Adapting Large Language Models for Document-Level Machine Translation

Anonymous ACL submission

Abstract

Large language models (LLMs) have significantly advanced various natural language processing (NLP) tasks. Recent research indicates that moderately-sized LLMs often outperform larger ones after task-specific fine-tuning. This study focuses on adapting LLMs for document-level machine translation (DocMT) for specific language pairs. We first investigate the impact of prompt strategies on translation performance and then conduct extensive experiments using two fine-tuning methods, three LLM backbones, and 18 translation tasks across nine language pairs. Our results show that specialized models can sometimes surpass GPT-4 in translation performance but still face issues like *off-target translation* due to error propagation in decoding. We provide an in-depth analysis of these LLMs tailored for DocMT, examining translation errors, discourse phenomena, training strategies, the scaling law of parallel documents, recent test set evaluations, and zero-shot crosslingual transfer. Our findings highlight the strengths and limitations of LLM-based DocMT models and provide a foundation for future research.

1 Introduction

Large language models (LLMs) demonstrate impressive proficiency in a wide range of applications (Ouyang et al., 2022; Wei et al., 2022a; Sanh et al., 2022; Chung et al., 2022; OpenAI, 2023; Anil et al., 2023; Touvron et al., 2023a,b; Jiang et al., 2023). However, in the realm of translation tasks, only few very large models, such as GPT-3.5-TURBO and GPT-4-TURBO, can match or surpass the performance of state-of-the-art supervised encoder-decoder models like NLLB (Costa-jussà et al., 2022), while they still under-perform in translating low-resource languages (Robinson et al., 2023; Jiao et al., 2023; Hendy et al., 2023). Consequently, a number of recent works attempt to bridge the gap between LLMs and supervised encoder-decoder models in translation tasks (Zhu et al., 2023; Yang et al., 2023; Zhang et al., 2023; Moslem et al., 2023; Xu et al., 2023; Kudugunta et al., 2023). Recently, research suggests that smaller, specialized models can outperform larger, general-purpose models in specific tasks (Gunasekar et al., 2023; Luo et al., 2023; Azerbayev et al., 2023). Therefore, we explore adapting LLMs for document-level machine translation (DocMT) in this study.

In this study, we analyze moderately-sized LLMs (with $7B$ parameters) across 18 translation tasks involving nine language pairs. We fine-tune three LLMs using Parameter-Efficient Fine-Tuning (PEFT) and Fully Fine-Tuning (FFT). Comparisons with state-of-the-art translation models, using metrics like sBLEU, dBLEU, and COMET, confirm the superior translation capabilities of LLMs after fine-tuning. However, we identify a significant issue of *off-target translations*, observed even after exclusive fine-tuning on bilingual corpora. Our further analysis shows that the high off-target translation rate results from the error propagation in the decoding process. Additionally, we present an in-depth analysis of our LLM-based DocNMT models from various perspectives: translation error distribution, discourse phenomena, training strategy, the scaling law of parallel documents, additional evaluations on WMT2023 test sets (Koehn et al., 2023), and zero-shot cross-lingual transfer, aiming to enhance understanding and efficacy of LLMs in DocMT tasks.

We present extensive empirical evidence that highlights both the superior translation capabilities and limitations of the LLM-based DocMT models in this study, making several significant discoveries. Here are the main takeaways:

- **Selective Excellence in Translation Tasks**: Our findings show that our moderately-sized LLMs outperform GPT-4-TURBO in certain translation tasks, but struggle in others due to the *off-target translation* issue due to error propagation in decoding. Despite this, our DocMT models exhibit better context awareness and fewer errors, while maintaining comparable performance.

- **Fine-Tuning Strategies**: Our research indicates that the PEFT approach outperforms the FFT approach overall. However, the FFT approach shows greater data efficiency, needing only about 1% of the total dataset to reach the performance level of models trained on the entire dataset. In contrast, the PEFT approach requires 10% of the total dataset for comparable results.
- **Evaluation on Recent Test Sets**: We evaluate our models on recent test sets between English and German from WMT2023 (Koehn et al., 2023). Our empirical results show that, when the data leakage risks are mitigated, the LLM-based DocMT models generalize better on out-of-domain text, compared to the conventional DocMT models.
- **Advantage of Base LLMs for Task-Specific Supervised Fine-Tuning**: Our study shows that base LLMs, when used as the backbone for task-specific supervised fine-tuning, perform better than instruction-tuned LLMs. They demonstrate more effective zero-shot cross-lingual transfer.

2 Related Work

Document-Level Machine Translation In recent years, numerous approaches have been proposed for document-level machine translation (DocMT). There exist other approaches to DocMT, including document embedding (Macé and Servan, 2019; Huo et al., 2020), multiple encoders (Wang et al., 2017; Bawden et al., 2018; Voita et al., 2018; Zhang et al., 2018), attention variations (Miculicich et al., 2018; Zhang et al., 2020; Maruf et al., 2019; Wong et al., 2020; Wu et al., 2023), and translation caches (Maruf and Haffari, 2018; Tu et al., 2018; Feng et al., 2022). Furthermore, Maruf et al. (2022) present a comprehensive survey of DocMT.

Large Language Models Large language models (LLMs) have demonstrated remarkable proficiency across a wide range of Natural Language Processing (NLP) tasks (Brown et al., 2020; Chowdhery et al., 2022; Scao et al., 2022; Anil et al., 2023; Touvron et al., 2023a,b). Furthermore, recent research has shown that supervised fine-tuning (SFT) and Reinforcement Learning from Human Feedback (RLHF) can significantly enhance their performance when following general language instructions (Mishra et al., 2022; Wang et al., 2022; Shen et al., 2023; Li et al., 2023; Wu and Aji, 2023; Wu et al., 2024a). More recently, there is a growing body of work exploring the translation capabilities of LLMs (Lu et al., 2023; Zhang et al., 2023; Xu et al., 2023; Robinson et al., 2023; Wu et al., 2024d). However, it is important to note that these efforts have primarily focused on sentence-level machine translation (SenMT) and have not delved into document-level machine translation (DocMT). A noteworthy study in DocMT is conducted by Wang et al. (2023b), where they investigate the document-level translation capabilities of gpt-3.5-turbo, making it the most closely related work to our work.

Ours In contrast to the work of Wang et al. (2023b), who primarily investigate the use of gpt-3.5-turbo for DocMT through prompting techniques, our study concentrates on analyzing the effectiveness of parameter-efficient fine-tuning (PEFT) and full fine-tuning (FFT) methods on moderately-sized LLMs in the context of DocMT.

3 Experimental Setup

In this study, we aim to adapt multilingual pre-trained large language models (LLMs) into a *bilingual* document-level machine translation (DocMT) model. In this section, we describe our experimental setup of this work, including training strategy (Section 3.1), datasets (Section 3.2), models (Section 3.3), and evaluation (Section 3.4).

3.1 Two-Stage Training

DocMT approaches typically start by pre-training the translation model on parallel sentences and then fine-tuning it on parallel documents (Voita et al., 2019; Ma et al., 2020; Wu et al., 2023). Recently, Xu et al. (2023) proposed a two-stage training strategy. This involves initially fine-tuning a LLM on monolingual text, followed by a second fine-tuning phase on parallel text. Following Xu et al. (2023), our approach begins with fine-tuning an LLM on monolingual documents, followed by fine-tuning on parallel documents.

Fine-tuning on Monolingual Documents Existing LLMs are typically pre-trained on English-centric corpora. Recent research highlights that these LLMs often exhibit sub-optimal performance on multilingual benchmarks (Li et al., 2023; Chen et al., 2023; Scao et al., 2022). To address this limitation, our initial step involves fine-tuning all the parameters of LLMs using monolingual data from the target languages.

Fine-tuning on Parallel Documents We fine-tune the model on document-level parallel corpora in this stage. Following Wang et al. (2023a), we condition each sentence pair on its context, consisting of the three preceding consecutive sentence pairs. As demonstrated by Wang et al. (2023b), the prompting strategy plays a significant role in translating documents using LLMs. However, they only investigate how the prompting strategies affect GPT-3.5-TURBO and GPT-4-TURBO at the inference stage. In our study, we first delve into how these prompting strategies impact the fine-tuning process, as shown in Figure 1, and we present our findings in Section 4.

3.2 Datasets

Parallel Documents Following Zhang et al. (2022), we conduct experiments on IWSLT2017 translation tasks (Cettolo et al., 2017). IWSLT2017 comprises translation datasets sourced from TED talks, encompassing translations between English and nine other languages, including Arabic, German, French, Italian, Japanese, Korean, Dutch, Romanian, and Chinese. There are approximately $1.9K$ sentence-aligned parallel documents with about $240K$ sentences for each language pair. The dataset statistics can be found in Appendix A.

Monolingual Documents We gather monolingual documents for all the target languages in our translation tasks, totaling ten languages. To manage computational limitations and address concerns about catastrophic forgetting that might result from excessive continued training, we leverage the data pruning technique suggested by Marion et al. (2023) to select $100M$ tokens for each language, including English, from the CulturaX corpus (Nguyen et al., 2023), totaling $1B$ tokens.

3.3 Models

Baselines The baseline models in this study can be classified into three categories, including state-of-the-art LLMs and SENMT models, and our re-implemented DOCMT models:

- **State-of-the-art SENMT models**: Our selection includes models such as NLLB, which are available with three different sets of parameters: 600M, 1.3B, and 3.3B.[1] We also incorporate the widely-used commercial translation system, Google Translate.
- **State-of-the-art LLMs**: For our baseline LLMs in the context of DOCMT, we utilize GPT-3.5-TURBO and GPT-4-TURBO.[2] We use the Prompt 4 as detailed in Figure 1d during the translation process.
- **Our re-implemented DOCMT models**: We conduct full fine-tuning on the concatenation-based DOCMT model (Tiedemann and Scherrer, 2017), as well as several recent DOCMT baselines (Sun et al., 2022; Wu et al., 2023, 2024b), initialized with MT5 (Xue et al., 2021). These models are available with parameters of 300M, 580M, and 1.2B, representing the strong DOCMT baseline.

