

Fine-Tuning Large Language Models to Translate: Will a Touch of Noisy Data in Misaligned Languages Suffice?

Dawei Zhu¹ Pinzhen Chen² Miaoran Zhang¹
Barry Haddow² Xiaoyu Shen^{3*} Dietrich Klakow¹

¹Saarland University, Saarland Informatics Campus ²University of Edinburgh

³Digital Twin Institute, Eastern Institute of Technology, Ningbo

{dzhu,mzhang}@lsv.uni-saarland.de pinzhen.chen@ed.ac.uk

Abstract

Traditionally, success in multilingual machine translation can be attributed to three key factors in training data: large volume, diverse translation directions, and high quality. In the current practice of fine-tuning large language models (LLMs) for translation, we revisit the importance of these factors. We find that LLMs display strong translation capability after being fine-tuned on as few as 32 parallel sentences and that fine-tuning on a single translation direction enables translation in multiple directions. However, the choice of direction is critical: fine-tuning LLMs with only English on the target side can lead to task misinterpretation, which hinders translation into non-English languages. Problems also arise when noisy synthetic data is placed on the target side, especially when the target language is well-represented in LLM pre-training. Yet interestingly, synthesized data in an under-represented language has a less pronounced effect. Our findings suggest that when adapting LLMs to translation, the requirement on data quantity can be eased but careful considerations are still crucial to prevent an LLM from exploiting unintended data biases.¹

1 Introduction

Large language models (LLMs) have reached new heights in various NLP tasks (Radford et al., 2019; Brown et al., 2020; Touvron et al., 2023; Jiang et al., 2023). Supervised fine-tuning (SFT, Ouyang et al., 2022, alternatively, instruction tuning or simply fine-tuning in some literature) further prepares these models for better generalization and reliability in downstream tasks by training on task input-output data combined with instructions in natural languages (Sanh et al., 2022; Wei et al., 2022; Mishra et al., 2022). In this research direction, various works have studied the “scaling up” of SFT

data size, number of languages, etc (Chung et al., 2024; Muennighoff et al., 2023). On the other hand, recent papers also embraced the philosophy of “less is more” by achieving strong results with a small set of high-quality training instances, claiming a “superficial alignment hypothesis” (Zhou et al., 2023) with similar findings by others.

This work investigates the role of SFT *data* in aligning LLMs to machine translation (MT), a cross-lingual generation task with high demands in practical domains. Prior research has found fine-tuning to improve translation performance (Zhang et al., 2023c) and more recent works also integrated continued pre-training with more data to provide further improvement (Xu et al., 2024a; Alves et al., 2024). For encoder-decoder models, Wu et al. (2024a) used little data to enable an English-centric model to translate between any two languages. Nonetheless, the feasibility of “less is more” in LLM translation fine-tuning is rather under-explored. In translation prompting, researchers have suggested that a model’s translation capability can be attributed to the bilingual signals exposed during pre-training (Briakou et al., 2023) and task recognition in LLM layers (Sia et al., 2024), hinting that the translation capability has been picked up during pre-training. A natural question follows: *Can we put reduced effort into data?*

From a data efficiency perspective, we squeeze the translation SFT data to a mere size of 32 or the translation direction to 1 for multilingual translation, for which we believe LLMs already possess a strong pre-trained foundation in multilingual understanding and generation. Beyond quantity and language diversity, we perform SFT on synthesized data via machine translation, which is a common data augmentation practice for under-served languages. To summarize, our analysis is grounded in the task of MT, with “scaling down” in mind. In multiple dimensions—data size (§3.2), translation direction (§3.3 and §3.4), and data synthesis

*Corresponding author (xyshen@eitech.edu.cn)

¹Code available at: github.com/uds-lsv/mt-sft.

(§3.5)—our findings verify, complement, and refine the existing superficial alignment hypothesis for fine-tuning LLMs for translation tasks:

1. 32 data instances successfully enable an LLM to translate in 11 directions. More data still helps but the return diminishes.
2. Data in a single translation direction can effectively align an LLM to translate to and from multiple directions. Yet, it is crucial to pick the right direction—we recommend not placing English on the target side.
3. When fine-tuning on lower-quality synthetic data, LLMs are affected if the data is placed on the target side, but they show greater resilience against such flaws in low-resource languages, which are less represented during pre-training.

2 Preliminaries

2.1 Supervised fine-tuning

In this work, we perform SFT to prepare pre-trained LLMs for MT. Let S denote a source input and $T = [t_1, t_2, \dots, t_{|T|}]$ denote a target-side reference. We start with placing the input into a prompt template by applying $\mathcal{I}(\cdot)$ to S . For each training instance, the instruction template is randomly selected from a pre-defined pool. We fine-tune an LLM parameterized by θ by optimizing the log-likelihood:

$$\begin{aligned} \mathcal{L}_{SFT}(\mathcal{I}(S), T; \theta) &= -\log P(T|\mathcal{I}(S); \theta) \\ &= -\log \prod_{k=1}^{|T|} P(t_k | t_{<k}, \mathcal{I}(S); \theta) \\ &= -\sum_{k=1}^{|T|} \log P(t_k | t_{<k}, \mathcal{I}(S); \theta) \end{aligned}$$

2.2 Superficial alignment hypothesis

Zhou et al. (2023) claim that a model’s knowledge and capabilities are acquired almost entirely during pre-training, and the effect of alignment tuning might be “superficial”, in that it teaches the model the format for interacting with users. This idea is further supported by recent works (Lin et al., 2024; Ghosh et al., 2024). However, to what extent this applies to multilingual translation in LLMs is little known. To bridge this gap, we conduct a series of controlled experiments on fine-tuning LLMs for translation, complementing previous research across three dimensions. First, we study the parallel

data efficiency in the era of LLMs, aiming to determine the minimum data needed for effective model alignment to the translation task. Next, we explore the scope of alignment by probing whether aligning one translation direction influences other directions. Finally, we investigate how synthesized fine-tuning data quality impacts the LLMs’ behaviour in generating translations.

3 Experiments and Results

3.1 Experimental setup

Training. By default, we take the test sets from WMT17 to WMT20 as our parallel training data (Bojar et al., 2017, 2018; Barrault et al., 2019, 2020); we also use the development sets in WMT21 (Akhbardeh et al., 2021) for training if a language pair of interest is not available in earlier years. The specific training data configurations will be detailed in the subsequent sections. The test sets from WMT21 are used for validation. Detailed data statistics can be found in Appendix F.1. The LLM we use for SFT is the base version of Llama-2 7B (Touvron et al., 2023). When performing SFT, we use a learning rate of 5e-6, an effective batch size of 64, and a linear learning rate scheduling with a warmup ratio of 0.1. We select the model checkpoint based on COMET scores on the validation sets.² To form the model input for SFT, we feed the source sentence into the Alpaca prompt template (Taori et al., 2023), supplementing it with a translation instruction that is randomly selected from a pool of 31 diverse instructions. Refer to Table 4 in the appendix for a complete list of templates.

Evaluation. We primarily evaluate the models on the WMT22 test sets (Kocmi et al., 2022) covering 11 translation directions: en↔cs, en↔de, en↔jp, en↔ru, en↔zh, and en→hr.³ Languages in these 11 directions are explicitly included in Llama-2’s pre-training corpus. In Section 3.4, we extend our evaluation to translation directions involving medium and low resource languages: Icelandic and Hausa (i.e., en↔is, en↔ha), which comes from WMT21’s test set. At inference time, a fixed translation instruction is applied (Table 4 row 1). We

²In our preliminary experiments, we found that validation perplexity has a relatively weak correlation with COMET scores measured on the validation set, similar to earlier findings (Ouyang et al., 2022).

³Language codes: cs=Czech, de=German, hr=Croatian, jp=Japanese, ru=Russian, zh=Chinese. “↔” means that both translation directions are covered. Note that only en→hr is available in WMT22 but not hr→en.

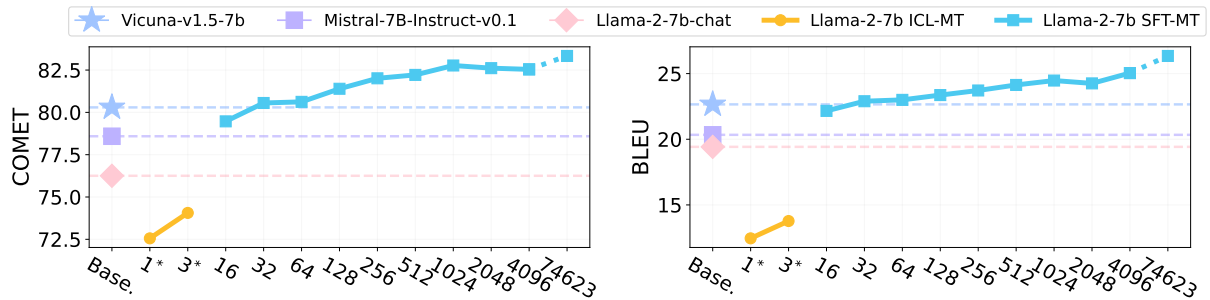


Figure 1: Performance comparison between instruction-tuned baselines and Llama-2 fine-tuned with different training data sizes. Average COMET (left) and BLEU (right) scores across 11 translation directions are presented. For training data sizes of 1 and 3, ICL is applied, marked with an asterisk “*”; otherwise, we perform SFT. With only 32 training examples for SFT, Llama-2 outperforms general-purpose, instruction-tuned baselines. Base.: instruction-tuned baseline models. See individual performance for the 11 translation directions in Appendix A.

use beam search with a beam size of 4 for generation, as our preliminary results indicate that it offers better translation quality than sampling-based generation, an observation consistent with recent works (Jiao et al., 2023; Zeng et al., 2024). The maximum generation length is set to 256 tokens. We used a reference-based COMET22 checkpoint⁴ (Rei et al., 2020) and BLEU (Papineni et al., 2002) as the evaluation metrics. See Appendix F.3 for detailed software configurations.

