Multilingual Contextualization of Large Language Models for Document-Level Machine Translation

Miguel Moura Ramos[1,2] **Patrick Fernandes**[1,2,3] **Sweta Agrawal**[2]
André F. T. Martins[1,2,4]
[1]Instituto Superior Técnico, Universidade de Lisboa (ELLIS Unit Lisbon)
[2]Instituto de Telecomunicações [3]Carnegie Mellon University [4]Unbabel

Abstract

Large language models (LLMs) have demonstrated strong performance in sentence-level machine translation, but scaling to document-level translation remains challenging, particularly in modeling long-range dependencies and discourse phenomena across sentences and paragraphs. In this work, we propose a method to improve LLM-based long-document translation through targeted fine-tuning on high-quality document-level data, which we curate and introduce as DocBlocks. Our approach supports multiple translation paradigms, including direct document-to-document and chunk-level translation, by integrating instructions both with and without surrounding context. This enables models to better capture cross-sentence dependencies while maintaining strong sentence-level translation performance. Experimental results show that incorporating multiple translation paradigms improves document-level translation quality and inference speed compared to prompting and agent-based methods.

1 Introduction

Large language models (LLMs) have demonstrated strong performance across a wide range of natural language processing tasks (Ouyang et al., 2022; Sanh et al., 2021; Wei et al., 2021; Touvron et al., 2023; Jiang et al., 2023), including machine translation (MT) (Zhang et al., 2023; He et al., 2024; Hendy et al., 2023). Recent work (Alves et al., 2024; Xu et al., 2023) has convincingly shown that LLM-based systems consistently outperform specialized encoder-decoder MT systems for many language pairs (Kocmi et al., 2024; Deutsch et al., 2025). However, the focus of these studies has been primarily on sentence- and paragraph-level translation, overlooking the complexities of translating entire documents where maintaining coherence, consistency, and discourse structure is fundamental. For document-level translation, various techniques have emerged, such as context-aware prompting (Wang et al., 2023a; Wu et al., 2024) and agent-based translation strategies (Wu et al., 2024b; Wang et al., 2024). On the other hand, supervised fine-tuning (SFT) has proven highly effective for improving sentence-level MT (Alves et al., 2024; Xu et al., 2023), but its adaptation to document-level translation, and its comparison with other techniques, remain open questions.

In this work, we take a fresh perspective on the problem and explore whether we can employ SFT on strong sentence-level LLMs (Alves et al., 2024; Xu et al., 2023; Martins et al., 2024) to adapt them for multilingual contextualization. Specifically, we tackle the underexplored challenge of training LLMs to leverage surrounding context for document-level translation. Towards this end, we introduce DocBlocks, an MT contextualization dataset carefully curated and filtered to incorporate high-quality document-level data. Unlike prior datasets (NLLB Team et al., 2022; Kocmi et al., 2023; 2024) that focus primarily on isolated sentence pairs or limited context windows, DocBlocks consists of full documents and contextual segments sourced from trusted parallel corpora across diverse domains, including news, TED talk transcripts (Cettolo et al., 2017), literary texts, and parliamentary proceedings. We then introduce a simple yet effective contextualization method that fine-tunes existing LLM-based MT systems on DocBlocks by integrating the data in multiple formats. These include full documents, contextualized document chunks with preceding sections (Wang

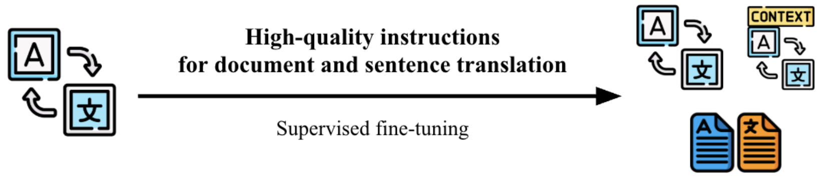

Figure 1: Illustration of our approach for adapting an LLM to document translation.

et al., 2023a) and standalone sentence-level examples. This multi-granular training strategy enables the model to capture both document structure and inter-sentence relationships, improving document-level translation performance while maintaining its robust sentence-level capabilities (Figure 1). Finally, to evaluate the dependence of our approach on the base MT model, we experiment with three different LLM-based MT models: Tower (Alves et al., 2024), EuroLLM (Martins et al., 2024), and Qwen2.5 (Bai et al., 2023). This allows us to assess the generalizability of our method across various model architectures.