Ours In this work, we utilize LLAMA2-7B, BLOOM-7B, and VICUNA-7B, as our backbones.[3] The LLAMA2 models are predominantly pre-trained on English text, while the BLOOM models are pre-trained on multilingual text. The use of VICUNA models allows us to compare the differences between base models and instruction-tuned models (LLAMA2 vs. VICUNA). We denote those fully fine-tuned models as L-7B-FFT, B-7B-FFT, and V-7B-FFT. We denote those models fine-tuned with LORA (Hu et al., 2022) as L-7B-LORA, B-7B-LORA, and V-7B-LORA. The optimization details can be found in Appendix B.

3.4 Evaluation

Evaluation Metrics We evaluate the translation quality using sentence-level BLEU (Papineni et al., 2002) and document-level BLEU (Liu et al., 2020) using SacreBLEU (Post, 2018), denoted as sBLEU and dBLEU.[4] Furthermore, as conventional MT metrics like BLEU demonstrate poor correlation to human judgments (Freitag et al., 2022), we also evaluate the translation quality with the state-of-the-art neural evaluation metric COMET (Rei et al.,

[1] Model signatures: `facebook/nllb-200-distilled-600M`, `facebook/nllb-200-1.3B`, and `facebook/nllb-200-3.3B`.

[2] Model signatures: `gpt-3.5-turbo-1106` and `gpt-4-1106-preview`.

[3] LLAMA2 signature: `meta-llama/Llama-2-7b-hf`, BLOOM signature: `bigscience/bloom-7b1`, and VICUNA signature: `lmsys/vicuna-7b-v1.5`. Note that VICUNA-v1.5 models are fine-tuned from LLAMA2.

[4] BLEU signature: `nrefs:1|case:mixed|eff:no|tok:[13a|ja-mecab-0.996-IPA|ko-mecab-0.996/ko-0.9.2-KO|zh]|smooth:exp|version:2.3.1`.

```
[<src_lang> Context]: <src1> <src2> <src3>
[<tgt_lang> Context]: <tgt1> <tgt2> <tgt3>
[<src_lang> Sentence]: <src4>
[<tgt_lang> Sentence]: <tgt4>
```

(a) Prompt 1

```
[<src_lang>]: <src1> [<tgt_lang>]: <tgt1>
[<src_lang>]: <src2> [<tgt_lang>]: <tgt2>
[<src_lang>]: <src3> [<tgt_lang>]: <tgt3>
[<src_lang>]: <src4> [<tgt_lang>]: <tgt4>
```

(b) Prompt 2

```
[<src_lang> Context]: <src1> <src2> <src3>
[<tgt_lang> Context]: <tgt1> <tgt2> <tgt3>
Given the provided parallel context, translate the following
↪  <src_lang> sentence to <tgt_lang>:
[<src_lang> Sentence]: <src4>
[<tgt_lang> Sentence]: <tgt4>
```

(c) Prompt 3

```
[<src_lang>]: <src1> [<tgt_lang>]: <tgt1>
[<src_lang>]: <src2> [<tgt_lang>]: <tgt2>
[<src_lang>]: <src3> [<tgt_lang>]: <tgt3>
Given the provided parallel sentence pairs, translate the following
↪  <src_lang> sentence to <tgt_lang>:
[<src_lang>]: <src4> [<tgt_lang>]: <tgt4>
```

(d) Prompt 4

Figure 1: Prompt types used in the preliminary study. <src_lang> and <tgt_lang> indicate the source and target languages. <src*> and <tgt*> indicate the source and target sentences. **Note that the target sentences <tgt*> are only used during training and are replaced with the hypotheses <hyp*> generated by the model during inference.** Concrete examples for each prompt variation can be found in Appendix C.

	PID	$\mu_{s\text{BLEU}}$	$\mu_{d\text{BLEU}}$	μ_{COMET}
L-7B-LoRA	1	15.5	18.2	67.5
	2	19.0	21.9	70.7
	3	15.8	18.3	69.8
	4	**20.2**	**23.4**	**72.7**
B-7B-LoRA	1	19.3	20.5	70.5
	2	20.6	23.5	73.6
	3	19.8	20.8	73.9
	4	**23.1**	**27.3**	**76.8**
V-7B-LoRA	1	19.0	22.4	74.2
	2	20.4	23.5	71.6
	3	18.3	21.4	70.0
	4	**22.4**	**25.7**	**76.2**

Table 1: Overall performance given by L-7B-LoRA, B-7B-LoRA, and V-7B-LoRA on different prompt variations, across four English-centric translation tasks involving German and Chinese. PID indicates the prompt ID in Figure 1. Best results are highlighted in **bold**.

2020).[5] Moreover, we use the average sentence-level BLEU $\mu_{s\text{BLEU}}$, the average document-level BLEU $\mu_{d\text{BLEU}}$, and the average COMET μ_{COMET} for the overall performance.

Inference We use beam search with the beam size of 5 during translation. As shown in Figure 1d, previous translations serve as the context for the current translation, so the test examples are translated in their original order, beginning with the first sentence free from context.

4 A Preliminary Study on Prompts

The prompt plays a crucial role in LLM research. Recent studies show that an optimal prompt can greatly enhance model performance and reveal unexpected model capabilities (Kojima et al., 2022; Wei et al., 2022b). Hence, our initial focus is on investigating the prompt's impact during fine-tuning.

Prompt Variations Displayed in Figure 1, our preliminary study features four prompt types. These designs aim to tackle two research questions: *How does context structure impact translation quality?* (Prompt 1 vs. Prompt 2) and *How do natural language instructions influence translation quality?* (Prompt 1 vs. Prompt 3). We also investigate the combined effect of these aspects in Prompt 4.

Results Our investigation analyzes prompt variations using three PEFT models (L-7B-LoRA, B-7B-LoRA, and V-7B-LoRA) on four English-centric translation tasks involving German and Chinese. Overall results are presented in Table 1. Comparing Prompt 1 (Figure 1a) and Prompt 2 (Figure 1b), we find that models fine-tuned with Prompt 2 generally outperform those with Prompt 1, indicating Prompt 2's effectiveness in enhancing LLM performance. Regarding our second research question (Figure 1a vs. Figure 1c), we observe varied performance. L-7B-LoRA and B-7B-LoRA perform better with Prompt 3, while V-7B-LoRA performs better with Prompt 1. These results highlight varying impacts of prompt variations across models and suggest natural language instructions are less effective when using instruction-tuned language models as model backbones. Finally, LLMs with Prompt 4 (Figure 1d) achieve the best overall performance, suggesting a positive compound effect of context structure and instructions.

Conclusion As expected, the prompt plays a significant role in LLM performance. A well-structured prompt, combining appropriate con-

[5] COMET signature: `Unbabel/wmt22-comet-da`.

	# of param.	# of train. param.	En-X			X-En		
			$\mu_{s\text{BLEU}}$	$\mu_{d\text{BLEU}}$	μ_{COMET}	$\mu_{s\text{BLEU}}$	$\mu_{d\text{BLEU}}$	μ_{COMET}
State-of-the-art SENMT *baselines*								
NLLB	600M	—	23.6	27.3	82.3	18.2	22.1	72.8
	1.3B	—	25.7	29.5	83.5	25.0	28.7	78.1
	3.3B	—	26.8	30.5	84.3	25.8	29.4	78.9
GOOGLETRANS	—	—	24.5	28.4	81.6	25.0	28.5	81.2
State-of-the-art LLMs								
GPT-3.5-TURBO	—	—	26.3	30.1	85.3	30.7	34.1	85.5
GPT-4-TURBO	—	—	**27.0**	**30.7**	**86.3**	**31.7**	**35.1**	**86.0**
LLM backbones								
LLAMA2-7B	—	—	2.7	3.5	40.1	4.2	4.4	52.2
BLOOM-7B	—	—	2.5	2.9	35.5	6.7	7.3	49.4
VICUNA-7B	—	—	10.2	12.4	64.7	9.5	9.8	62.9
Re-implemented DOCMT *baselines*								
DOC2DOC-MT5 (2017)	300M	300M	17.2	20.2	75.1	19.4	21.2	75.1
	580M	580M	18.6	21.5	78.3	20.7	22.5	77.4
	1.2B	1.2B	18.4	21.4	79.2	21.5	23.4	78.7
MR-DOC2SEN-MT5 (2022)	1.2B	1.2B	18.8	21.9	79.9	22.0	23.8	79.3
MR-DOC2DOC-MT5 (2022)	1.2B	1.2B	—	22.5	—	—	24.0	—
DOCFLAT-MT5 (2023)	1.2B	1.2B	19.2	22.4	80.2	22.2	24.3	79.3
IADA-MT5 (2024c)	1.2B	1.2B	19.3	22.4	80.4	22.1	24.0	79.5
LLM-based DOCMT *models (Ours)*								
L-7B-LORA	7B	8M	17.2	20.2	70.8	23.8	25.7	73.7
L-7B-FFT	7B	7B	13.7	16.2	67.4	22.4	24.1	74.0
B-7B-LORA	7B	8M	17.7	20.5	68.5	29.9	33.6	81.4
B-7B-FFT	7B	7B	12.0	13.8	59.6	22.3	24.5	69.9
V-7B-LORA	7B	8M	15.8	18.6	68.8	21.6	23.3	71.4
V-7B-FFT	7B	7B	14.3	16.8	65.0	21.8	23.5	74.3

Table 2: Overall performance on IWSLT2017. # of param. indicates the number of parameters of the model. # of train. param. indicates the number of trainable parameters of the model. All the LLM approaches use Prompt 4 (Figure 1d) during inference. Best results are highlighted in **bold**. Best results in each group are underlined.

text and clear instructions, can significantly boost model performance. In this work, we use Prompt 4 (Figure 1d) in our experiments.