3.2 How much SFT data enables LLMs to translate?

Recent works in machine translation suggest that pre-trained LLMs require significantly less parallel data for fine-tuning (via SFT), compared to training conventional translation models from scratch. However, the SFT process in these works still operates with an order of 10^5 parallel samples (Jiao et al., 2023; Zhang et al., 2023c; Zeng et al., 2024; Xu et al., 2024a, i.a.), without a clear justification for selecting this specific data size and source. This raises a pivotal question, inspired by the recently proposed “superficial alignment hypothesis” (Zhou et al., 2023): Is SFT mainly a method for superficially aligning LLMs for translation tasks? If so, what is the actual minimal amount of data required to achieve effective “alignment”?

Setup. We fine-tune Llama-2 7B using different numbers of training samples and evaluate the multilingual translation performance of the resulting models. We collect training data covering 10 translation directions: $en \leftrightarrow \{cs, de, jp, ru, zh\}$. The training data sourced from WMT17-20 contains a

total of 74,623 parallel examples. Note that the training samples across translation directions are not evenly distributed. To create training sets of varying sizes, we subsample the original data into subsets that are powers of 2, starting from 16 (2^4) and ending with 4096 (2^{12}); larger subsets always contain smaller ones. To ensure balanced language representation in our subsets, we distribute samples as evenly as possible among the language pairs.⁵

We refer to the fine-tuned model as **SFT-MT**. Considering LLMs can also perform translation through prompting, we compare SFT-MT with 1- and 3-shot in-context learning (ICL), denoted as **ICL-MT**. For ICL, we randomly select demonstrations from the training set in the test direction for each test sentence. We do not consider Llama-2’s zero-shot performance because, although it sometimes produces acceptable translations in the beginning, it often continues generating, which makes it difficult to accurately estimate its performance. Lastly, since LLMs fine-tuned on diverse tasks also serve as strong translation systems (Zhu et al., 2024), we compare our models with open-source general-purpose instruction-tuned LLMs, which we denote as **IT-LLM**. These include Vicuna-v1.5-7b (Chiang et al., 2023), Mistral-7b-Instruct (Jiang et al., 2023), and Llama-2-7b-chat (Touvron et al., 2023).⁶

Results. Figure 1 illustrates the effect of varying training sizes on translation performance. In both 1- and 3-shot cases, ICL-MT underperforms IT-LLM baselines like Llama-2-7b-chat despite sharing the

⁴Specifically, COMET is reported on a scale of 0 to 100 as opposed to its raw 0 to 1 range.

⁵For example, the data size distribution for our 32-example training set is [4, 4, 3, 3, 3, 3, 3, 3, 3, 3].

⁶lmsys/vicuna-7b-v1.5, Mistral-7B-Instruct-v0.1, and meta-llama/Llama-2-7b-chat-hf.

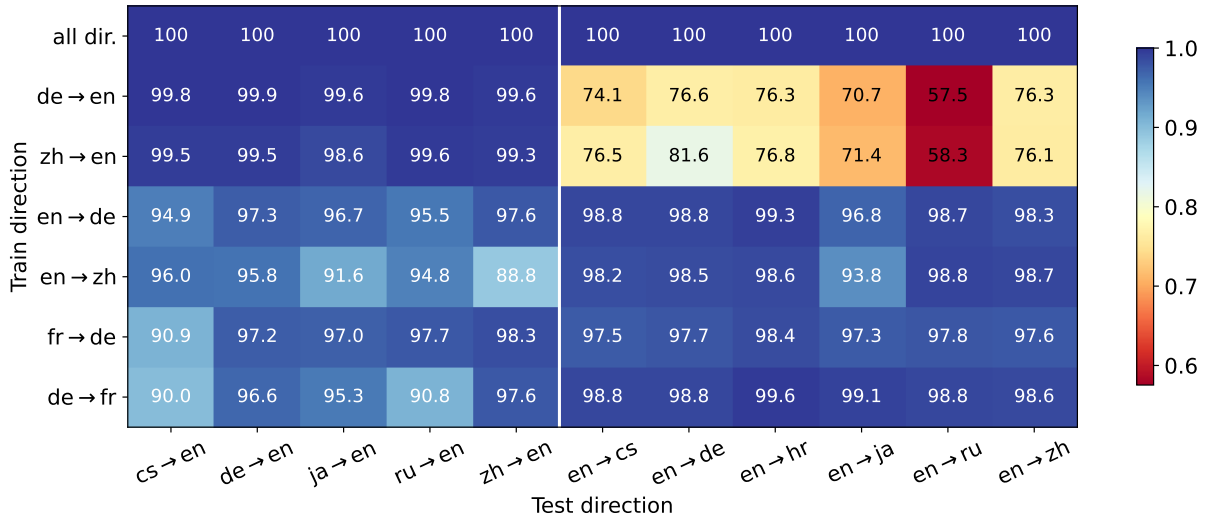


Figure 2: Normalized COMET score (as a % of performance from fine-tuning on an equivalent sized dataset of all 10 directions) resulted from varying combinations of train and test translation directions. In most cases, Llama-2 fine-tuned on a single translation direction can effectively translate across other directions, achieving performance comparable to models trained on all directions, with a few exceptions when trained on $X \rightarrow en$ but tested on $en \rightarrow X$. Performance measured in BLEU score is provided in Appendix B.

same foundation model, indicating that a few in-context demonstrations may not effectively align Llama-2 for translation.

However, performance significantly improves when Llama-2 is fine-tuned with just 16 samples. With further increases in the training size to 32 samples, Llama-2 performs on par with or surpasses all three IT-LLM baselines in both COMET and BLEU metrics. This suggests that a handful of high-quality parallel data can effectively specialize the model into a performant translation system. Increasing parallel data further boosts performance, though with diminishing returns: the COMET score rises by an average of 2 points when expanding from 32 to 1024 samples, but only by 0.5 points when increasing further from 1024 to 75K samples (full training set). Given that it is unlikely that these 32 training samples “teach” Llama-2 new translation skills, this shows strong evidence that superficial alignment applies to MT. We observe a similar trend in Mistral-7B and Llama-2-13B. Refer to Appendix A for their performance across varying data sizes. In summary, effective translation alignment begins with minimal training data, revealing **less is good alignment and more is better with diminishing gains**.

3.3 Do we need to include all directions?

In the preceding section, we follow the traditional practice in multilingual MT by including multiple

translation directions during training. However, the observation that only a few dozen examples make Llama-2 translate well leads us to reconsider the necessity of including samples from all directions of interest. Specifically, will training on just a single translation direction be sufficient to help LLMs perform multilingual translation?

Setup. We explore six training configurations, each focusing on a single translation direction: $de \rightarrow en$, $zh \rightarrow en$, $en \rightarrow de$, $en \rightarrow zh$, $fr \rightarrow de$, and $de \rightarrow fr$. These configurations include cases where English appears on the source side, the target side, as well as settings with English excluded, to investigate if specific languages have a different impact on the overall performance. The training size is set to 1024 for SFT. Evaluations are conducted across the same 11 test directions as used in the previous section. Additionally, we explore similar settings in ICL, where we present demonstrations with translation directions that do not match those used in evaluations, to determine if the mechanisms of both SFT and ICL exhibit similarities. Lastly, we conduct a joint evaluation, progressively expanding both the training size and the range of covered translation directions to understand the combined effect of these factors.

SFT results. Figure 2 demonstrates the normalized performance of Llama-2 when fine-tuned in various single directions. Remarkably, training

Evaluation on de→en				
demo lang	1-shot		3-shot	
	COMET	BLEU	COMET	BLEU
de→en	73.47	19.7	75.04	22.4
en→de	55.96	7.3	44.39	3.5
de→fr	66.35	12.1	64.61	17.6
fr→de	58.06	7.8	57.13	10.5
zh→en	56.66	10.7	54.82	7.1
en→zh	51.30	7.8	56.87	1.8

Evaluation on en→de				
demo lang	1-shot		3-shot	
	COMET	BLEU	COMET	BLEU
en→de	67.37	10.5	69.80	14.3
de→en	57.83	8.7	45.54	5.0
en→zh	59.76	9.5	59.53	8.4
zh→en	47.31	4.5	49.24	5.0
fr→de	59.36	8.6	66.01	12.9
de→fr	60.70	11.0	61.76	11.3

Table 1: ICL-MT performance with aligned vs. misaligned demonstrations, evaluated on de→en and en→de. 1-shot/3-shot: using 1 or 3 demonstrations randomly sampled from the training set. Misaligned demonstrations consistently cause a substantial performance drop.

with just one direction enables Llama-2 to translate between multiple languages. For instance, after fine-tuning on de→en or zh→en, the model can translate from all considered languages to English, scoring at least 98.6% of the original COMET scores for training on all directions. Similarly, the model fine-tuned on en→de, en→zh, fr→de or de→fr also demonstrates only a slight performance decline when translating from English.

Notable declines are observed in two scenarios: (1) trained to translate to English and evaluated on translating to non-English; and (2) trained to translate to non-English and evaluated on translating to English.⁷ Of these two scenarios, scenario 1 exhibits a much larger performance drop. The fact that both scenarios involve a mismatch between using English and non-English suggests that *Llama-2, as an English-centric LLM, may process English differently compared to other languages*. When fine-tuned for English generation, the model may misinterpret the task as only generating in English. Generalization among non-English languages is much easier than generalization between English and non-English languages, as evidenced by the negligible performance drop when fine-tuning and testing on two vastly different language pairs such as de→fr and en→zh. Overall, the findings suggest that **SFT in one translation direction effectively enables the many directions, though avoiding misinterpretation is crucial**.

ICL results. We also provide results of performing ICL with misaligned translation directions between demonstration and test in Table 1. It can be seen that misaligned demonstrations significantly degrade translation performance, with 3-shot be

⁷Analysis of model outputs reveals that they often merely echo the source sentence, ignoring the translation instruction.