Our key contributions are summarized as follows:

- **DocBlocks dataset**[1]: We release a document-level parallel dataset for tailoring LLMs to document translation, designed to support long-range dependency modeling and discourse coherence.
- **Fine-tuning for Multilingual Contextualization**: We propose an efficient fine-tuning approach that improves the document-level translation capabilities of LLMs by training on diverse, high-quality instructions across various translation paradigms—improving contextual understanding while preserving sentence-level performance.
- **Comprehensive Evaluation**: We conduct a systematic comparison of fine-tuning, prompting, and agent-based methods for document-level MT. Additionally, we perform ablation studies to evaluate the individual components of our approach, demonstrating its effectiveness in tailoring LLMs for document translation.

Our experiments across multiple datasets and language pairs demonstrate that our fine-tuning approach, combined with lightweight prompt-based context modeling, significantly improves document-level translation performance across various paradigms, outperforming vanilla LLM-based MT models, as well as both prompting and agent-based techniques.

2 Multilingual Contextualization of LLMs for Document-level MT

In this section, we present our contributions, starting with DocBlocks, a high-quality parallel dataset curated to address data scarcity in document-level MT (§2.1). Next, we outline the fine-tuning process of LLMs using this dataset to enhance the performance of document-level machine translation (§2.2). Finally, we investigate various inference methods and assess their impact on document translation quality (§2.3).

2.1 DocBlocks: A High-Quality Document-level MT Parallel Corpora

Corpora Collection. The DocBlocks corpus is constructed from a collection of publicly available document-level datasets, selected to represent a broad range of document types and content domains. Figure 2 summarizes the source datasets and their key characteristics. The primary objective of developing DocBlocks was to compile a versatile corpus that enables effective training of LLMs for document-level MT tasks. The ***News Commentary*** corpus contains political and economic news texts and was included as a representative of

[1] Available on Hugging Face.

Domain	Corpus	Pre-filtering				Post-filtering			
		$\|D\|$	$\|S\|$	$\|W\|$	$\|W\|/\|D\|$	$\|D\|$	$\|S\|$	$\|W\|$	$\|W\|/\|D\|$
news	News Commentary v18.1	125.0K	4.9M	119.0M	1.0K	110.0K	4.4M	96.4M	0.9K
TED Talks	IWSLT2016/2017	33.5K	3.3M	82.4M	2.5K	29.7K	2.6M	64.0M	2.1K
Parliamentary	Europarl v10	44.8K	22.3M	542.0M	12.1K	21.3K	10.5M	264.8M	12.4K
Novels	GuoFeng	22.6K	1.9M	51.4M	2.3K	18.2K	1.7M	28.5M	1.6K
	BWB	196.0K	10.3M	450.0M	2.3K	62.3K	4.2M	110.3M	1.8K

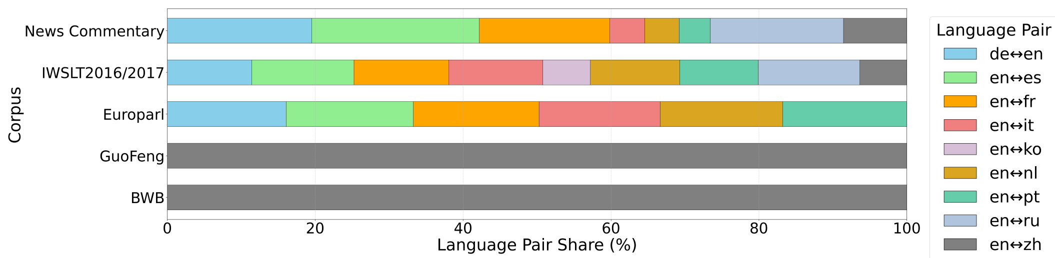

Figure 2: Statistics of datasets used to create DocBlocks, including domain, corpus name, number of documents $|D|$, sentences $|S|$, words $|W|$, and average document length $|W|/|D|$ as a proxy for discourse complexity (Wang et al., 2023a). We also illustrate the distribution of language pairs across the filtered datasets, highlighting their multilingual composition.

