICL is to enhance the model's understanding of the translation behavior instead of source-targ et alignment knowledge.

However, after adding pivot parallel sentences, the model's performance (+*pivot*) is significantly boosted. This demonstrates the potential of mFTI to leverage pivot languages to boost direct alignment between languages and improve translation performance.

	Ru→Fr	**Bg→Ar**	**Ca→Ta**
8-shot ICL	27.0	11.6	9.7
mFTI-all	**28.2**	**13.2**	<u>10.6</u>
mFTI-150	27.5	11.6	9.2
+ *pivot*	<u>27.9</u>	<u>13.0</u>	**10.8**

Table 6: Translation performance on test sentences without instruction-following errors. Best performances are in **bold**. The second-best performances are underlined.

5 Related Work

5.1 LLMs for MT

Machine translation researchers have widely recognized the potential of utilizing LLMs for MT, as these models acquire advanced language understanding skills during pretraining. The prevailing paradigm for leveraging LLMs for MT is in-context learning (ICL). For instance, Lin et al. (2022) demonstrated that providing 32 examples during translation can outperform GPT-3 and a supervised multilingual translation model. Other studies such as Vilar et al. (2023), Agrawal et al. (2023), and Zhu et al. (2023) have investigated different factors that affect ICL's performance, including example quality, example selection strategy, and template sensitivity. Moreover, works such as Hendy et al. (2023) and Jiao et al. (2023b) have studied the translation quality of various GPT-3 models and found their performances to be comparable to commercial translation systems on high-resource language pairs. In contrast to these works, our research focuses on exploring existing LLMs' translation ability by directly tuning them to follow translation instructions.

The most similar work to ours is Jiao et al. (2023a), which finetunes an open-source LLM LLaMA (Touvron et al., 2023) on the mixes translation data and the *alpaca* instruction dataset (Taori et al., 2023) to make it a better translator. However, they mainly focus on the bilingual translation setting while our work investigates the multilingual generalization when finetuning LLMs to carry out translation instructions.

5.2 Generalization On Unseen Language Pairs

Our work also has a close relation to zero-shot translation in the multilingual translation setting, where there are no direct parallel sentences between the source and target language. There are two major problems for zero-shot translation: generating correct languages and learning universal language representations.

For the first problem, Gu et al. (2019) and Zhang et al. (2020) leverage back-translation to add more target-language-related training data. Arivazhagan et al. (2019) and Liu et al. (2021) impose regularization on the encoder/decoder to make the model more aware of the target language. Unlike their work, we discuss the off-target problem in the context of LLMs, and find that adding both irrelevant language pairs and additional monolingual sentences can ease the problem to a great extent.

For the second problem, previous studies focus on learning language-agnostic representations through additional regularization of model representations (Arivazhagan et al., 2019; Pan et al., 2021), and consistency between semantic equivalent sentences (Al-Shedivat and Parikh, 2019; Yang et al., 2021). Instead, our work mainly aims to reveal the helpfulness of multilingual finetuning LLMs for unseen language pairs by internalizing the pivot language information.

Furthermore, our discussion encompasses a more stringent version of zero-shot translation, where neither source nor target language sentences are present in the finetuning corpus. This demands a stronger generalization ability, as the model must effectively utilize the language knowledge acquired during pretraining and the translation task knowledge acquired during finetuning to generate high-quality translations.

5.3 Instruction Finetuning

Our work focuses on finetuning LLMs with instructions to improve zero-shot translation performance. Prior studies have demonstrated that LLMs face great difficulty in achieving good performance in zero-shot settings when lacking few-shot examples. Nevertheless, finetuning LLMs on a variety of tasks can significantly improve zero-shot performance on several tasks. For instance, Wei et al. (2022) aims to improve generalization in unseen tasks by performing instruction tuning. Muennighoff et al. (2023) further extend to finetune LLM by multilingual data instead of English data and find that multilingual finetuning leads to better performance on unseen tasks and unseen languages. Chung et al. (2022) explore instruction tuning from the perspective of the number of tasks in finetuning corpus and LLM size. Chung et al. (2022) found that scaling these factors can dramatically improve zero-shot performance.

In our work, we primarily focus on the translation performance of LLMs. We adopt a comprehensive approach to consider the factors mentioned above, including the scale of the finetuning corpus, the size of model parameters, and the language selection within the fine-tuning corpus, for a comprehensive analysis of the translation performance of the LLMs. Additionally, we conduct a detailed analysis of the model's understanding and execution capabilities in translation tasks after instruction finetuning.