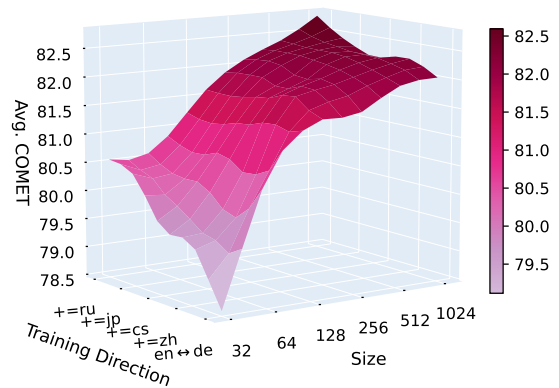


Figure 3: Average performance (in COMET) across 11 test directions for models trained with varying data sizes and directions. Both factors positively impact performance. +=: training directions added on top of previous directions; two directions are added at each time. For example, “+=ru” covers 10 directions: en ↔ {de, zh, cs, jp, ru}. Performance on individual test directions is provided in Appendix C.

often worse than 1-shot. We observe that the model may output Chinese characters, emojis, time, etc., but no clear error patterns are observed. This contrasts sharply with findings from SFT: **while SFT can recognize the format of translation, ICL requires language-aligned demonstrations**.

Joint evaluation. Figure 3 presents a joint evaluation of size and translation direction. For small training sizes, covering diverse translation directions in training proves to be beneficial. However, the benefits of such diversity level off as the training size increases. With a training size of 1024, models trained exclusively on two directions, en↔de, perform on par with those trained on all directions.

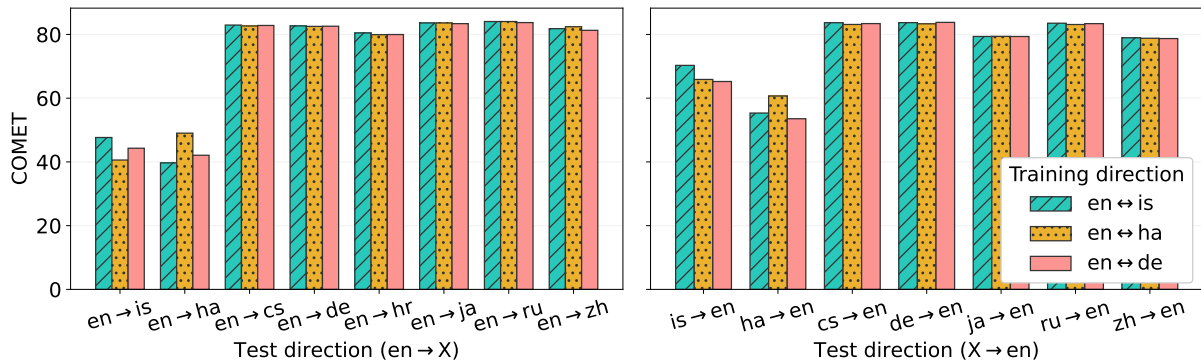


Figure 4: Model performance (in COMET) across 15 translation directions under different training configurations. Training models on *unseen* languages (en↔is, en↔ha) results in slight improvements in translating these languages compared to models trained on en↔de. The differences in performance when translating between *seen* languages are minimal across all training configurations. Performance measured in BLEU score is provided in Appendix D.

3.4 Can alignment be achieved for unseen languages?

Previous sections focus on translation directions involving languages explicitly included in Llama-2’s pre-training corpus. We now extend our investigation to languages that do not have an identified presence of over 0.005% in the pre-training data (c.f. [Touvron et al., 2023](#), p22), referred to as *unseen* languages. Here we seek answers to two questions: (1) Can we effectively make Llama-2 translate both from and to unseen languages by fine-tuning it with a small amount of data? (2) How well can this fine-tuned model translate from and to languages *seen* in Llama?

Setup. We consider three training configurations: en↔is, en↔ha, and en↔de, with Icelandic (is) and Hausa (ha) being unseen languages. en↔de serves as a control to assess Llama-2’s initial translation capabilities into unseen languages without specific fine-tuning. The training size is fixed at 1024 (512 samples for each direction). The test directions include the 11 directions as before, plus en↔is and en↔ha coming from the WMT21 test.

Results. The results are presented in Figure 4. It can be seen that fine-tuning on Icelandic and Hausa enhances a model’s translation quality on these languages compared to the control setup, yet the gains are modest. We observe that Llama-2 manages to produce tokens in these languages, however, the translations often largely deviate from the original meanings. This suggests that it is difficult to teach models new translation directions via SFT with limited data. Interestingly, we find fine-tuning on Icelandic or Hausa does not hinder Llama-2’s

ability to translate from and to all seen languages, maintaining performance levels comparable to the control scenario with en↔de. Based on these results, we propose a complement to the superficial alignment hypothesis in MT: **LLMs may learn the essence of the translation task without requiring input-output mappings in languages it “understands” well.**

3.5 Can we use synthesized data?

We have observed that LLMs quickly recognize the translation task with minimal high-quality, manually curated data, but what if the quality of the training data is subpar? This situation may occur, for example when parallel data is web-crawled or machine-generated. Can LLMs still adapt to the translation task or will they overfit to the imperfections in lower-quality data, leading to degraded translation performance?

Setup. We replace either the source or target sentences in the original training set with lower-quality synthesized ones. We try two types of data synthesis: one by translating entire sentences on the other side and another by concatenating word-to-word translations. Pleasingly, these correspond to back-translation ([Sennrich et al., 2016](#)) using translation engines or bilingual word dictionaries which are practical at different levels of resource availability. Specifically, we use the OPUS-MT suite ([Tiedemann and Thottingal, 2020](#)) to translate from English to a target non-English language.⁸

⁸E.g. for de→en, the process is run in en→de with the created data reversed, hence the translated content is on the source side. Checkpoints are available on Hugging Face: `Helsinki-NLP/opus-mt-en- $\{trg\}$` .

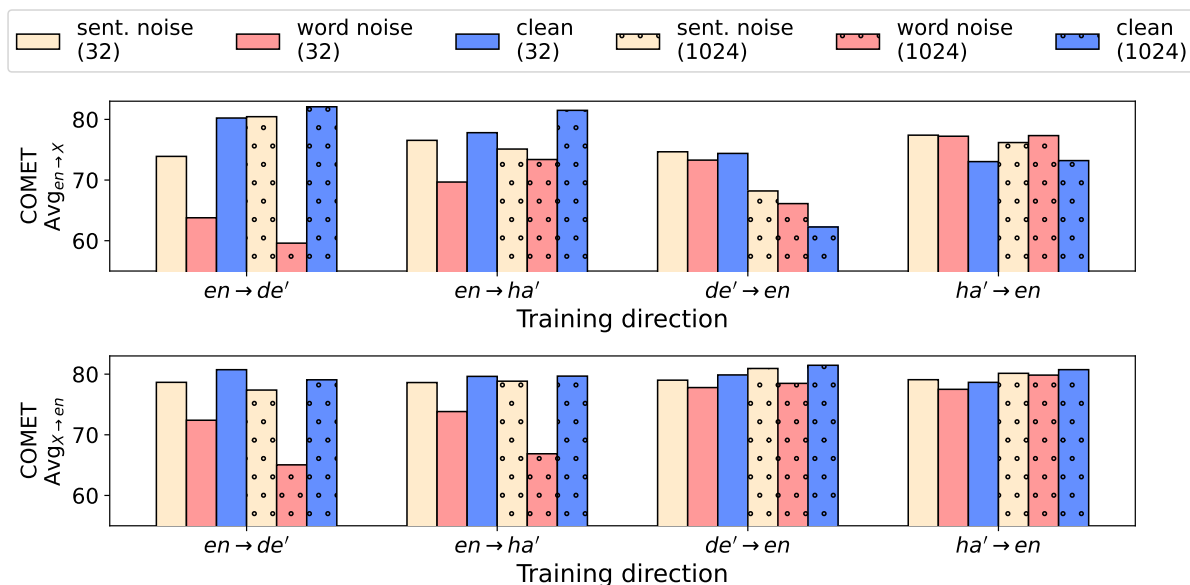


Figure 5: Model performance in COMET score varying training sizes, directions, and noise types. Top (Bottom): score averaged across all $en \rightarrow X$ ($X \rightarrow en$) test directions. Training sizes considered are 32 and 1024. Generally, introducing noise on the target side tends to degrade model performance more, with the extent of impact also depending on the particular language involved. Performance measured in BLEU score is provided in Appendix E.

Source	Ref./Data config.	Model output
Das finde ich ehrlich gesagt sehr ärgerlich.	reference literal $en \rightarrow de$ clean $en \rightarrow de$ sent. noise $en \rightarrow de$ word noise	That really bothers me, I must say. The find I honest said very annoying. I find that really annoying. I find that honestly very annoying. The find I honestly said very annoying.
以免再次发生这样的事情	reference literal $en \rightarrow de$ clean $en \rightarrow de$ sent. noise $en \rightarrow de$ word noise	So that such a thing won't happen again. in order to avoid again happen such thing. Let's not let it happen again. In order not to happen again. Avoid again happen this way.

Table 2: Examples of testing Llama-2 trained on $en \rightarrow de$ with 1024 clean and noisy target sentences. The test directions are $de \rightarrow en$ (Top) and $zh \rightarrow en$ (Bottom). The reference translation is provided by the WMT22 test set. Word-to-word references were created by the authors in consultation with native speakers. Word-level noise makes Llama-2 degenerate into a literal translator.

For word-level translation, we translate each space-delimited source word by feeding it into the MT model one at a time. Naturally, the synthesized versions introduce translation errors, adding “noise” to the training process. We investigate the impact of such noise in four translation directions: $en \rightarrow de'$, $de' \rightarrow en$, $en \rightarrow ha'$, and $ha' \rightarrow en$, where the prime (') notation denotes the side that is created using translation (noised). We consider two training sizes: 32 and 1024. In this section, our evaluation focuses on the 11 translation directions described in Section 3.1. Note that although Hausa is included in the current training setup, translation directions involving Hausa are excluded from our evaluation—because performance is sub-par for

unseen languages as demonstrated in Section 3.4.