5 Main Results

Overall Performance In our results presented in Table 2, we observe that GPT-4-TURBO and GPT-3.5-TURBO significantly outperform all other models. Notably, the NLLB variants also demonstrate superior performance among specialized machine translation (MT) models. In the context of DOCMT, conventional DOCMT models still outperform our LLM-based DOCMT models for translations from English to other languages when evaluated using standard MT metrics. Conversely, for translations from other languages to English, our LLM-based DOCMT models perform on par or better than conventional DOCMT models in $\mu_{s\text{BLEU}}$ and $\mu_{d\text{BLEU}}$ metrics, while conventional DOCMT models maintain superior performance in μ_{COMET}.

LLM-based DOCMT Models As shown in Table 2, models using LORA generally outperform fully fine-tuned (FFT) LLMs. However, V-7B-FFT surpasses V-7B-LORA in translating other languages to English, likely due to overfitting. Extensive fine-tuning with a large corpus often leads to rapid overfitting, whereas LORA, which updates only a few parameters, helps prevent this issue. Additionally, L-7B and V-7B models perform similarly, indicating that further fine-tuning with instruction-tuned models does not always improve task-specific performance.

Breakdown Performance We present the results for translation tasks from other languages to English in Figure 2. For clarity, only the results from our models using LORA are shown. Our LLM-based DOCMT models generally perform better, sometimes even surpassing GPT-4-TURBO in certain tasks, though they fail completely in others. A manual review indicates that the primary cause of these failures is *off-target translation*. An in-depth analysis of this issue is provided in Section 6, and a complete breakdown of the results is available in Appendix E.

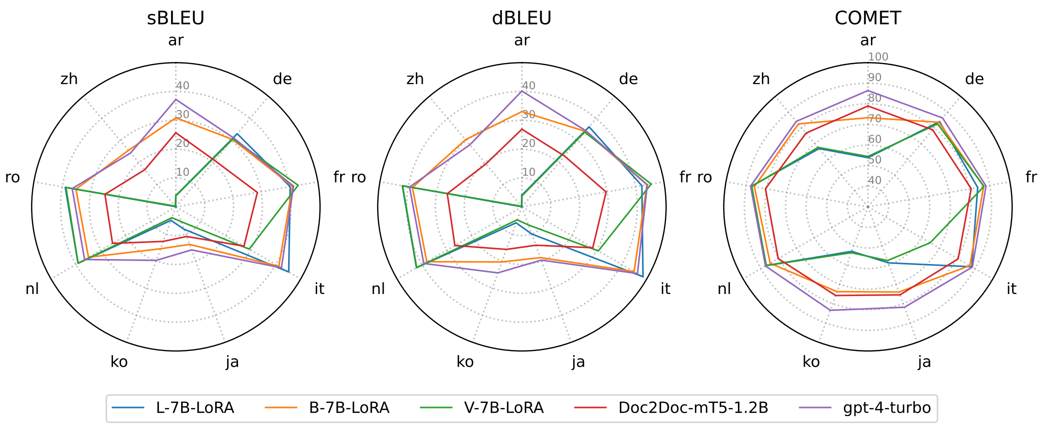

Figure 2: Breakdown results on sBLEU, dBLEU, and COMET given by L-7B-LoRA, V-7B-LoRA, B-7B-LoRA, Doc2Doc-mT5-1.2B, and GPT-4-TURBO for the translation tasks **from other languages to English**.

	$\mu_{\%}$	Ar	Ja	Ko	Zh
L-7B-LoRA	29.2	87.9	25.5	44.2	93.1
L-7B-FFT	40.2	87.9	75.5	92.3	93.6
B-7B-LoRA	2.8	2.9	4.0	8.4	1.6
B-7B-FFT	28.0	54.1	43.8	70.4	76.4
V-7B-LoRA	32.3	88.2	40.4	35.7	90.5
V-7B-FFT	44.7	94.1	98.3	96.6	94.6

Table 3: Off-target rate (%) provided by our LLM-based DocMT models for translation tasks from selective languages to English. $\mu_{\%}$ indicates the average off-target rate across all nine language pairs. **A lower off-target rate indicates better performance**. The complete results are presented in Appendix F.

	sBLEU	dBLEU	COMET
L-7B-LoRA + ReUse	3.9	4.1	53.9
L-7B-LoRA + ReGen	17.5	19.6	69.7
L-7B-FFT + ReUse	2.5	2.6	51.6
L-7B-FFT + ReGen	15.9	17.4	65.0

Table 4: Results on Arabic-English translation given by different decoding strategies.

6 Analysis

In this section, we examine off-target issues with decoding strategy and use GPT-4 to analyze translation errors. We also explore discourse phenomena, training strategies, the scaling law of parallel documents, and additional evaluations on recent WMT2023 test sets. Additionally, we evaluate the zero-shot crosslingual transfer capabilities of our models and present the results in Appendix I.

Off-Target Translation In Figure 2, our LLM-based DocMT models excel in some translation tasks but struggle in others due to off-target translation issues. We investigate this problem using the *fasttext* library (Bojanowski et al., 2017) to identify translation languages and quantify off-target rates, which represent the proportion of translations that are off-target. Results are presented in Table 3, with off-target rates reaching up to 98.3% in failing tasks. Notably, only B-7B-LoRA consistently maintains low off-target rates, likely due to Bloom-7B's multilingual pre-training. Detailed off-target rates are provided in Appendix F.

Decoding Strategy As described in Section 3.4, previous translations provide context for the current translation. We hypothesize that the high off-target translation rate in Table 3 is due to error propagation during the decoding stage. Therefore, we refer to the decoding strategy in Section 3.4 as ReUse, and introduce an alternative strategy where all context translations are re-generated, called ReGen. As shown in Table 4, the ReGen strategy significantly improves translation quality, confirming our hypothesis on decoding error propagation. However, the inference cost of ReGen is four times higher than that of ReUse. These findings highlight the main reason for translation failures in LLM-based DocMT models, offering insights for future research.

Translation Errors To understand the translation capabilities of our LLM-based DocMT models, we select specific error types from the Multidimensional Quality Metrics (MQM) framework (Burchardt, 2013). Kocmi and Federmann (2023) demonstrate that GPT-4 can identify error spans and achieve state-of-the-art MT evaluation accuracy, so we use GPT-4-TURBO to analyze translation errors in texts translated by these models.

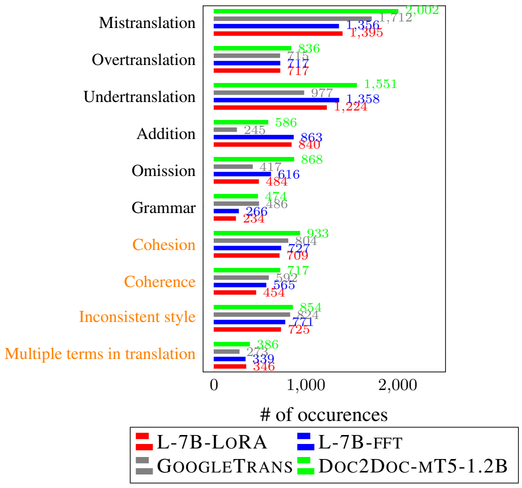

Figure 3: Error type analysis given by GPT-4-TURBO for translations from English to German, Romanian, and Chinese. The error types in orange are context-dependent. We omit those error types that are rare.

	En-De	En-Fr
Doc2Doc-mT5-1.2B	77.0	89.9
L-7B-LoRA	83.1	**95.1**
L-7B-FFT	81.1	94.6
B-7B-LoRA	75.5	91.9
B-7B-FFT	68.3	90.8
V-7B-LoRA	**84.9**	94.8
V-7B-FFT	84.4	94.8

Table 5: Accuracy (in %) on the English-German and English-French contrastive test sets. Best results are highlighted in **bold**.

Due to resource constraints, we focus on four models: L-7B-LoRA, L-7B-FFT, Doc2Doc-mT5-1.2B, and GoogleTrans, assessing translations from English to German, Romanian, and Chinese. The error identification prompt is detailed in Appendix D, and we present the frequency of error types in Figure 3. Most errors are limited to individual sentences. Despite similar scores in metrics such as sBLEU, dBLEU, and COMET among the models, our LLM-based DocMT models (L-7B-LoRA and L-7B-FFT) exhibit fewer context-independent and context-dependent errors. This suggests that current evaluation metrics may not sufficiently assess document-level translations and indicates that fine-tuning LLMs holds promise for enhancing DocMT performance.

Discourse Phenomena To evaluate our LLM-based DocMT model's ability to leverage contextual information, we assessed it using the English-German contrastive test set by Müller et al. (2018) and the English-French contrastive

	sBLEU	dBLEU	COMET
One-Stage			
Nl-En	30.2	32.4	71.2
Ro-En	28.3	30.1	72.3
Ar-En	1.3	2.1	50.1
Zh-En	0.1	0.1	60.9
Two-Stage			
Nl-En	38.9	41.9	87.0
Ro-En	38.2	41.4	87.3
Ar-En	2.5	2.6	51.6
Zh-En	0.1	0.1	67.1
Three-Stage			
Nl-En	39.1	42.1	87.0
Ro-En	38.4	41.6	87.3
Ar-En	2.3	2.4	52.4
Zh-En	0.3	0.3	67.4

Table 6: Results from one-stage, two-stage, and three-stage training strategies. The one-stage strategy involves directly fine-tuning LLaMA2-7B on parallel documents. The two-stage results are produced by L-7B-FFT. In the three-stage strategy, all model parameters of LLaMA2-7B are fine-tuned across all three stages.

test set by Lopes et al. (2020). This evaluation tests the model's accuracy in selecting the correct pronoun from multiple translation options. Results, shown in Table 5, reveal that models initialized with LLaMA2-7B and Vicuna-7B outperform Doc2Doc-mT5-1.2B, while Bloom-7B-initialized models perform worse, indicating that contextual understanding is mostly acquired during pre-training, as detailed by Scao et al. (2022) due to the lack of German text in Bloom pre-training. The *generative accuracy* results following Post and Junczys-Dowmunt (2023) are presented in Appendix G.