6 Conclusion

In this paper, we explore Multilingual Finetuning with Translation Instructions (mFTI), to better unleash the translation ability of multilingual LLMs. Through extensive experiments, we demonstrate that by training on a mixture of 1000 sentences per language pair, mFTI achieves better performance than 8-shot ICL, indicating the untapped potential of translation ability in LLMsv by previous works.

Moreover, we systematically discuss the working mechanism of mFTI by analyzing it from the view of instruction-following. Our experiments demonstrate that mFTI helps the model better follow the instruction by introducing more language pairs and monolingual sentences, and enhances the direct language alignment by learning from pivot language pairs.

Our paper also unveils remaining translation issues when adopting LLMs for zero-shot machine translation, i.e., over/under translation, oscillatory hallucination, and mistranslation caused by incorrect alignments. Future work should focus on acquiring more language knowledge from the pretraining phase and designing better regularization terms to solve these problems.

Acknowledgments

We would like to thank the anonymous reviewers and the editor for their insightful comments. This work is supported by National Science Foundation of China (No. 62376116, 62176120), the Liaoning Provincial Research Foundation for Basic Research (No. 2022-KF-26-02).

References

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2023. In-context examples selection for machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8857–8873, Toronto, Canada. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2023.findings-acl.564`

Maruan Al-Shedivat and Ankur Parikh. 2019. Consistency by agreement in zero-shot neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1184–1197, Minneapolis, Minnesota. Association for Computational Linguistics. `https://doi.org/10.18653/v1/N19-1121`

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Roee Aharoni, Melvin Johnson, and Wolfgang Macherey. 2019. The missing ingredient in zero-shot neural machine translation. *CoRR*, cs.CL/1903.07091v1.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskeve, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, 2005.14165.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *CoRR*, cs.LG/2210.11416v5.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *CoRR*, cs.CL/2207.04672v3.

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzman, and Philipp Koehn. 2020. CCAligned: A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*, pages 5960–5969, Online. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2020.emnlp-main.480`

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond English-centric multilingual machine translation.

Xavier Garcia, Yamini Bansal, Colin Cherry, George Foster, Maxim Krikun, Fangxiaoyu

Feng, Melvin Johnson, and Orhan Firat. 2023. The unreasonable effectiveness of few-shot learning for machine translation. *CoRR*, cs.CL/2302.01398v1.

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. https://doi.org/10.1162/tacl_a_00474

Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor O. K. Li. 2019. Improved zero-shot neural machine translation via ignoring spurious correlations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1258–1268, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1121

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. *CoRR*, cs.CL/2302.09210v1.

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*.

Wenxiang Jiao, Jen tse Huang, Wenxuan Wang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023b. ParroT: Translating during chat using large language models. *CoRR*, cs.CL/2304.02426v4.

Wenxiang Jiao, Wenxuan Wang, Jen tse Huang, Xing Wang, and Zhaopeng Tu. 2023b. Is ChatGPT a good translator? Yes with GPT-4 as the engine. *CoRR*, cs.CL/2301.08745v3.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *CoRR*, cs.LG/2001.08361v1.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, cs.LG/1412.6980v9.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.616

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14, Valencia, Spain. Association for Computational Linguistics. https://doi.org/10.18653/v1/E17-2002

Hui Liu, Danqing Zhang, Bing Yin, and Xiaodan Zhu. 2021. Improving pretrained models for zero-shot multi-label text classification through reinforced label hierarchy reasoning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1051–1062, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.83

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M. Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel.

2023. Crosslingual generalization through multitask finetuning. https://doi.org/10.18653/v1/2023.acl-long.891

OpenAI. 2023. Chatgpt (mar 23 version) [large language model].

Lin Pan, Chung-Wei Hang, Haode Qi, Abhishek Shah, Saloni Potdar, and Mo Yu. 2021. Multilingual BERT post-pretraining alignment. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 210–219, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.20

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. https://doi.org/10.18653/v1/W18-6319

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.

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:2211.05100v4*.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. 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. https://doi.org/10.18653/v1/2021.eacl-main.115

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford Alpaca: An instruction-following LLaMA model. https://github.com/tatsu-lab/stanford_alpaca.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothee Lacroix, Baptiste Roziere, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and efficient foundation language models. *CoRR*, cs.CL/2302.13971v1.

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. https://doi.org/10.18653/v1/2023.acl-long.859

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Jian Yang, Yuwei Yin, Shuming Ma, Haoyang Huang, Dongdong Zhang, Zhoujun Li, and Furu Wei. 2021. Multilingual agreement for multilingual neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 233–239, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-short.31

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study. *CoRR*, cs.CL/2301.07069v1.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.148

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. LIMA: Less is more for alignment. *CoRR*, cs.CL/2305.11206v1.