Results. According to Figure 5, it can be seen that both types of data synthesis generally cause a drop in performance. However, The degree of degradation significantly varies depending on whether the noise appears on the source or target side of the translation as well as the language. Specifically, when noise is introduced to the target side, models fine-tuned on $en \rightarrow de'$ and $en \rightarrow ha'$ translations exhibit a sharp decline in performance. The impact of word noise is more severe than that of sentence noise. In the case of $en \rightarrow de'$, word-level synthesis causes the model to largely degenerate, leading to literal translations across many test

cases across translation directions. An example of this behaviour is presented in Table 2. In contrast, the performance drop caused by word noise is less pronounced with en→ha', particularly when evaluated on en→X.

Conversely, when noise is introduced on the source side, the negative impact is much smaller, and the disparity in performance degradation between the two types of noise diminishes. Even more strikingly, when evaluated on en→X, having noise at the source side often outperforms the clean settings. Notably, in Section 3.3, we show that fine-tuning models purely on X→en risks task misinterpretation, leading to low performance on en→X. However, adding noise appears to mitigate this issue, resulting in improvements in both COMET and BLEU scores, especially for the ha' →en case.

Summarizing the observations, Llama-2 is much more robust against the noise introduced in Hausa, likely because it has limited familiarity with the language, making it more difficult to detect and imitate imperfections present in the training data. As a result, Llama-2 tends to just recognize the essence of the translation task instead of overfitting to the biases present in low-quality data. In contrast, with German, Llama-2's understanding leads to a misinterpretation of the training objectives, such as fitting the word-level noise with a directive for literal translations. Overall, **LLMs may quickly fit translation imperfections in the training data, especially for seen languages; the resulting performance drop may be observable with just 32 training samples.**

4 Related Work

4.1 What does LLM SFT bring us?

Foundational language models become more robust and follow instructions better after being fine-tuned on task-oriented supervised data formulated as natural language text (Mishra et al., 2022; Sanh et al., 2022; Wei et al., 2022). We observe diverging trends in research on instruction tuning nowadays: (1) Many works attempt to scale up instruction data in terms of the number of tasks, languages, data size, and thus implicitly increasing training updates (Chung et al., 2024; Muennighoff et al., 2023; Wu et al., 2024c; Li et al., 2023; Üstün et al., 2024; Zhang et al., 2024). (2) Another stream of papers, argue that instruction tuning mainly alters a base model's response style but not content or knowledge—data quality and diversity outweigh

quantity (Zhou et al., 2023; Mitchell et al., 2024; Lin et al., 2024; Chen et al., 2024a). This work is a continued exploration of the latter, focusing on the machine translation task. We verify the effect of size variations and include two new factors—language directions and quality—aiming to provide practical and cost-effective guidance on this matter.

Specifically, language transfer has been demonstrated in smaller pre-trained models before LLMs (Wu and Dredze, 2019; Artetxe et al., 2020). For (sufficiently) multilingual models, training on certain languages might still benefit other languages at the test time (Choenni et al., 2023). In LLM instruction tuning, recent papers revealed cross-lingual transfer and improved robustness in unseen languages via multilingual instruction tuning with a small data sample (Chen et al., 2024c; Kew et al., 2023; Shaham et al., 2024). Furthermore, it has been claimed that even monolingual instruction tuning is sufficient to elicit multilingual responses in the correct languages with a key ingredient being the right learning rate (Chirkova and Nikoulina, 2024a,b). In relation to our experiments, language transfer to unseen languages might account for improved performance in language directions that are not directly fine-tuned.

4.2 How can we use LLMs for translation?

In the field of machine translation, earlier works provided analysis of general-purpose prompting (Vilar et al., 2023; Agrawal et al., 2023; Zhang et al., 2023a) followed by a blossom of strategies focusing on specific aspects of the translation process (Sarti et al., 2023; Ghazvininejad et al., 2023; He et al., 2024; Moslem et al., 2023; Chen et al., 2024b; Raunak et al., 2023). Nonetheless, as shown in our experimental results, few-shot prompting is not on par with using instruction-tuned models, illustrating the importance of further understanding the role of instruction tuning in translation tasks.

In terms of fine-tuning LLMs for translation, previous works have explored a wide range of sub-tasks: disambiguation, low-resource, document-level, and adaptive translation, etc (Li et al., 2024; Zhang et al., 2023b; Alves et al., 2023; Iyer et al., 2023; Mao and Yu, 2024; Wu et al., 2024b). These works focus on improving translation performance and specific applications. Stap et al. (2024) show that while fine-tuning improves translation quality, it can degrade certain key LLMs' advantages, such as the contextualization ability on document-level input. Some recent research aims to enhance

the translation capabilities of LLMs by incorporating human preference data (Jiao et al., 2023; Zeng et al., 2024; Zhu et al., 2024) or by extending the pre-training phase before fine-tuning (Xu et al., 2024a,b; Alves et al., 2024), yet these approaches require significantly more data or computing resources. The aim of this paper is not to pursue the state of the art but to investigate the opportunities of extending instruction-tuned LLMs’ translation capabilities in desirable compute-efficient scenarios. It is still worth noting that our investigation is orthogonal to previous works which employ relatively large monolingual and parallel data for continued pre-training.

5 Conclusion and Future Work

In this work, we conduct an in-depth analysis of fine-tuning LLMs for translation. We demonstrate that LLMs is capable of translating in multiple directions after being fine-tuned with *minimal low-quality training data in a single direction*. While this suggests pre-trained LLMs inherently possess multilingual translation capabilities which only need to be unlocked by aligning with the correct task format, we discover pitfalls and lessons in aligning LLMs; while LLMs make efforts to adjust to the translation task, they are good at imitating other patterns such as the noise in the parallel data. Future work could explore robust training methods that align LLMs with translation while minimizing the risk of overfitting to low-quality data.

Limitations

This work offers a range of insights into fine-tuning LLMs for translation. However, our study is not exhaustive and is subject to the following limitations.

Model size and diversity. Throughout our systematic study, we fine-tuned Llama-2 7B, Llama-2 12B, and Mistral 7B. These are strong and feasible options when the work is carried out. It is important to verify the generalizability of our findings to models with different capabilities or of different sizes.

Non-English centric MT. Our evaluation is English-centric, which is the condition of most LLM pre-training. Findings will be more comprehensive if future work can extend it to translation directions not involving English.

State-of-the-art performance. Our research primarily explores how SFT enables LLM to trans-

late to uncover data-efficient strategies in SFT and identify associated pitfalls. Recent studies have demonstrated that translation capabilities can be further enhanced through techniques such as continual pre-training (Xu et al., 2024a; Alves et al., 2024) and preference learning (Xu et al., 2024b; Zhu et al., 2024). However, these methods require significantly more training resources, which may pose challenges when applied to large models.

Fine-tuning methods. Throughout this work, we perform SFT with full-parameter updates. It is worthwhile to explore parameter-efficient methods which bring in heavier regularization to understand whether they exhibit patterns similar to those observed in our work.

Ethical considerations

Our work’s sole aim is to study the influence of data factors in applying supervised fine-tuning to large language models. We expect minimal social risks to be associated with our efforts.

Acknowledgments

We sincerely thank the reviewers of this work for their constructive and insightful feedback.

Pinzhen Chen and Barry Haddow received funding from UK Research and Innovation (UKRI) under the UK government’s Horizon Europe funding guarantee [grant number 10052546]. Miaoran Zhang received funding from the DFG (German Research Foundation) under project 232722074, SFB 1102. We thank EIT and IDT High Performance Computing Center for providing computational resources for this project.

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*.
- Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki

- Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. [Findings of the 2021 conference on machine translation \(WMT21\)](#). In *Proceedings of the Sixth Conference on Machine Translation*.
- Duarte M. Alves, Nuno M. Guerreiro, João Alves, José Pombal, Ricardo Rei, José de Souza, Pierre Colombo, and Andre Martins. 2023. [Steering large language models for machine translation with finetuning and in-context learning](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*.
- Duarte M. Alves, José Pombal, Nuno M Guerreiro, Pedro H Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, et al. 2024. [Tower: An open multilingual large language model for translation-related tasks](#). *arXiv preprint*.
- Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. [On the cross-lingual transferability of monolingual representations](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.
- Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. [Findings of the 2020 conference on machine translation \(WMT20\)](#). In *Proceedings of the Fifth Conference on Machine Translation*.
- Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. [Findings of the 2019 conference on machine translation \(WMT19\)](#). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*.
- Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. [Findings of the 2017 conference on machine translation \(WMT17\)](#). In *Proceedings of the Second Conference on Machine Translation*.
- Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. 2018. [Findings of the 2018 conference on machine translation \(WMT18\)](#). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*.
- 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)*.
- 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](#). In *Advances in Neural Information Processing Systems*.
- Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Sriniwasan, Tianyi Zhou, Heng Huang, et al. 2024a. [AlpagaSUS: Training a better Alpaca model with fewer data](#). In *The Twelfth International Conference on Learning Representations*.
- Pinzhen Chen, Zhicheng Guo, Barry Haddow, and Kenneth Heafield. 2024b. [Iterative translation refinement with large language models](#). In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*.
- Pinzhen Chen, Shaoxiong Ji, Nikolay Bogoychev, Andrey Kutuzov, Barry Haddow, and Kenneth Heafield. 2024c. [Monolingual or multilingual instruction tuning: Which makes a better Alpaca](#). In *Findings of the Association for Computational Linguistics: EACL 2024*.
- Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. [Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality](#). lmsys.org.
- Nadezhda Chirkova and Vassilina Nikoulina. 2024a. [Key ingredients for effective zero-shot cross-lingual knowledge transfer in generative tasks](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*.
- Nadezhda Chirkova and Vassilina Nikoulina. 2024b. [Zero-shot cross-lingual transfer in instruction tuning of large language models](#). In *Proceedings of the 17th International Natural Language Generation Conference*.
- Rochelle Choenni, Dan Garrette, and Ekaterina Shutova. 2023. [How do languages influence each other? studying cross-lingual data sharing during LM fine-tuning](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. [Scaling instruction-finetuned language models](#). *Journal of Machine Learning Research*.