Training Strategy In this study, we follow the two-stage approach of Xu et al. (2023). Unlike traditional DocMT methods that typically start with parallel sentence training, we investigate the effectiveness of this strategy on LLM-based DocMT models. Specifically, we attempt to directly fine-tune the LLMs on parallel documents (one-stage) and add an extra fine-tuning stage using parallel sentences to the training strategy in Section 3.1 (three-stage). The results in Table 6 indicate that both the one-stage and three-stage training strategies are sub-optimal for both high-performing languages (Dutch and Romanian) and low-performing languages (Arabic and Chinese) with LLM-based DocMT models.

Scaling Law of Parallel Documents We explore the scaling law for fine-tuning parallel documents.

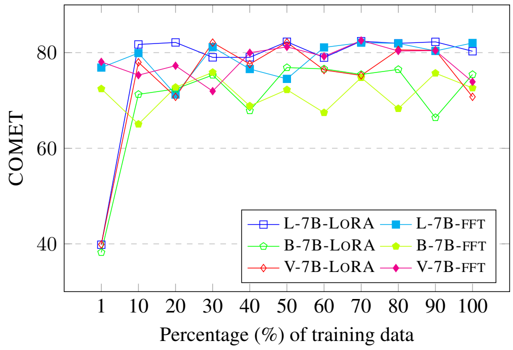

Figure 4: COMET-Percentage (%) of training data for the translations from English to German.

Results for English-German translation are presented in Figure 4, and for English-Romanian and English-Chinese in Appendix H. While LLMs typically excel with minimal training data, different fine-tuning strategies show distinct scaling behaviors. Our LoRA models match full training set performance with just 10% of the data (around $20K$ examples), while fully fine-tuned models achieve near-equivalent performance with only about 1% of the data (approximately $2K$ examples). These are crucial for low-resource languages, as recent LLMs are mainly pre-trained on English text.

Evaluation on Recent Test Sets Given their extensive pre-training on large text corpora, LLMs are susceptible to data leakage risks. We evaluate our models using recent test sets from WMT2023 (Koehn et al., 2023). These tests, conducted between English and German, assess the out-of-domain generalization of our models and help mitigate data leakage risks. We use spaCy to segment documents and discard any parallel documents where the source and target sides have a differing number of sentences. Our findings, presented in Table 7, reveal that while DocNMT5 models outperform LLM-based models in Table 2, LLM-based models excel in translating out-of-domain text on the WMT2023 test sets. These results highlight the ability of LLM-based DocNMT to generalize well to out-of-domain translation tasks.

Zero-Shot Crosslingual Transfer We also explore the transferability of translation capabilities acquired from one language pair to others. We assess our English-German LLM-based DocMT models on English-to-other-language test sets, comparing their COMET scores to their base models in Table 8. Our results indicate that models with fine-tuned instructions (LLAMA2-7B and BLOOM-7B) consistently exhibit positive transfer effects across all language pairs, while those with instruction-tuned models (VICUNA-7B) benefits only few languages.

	En-De		De-En	
	dBLEU	COMET	dBLEU	COMET
DOC2DOC-MT5-1.2B	20.2	74.4	20.0	76.5
MR-DOC2SEN-MT5	20.5	74.9	21.0	76.5
MR-DOC2DOC-MT5	21.2	75.6	21.5	76.5
DOCFLAT-MT5	20.9	75.1	21.8	76.5
IADA-MT5	21.2	75.4	22.0	76.5
L-7B-LoRA	28.9	76.4	35.5	83.2
L-7B-FFT	**29.0**	**77.0**	**36.1**	**84.0**
B-7B-LoRA	23.7	73.0	30.5	80.8
B-7B-FFT	21.0	69.0	30.0	80.5
V-7B-LoRA	20.5	63.8	33.9	81.8
V-7B-FFT	27.8	75.0	34.7	83.1

Table 7: dBLEU and COMET on WMT2023 test sets. Best results are highlighted in **bold**.

	μ_Δ	Ar	De	Fr	Zh
L-7B-LoRA	+29.4	+36.3	+38.8	+37.2	+29.5
L-7B-FFT	+29.0	+41.2	+40.5	+37.1	+37.5
B-7B-LoRA	+20.3	+7.5	+40.7	+20.7	+9.8
B-7B-FFT	+27.3	+14.8	+37.8	+28.9	+19.5
V-7B-LoRA	-8.9	-12.6	+22.1	+18.9	-34.1
V-7B-FFT	-1.4	+7.3	+25.2	+17.7	-3.5

Table 8: The difference (Δ) in COMET scores on the test sets from English to other languages between our English-German LLM-based DocMT models and their backbones. μ_Δ indicates the average difference across all the languages in this table. The complete results are presented in Appendix I.

7 Conclusion

This study investigates the adaptation of LLMs for DocMT through extensive experimentation. The research involves two fine-tuning methods, three LLM backbones, and 18 translation tasks across nine language pairs. Our findings indicate that while specialized models sometimes surpass GPT-4 in translation accuracy, they continue to face challenges such as off-target translation due to error propagation during decoding. We present a detailed analysis of these LLMs tailored for DocMT, exploring aspects such as translation errors, discourse phenomena, training strategies, the scaling law of parallel documents, recent test set evaluations, and zero-shot crosslingual transfer. The results highlight both the strengths and limitations of LLM-based DocMT models, offering a foundation for future research in this field.

8 Limitations

Constraints on Model Scale Our research is confined to language models of a moderate size, specifically those with $7B$ parameters. This limitation is due to the constraints of our available resources. Consequently, it is crucial to acknowledge that the outcomes of our study might vary if conducted with larger models.

Instability in Training The process of supervised fine-tuning for LLMs shows instability in our observations. As detailed in Figure 4, there are noticeable inconsistencies in performance. These variations are too significant to attribute solely to the randomness inherent in training. In some cases, the fine-tuning of LLMs fails to reach convergence. Unfortunately, our limited resources restrict us from investigating these failures in depth or devising potential remedies.

Influence of Prompting Techniques Section 4 of our study highlights the significant role of prompting methods in fine-tuning. We experiment with four different prompting techniques. It is important to note that the prompt we recommend may not be the most effective, potentially leading to suboptimal performance of our models.

We acknowledge these limitations and leave them to the future work.

References

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernández Ábrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan A. Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vladimir Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, and et al. 2023. Palm 2 technical report. *CoRR*, abs/2305.10403.

Zhangir Azerbayev, Hailey Schoelkopf, Keiran Paster, Marco Dos Santos, Stephen McAleer, Albert Q. Jiang, Jia Deng, Stella Biderman, and Sean Welleck. 2023. Llemma: An open language model for mathematics. *CoRR*, abs/2310.10631.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313, New Orleans, Louisiana. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Aljoscha Burchardt. 2013. Multidimensional quality metrics: a flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*, London, UK. Aslib.

Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Jan Niehues, Sebastian Stüker, Katsuhito Sudoh, Koichiro Yoshino, and Christian Federmann. 2017. Overview of the IWSLT 2017 evaluation campaign. In *Proceedings of the 14th International Conference on Spoken Language Translation*, pages 2–14, Tokyo, Japan. International Workshop on Spoken Language Translation.

Pinzhen Chen, Shaoxiong Ji, Nikolay Bogoychev, Barry Haddow, and Kenneth Heafield. 2023. Monolingual or multilingual instruction tuning: Which makes a better alpaca. *CoRR*, abs/2309.08958.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways. *CoRR*, abs/2204.02311.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. 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*, abs/2210.11416.

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, Loïc 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*, abs/2207.04672.

Yukun Feng, Feng Li, Ziang Song, Boyuan Zheng, and Philipp Koehn. 2022. Learn to remember: Transformer with recurrent memory for document-level machine translation. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1409–1420, Seattle, United States. Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. 2023. Textbooks are all you need. *CoRR*, abs/2306.11644.

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*, abs/2302.09210.

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 *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Jingjing Huo, Christian Herold, Yingbo Gao, Leonard Dahlmann, Shahram Khadivi, and Hermann Ney. 2020. Diving deep into context-aware neural machine translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 604–616, Online. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is chatgpt A good translator? A preliminary study. *CoRR*, abs/2301.08745.

Tom Kocmi and Christian Federmann. 2023. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore. Association for Computational Linguistics.

Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz, editors. 2023. *Proceedings of the Eighth Conference on Machine Translation*. Association for Computational Linguistics, Singapore.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *NeurIPS*.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Christopher A. Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. MADLAD-400: A multilingual and document-level large audited dataset. *CoRR*, abs/2309.04662.

Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. 2023. Bactrian-x : A multilingual replicable instruction-following model with low-rank adaptation. *CoRR*, abs/2305.15011.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pretraining for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

António Lopes, M. Amin Farajian, Rachel Bawden, Michael Zhang, and André F. T. Martins. 2020. Document-level neural MT: A systematic comparison. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 225–234, Lisboa, Portugal. European Association for Machine Translation.

Hongyuan Lu, Haoyang Huang, Dongdong Zhang, Haoran Yang, Wai Lam, and Furu Wei. 2023. Chain-of-dictionary prompting elicits translation in large language models. *CoRR*, abs/2305.06575.

Ziyang Luo, Can Xu, Pu Zhao, Qingfeng Sun, Xiubo Geng, Wenxiang Hu, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. 2023. Wizardcoder: Empowering code large language models with evol-instruct. *CoRR*, abs/2306.08568.

Shuming Ma, Dongdong Zhang, and Ming Zhou. 2020. A simple and effective unified encoder for document-level machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3505–3511, Online. Association for Computational Linguistics.