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. *CoRR*, cs.CL/2304.04675v2.

Appendix

A Full Results

Table 8 shows all 156 language pair results of 8-shot ICL and mFTI on XGLM, evaluated by both BLEU and COMET.

B Training and Evaluation Details of mFTI-156 in Different LLMs

The distribution of pretraining corpus varies across different LLMs, hence we adopt diverse hyperparameters in Table 7. Moreover, we conduct evaluations on LLMs at regular steps (250 for Full-finetuning and 500 for LoRA) during the training phase, selecting the best-performing result in the end.

C Comparison of mFTI and Supervised Machine Translation Models Evaluated by COMET

We present the comparison of mFTI and supervised MT models evaluated by COMET in Figure 7. It can be seen that when evaluated by COMET, mFTI's performance is comparable to M2M-1.2B, yet still substantially underperforms NLLB-3B.

D Additional Language Pairs for mFTI

We construct the additional language pairs in Section 4.3 from the other 17 languages covered in the pretraining corpus of XGLM, including Spanish, Greek, Portuguese, Japanese, Vietnamese, Urdu, Thai, Turkish, Telugu, Italian, Haitian, Creole, Basque, Indonesian, Estonian, and Bangali.

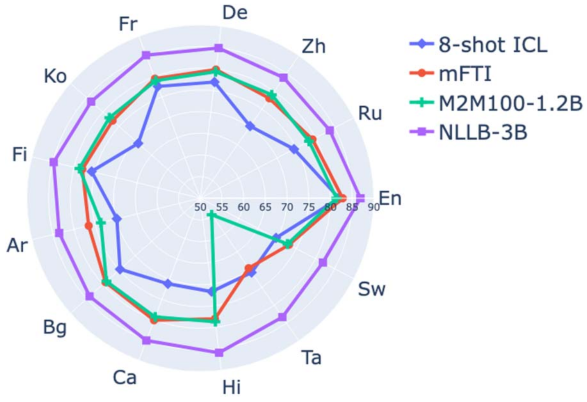

Figure 7: Comparison of mFTI with conventional supervised machine translation models. Performances are evaluated in COMET.

			BLOOM-7B and LLaMA-7B	XGLM-7.5B
Method	Hyperparameter		Value	Value
LoRA	LoRA	modules	query, key, value	query, key, value
		rank	4	4
		scaling factor	32	32
		dropout	0.1	0.1
	learning rate		5e-5	5e-4
	batch size		80	80
	training steps		5000	5000
	evaluation frequency		500	500
Full-finetuning	learning rate		1e-5	5e-6
	batch size		80	80
	training steps		2500	2000
	evaluation frequency		250	250

Table 7: Hyperparameter configurations of LoRA and Full-fine-tuning in LLMs.