- Marjan Ghazvininejad, Hila Gonen, and Luke Zettlemoyer. 2023. [Dictionary-based phrase-level prompting of large language models for machine translation](#). *arXiv preprint*.
- Sreyan Ghosh, Chandra Kiran Reddy Evuru, Sonal Kumar, Ramaneswaran S, Deepali Aneja, Zeyu Jin, Ramani Duraiswami, and Dinesh Manocha. 2024. [A closer look at the limitations of instruction tuning](#). In *Proceedings of the 41st International Conference on Machine Learning*.
- Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2024. [Exploring human-like translation strategy with large language models](#). *Transactions of the Association for Computational Linguistics*.
- Vivek Iyer, Pinzhen Chen, and Alexandra Birch. 2023. [Towards effective disambiguation for machine translation with large language models](#). In *Proceedings of the Eighth Conference on Machine Translation*.
- Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. [Mistral 7B](#). *arXiv preprint*.
- Wenxiang Jiao, Jen-tse Huang, Wenxuan Wang, Zhiwei He, Tian Liang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. [Parrot: Translating during chat using large language models tuned with human translation and feedback](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*.
- Tannon Kew, Florian Schottmann, and Rico Sennrich. 2023. [Turning english-centric LLMs into polyglots: How much multilinguality is needed?](#) *arXiv preprint*.
- Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thammie 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)*.
- Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. 2023. [Bactrian-X: Multilingual replicable instruction-following models with low-rank adaptation](#). *arXiv preprint*.
- Jiahuan Li, Hao Zhou, Shujian Huang, Shanbo Cheng, and Jiajun Chen. 2024. [Eliciting the translation ability of large language models via multilingual finetuning with translation instructions](#). *Transactions of the Association for Computational Linguistics*.
- Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, and Yejin Choi. 2024. [Urial: Aligning untuned LLMs with just the 'write' amount of in-context learning](#). In *The Twelfth International Conference on Learning Representations*.
- Zhuoyuan Mao and Yen Yu. 2024. [Tuning LLMs with contrastive alignment instructions for machine translation in unseen, low-resource languages](#). In *Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024)*.
- Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. [Cross-task generalization via natural language crowdsourcing instructions](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.
- Eric Mitchell, Rafael Rafailov, Archit Sharma, Chelsea Finn, and Christopher D Manning. 2024. [An emulator for fine-tuning large language models using small language models](#). In *The Twelfth International Conference on Learning Representations*.
- Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023. [Adaptive machine translation with large language models](#). In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*.
- 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](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. [Training language models to follow instructions with human feedback](#). In *Advances in Neural Information Processing Systems* 35.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*.
- Matt Post. 2018. [A call for clarity in reporting BLEU scores](#). In *Proceedings of the Third Conference on Machine Translation: Research Papers*.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. [Language models are unsupervised multitask learners](#). OpenAI blog.
- Vikas Raunak, Amr Sharaf, Hany Hassan Awadallah, and Arul Menezes. 2023. [Leveraging GPT-4 for automatic translation post-editing](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*.

- 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)*.
- 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 *International Conference on Learning Representations*.
- Gabriele Sarti, Phu Mon Htut, Xing Niu, Benjamin Hsu, Anna Currey, Georgiana Dinu, and Maria Nadejde. 2023. [RAMP: Retrieval and attribute-marking enhanced prompting for attribute-controlled translation](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. [Improving neural machine translation models with monolingual data](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.
- Uri Shaham, Jonathan Herzig, Roei Aharoni, Idan Szepktor, Reut Tsarfaty, and Matan Eyal. 2024. [Multilingual instruction tuning with just a pinch of multilinguality](#). In *Findings of the Association for Computational Linguistics ACL 2024*.
- Suzanna Sia, David Mueller, and Kevin Duh. 2024. [Where does in-context translation happen in large language models](#). *arXiv preprint*.
- David Stap, Eva Hasler, Bill Byrne, Christof Monz, and Ke Tran. 2024. [The fine-tuning paradox: Boosting translation quality without sacrificing LLM abilities](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.
- 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](#). GitHub repository.
- Jörg Tiedemann and Santhosh Thottingal. 2020. [OPUS-MT – building open translation services for the world](#). In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shrusti Bhosale, et al. 2023. [Llama 2: Open foundation and fine-tuned chat models](#). *arXiv preprint*.
- Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D’souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, et al. 2024. [Aya model: An instruction finetuned open-access multilingual language model](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.
- 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)*.
- 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*.
- Di Wu, Shaomu Tan, Yan Meng, David Stap, and Christof Monz. 2024a. [How far can 100 samples go? unlocking zero-shot translation with tiny multi-parallel data](#). In *Findings of the Association for Computational Linguistics ACL 2024*.
- Minghao Wu, Thuy-Trang Vu, Lizhen Qu, George Foster, and Gholamreza Haffari. 2024b. [Adapting large language models for document-level machine translation](#). *arXiv preprint*.
- Minghao Wu, Abdul Waheed, Chiyu Zhang, Muhammad Abdul-Mageed, and Alham Aji. 2024c. [LaMini-LLM: A diverse herd of distilled models from large-scale instructions](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*.
- Shijie Wu and Mark Dredze. 2019. [Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT](#). 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)*.
- Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024a. [A paradigm shift in machine translation: Boosting translation performance of large language models](#). In *The Twelfth International Conference on Learning Representations*.
- Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024b. [Contrastive preference optimization: Pushing the boundaries of LLM performance in machine translation](#). In *Proceedings of the 41st International Conference on Machine Learning*.
- Jiali Zeng, Fandong Meng, Yongjing Yin, and Jie Zhou. 2024. [Teaching large language models to translate with comparison](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*.
- Biao Zhang, Barry Haddow, and Alexandra Birch. 2023a. [Prompting large language model for machine translation: a case study](#). In *Proceedings of the 40th International Conference on Machine Learning*.

- Biao Zhang, Zhongtao Liu, Colin Cherry, and Orhan Firat. 2024. [When scaling meets LLM finetuning: The effect of data, model and finetuning method](#). In *The Twelfth International Conference on Learning Representations*.
- Shaolei Zhang, Qingkai Fang, Zhuocheng Zhang, Zhengrui Ma, Yan Zhou, Langlin Huang, Mengyu Bu, Shangdong Gui, Yunji Chen, Xilin Chen, et al. 2023b. [BayLing: Bridging cross-lingual alignment and instruction following through interactive translation for large language models](#). *arXiv preprint*.
- Xuan Zhang, Navid Rajabi, Kevin Duh, and Philipp Koehn. 2023c. [Machine translation with large language models: Prompting, few-shot learning, and fine-tuning with QLoRA](#). In *Proceedings of the Eighth Conference on Machine Translation*.
- Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2023. [LIMA: Less is more for alignment](#). In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Dawei Zhu, Sony Trenous, Xiaoyu Shen, Dietrich Klakow, Bill Byrne, and Eva Hasler. 2024. [A preference-driven paradigm for enhanced translation with large language models](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

A Model Performance with Varying Training Sample Sizes

In Figure 6 and Figure 7, we present the performance for instruction-tuned baselines and our models on different evaluation directions. For most directions, using only 32 training samples can achieve competitive performance and beat all three instruction-tuned baselines. There are several exceptional cases, including $en \rightarrow zh$ and $en \rightarrow ja$, in which the COMET score of SFT with a limited number of samples (32 or 64) is worse than 1-shot in-context learning.

While we primarily report the results with Llama-2 7B in our experiments, we hypothesize that state-of-the-art LLMs are largely homogeneous in terms of language distribution and inherent translation capability making our findings applicable to other LLMs. To support this hypothesis, we conduct fine-tuning experiments with Mistral 7B and Llama-2 13B using varying data sizes: 32, 1024, and 70K. As shown in Figure 8, the general trend is quite similar to the Llama-2 7B case: fine-tuning with 32 examples results in competitive performance, matching or surpassing general-purpose instruction-tuned models. Furthermore, increasing the number of training examples leads to diminishing returns.

B Model Performance with Varying Training Directions

Figure 9 shows normalized BLEU scores for different combinations of train and test translation directions. Similar to the COMET scores in Figure 2, we observe that when training the model on a single direction, its translation ability across other non-targeted directions is also elicited to a certain degree. It is worth noting that when the training direction is $X \rightarrow en$, the performance on directions $en \rightarrow X$ is significantly worse than training on all directions.

C Combined Effect of Training Size and Direction

Figure 12 illustrates the model performance across varying training sizes and translation directions, evaluated on $en \rightarrow cs$, de , zh . Similarly, Figure 13 presents the results on $en \rightarrow cs$, de , zh , and $en \rightarrow hr$. Consistently across all plots, we observe a positive impact on performance with an increasing number of training directions, particularly with smaller training sizes.

D Model Performance with Unseen Languages

In Figure 10, we find similar patterns as the COMET score, where fine-tuning on unseen languages can elicit the model’s ability to translate from and to all seen languages. However, the translation performance on unseen languages themselves remains subpar, suggesting that SFT primarily reveals the knowledge LLMs have possessed during pre-training.

E Model Performance with Noisy Data

Figure 11 shows the BLEU score of different translation directions with two noise types. We can find that models are more sensitive to word-level noise than sentence-level noise. Also, the performance degradation is more noticeable when injecting noise into the source translation side. In comparison to the results of size 1024, using 32 training examples still achieves comparable or even better performance in the noisy condition.