Valentin Macé and Christophe Servan. 2019. Using whole document context in neural machine translation. In *Proceedings of the 16th International Conference on Spoken Language Translation*, Hong Kong. Association for Computational Linguistics.

Max Marion, Ahmet Üstün, Luiza Pozzobon, Alex Wang, Marzieh Fadaee, and Sara Hooker. 2023. When less is more: Investigating data pruning for pretraining llms at scale. *CoRR*, abs/2309.04564.

Sameen Maruf and Gholamreza Haffari. 2018. Document context neural machine translation with memory networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1275–1284, Melbourne, Australia. Association for Computational Linguistics.

Sameen Maruf, André F. T. Martins, and Gholamreza Haffari. 2019. Selective attention for context-aware 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 3092–3102, Minneapolis, Minnesota. Association for Computational Linguistics.

Sameen Maruf, Fahimeh Saleh, and Gholamreza Haffari. 2022. A survey on document-level neural machine translation: Methods and evaluation. *ACM Comput. Surv.*, 54(2):45:1–45:36.

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-level neural machine translation with hierarchical attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2947–2954, Brussels, Belgium. Association for Computational Linguistics.

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)*, pages 3470–3487, Dublin, Ireland. Association for Computational Linguistics.

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

Mathias Müller, Annette Rios, Elena Voita, and Rico Sennrich. 2018. A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 61–72, Brussels, Belgium. Association for Computational Linguistics.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2023. Culturax: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. *CoRR*, abs/2309.09400.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*.

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*, pages 311–318, Philadelphia, Pennsylvania, USA. 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*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Matt Post and Marcin Junczys-Dowmunt. 2023. Escaping the sentence-level paradigm in machine translation. *CoRR*, abs/2304.12959.

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

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

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal V. Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Févry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, and et al. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100.

Sheng Shen, Le Hou, Yanqi Zhou, Nan Du, Shayne Longpre, Jason Wei, Hyung Won Chung, Barret Zoph, William Fedus, Xinyun Chen, Tu Vu, Yuexin Wu, Wuyang Chen, Albert Webson, Yunxuan Li, Vincent Zhao, Hongkun Yu, Kurt Keutzer, Trevor Darrell, and Denny Zhou. 2023. Flan-moe: Scaling instruction-finetuned language models with sparse mixture of experts. *CoRR*, abs/2305.14705.

Zewei Sun, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Shujian Huang, Jiajun Chen, and Lei Li. 2022. Rethinking document-level neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3537–3548, Dublin, Ireland. Association for Computational Linguistics.

Jörg Tiedemann and Yves Scherrer. 2017. Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 82–92, Copenhagen, Denmark. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

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

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

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212, Florence, Italy. Association for Computational Linguistics.

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. Context-aware neural machine translation learns anaphora resolution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1264–1274, Melbourne, Australia. Association for Computational Linguistics.

Longyue Wang, Siyou Liu, Mingzhou Xu, Linfeng Song, Shuming Shi, and Zhaopeng Tu. 2023a. A survey on zero pronoun translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3325–3339, Toronto, Canada. Association for Computational Linguistics.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023b.

Document-level machine translation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16646–16661, Singapore. Association for Computational Linguistics.

Longyue Wang, Zhaopeng Tu, Andy Way, and Qun Liu. 2017. Exploiting cross-sentence context for neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2826–2831, Copenhagen, Denmark. Association for Computational Linguistics.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. 2022. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022a. Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022b. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*.

KayYen Wong, Sameen Maruf, and Gholamreza Haffari. 2020. Contextual neural machine translation improves translation of cataphoric pronouns. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5971–5978, Online. Association for Computational Linguistics.

Minghao Wu and Alham Fikri Aji. 2023. Style over substance: Evaluation biases for large language models. *CoRR*, abs/2307.03025.

Minghao Wu, George Foster, Lizhen Qu, and Gholamreza Haffari. 2023. Document flattening: Beyond concatenating context for document-level neural machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 448–462, Dubrovnik, Croatia. Association for Computational Linguistics.

Minghao Wu, Abdul Waheed, Chiyu Zhang, Muhammad Abdul-Mageed, and Alham Aji. 2024a. LaMini-LM: 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 (Volume 1: Long Papers)*, pages 944–964, St. Julian's, Malta. Association for Computational Linguistics.

Minghao Wu, Yufei Wang, George Foster, Lizhen Qu, and Gholamreza Haffari. 2024b. Importance-aware data augmentation for document-level neural machine translation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 740–752, St. Julian's, Malta. Association for Computational Linguistics.

Minghao Wu, Yufei Wang, George Foster, Lizhen Qu, and Gholamreza Haffari. 2024c. Importance-aware data augmentation for document-level neural machine translation. *arXiv preprint arXiv:2401.15360*.

Minghao Wu, Yulin Yuan, Gholamreza Haffari, and Longyue Wang. 2024d. (perhaps) beyond human translation: Harnessing multi-agent collaboration for translating ultra-long literary texts. *arXiv preprint arXiv:2405.11804*.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2023. A paradigm shift in machine translation: Boosting translation performance of large language models. *CoRR*, abs/2309.11674.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Wen Yang, Chong Li, Jiajun Zhang, and Chengqing Zong. 2023. Bigtrans: Augmenting large language models with multilingual translation capability over 100 languages. *CoRR*, abs/2305.18098.

Biao Zhang, Ankur Bapna, Melvin Johnson, Ali Dabirmoghaddam, Naveen Arivazhagan, and Orhan Firat. 2022. Multilingual document-level translation enables zero-shot transfer from sentences to documents. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4176–4192, Dublin, Ireland. Association for Computational Linguistics.

Jiacheng Zhang, Huanbo Luan, Maosong Sun, Feifei Zhai, Jingfang Xu, Min Zhang, and Yang Liu. 2018. Improving the transformer translation model with document-level context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 533–542, Brussels, Belgium. Association for Computational Linguistics.

Pei Zhang, Boxing Chen, Niyu Ge, and Kai Fan. 2020. Long-short term masking transformer: A simple but effective baseline for document-level neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1081–1087, Online. Association for Computational Linguistics.

Shaolei Zhang, Qingkai Fang, Zhuocheng Zhang, Zhengrui Ma, Yan Zhou, Langlin Huang, Mengyu Bu, Shangtong Gui, Yunji Chen, Xilin Chen, and Yang Feng. 2023. Bayling: Bridging cross-lingual alignment and instruction following through interactive translation for large language models. *CoRR*, abs/2306.10968.

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*, abs/2304.04675.

	Train		Validation		Test	
	# of sen.	# of doc.	# of sen.	# of doc.	# of sen.	# of doc.
En-Ar	232K	1907	2453	19	1460	12
En-De	206K	1705	2456	19	1138	10
En-Fr	233K	1914	2458	19	1455	12
En-It	232K	1902	2495	19	1147	10
En-Ja	223K	1863	2420	19	1452	12
En-Ko	230K	1920	2437	19	1429	12
En-Nl	237K	1805	2780	19	1181	10
En-Ro	221K	1812	2592	19	1129	10
En-Zh	231K	1906	2436	19	1459	12

Table 9: Dataset statistics of parallel documents.

A Statistics of Parallel Documents

We present the dataset statistics of parallel documents in Table 9.

B Optimization and Hyperparameters

Fine-tuning on Monolingual Documents We fine-tune all the parameters of large language models (LLMs) using a learning rate of 5×10^{-5} and a batch size of 256. During the training process, we apply the linear learning rate schedule, which includes a warm-up phase comprising 10% of the total training steps.

Fine-tuning on Parallel Documents When fine-tuning L-7B-LoRA and V-7B-LoRA on parallel documents, we use a learning rate of 5×10^{-5} and a batch size of 64. We apply a linear learning rate schedule with a 10% warm-up phase. The LoRA rank is set to 16, affecting only 0.1% of the parameters (about 8M parameters). The same hyperparameters are used for fine-tuning Doc2Doc-mT5 models, except for a learning rate of 5×10^{-4}. L-7B-LoRA and V-7B-LoRA are fine-tuned for up to 3 epochs, and Doc2Doc-mT5 models for up to 10 epochs. Early stopping is based on validation loss.

C Prompt Types

We present concrete examples of prompt variations in Figure 5.

D GPT-4 Prompts

We present the prompts used for error type analysis in Figure 6.

E Breakdown Results

We provide detailed breakdowns of the translation tasks from English to other languages, evaluated using sBLEU, dBLEU, and COMET. These are presented in Table 10, Table 11, and Table 12, respectively. Additionally, we present similar breakdowns for translations from other languages to English, assessed using the same metrics. These results can be found in Table 13, Table 14, and Table 15.

F Off-Target Translation

We present the complete results on the off-target translation problem in Table 16 and Table 17.

G Discourse Phenomena

Post and Junczys-Dowmunt (2023) propose to evaluate the accuracy on the constrastive test sets in a generative way. Hence, we present the generative accuracy in Table 18.

H Scaling Law of Parallel Documents from English to Romanian and Chinese

In Section 6, we find that our LLM-based DocMT models are highly efficient in terms of the amount of training data. To confirm our findings in Section 6, we conduct additional experiments on the translation tasks from English to Romanian and Chinese. As shown in Figure 7, we can confirm the superiority of LLM-based DocMT models with regard to data efficiency.

I Zero-Shot Crosslingual Transfer

We present the complete results on zero-shot crosslingual transfer in Table 19.