well-structured written content from the journalistic domain. The ***IWSLT*** (Cettolo et al., 2017) corpus, based on TED Talk transcripts, provides examples of conversational and spoken language across diverse topics. The ***Europarl*** (Koehn, 2005) corpus consists of European Parliament proceedings and represents formal writing with longer sentence structures and extended document length, typical of institutional discourse. The ***BWB*** (Jiang et al., 2022) and ***GuoFeng Webnovel*** (Wang et al., 2023b) corpora were included for their expert translations of Chinese novels and web fiction, which feature complex discourse structures and genre diversity. We also considered the ***United Nations Parallel Corpus*** (Ziemski et al., 2016) and ***OpenSubtitles*** (Lison & Tiedemann, 2016) for their multilingual scope and size, but document segmentation, alignment issues, and increased computational costs led us to prioritize better-curated datasets.

Data Curation: Preprocessing, Cleaning, and Augmentation. All datasets consist of document-level parallel corpora. We do not enforce strict sentence-to-sentence alignment. Instead, alignment is maintained at the document level, allowing for variation in sentence count and structure – reflecting realistic conditions in document translation. This approach enables the model to rely on broader semantic context rather than rigid sentence mappings. For datasets containing paragraph-level content with document boundary metadata, we grouped paragraphs into full documents and concatenated them up to a maximum length of $32,768$ tokens. To ensure high-quality data, we applied a rigorous cleaning pipeline to DocBlocks. Low-quality translations were identified using automatic filters, including Bicleaner (Ramírez-Sánchez et al., 2020) (threshold: 0.5) and CometKiwi-23 (Rei et al., 2023) (threshold: 0.65). We removed entire documents if over 20% of their sentences fell below these thresholds. We further applied language identification with `langid` (Lui & Baldwin, 2012) to exclude misaligned language pairs. Additional filters removed documents with fewer than 50 words, more than 10% identical characters/words/numbers, or a source–target length ratio above 1.3. We also deduplicated all entries.

These steps yielded a clean, reliable dataset optimized for document-level translation. To further enhance learning, we incorporated two complementary training techniques:

1. *Multi-Resolutional Document-to-Document Training (MRD2D)* (Sun et al., 2022): We split each document into k parts where $k \in \{1, 2, 4\}$ to create sequences of varying lengths, improving computational efficiency.

2. *Context-Aware Prompt Tuning (CAPT)* (Wang et al., 2023a): Building on the same preprocessed chunks, we incorporate a context window of up to the previous 3 segments in each training prompt. This allows the model to better capture document-level dependencies during training.

We applied MRD2D and CAPT selectively to datasets with longer average document lengths and moderate size, specifically, IWSLT, Europarl, and GuoFeng, to manage data growth and training costs. Experiments show that this curation and training strategy significantly improves translation performance (see Table 6).

2.2 Document-Level Fine-Tuning with DocBlocks

Model Training. Following prior work (Zhang et al., 2018; Wu et al., 2024), we adopt a two-step learning strategy, where LLMs are first tuned for sentence-level MT and then adapted for document-level MT tasks. Given the availability of sentence-level instruction-tuned LLMs (Alves et al., 2024; Martins et al., 2024), we focus only on the second step, which further fine-tunes these models on document-level translation data. To ensure high-quality translations, we calculate the loss exclusively on target tokens, excluding prompt tokens (source and instruction tokens), thereby preventing penalties for prompt format adherence.

DocBlocks instruction format. We support three instruction formats using the `chatml` (OpenAI, 2023) template. Following Wu et al. (2024), we use a contextual block of up to $N = 3$ previous chunks and format the texts as shown in Table 1. In addition to the chunk-level format, we also include instruction examples for document-to-document and sentence-to-sentence translation tasks. These tasks follow a simpler instruction structure, without the contextual block.

User	`<|im_start|>user`
	Context:
	{source_lang}: {source$_1$} {target_lang}: {target$_1$}
	{source_lang}: {source$_2$} {target_lang}: {target$_2$}
	{source_lang}: {source$_3$} {target_lang}: {target$_3$}
	Translate the following source text from **{source_lang}** into **{target_lang}**.
	{source_lang}: **{source}**.
	{target_lang}: `<|im_end|>`
	`<|im_start|>assistant`

Model	**{target}**. `<|im_end|>`

Table 1: Illustration of a DocBlocks instance, with the contextual block in gray.