F Technical Details

F.1 Datasets

Our parallel data is derived from the development and test sets of WMT17 through WMT22. Detailed dataset statistics are available in Table 3. For most experiments, we use the test sets from WMT17 to WMT20 for training. The test set from WMT22 is used specifically for testing. An exception is noted in Section section 3.4, where models are trained using the $en \leftrightarrow ha$ and $en \leftrightarrow is$ language pairs from WMT21’s development set. Subsequently, these models are evaluated using the corresponding test sets from WMT21.

F.2 Translation instructions

The collection of translation instruction templates used in this work can be found in Table 4.

F.3 Evaluation packages

To obtain COMET scores, we use `Unbabel/wmt22-comet-da`⁹ and for BLEU scores, we use `sacreBLEU`¹⁰ (Post, 2018). The signature from the `sacreBLEU` package is `nrefs:1, case:mixed, eff:no, tok:13a, smooth:exp, version:2.0.0` for all language pairs, except for tokenization for $en \rightarrow zh$ and $en \rightarrow jp$, where we use `tok:zh` and `tok:jp-mecab`, respectively.

⁹<https://github.com/Unbabel/COMET>

¹⁰<https://github.com/mjpost/sacrebleu>

Direction	Training					Validation*	Test
	WMT17	WMT18	WMT19	WMT20	WMT21dev	WMT21	WMT22
en-cs	3005	2983	1997	1418	0	1002	2037
en-de	3004	2998	1997	1418	0	1002	2037
en-hr	0	0	0	0	0	0	1671
en-ja	0	0	0	1000	0	0	2037
en-ru	3001	3000	1997	2002	0	1002	2037
en-zh	2001	3981	1997	1418	0	1002	2037
cs-en	3005	2983	0	664	0	1000	1448
de-en	3004	2998	2000	785	0	1000	1984
ja-en	0	0	0	993	0	1005	2008
ru-en	3001	3000	2000	991	0	1000	2016
zh-en	2001	3981	2000	2000	0	1948	1875
en-ha	0	0	0	0	2000	1000	0
ha-en	0	0	0	0	2000	997	0
en-is	0	0	0	0	2004	1000	0
is-en	0	0	0	0	2004	1000	0
de-fr	0	0	1701	1619	0	⊗	1984
fr-de	0	0	1701	1619	0	⊗	2006

Table 3: Data statistics. *Generally, WMT21 test is used for validation purposes; exceptions are en↔ha and en↔is, which are used for testing. ⊗Although WMT21 includes data for de↔fr, these language pairs are excluded from experiments.

F.4 Hardware specifications and runtime

Our experiments are conducted on a computing node with either 8 NVIDIA A100-40GB GPUs or 8 H100-80GB GPUs. DeepSpeed¹¹ with zero-stage 1 and mixed precision bfloat16 is used for performing SFT. Given the limited dataset size, typically fewer than 1024 samples, each SFT experiment can be completed within a mere 15 minutes using four H100 GPUs. However, given the necessity to evaluate the models across more than ten translation directions, the evaluation process may require up to four hours when performed on a single A100-40GB GPU.

¹¹<https://github.com/microsoft/DeepSpeed>

Instruction pool
Please provide the [TGT] translation for the following text
Convert the subsequent sentences from [SRC] into [TGT] :
Render the listed sentences in [TGT] from their original [SRC] form:
Transform the upcoming sentences from [SRC] language to [TGT] language:
Translate the given text from [SRC] to [TGT] :
Turn the following sentences from their [SRC] version to the [TGT] version:
Adapt the upcoming text from [SRC] to [TGT] :
Transpose the next sentences from the [SRC] format to the [TGT] format.
Reinterpret the ensuing text from [SRC] to [TGT] language.
Modify the forthcoming sentences, converting them from [SRC] to [TGT] .
What is the meaning of these sentences when translated to [TGT] ?
In the context of [TGT] , what do the upcoming text signify? The text is:
How would you express the meaning of the following sentences in [TGT] ?
What is the significance of the mentioned sentences in [TGT] ?
In [TGT] , what do the following text convey?
When translated to [TGT] , what message do these sentences carry?
What is the intended meaning of the ensuing sentences in [TGT] ?
How should the following sentences be comprehended in [TGT] ?
In terms of [TGT] , what do the next sentences imply?
Kindly furnish the [TGT] translation of the subsequent sentences.
Could you supply the [TGT] translation for the upcoming sentences?
Please offer the [TGT] rendition for the following statements.
I'd appreciate it if you could present the [TGT] translation for the following text:
Can you deliver the [TGT] translation for the mentioned sentences?
Please share the [TGT] version of the given sentences.
It would be helpful if you could provide the [TGT] translation of the ensuing sentences.
Kindly submit the [TGT] interpretation for the next sentences.
Please make available the [TGT] translation for the listed sentences.
Can you reveal the [TGT] translation of the forthcoming sentences?
Translate from [SRC] to [TGT] :

Table 4: A collection of 31 translation prompts. Each instruction is randomly selected to form a training sample. At inference time, the first instruction is always selected. The placeholders [SRC] and [TGT] represent the source and target languages, respectively, and will be replaced with the appropriate languages depending on the specific example at hand.

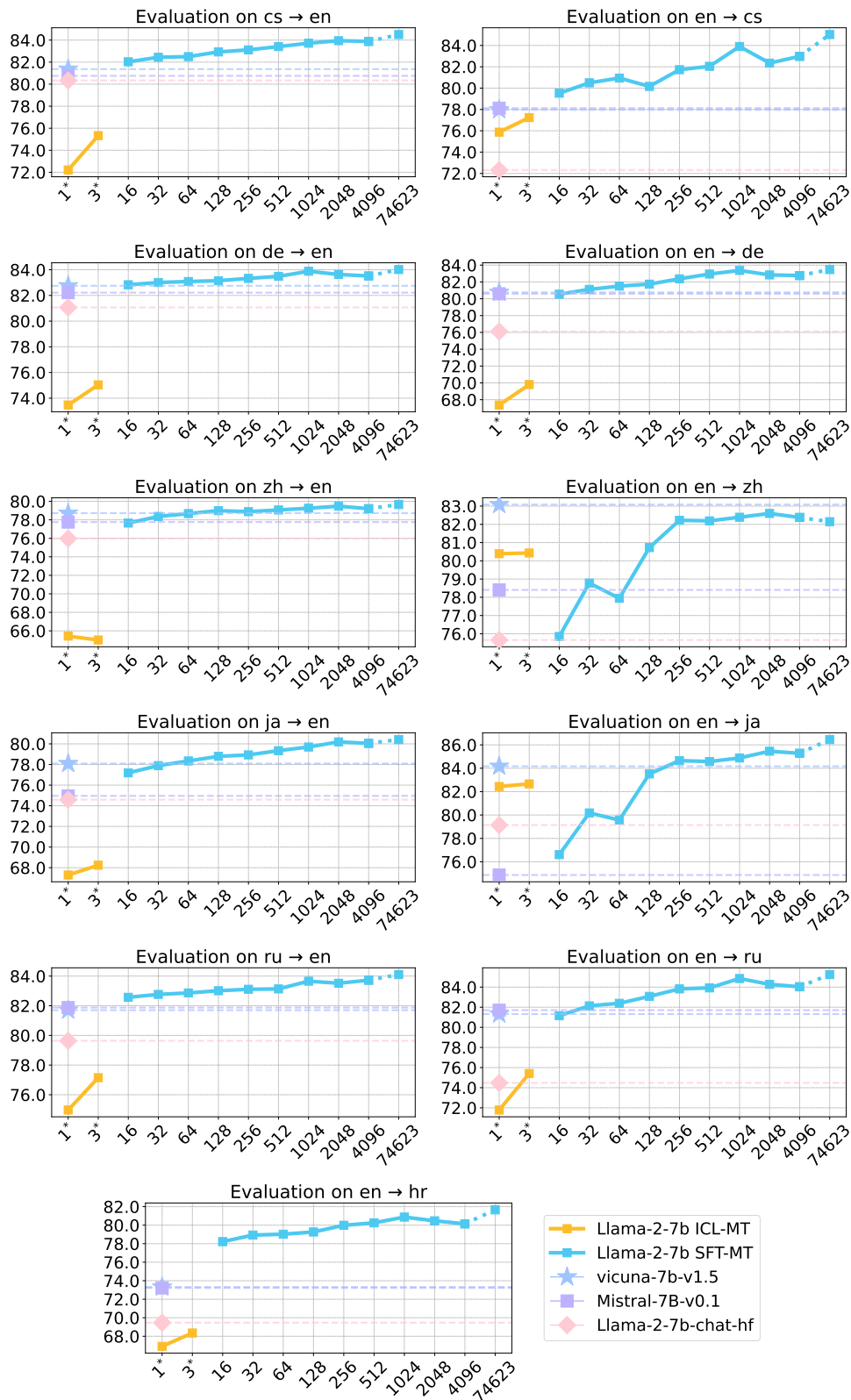


Figure 6: COMET scores between instruction-tuned baselines and our models at different training data sizes, evaluated on individual translation directions. ICL is used for training sizes at or below 3, indicated with "*"; otherwise, we perform SFT. With only 32 examples for SFT, Llama-2 outperforms general-purpose, instruction-tuned baselines. Base.: instruction-tuned baseline models.