			en	de	fr	ca	fi	ru	bg	zh	ko	ar	sw	hi	ta	avg
en	bleu	8-shot ICL	–	29.4	38.5	37.2	25.1	25.6	34.8	**17.5**	15.6	15.7	**22.2**	19.9	**10.1**	24.3
		mFTI	–	**30.4**	**39.7**	**37.5**	**25.5**	**26.3**	**35.2**	16.6	**16.2**	**16.6**	22.1	**21.7**	9.1	**24.7**
	comet	8-shot ICL	–	82.7	83.2	84.5	89.8	85.4	87.7	**81.4**	**82.7**	80.2	79.1	70.6	**77.1**	82.0
		mFTI	–	**84.3**	**85.1**	**85.2**	**90.8**	**85.8**	**88.4**	81.1	81.8	**81.8**	**80.8**	**73.0**	75.2	**82.8**
de	bleu	8-shot ICL	38.0	–	28.8	17.6	**20.9**	19.7	25.3	12.4	7.7	8.9	**14.8**	**16.4**	**8.9**	18.3
		mFTI	**41.3**	–	**31.0**	**27.7**	20.2	**21.0**	**26.6**	**12.9**	**14.3**	**16.9**	11.7	16.0	5.2	**20.4**
	comet	8-shot ICL	86.3	–	80.0	77.7	86.7	82.6	83.9	77.8	69.4	69.3	73.7	65.0	**71.6**	77.0
		mFTI	**90.1**	–	**85.3**	**82.4**	**88.5**	**87.2**	**86.1**	**78.4**	**77.8**	**78.6**	**74.7**	**66.1**	63.7	**79.9**
fr	bleu	few-shot	39.3	15.8	–	29.0	**19.5**	20.5	27.2	12.3	6.5	10.7	**18.1**	**14.3**	**8.9**	18.5
		mFTI	**41.7**	**23.8**	–	**33.9**	18.6	**22.0**	**27.5**	**16.7**	**11.6**	**13.7**	15.0	14.0	6.6	**20.4**
	comet	8-shot ICL	86.1	75.4	–	81.7	**86.5**	83.4	86.3	77.0	68.3	73.3	**76.6**	63.5	**73.5**	77.6
		mFTI	**86.7**	**84.8**	–	**86.2**	85.5	**85.1**	**88.6**	**78.1**	**77.7**	**78.7**	74.7	**64.1**	63.9	**79.5**
ca	bleu	8-shot ICL	41.0	19.5	33.5	–	9.2	9.0	23.5	11.1	4.2	1.8	13.7	10.8	8.4	15.5
		mFTI	**42.6**	**22.4**	**35.6**	–	**18.2**	**21.4**	**27.2**	**12.6**	**11.4**	**13.5**	**16.3**	**14.2**	**9.6**	**20.4**
	comet	8-shot ICL	86.4	76.3	82.2	–	70.1	61.6	81.4	77.5	60.5	57.2	69.8	59.2	**71.4**	71.1
		mFTI	**87.5**	**81.4**	**85.7**	–	**88.7**	**85.3**	**86.3**	**79.7**	**77.3**	**78.6**	**77.6**	**66.0**	66.9	**80.1**
fi	bleu	8-shot ICL	29.0	**17.9**	23.9	18.4	–	15.9	12.7	**12.2**	10.0	8.9	9.2	10.3	3.9	14.4
		mFTI	**30.6**	17.6	**24.4**	**23.2**	–	**17.4**	**20.7**	12.0	**11.7**	**9.7**	**14.9**	**12.9**	**4.1**	**16.6**
	comet	8-shot ICL	87.1	**80.8**	81.2	80.4	–	82.2	74.7	**79.1**	76.6	74.8	70.9	60.5	61.7	75.8
		mFTI	**88.0**	79.4	**82.0**	**81.7**	–	**85.2**	**84.3**	78.3	**76.7**	**76.0**	**73.4**	**66.6**	**64.8**	**78.0**
ru	bleu	8-shot ICL	30.9	19.3	25.7	15.6	4.2	–	25.5	12.0	7.7	7.9	7.8	12.2	**6.8**	14.6
		mFTI	**32.5**	**20.4**	**28.0**	**25.7**	**17.6**	–	**29.5**	**12.4**	**10.1**	**14.9**	**12.2**	**13.2**	6.2	**18.6**
	comet	8-shot ICL	83.7	78.5	79.1	76.9	72.6	–	88.4	**77.1**	72.1	70.4	64.8	62.0	**66.7**	74.4
		mFTI	**86.2**	**80.6**	**84.1**	**83.0**	**88.1**	–	**91.7**	76.3	**76.4**	**78.4**	**75.6**	**67.0**	63.4	**79.2**
bg	bleu	8-shot ICL	35.7	20.4	26.6	14.5	14.5	22.8	–	12.5	8.9	10.7	10.0	6.8	1.6	15.4
		mFTI	**37.6**	**21.2**	**30.2**	**28.1**	**17.1**	**24.7**	–	12.5	**10.2**	**12.7**	**12.9**	**14.4**	**4.6**	**19.2**
	comet	8-shot ICL	85.8	78.7	79.8	78.2	81.0	86.8	–	77.1	74.4	74.8	69.5	53.6	57.1	74.7
		mFTI	**88.0**	**83.4**	**83.0**	**84.8**	**87.3**	**88.9**	–	**77.4**	**74.5**	**78.4**	**74.8**	**66.6**	**63.5**	**79.2**