	$\mu_{s\text{BLEU}}$	Ar	De	Fr	It	Ja	Ko	Nl	Ro	Zh
NLLB-600M	23.6	14.2	22.2	38.5	36.0	11.4	16.2	30.2	25.6	17.6
NLLB-1.3B	25.7	16.2	27.6	40.6	37.7	12.6	18.6	32.2	27.3	18.3
NLLB-3.3B	26.8	17.4	28.8	41.3	39.2	14.1	19.5	33.7	28.1	18.7
GOOGLETRANS	24.5	14.2	25.3	38.0	35.0	11.6	16.5	29.6	24.0	26.4
GPT-3.5-TURBO	26.3	14.9	27.2	40.5	36.6	13.2	15.9	31.5	26.6	30.4
GPT-4-TURBO	27.0	16.1	27.4	40.0	35.8	14.1	18.3	32.2	27.3	31.6
LLAMA2-7B	2.8	0.4	6.6	4.5	1.0	1.0	1.9	0.2	1.9	7.4
BLOOM-7B	2.5	1.0	1.0	12.1	1.4	0.1	3.1	0.7	0.1	3.4
VICUNA-7B	10.2	4.5	6.4	6.4	8.6	10.2	9.8	13.9	6.8	25.4
DOC2DOC-MT5-300M	17.2	9.4	16.8	24.0	21.0	11.0	13.7	20.5	17.1	21.6
DOC2DOC-MT5-580M	18.6	10.8	18.2	24.9	23.0	12.9	15.2	21.7	17.8	22.9
DOC2DOC-MT5-1.2B	18.4	10.3	18.1	24.9	22.4	13.9	15.4	19.6	18.8	22.6
MR-DOC2SEN-MT5 -1.2B	18.8	10.2	18.8	25.6	22.3	14.5	16.2	19.6	19.3	22.8
MR-DOC2DOC-MT5 -1.2B	—	—	—	—	—	—	—	—	—	—
DOCFLAT-MT5 -1.2B	19.2	11.0	19.2	25.7	22.6	14.7	16.5	20.3	19.2	23.8
IADA-MT5 -1.2B	19.3	11.7	19.4	26.3	23.9	15.2	16.9	20.9	19.6	23.4
L-7B-LORA	17.2	13.0	25.1	34.9	6.8	8.7	13.0	3.7	22.7	27.3
L-7B-FFT	13.7	13.1	25.3	19.5	2.6	7.9	7.2	4.1	21.1	22.8
B-7B-LORA	17.7	12.1	20.6	32.6	32.9	3.6	1.4	28.1	12.2	15.7
B-7B-FFT	12.0	10.1	19.6	38.5	0.1	1.9	2.4	1.5	19.9	14.5
V-7B-LORA	16.4	13.3	20.1	20.7	13.6	9.1	14.3	5.5	23.0	28.1
V-7B-FFT	14.3	13.5	23.3	21.1	4.8	3.8	15.9	3.3	17.4	25.8

Table 10: Breakdown sBLEU results for the translation tasks from English to other languages.

	$\mu_{d\text{BLEU}}$	Ar	De	Fr	It	Ja	Ko	Nl	Ro	Zh
NLLB-600M	27.3	15.4	26.0	42.1	39.0	17.1	23.8	33.8	28.5	20.2
NLLB-1.3B	29.5	17.4	31.8	44.0	40.8	18.1	26.6	35.8	30.2	21.0
NLLB-3.3B	30.5	18.6	32.9	44.6	42.2	19.8	27.4	37.0	30.9	21.2
GOOGLETRANS	28.4	16.0	29.3	41.3	38.5	15.7	23.4	32.8	26.7	32.1
GPT-3.5-TURBO	30.1	16.4	30.9	43.7	39.7	17.5	22.7	34.5	29.0	36.3
GPT-4-TURBO	30.7	17.4	31.1	43.2	39.0	18.4	25.3	35.3	29.8	37.2
LLAMA2-7B	3.5	0.5	7.4	4.9	1.1	1.8	3.7	0.2	2.2	9.6
BLOOM-7B	2.8	1.0	1.3	12.9	1.7	0.3	2.7	1.0	0.1	4.4
VICUNA-7B	12.4	5.7	6.4	7.1	8.5	15.0	16.1	14.4	7.4	31.2
DOC2DOC-MT5-300M	20.2	10.3	18.8	26.1	21.9	16.8	21.5	21.5	18.4	26.6
DOC2DOC-MT5-580M	21.5	11.7	20.0	27.0	23.9	18.4	23.2	22.6	18.8	28.0
DOC2DOC-MT5-1.2B	21.4	11.2	20.1	27.1	23.4	19.7	23.0	20.7	20.2	27.1
MR-DOC2SEN-MT5 -1.2B	21.9	11.9	20.7	27.9	23.8	19.8	23.3	21.5	20.7	27.9
MR-DOC2DOC-MT5 -1.2B	22.5	12.1	20.8	28.0	24.7	20.9	24.3	21.9	21.7	27.9
DOCFLAT-MT5 -1.2B	22.4	12.2	21.3	28.4	24.0	20.8	24.1	21.1	21.4	28.1
IADA-MT5 -1.2B	22.4	12.7	21.7	28.7	24.2	21.4	24.3	21.6	21.7	28.0
L-7B-LORA	20.2	14.7	29.1	37.5	7.3	13.9	19.5	4.2	22.9	33.1
L-7B-FFT	16.2	14.7	29.4	20.6	2.7	12.5	12.3	4.5	21.6	27.6
B-7B-LORA	20.5	13.7	24.8	36.1	36.3	6.9	2.6	32.2	12.2	19.7
B-7B-FFT	13.8	11.2	23.6	41.7	0.1	3.7	3.9	1.7	20.1	18.1
V-7B-LORA	19.3	14.9	23.1	21.8	14.7	14.3	21.6	5.9	23.3	34.2
V-7B-FFT	16.8	15.2	26.9	22.3	4.9	6.2	23.6	3.7	17.5	31.1

Table 11: Breakdown dBLEU results for the translation tasks from English to other languages.

```
[English Context]: And it's truly a great
↪  honor to have the opportunity to come to
↪  this stage twice; I'm extremely grateful.
↪  I have been blown away by this conference,
↪  and I want to thank all of you for the
↪  many nice comments about what I had to say
↪  the other night. And I say that sincerely,
↪  partly because  I need that.
[German Context]: Es ist mir wirklich eine
↪  Ehre, zweimal auf dieser Bühne stehen zu
↪  dürfen. Tausend Dank dafür. Ich bin
↪  wirklich begeistert von dieser Konferenz,
↪  und ich danke Ihnen allen für die vielen
↪  netten Kommentare zu meiner Rede
↪  vorgestern Abend. Das meine ich ernst,
↪  teilweise deshalb -- weil ich es wirklich
↪  brauchen kann!
[English Sentence]: Put yourselves in my
↪  position.
[German Sentence]: Versetzen Sie sich mal in
↪  meine Lage!
```

(a) Prompt 1

```
[English]: And it's truly a great honor to
↪  have the opportunity to come to this stage
↪  twice; I'm extremely grateful.
[German]: Es ist mir wirklich eine Ehre,
↪  zweimal auf dieser Bühne stehen zu dürfen.
↪  Tausend Dank dafür.
[English]: I have been blown away by this
↪  conference, and I want to thank all of you
↪  for the many nice comments about what I
↪  had to say the other night.
[German]: Ich bin wirklich begeistert von
↪  dieser Konferenz, und ich danke Ihnen
↪  allen für die vielen netten Kommentare zu
↪  meiner Rede vorgestern Abend.
[English]: And I say that sincerely, partly
↪  because  I need that.
[German]: Das meine ich ernst, teilweise
↪  deshalb -- weil ich es wirklich brauchen
↪  kann!
[English]: Put yourselves in my position.
[German]: Versetzen Sie sich mal in meine
↪  Lage!
```

(b) Prompt 2

```
[English Context]: And it's truly a great
↪  honor to have the opportunity to come to
↪  this stage twice; I'm extremely grateful.
↪  I have been blown away by this conference,
↪  and I want to thank all of you for the
↪  many nice comments about what I had to say
↪  the other night. And I say that sincerely,
↪  partly because  I need that.
[German Context]: Es ist mir wirklich eine
↪  Ehre, zweimal auf dieser Bühne stehen zu
↪  dürfen. Tausend Dank dafür. Ich bin
↪  wirklich begeistert von dieser Konferenz,
↪  und ich danke Ihnen allen für die vielen
↪  netten Kommentare zu meiner Rede
↪  vorgestern Abend. Das meine ich ernst,
↪  teilweise deshalb -- weil ich es wirklich
↪  brauchen kann!
Given the provided parallel context, translate
↪  the following English sentence to German:
[English Sentence]: Put yourselves in my
↪  position.
[German Sentence]: Versetzen Sie sich mal in
↪  meine Lage!
```

(c) Prompt 3

```
[English]: And it's truly a great honor to
↪  have the opportunity to come to this stage
↪  twice; I'm extremely grateful.
[German]: Es ist mir wirklich eine Ehre,
↪  zweimal auf dieser Bühne stehen zu dürfen.
↪  Tausend Dank dafür.
[English]: I have been blown away by this
↪  conference, and I want to thank all of you
↪  for the many nice comments about what I
↪  had to say the other night.
[German]: Ich bin wirklich begeistert von
↪  dieser Konferenz, und ich danke Ihnen
↪  allen für die vielen netten Kommentare zu
↪  meiner Rede vorgestern Abend.
[English]: And I say that sincerely, partly
↪  because  I need that.
[German]: Das meine ich ernst, teilweise
↪  deshalb -- weil ich es wirklich brauchen
↪  kann!
Given the provided parallel sentence pairs,
↪  translate the following English sentence
↪  to German:
[English]: Put yourselves in my position.
[German]: Versetzen Sie sich mal in meine
↪  Lage!
```