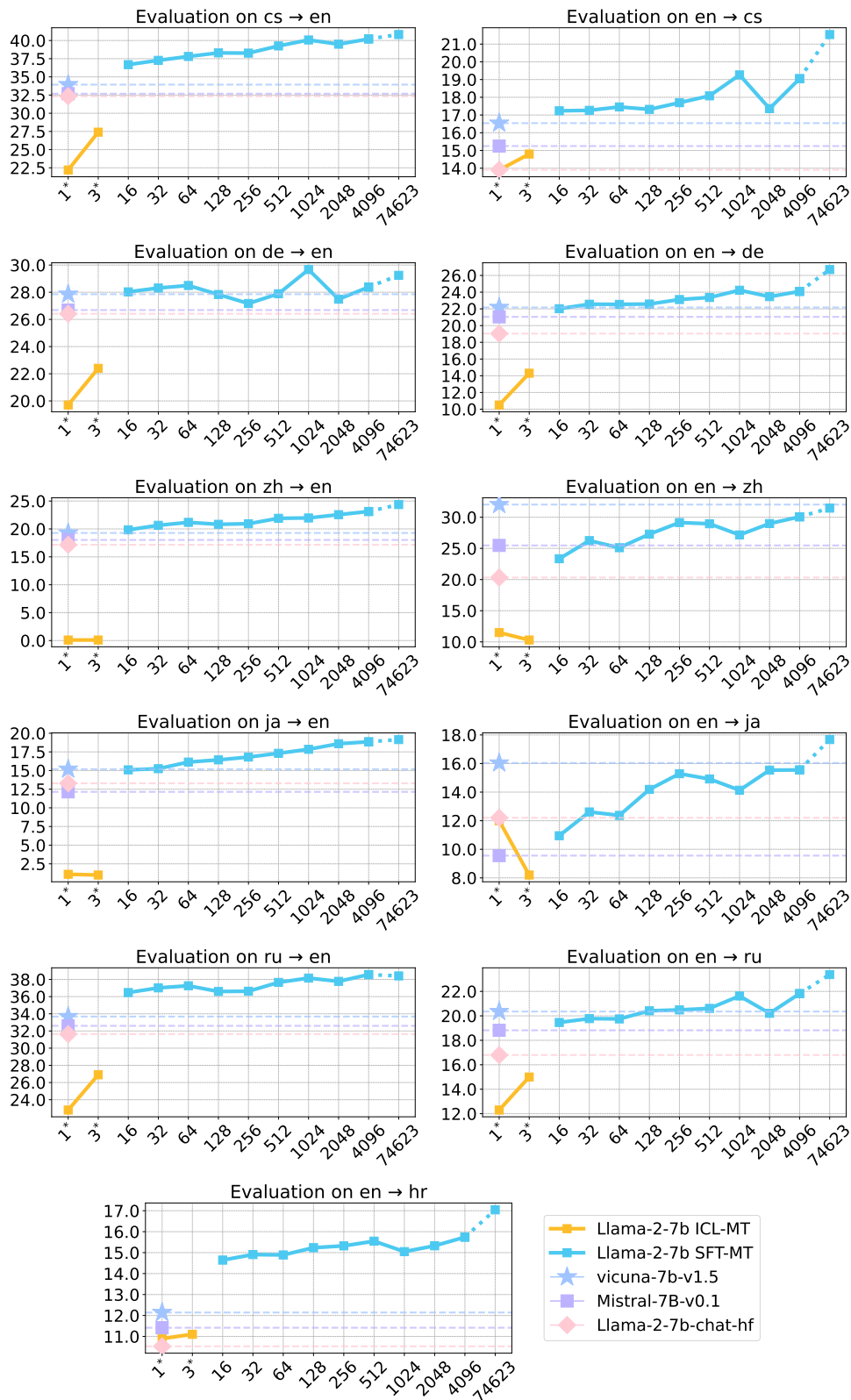


Figure 7: BLEU scores between instruction-tuned baselines and our models at different training data sizes, evaluated on individual translation directions. ICL is used for training sizes at or below 3, indicated with "*"; otherwise, we perform SFT. With only 32 examples for SFT, Llama-2 outperforms general-purpose, instruction-tuned baselines. Base.: instruction-tuned baseline models.

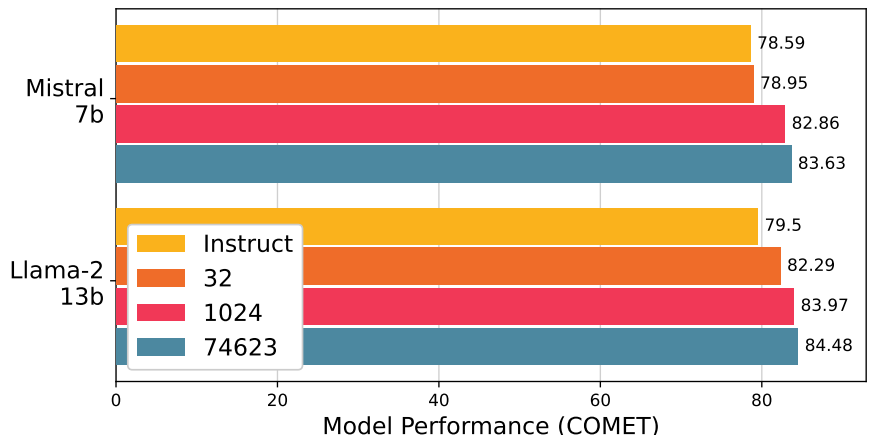


Figure 8: Performance comparison between instruction-tuned baselines and fine-tuned models with different training data sizes. “Instruct” refers to the instruction-tuned baselines, specifically [Mistral-7B-Instruct-v0.1](#) and [Llama-2-13b-chat](#). “32/1024/74623” represents models fine-tuned on 32, 1024, and 74623 examples, using pre-trained only models: [Mistral-7B-v0.1](#) and [Llama-2-13b](#).

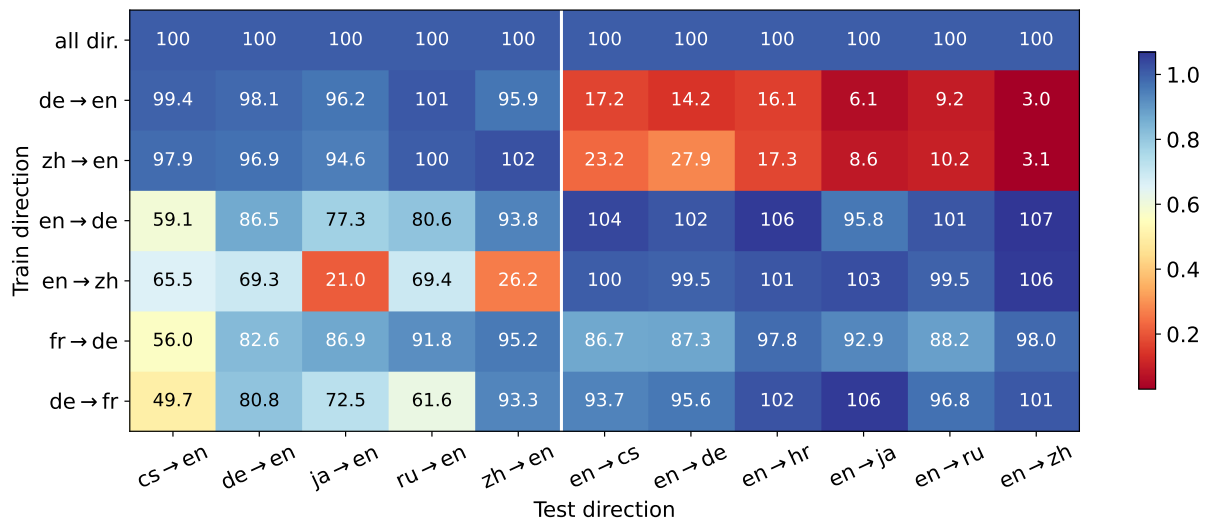


Figure 9: Model performance (%) in BLEU score resulted from varying combinations of train and test translation directions. The scores are normalized according to Llama-2 fine-tuned on all 10 training directions.

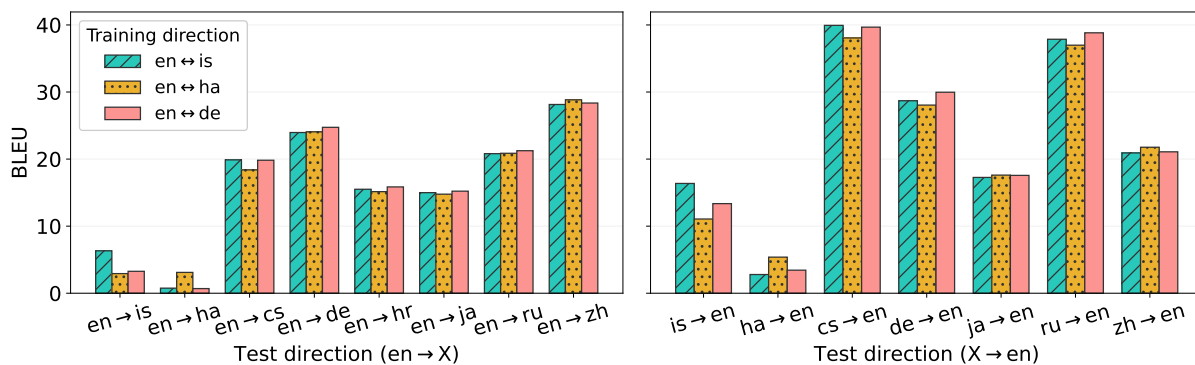


Figure 10: Model performance evaluated across 15 translation directions. While models trained on *unseen* languages ($en \leftrightarrow is$, $en \leftrightarrow ha$) exhibit moderate improvements in translating these languages, they demonstrate accurate translations from and to *seen* languages.