2.3 Inference

We investigate the impact of different inference methods on document translation performance, as our models are designed to support a variety of translation strategies. We explore two methods: **Document-to-Document (Doc2Doc)** and **Chunking**.

The *Doc2Doc* method translates the entire document in a single pass, leveraging the ability of the LLM to capture long-range context. When the document fits within the model context window, it typically yields the most coherent and consistent translations (Wang et al., 2023a).

The *chunking* method translates a document chunk by chunk, with each chunk comprising a fixed number of sentences, paragraphs, or tokens. It can be applied on its own or enhanced in two complementary ways that address different aspects of the translation process. First, *contextual chunking* improves coherence by conditioning the translation of each chunk on previously translated source-target chunk pairs. Second, *quality-aware chunking* focuses on selecting the best translation using Minimum Bayes Risk (MBR) decoding (Kumar & Byrne, 2004; Eikema & Aziz, 2022), guided by quality estimation metrics such as COMET (Rei et al., 2022) and ContextCOMET (Vernikos et al., 2022). These metrics may incorporate contextual information and help choose the final output by maximizing expected utility:

$\hat{y}_{mbr} = \arg\max_{y \in \bar{\mathcal{Y}}} \mathbb{E}_Y [u(Y, y)] \approx \arg\max_{y \in \bar{\mathcal{Y}}} \frac{1}{|\bar{\mathcal{Y}}|} \sum_{y_t \in \bar{\mathcal{Y}}} \mathcal{M}([c_t, y_t], [c, y])$, where $\bar{\mathcal{Y}}$ is a list of candidates sampled by the model and $\mathcal{M}$ is the reference-based quality metric, which can optionally use context c.

3 Experiments

3.1 Experimental Setup

Datasets. We evaluate translation performance using several benchmarks. For document-level MT, we use IWSLT2017 (Cettolo et al., 2017) and GuoFeng (Wang et al., 2023b) test sets. IWSLT2017 contains parallel TED talk documents across six language pairs: en↔{de, fr, it, ko, nl, zh}. GuoFeng, a discourse-rich web novel corpus, is used for zh→en experiments on its combined simple and difficult test sets. For sentence-level MT, we use FLORES-200 (en↔{de, fr, it, ko, nl, zh, ru, pt, es}) (NLLB Team et al., 2022), WMT23 (en↔{de, ru, zh}) (Kocmi et al., 2023), and TICO-19 (en→{es, fr, pt, ru, zh}) (Anastasopoulos et al., 2020) to assess how document-level training affects sentence-level MT quality.

Baselines. In this work, we employ the closed-source GPT-4o (OpenAI et al., 2023) and the open-source models QWEN2.5-72B-INSTRUCT (Bai et al., 2023) and LLAMA-3.3-70B-INSTRUCT (Grattafiori et al., 2024) as our baselines. These three state-of-the-art LLMs serve as the primary comparisons for evaluating our approach to adapting sentence-level LLMs to document-level MT. We use greedy decoding for all inference methods, except for quality-aware chunking, where we generate 32 candidates using p-nucleus sampling (Holtzman et al., 2019) with $p = 0.6$.

Base sentence-level LLMs. To assess the effectiveness of our method across different base models, we experiment with three instruction-tuned LLMs: TOWERINSTRUCT-MISTRAL-7B (Alves et al., 2024), EUROLLM-9B-INSTRUCT (Martins et al., 2024), and QWEN2.5-7B-INSTRUCT (Bai et al., 2023). While TOWERINSTRUCT-MISTRAL-7B offers a 32768 context window, it struggles with document-level MT due to its pretraining and instruction tuning, which are focused on short sequences. EUROLLM-9B-INSTRUCT has a 4096 context window, limiting its performance in document-level MT. In contrast, QWEN2.5-7B-INSTRUCT performs better with its 32768 context window (see Tables 2 and 3).