zh	bleu	8-shot ICL	21.9	6.5	10.4	6.4	3.6	1.9	14.7	–	9.6	3.0	7.6	**13.3**	**9.6**	9.0
		mFTI	**23.4**	**11.5**	**18.7**	**25.2**	**12.3**	**12.8**	**15.4**	–	**12.4**	**8.5**	**9.5**	12.9	4.6	**13.9**
	comet	8-shot ICL	82.5	66.4	70.9	70.1	65.4	56.7	81.0	–	77.7	62.1	69.3	64.6	**75.6**	70.2
		mFTI	**86.1**	**76.4**	**78.9**	**80.4**	**84.7**	**83.7**	**84.3**	–	**79.9**	**76.2**	**73.0**	**67.3**	64.8	**78.0**
ko	bleu	8-shot ICL	21.1	**11.2**	15.0	7.3	2.8	3.1	9.9	11.9	–	3.0	2.4	0.7	0.4	7.4
		mFTI	**22.9**	10.3	**16.0**	**15.5**	**10.8**	**10.4**	**13.4**	**12.1**	–	**9.8**	**9.1**	**13.2**	**6.3**	**12.5**
	comet	8-shot ICL	83.6	74.3	**75.5**	73.0	66.7	58.4	74.5	78.8	–	62.4	58.8	56.4	66.6	69.1
		mFTI	**86.2**	**75.9**	74.8	**78.6**	**84.6**	**82.1**	**82.1**	**79.4**	–	**72.9**	**73.8**	**64.3**	**69.8**	**77.0**
ar	bleu	8-shot ICL	28.2	12.7	21.3	8.8	4.3	10.0	19.0	**9.4**	1.7	–	4.3	10.4	**7.2**	11.4
		mFTI	**30.6**	**13.7**	**23.2**	**24.8**	**14.0**	**14.9**	**21.9**	9.3	**8.3**	–	**11.9**	**11.4**	4.1	**15.7**
	comet	8-shot ICL	82.5	72.5	76.0	73.4	65.8	69.6	80.0	**74.4**	56.2	–	61.0	58.1	**67.9**	69.8
		mFTI	**86.6**	**75.4**	**79.4**	**82.9**	**84.0**	**83.9**	**84.5**	73.1	**71.6**	–	**72.4**	**61.6**	62.5	**76.5**
sw	bleu	8-shot ICL	**32.2**	14.2	**23.2**	17.5	**10.5**	**13.2**	12.4	**9.5**	7.1	9.7	–	8.5	2.7	13.4
		mFTI	32.1	**14.9**	21.0	**20.4**	10.4	12.5	**16.3**	8.6	**9.0**	10.4	–	**12.7**	**4.8**	**14.4**
	comet	8-shot ICL	80.9	**70.9**	**73.2**	73.6	73.4	**76.6**	68.0	**72.2**	70.2	71.3	–	54.2	51.6	69.7
		mFTI	**82.7**	70.0	72.6	**75.8**	**76.6**	76.2	**81.0**	70.0	**72.5**	**74.6**	–	**62.1**	**63.0**	**73.1**
hi	bleu	8-shot ICL	23.7	12.8	15.6	8.4	9.5	11.3	9.9	**11.3**	11.0	6.0	5.0	–	0.2	10.4
		mFTI	**28.0**	12.8	**17.8**	**16.9**	**12.1**	**12.0**	**16.3**	10.8	**12.2**	**8.9**	**10.9**	–	**10.1**	**14.1**
	comet	8-shot ICL	82.5	74.6	74.9	74.1	78.1	78.5	66.4	**76.9**	76.7	68.7	62.8	–	45.7	71.7
		mFTI	**86.1**	**76.2**	**75.3**	**79.9**	**82.2**	**81.3**	**81.1**	75.9	**77.9**	**75.0**	**73.8**	–	**71.7**	**78.0**
ta	bleu	8-shot ICL	**16.1**	**8.8**	**11.2**	5.7	**6.9**	**9.0**	**8.9**	**8.3**	**8.7**	**5.5**	3.1	6.6	–	8.2
		mFTI	16.0	7.7	9.5	**10.4**	5.9	6.3	8.1	6.6	7.7	5.4	**6.4**	**13.3**	–	**8.6**
	comet	8-shot ICL	78.5	**70.7**	**71.3**	**70.6**	74.2	**77.0**	72.3	**73.2**	**75.6**	**71.1**	61.4	53.3	–	**70.8**
		mFTI	**78.6**	64.3	66.0	70.1	**74.4**	71.6	**72.9**	68.4	69.0	69.1	**67.1**	**63.4**	–	69.6
avg	bleu	8-shot ICL	29.8	15.7	22.8	15.5	10.9	13.5	18.7	11.7	8.2	7.7	9.9	10.9	5.7	13.9
		mFTI	**31.6**	**17.2**	**24.6**	**24.1**	**15.2**	**16.8**	**21.5**	**11.9**	**11.3**	**12.1**	**12.7**	**14.1**	**6.3**	**16.9**
	comet	8-shot ICL	83.8	75.2	77.3	76.2	75.9	74.9	78.7	**76.9**	71.7	69.6	68.1	60.1	65.5	73.4
		mFTI	**86.1**	**77.7**	**79.4**	**80.9**	**84.6**	**83.0**	**84.3**	76.3	**76.1**	**76.5**	**74.3**	**65.7**	**66.1**	**77.7**

Table 8: Translation performance of 8-shot ICL and mFTI based on XGLM-7.5B on FLORES-101 (test). Source language in rows, target language in columns.