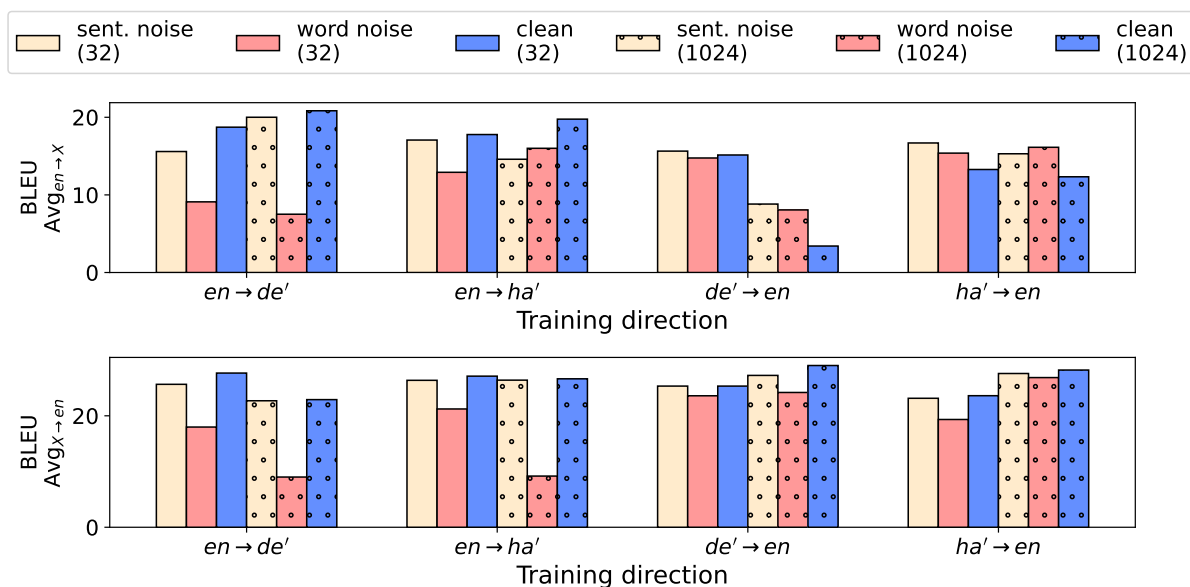


Figure 11: Model performance in BLEU score varying training sizes, directions, and noise types. Top (Bottom): score averaged across all $en \rightarrow X$ ($X \rightarrow en$) test directions. Training sizes considered are 32 and 1024.

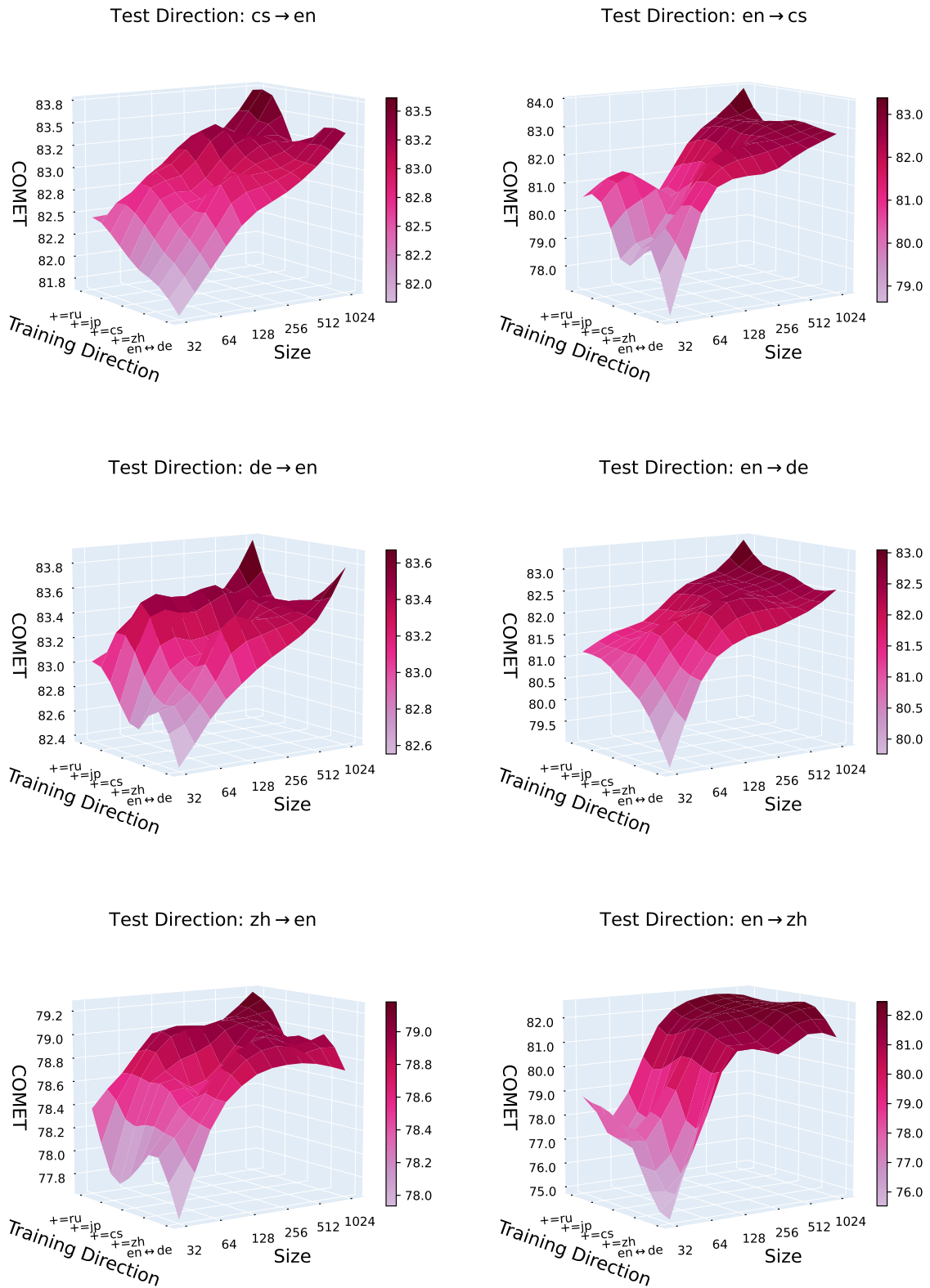


Figure 12: Model performance (in COMET) on individual directions for models trained with varying data sizes and directions. Both factors positively impact performance. +=: training directions added on top of previous directions; two directions (from and to English) at a time. For example, “+=ru” covers 10 directions: en ↔ {de, zh, cs, jp, ru}.

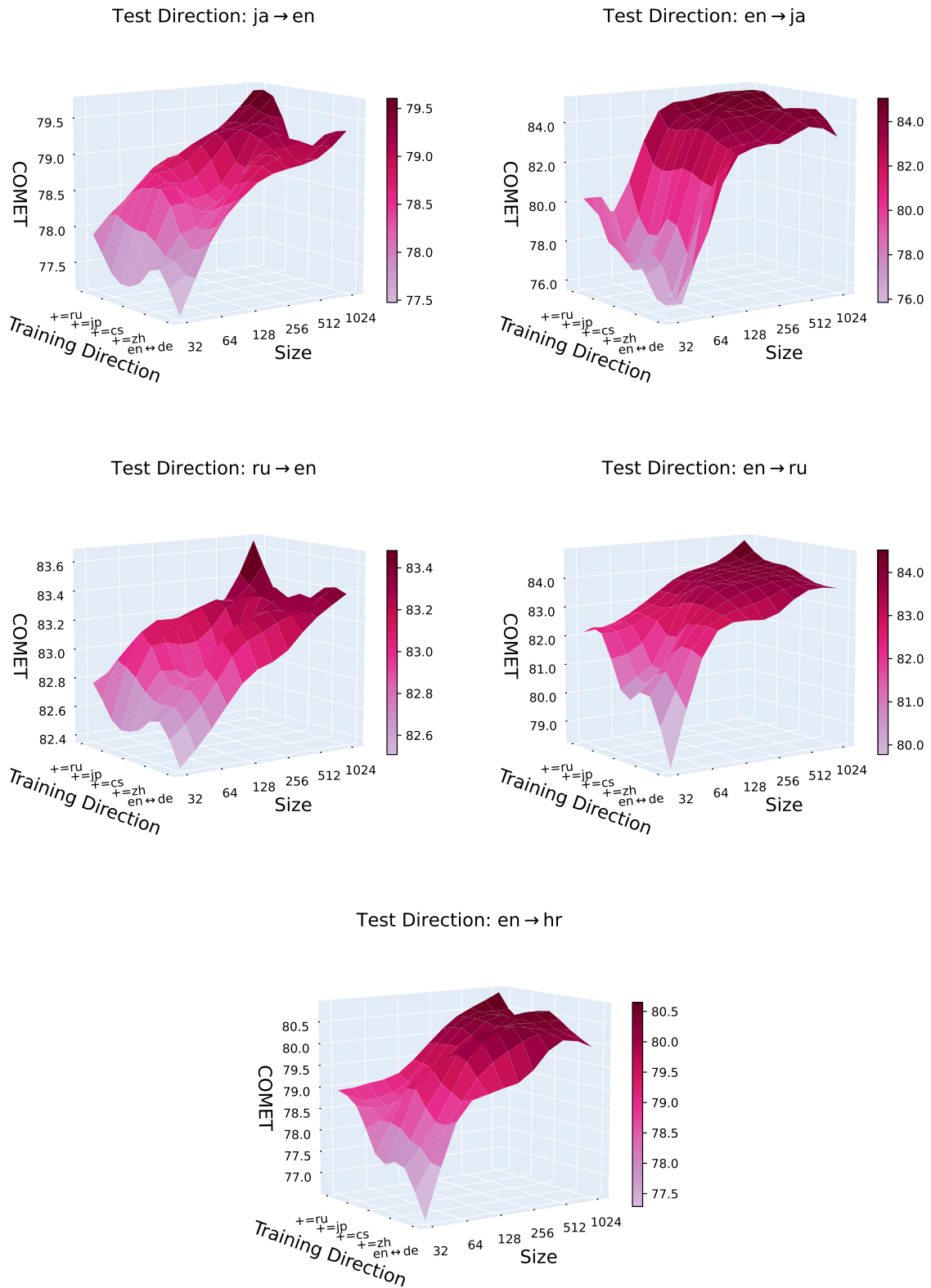


Figure 13: Model performance (in COMET) on individual directions for models trained with varying data sizes and directions. Both factors positively impact performance. +=: training directions added on top of previous directions; two directions (from and to English) at a time. For example, “+=ru” covers 10 directions: en ↔ {de, zh, cs, jp, ru}.