(d) Prompt 4

Figure 5: Prompt types used in the preliminary study. <src_lang> and <tgt_lang> indicate the language IDs. <src*> and <tgt*> indicate the source and target sentences. **Note that the target sentences <tgt*> are only used during training and are replaced with the hypotheses <hyp*> generated by the model during inference.**

```
[Context]:

[Source]: <src1>
[Reference]: <tgt1>
[Hypothesis]: <hyp1>
[Source]: <src2>
[Reference]: <tgt2>
[Hypothesis]: <hyp2>
[Source]: <src3>
[Reference]: <tgt3>
[Hypothesis]: <hyp3>

[Current Sentence]:

[Source]: <src4>
[Reference]: <tgt4>
[Hypothesis]: <hyp4>

[Error Types]:

- Mistranslation: Error occuring when the target content does not accurately represent the source
↪ content.
- Overtranslation: Error occuring in the target content that is inappropriately more specific than
↪ the source content.
- Undertranslation: Error occuring in the target content that is inappropriately less specific than
↪ the source content.
- Addition: Error occuring in the target content that includes content not present in the source.
- Omission: Error where content present in the source is missing in the target.
- Unjustfied euphemism: Target content that is potentially offensive in some way in the source
↪ language, but that has been inappropriately "watered down" in the translation.
- Do not translate: Error occuring when a text segment marked "Do not translate!" is translated in
↪ the target text.
- Untranslated: Error occuring when a text segment that was intended for translation is omitted in
↪ the target content.
- Retained factual error: Untrue statement or an incorrect data value present in the source content
↪ and retained in the target content.
- Completeness: Source text incomplete, resulting in instances where needed content is missing in
↪ the source language.
- Grammar: Error that occurs when a text string (sentence, phrase, other) in the translation
↪ violates the grammatical rules of the target language.
- Punctuation: Punctuation incorrect according to target language conventions.
- Spelling: Error occurring when a word is misspelled.
- Duplication: Content (e.g., a word or longer portion of text) repeated unintentionally.
- Unclear reference: Relative pronouns or other referential mechanisms unclear in their reference.
- Cohesion: Portions of the text needed to connect it into an understandable whole (e.g., reference,
↪ substitution, ellipsis, conjunction, and lexical cohesion) missing or incorrect.
- Coherence: Text lacking a clear semantic relationship between its parts, i.e., the different parts
↪ don't hang together, don't follow the discourse conventions of the target language, or don't
↪ "make sense."
- Inconsistent style: Style that varies inconsistently throughout the text, e.g., One part of a text
↪ is written in a clear, "terse" style, while other sections are written in a more wordy style.
- Multiple terms in translation: Error where source content terminology is correct, but target
↪ content terms are not used consistently.

Considering the provided context, please identify the errors of the translation from the source to
↪ the target in the current sentence based on a subset of Multidimensional Quality Metrics (MQM)
↪ error typology.
You should pay extra attention to the error types related to the relationship between the current
↪ sentence and its context, such as "Unclear reference", "Cohesion", "Coherence", "Inconsistent
↪ style", and "Multiple terms in translation".
You should list all the errors you find in the sentence, and provide a justification for each error.
Your output should always be in JSON format, formatted as follows: {'justification': '...',
↪ 'error_types': [...]}.
```

Figure 6: Prompt used for analyzing translation error types.

	μ_{COMET}	Ar	De	Fr	It	Ja	Ko	Nl	Ro	Zh
NLLB-600M	82.3	82.6	81.3	83.7	86.3	79.5	82.5	84.0	85.0	75.4
NLLB-1.3B	83.5	84.3	83.0	84.8	87.5	80.1	84.2	85.2	86.3	76.4
NLLB-3.3B	84.3	84.8	84.1	85.3	88.0	82.1	85.2	86.0	86.4	76.7
GOOGLETRANS	81.6	81.5	80.2	82.3	85.1	80.7	79.9	83.7	82.9	78.5
GPT-3.5-TURBO	85.3	83.8	84.6	85.9	87.7	84.7	84.2	86.3	86.5	83.9
GPT-4-TURBO	86.3	85.4	85.4	86.2	87.8	86.0	86.4	87.0	87.4	84.8
LLAMA2-7B	40.1	37.4	41.5	41.2	39.3	39.7	42.8	35.5	41.6	42.1
BLOOM-7B	35.5	34.9	34.8	45.3	33.0	33.9	35.5	34.2	28.6	39.0
VICUNA-7B	64.7	68.3	48.7	49.0	62.5	81.3	72.4	64.1	56.1	80.2
DOC2DOC-MT5-300M	75.1	77.2	71.0	72.4	74.1	78.5	77.8	74.3	74.0	77.0
DOC2DOC-MT5-580M	78.3	80.8	74.4	74.8	77.9	81.5	82.0	76.8	76.6	79.6
DOC2DOC-MT5-1.2B	79.2	81.1	75.9	75.9	78.9	82.9	82.4	76.5	79.3	80.3
MR-DOC2SEN-MT5 -1.2B	79.9	82.2	76.3	76.5	80.0	83.6	83.1	76.7	79.7	80.9
MR-DOC2DOC-MT5 -1.2B	—	—	—	—	—	—	—	—	—	—
DOCFLAT-MT5 -1.2B	80.4	81.8	77.3	76.9	80.3	83.6	83.4	78.0	80.7	81.8
IADA-MT5 -1.2B	80.7	82.4	77.0	77.3	80.9	84.1	83.7	77.8	80.9	81.8
L-7B-LORA	70.8	82.5	80.3	79.1	42.9	70.4	75.4	42.5	83.6	80.6
L-7B-FFT	67.4	83.1	82.0	59.4	38.8	65.8	69.7	49.2	82.7	75.5
B-7B-LORA	68.5	77.0	75.4	76.8	85.1	51.4	40.4	82.9	61.7	65.5
B-7B-FFT	59.6	68.4	72.6	83.8	45.3	40.3	45.2	46.8	71.6	62.2
V-7B-LORA	69.7	82.7	70.8	60.1	56.2	70.2	76.9	44.1	85.0	81.3
V-7B-FFT	65.0	83.1	73.9	58.8	41.5	54.8	81.1	42.4	69.6	79.4

Table 12: Breakdown COMET results for the translation tasks from English to other languages.

	$\mu_{s\text{BLEU}}$	Ar	De	Fr	It	Ja	Ko	Nl	Ro	Zh
NLLB-600M	18.2	26.8	11.0	31.0	22.7	10.9	13.6	17.1	13.2	17.9
NLLB-1.3B	25.0	35.9	18.8	37.4	35.3	13.2	15.7	26.0	23.1	19.9
NLLB-3.3B	25.8	36.5	22.3	36.8	33.5	12.4	18.5	28.3	25.2	19.1
GOOGLETRANS	25.0	28.7	26.1	34.7	35.1	10.2	13.3	30.8	29.6	16.6
GPT-3.5-TURBO	30.7	35.8	30.8	40.7	41.8	15.5	17.3	36.1	35.4	22.9
GPT-4-TURBO	31.7	37.2	31.2	41.4	42.3	15.9	19.8	36.6	36.5	24.4
LLAMA2-7B	4.2	0.1	6.2	1.3	4.4	0.0	0.1	15.1	10.3	0.1
BLOOM-7B	6.7	4.7	8.7	14.3	16.9	0.1	0.3	9.4	5.5	0.2
VICUNA-7B	9.5	1.1	14.4	24.9	14.1	4.7	0.4	17.2	8.5	0.0
DOC2DOC-MT5-300M	19.4	23.0	19.5	26.6	25.4	9.5	11.4	22.0	22.6	14.5
DOC2DOC-MT5-580M	20.7	24.2	20.3	28.0	26.6	11.0	11.6	24.4	23.8	16.1
DOC2DOC-MT5-1.2B	21.5	25.7	21.0	28.7	27.3	11.0	12.8	25.4	25.0	16.8
MR-DOC2SEN-MT5 -1.2B	22.0	26.9	22	29.9	27.7	11.8	13.9	26.5	26.1	18
MR-DOC2DOC-MT5 -1.2B	—	—	—	—	—	—	—	—	—	—
DOCFLAT-MT5 -1.2B	22.2	26.6	22.3	29.7	28.4	11.7	13.9	26.4	25.9	17.6
IADA-MT5 -1.2B	22.1	26.9	22.7	30.5	28.5	12.8	14.5	27.1	26.1	18.7
L-7B-LORA	23.8	3.9	33.1	40.3	45.2	8.3	5.0	39.2	39.0	0.1
L-7B-FFT	22.4	2.5	32.2	42.6	44.8	1.0	1.0	38.9	38.2	0.1
B-7B-LORA	29.9	30.9	30.5	41.1	41.2	13.9	15.5	35.0	35.2	25.6
B-7B-FFT	22.3	17.0	29.8	41.3	40.7	0.4	1.1	35.6	34.3	1.0
V-7B-LORA	21.6	3.8	30.8	43.1	29.4	5.5	4.0	39.1	38.7	0.3
V-7B-FFT	21.8	2.2	31.0	43.4	45.0	0.0	0.7	36.2	38.0	0.1

Table 13: Breakdown sBLEU results for the translation tasks from other languages to English.