Hyperparameters. We fine-tune all sentence-level LLMs on the instructional DOCBLOCKS using the standard cross-entropy loss and enable `bfloat16` mixed precision and packing (Raffel et al., 2020). We detail all hyperparameters in Appendix A. For TOWERINSTRUCT-MISTRAL-7B and QWEN2.5-7B-INSTRUCT, we kept the original RoPE theta (Su et al., 2021) since both were pretrained with a 32768 context window. For EuroLLM, initial tests showed minimal gains from extending RoPE, so we retained the original design.

Evaluation. We evaluate different approaches using both sentence-level and document-level metrics. Sentence-level evaluation includes BLEU (Papineni et al., 2002) and COMET (`wmt22-comet-da`) (Rei et al., 2022). Sentence alignments are first computed using `bleualign`, after which these metrics are applied. However, these metrics have known limitations in capturing document-level phenomena such as coherence, cohesion, and referential consistency (Castilho et al., 2020; Fernandes et al., 2021; Läubli et al., 2018; Toral, 2020; Freitag et al., 2021). To better evaluate document-level translation quality, we report document-level BLEU (d-BLEU) (Papineni et al., 2002; Liu et al., 2020) with corresponding brevity penalty (BP), and document-level COMET (d-COMET) (Rei et al., 2022), computed using the SLIDE approach (Raunak et al., 2024). SLIDE divides each document into overlapping 512-token chunks (the context limit for COMET), computes chunk-level scores independently, and averages them to obtain a document-level score. We also report the Lexical Translation Consistency Ratio (LTCR) (Wang et al., 2023a; Wang et al., 2024; Xiao et al., 2011; Lyu et al., 2021), which captures the consistency of repeated terminology across a document. Although recent work has proposed alternative document-level metrics (Vernikos et al., 2022; Jiang et al., 2022; Gong et al., 2015; Wong & Kit, 2012), there remains no universally accepted method for evaluating document-level translation quality. Developing more robust and comprehensive evaluation strategies remains an important open problem, complementary to ours, which we leave for future work.

3.2 Results and Analysis

DocMT-LLMs achieve superior document-to-document (Doc2Doc) translation quality. DocMT-LLMs (LLM-based document-level MT models) consistently outperform sentence-level models across all benchmarks. On both GuoFeng (Table 2) and IWSLT2017 (Table 3) datasets, DocMT-LLMs all outperform their sentence-level counterparts, with strong gains across both sentence- and document-level metrics. These lower scores for sentence-level baselines are expected, as they lack access to document-wide context during training, which is crucial for ensuring translation coherence. Despite being much smaller, our models outperform strong baselines like LLAMA-3.3-70B-INSTRUCT, QWEN2.5-72B-INSTRUCT and GPT-4O on GuoFeng and perform well on IWSLT2017. While these baselines slightly lead on d-COMET and on d-BLEU (only for IWSLT2017 en→xx), our models still achieve substantial gains over undertrained sentence-level LLMs. These results demonstrate that incorporating DOCBLOCKS enables sentence-level LLMs to generate coherent, high-quality document-level translations.

Models	GuoFeng zh→en				
	BLEU	COMET	d-BLEU (BP)	d-COMET	LTCR
`Llama-3.3-70B-Instruct`	12.12	80.16	15.32 (0.83)	71.37	65.24
`Qwen2.5-72B-Instruct`	17.87	**84.85**	20.99 (0.97)	72.13	65.72
`GPT-4o`	14.61	73.66	17.46 (1.00)	**73.73**	59.69
`TowerInstruct-Mistral-7B`	6.78	58.40	7.27 (1.00)	58.23	54.16
`DocMT-TowerInstruct-Mistral-7B`	**34.06**	78.72	**37.57 (1.00)**	72.84	**74.59**
`EuroLLM-9B-Instruct`	12.43	63.14	12.03 (0.84)	60.11	51.07
`DocMT-EuroLLM-9B-Instruct`	24.12	73.88	27.83 (1.00)	72.41	67.05
`Qwen2.5-7B-Instruct`	16.51	84.34	19.56 (0.96)	71.66	63.22
`DocMT-Qwen2.5-7B-Instruct`	30.67	77.28	34.80 (0.99)	71.88	73.48

Table 2: Doc2Doc translation results on the GuoFeng dataset. Best overall values are **bolded**, and group-specific bests are underlined.

Models	IWSLT2017 en→xx					IWSLT2017 xx→en				
	BLEU	COMET	d-BLEU (BP)	d-COMET	LTCR	BLEU	COMET	d-BLEU (BP)	d-COMET	LTCR
`Llama-3.3-70B-Instruct`	15.61	61.17	27.08 (0.85)	70.56	57.64	12.34	54.87	37.91 (0.90)	72.01	82.20
`Qwen2.5-72B-Instruct`	18.48	68.47	26.59 (0.93)	**75.15**	60.75	13.12	58.85	38.81 (0.93)	68.44	81.57
`GPT-4o`	**27.96**	74.79	**32.25 (0.97)**	74.73	59.05	28.53	76.55	33.47 (0.89)	**75.04**	77.88
`TowerInstruct-Mistral-7B`	1.77	41.67	2.41 (0.62)	29.68	27.53	10.10	59.64	2.36 (0.21)	31.79	28.88
`DocMT-TowerInstruct-Mistral-7B`	22.67	77.33	25.60 (0.94)	74.38	60.46	27.73	59.69	26.79 (1.00)	60.29	79.97
`EuroLLM-9B-Instruct`	6.51	73.64	3.75 (0.56)	24.97	35.20	8.83	78.94	3.97 (0.39)	29.81	33.79
`DocMT-EuroLLM-9B-Instruct`	20.94	72.73	20.24 (1.00)	61.74	55.47	28.75	69.66	28.46 (1.00)	62.71	78.20
`Qwen2.5-7B-Instruct`	17.32	70.57	20.86 (0.80)	71.73	56.36	23.94	75.95	34.59 (0.90)	73.29	80.50
`DocMT-Qwen2.5-7B-Instruct`	23.98	**78.48**	25.29 (0.98)	73.00	**62.05**	**37.64**	**81.91**	**40.56 (0.97)**	73.02	**84.09**

Table 3: Doc2Doc translation results on IWSLT2017, averaged over en→xx and xx→en. Best overall values are **bolded**, and group-specific bests are underlined.

DocMT-LLMs deliver high-quality chunking decoding. Chunking-based decoding enables efficient document translation by processing text in segments. Larger chunks provide more context, bridging sentence- and document-level translation. DocMT-LLMs consistently outperform sentence-level models on d-BLEU and d-COMET (Figure 3), especially at larger chunk sizes. In contrast, TOWERINSTRUCT-MISTRAL-7B and EUROLLM-9B-INSTRUCT degrade as the chunk size increases. Document-level models maintain strong performance and often surpass QWEN2.5-72B-INSTRUCT. DOCMT-QWEN2.5-7B-INSTRUCT improves quality across most chunk sizes, despite QWEN2.5-7B-INSTRUCT already supporting document-level MT. These results show that training on DOCBLOCKS enhances both document translation and chunking decoding, making DocMT-LLMs more robust.

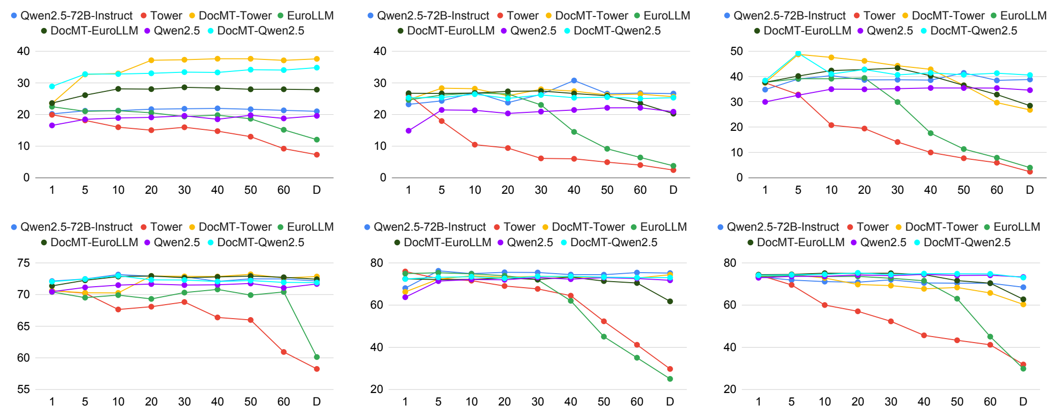

Figure 3: d-BLEU (top) and d-COMET (bottom) scores for the GuoFeng (left), IWSLT2017 en→xx (middle), and xx→en (right) datasets as the decoding chunk size increases. "D" indicates that the entire document is treated as a single chunk during decoding.