	$\mu_{d\text{BLEU}}$	Ar	De	Fr	It	Ja	Ko	Nl	Ro	Zh
NLLB-600M	22.0	30.5	14.7	34.1	26.3	14.8	18.1	21.0	16.8	22.2
NLLB-1.3B	28.6	39.2	22.6	40.1	38.7	17.1	20.3	29.4	26.7	23.7
NLLB-3.3B	29.4	39.7	26.1	39.6	37.0	16.5	23.2	31.3	28.8	22.8
GOOGLETRANS	28.5	32.0	29.8	37.8	38.9	13.3	17.7	33.7	33.1	20.4
GPT-3.5-TURBO	34.0	38.8	34.0	43.4	44.8	19.5	22.0	38.8	38.4	26.9
GPT-4-TURBO	35.1	40.2	34.5	44.2	46.3	19.7	24.4	39.3	39.6	28.1
LLAMA2-7B	4.4	0.1	6.9	1.5	4.7	0.0	0.1	15.7	10.9	0.1
BLOOM-7B	7.3	5.3	9.8	15.2	17.6	0.1	0.5	10.6	6.6	0.2
VICUNA-7B	9.8	1.1	14.5	24.6	14.4	5.9	0.5	18.5	8.9	0.0
DOC2DOC-MT5-300M	21.2	24.5	21.1	27.5	26.5	12.6	14.1	23.5	23.9	17.0
DOC2DOC-MT5-580M	22.5	25.5	22.1	28.9	27.8	14.0	14.6	25.8	25.2	18.5
DOC2DOC-MT5-1.2B	23.4	26.9	23.0	29.7	28.4	14.2	15.7	26.8	26.3	19.4
MR-DOC2SEN-MT5 -1.2B	23.8	27.4	24.2	30.3	29.4	14.9	16.1	27.5	26.8	19.8
MR-DOC2DOC-MT5 -1.2B	24.0	28.3	24.3	30.5	29.8	15.7	16.8	27.8	27.8	20.8
DOCFLAT-MT5 -1.2B	24.3	27.6	24.5	31.1	29.7	15.1	17.0	28.1	27.8	20.3
IADA-MT5 -1.2B	24.0	28.2	24.6	30.9	29.6	15.0	17.1	27.8	27.1	20.5
L-7B-LORA	25.7	4.1	36.2	42.2	48.5	10.0	5.9	42.2	42.1	0.1
L-7B-FFT	24.1	2.6	35.3	45.1	48.1	1.0	1.1	41.9	41.4	0.1
B-7B-LORA	33.6	33.0	34.2	44.0	44.9	18.8	20.4	38.1	38.6	30.4
B-7B-FFT	24.5	18.6	33.4	44.2	44.4	0.6	1.4	38.7	37.9	1.1
V-7B-LORA	23.3	3.8	33.9	45.5	30.6	6.7	4.8	42.1	42.0	0.3
V-7B-FFT	23.5	2.3	34.1	46.0	48.3	0.0	0.7	39.2	41.0	0.0

Table 14: Breakdown dBLEU results for the translation tasks from other languages to English.

	μ_{COMET}	Ar	De	Fr	It	Ja	Ko	Nl	Ro	Zh
NLLB-600M	72.8	76.6	63.0	79.8	72.4	74.2	76.3	68.4	67.0	77.6
NLLB-1.3B	78.1	82.3	71.9	83.8	80.6	75.3	78.4	77.5	75.4	77.9
NLLB-3.3B	78.9	82.6	76.1	83.7	80.2	74.9	80.1	78.7	76.5	77.7
GOOGLETRANS	81.2	81.1	82.3	84.6	84.5	75.6	76.9	84.2	83.8	78.1
GPT-3.5-TURBO	85.5	85.7	86.0	87.9	88.1	81.4	82.4	87.2	87.4	83.6
GPT-4-TURBO	86.0	86.5	86.3	88.2	88.5	81.9	83.5	87.5	87.9	84.2
LLAMA2-7B	52.2	50.3	47.1	42.6	51.6	56.1	55.7	56.5	53.0	56.9
BLOOM-7B	49.4	50.6	52.8	55.5	56.8	44.5	44.3	51.2	45.2	43.3
VICUNA-7B	62.7	51.3	65.2	70.5	57.3	69.5	56.3	68.0	58.8	67.5
DOC2DOC-MT5-300M	75.1	75.0	75.2	78.0	77.5	71.8	72.4	75.3	76.6	73.9
DOC2DOC-MT5-580M	77.4	77.4	77.0	79.7	79.8	74.7	74.3	78.5	79.1	75.8
DOC2DOC-MT5-1.2B	78.7	79.0	78.8	80.9	80.5	75.5	75.8	80.3	80.5	76.8
MR-DOC2SEN-MT5 -1.2B	79.8	80.3	79.8	82.3	81.1	76.4	76.6	81.5	81.8	78.2
MR-DOC2DOC-MT5 -1.2B	—	—	—	—	—	—	—	—	—	—
DOCFLAT-MT5 -1.2B	80.3	80.2	80.0	82.5	81.7	77.4	77.6	82.2	82.3	78.4
IADA-MT5 -1.2B	80.4	80.3	80.7	82.9	82.3	77.7	77.7	81.8	81.7	78.5
L-7B-LORA	73.7	53.9	84.0	84.1	88.2	59.0	53.0	87.0	87.6	66.9
L-7B-FFT	74.0	51.6	81.9	86.5	88.3	63.6	52.9	87.0	87.3	67.1
B-7B-LORA	81.4	73.3	83.6	87.0	87.1	74.0	73.8	84.8	86.0	82.6
B-7B-FFT	69.9	53.7	83.2	86.9	86.8	50.5	43.3	85.1	84.4	55.5
V-7B-LORA	71.4	54.5	82.6	87.2	64.9	57.8	53.7	87.1	87.3	67.7
V-7B-FFT	74.3	52.4	81.3	87.0	88.2	65.6	55.5	84.2	87.2	67.0

Table 15: Breakdown COMET results for the translation tasks from other languages to English.

	$\mu_{\%}$	Ar	De	Fr	It	Ja	Ko	Nl	Ro	Zh
L-7B-LoRA	6.2	0.4	4.6	1.2	0.5	17.1	0.4	0.8	1.0	29.7
L-7B-FFT	9.4	0.3	0.9	0.3	4.4	17.7	9.4	15.4	1.2	34.7
B-7B-LoRA	11.2	8.4	1.0	20.8	3.9	16.4	0.0	2.8	0.9	46.9
B-7B-FFT	31.8	36.6	15.8	2.7	90.1	10.7	0.1	82.0	0.2	47.7
V-7B-LoRA	10.6	0.2	15.4	0.3	13.3	15.9	0.5	20.3	0.9	28.9
V-7B-FFT	8.9	0.1	14.4	0.5	0.4	27.8	0.5	4.6	0.4	31.5

Table 16: Off-target rate (%) provided by our LLM-based DocMT models for translation tasks from English to other languages. $\mu_{\%}$ indicates the average off-target rate. **A lower off-target rate indicates better performance**.

	$\mu_{\%}$	Ar	De	Fr	It	Ja	Ko	Nl	Ro	Zh
L-7B-LoRA	29.2	87.9	2.0	4.9	1.4	25.5	44.2	1.9	1.8	93.1
L-7B-FFT	40.2	87.9	5.8	2.1	1.5	75.5	92.3	1.6	1.9	93.6
B-7B-LoRA	2.8	2.9	2.1	1.0	1.3	4.0	8.4	1.9	2.0	1.6
B-7B-FFT	28.0	54.1	2.0	1.0	1.1	43.8	70.4	1.6	1.9	76.4
V-7B-LoRA	32.3	88.2	2.6	1.2	28.0	40.4	35.7	1.9	1.9	90.5
V-7B-FFT	44.7	94.1	9.0	1.3	1.3	98.3	96.6	5.3	1.9	94.6

Table 17: Off-target rate (%) provided by our LLM-based DocMT models for translation tasks from other languages to English. $\mu_{\%}$ indicates the average off-target rate. **A lower off-target rate indicates better performance**.

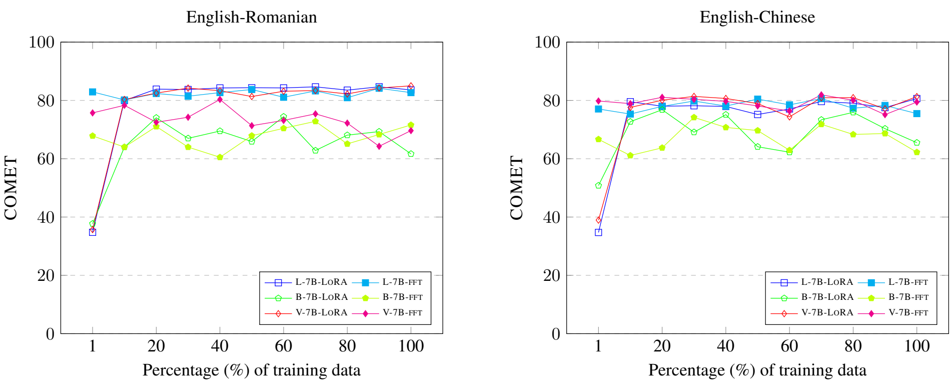

Figure 7: COMET-Percentage (%) of training data for the translations from English to Romanian, and Chinese.

	En-De	En-Fr
DOC2DOC-MT5-1.2B	39.8	26.2
L-7B-LORA	**64.4**	29.9
L-7B-FFT	62.4	28.4
B-7B-LORA	58.3	23.6
B-7B-FFT	49.8	25.1
V-7B-LORA	48.9	**30.2**
V-7B-FFT	43.5	27.2

Table 18: Generative accuracy (in %) on the English-German and English-French contrastive test sets. Best results are highlighted in **bold**.

	μ_Δ	Ar	De	Fr	It	Ja	Ko	Nl	Ro	Zh
L-7B-LoRA	+29.4	+36.3	+38.8	+37.2	+32.1	+15.9	+17.1	+21.7	+35.8	+29.5
L-7B-FFT	+29.0	+41.2	+40.5	+37.1	+18.0	+27.7	+29.4	+11.2	+18.5	+37.5
B-7B-LoRA	+20.3	+7.5	+40.7	+20.7	+21.9	+17.5	+15.9	+23.7	+25.3	+9.8
B-7B-FFT	+27.3	+14.8	+37.8	+28.9	+43.3	+13.1	+15.3	+38.5	+34.7	+19.5
V-7B-LoRA	-8.9	-12.6	+22.1	+18.9	-28.6	-27.8	-18.7	-11.8	+12.1	-34.1
V-7B-FFT	-1.4	+7.3	+25.2	+17.7	-14.6	-24.7	-5.3	-21.8	+7.6	-3.5

Table 19: The difference (Δ) in COMET scores on the test sets from English to other languages between our English-German LLM-based DocMT models and their backbones. μ_Δ indicates the average difference across all the languages in this table.