DocMT-LLMs can be coupled with contextual chunking and quality-aware chunking to improve translation quality. DocMT-LLMs improve translation quality by incorporating context during chunk-level decoding. Figure 4 shows that context-aware prompt tuning (CAPT) helps models better use surrounding context, leading to higher scores across document-level metrics. Quality-aware chunking (Figure 5) further improves performance by combining contextual chunking with paragraph-level metrics like COMET and CONTEXTCOMET, using a context window of three chunks within a 512-token limit. However, such metrics may not fully capture document-level quality and can lead to suboptimal optimization. As shown in Table 4, standard chunking is the most efficient due to fully parallelizable decoding. In contrast, contextual and quality-aware chunking require sequential decoding, with quality-aware chunking being slower due to candidate reranking with quality metrics. Nevertheless, DOCMT-TOWERINSTRUCT-MISTRAL-7B supports multiple translation paradigms, offering varying translation quality and efficiency trade-offs.

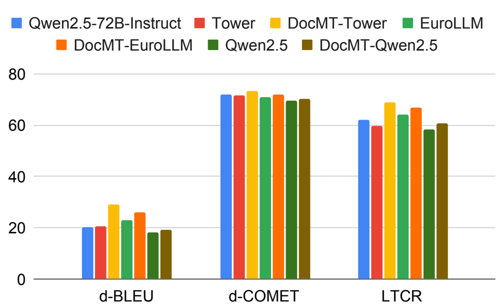

Figure 4: Evaluation scores for contextual chunking on GuoFeng using a context window of 3, consistent with the training setup. Improvements demonstrate CAPT's effect in enabling DocMT-LLMs to better leverage prior translated chunks during decoding.

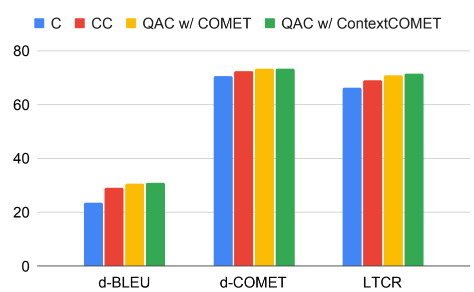

Figure 5: DOCMT-TOWERINSTRUCT-7B performance on GuoFeng using chunking (C), contextual chunking (CC), and quality-aware chunking (QAC), which combines contextual chunking with paragraph-level quality metrics.

Inference Method	Chunking	Contextual Chunking	Quality-aware Chunking	Document-to-Document
Throughput ↑	392.45	39.53	16.67	204.22

Table 4: Throughput comparison of inference methods on Guofeng with DOCMT-TOWERINSTRUCT-MISTRAL-7B, measured in tokens per second.

Sentence-level MT capabilities of DocMT-LLMs do not deteriorate despite document-level MT training on DocBlocks. Table 5 shows that DocMT-LLMs fine-tuned with DocBlocks retain strong sentence-level performance, with COMET score differences from their sentence-level counterparts typically within 0.5 points, indicating no catastrophic forgetting. Qwen2.5-7B-Instruct even improves sentence-level performance through document-level training, enhancing robustness without sacrificing sentence-level quality. Detailed results per language pair are provided in Appendix B.

Models	FLORES-200		WMT23		TICO-19
	en→xx	xx→en	en→xx	xx→en	en→xx
`TowerInstruct-Mistral-7B`	88.96	88.45	85.26	83.03	87.46
`DocMT-TowerInstruct-Mistral-7B`	88.54	88.21	84.93	82.97	87.17
`EuroLLM-9B-Instruct`	89.21	88.31	85.54	83.24	87.53
`DocMT-EuroLLM-9B-Instruct`	88.79	88.41	85.26	82.88	87.20
`Qwen2.5-7B-Instruct`	83.08	86.70	79.89	79.81	83.36
`DocMT-Qwen2.5-7B-Instruct`	87.77	88.06	83.91	81.80	86.51

Table 5: Evaluation of sentence-level LLMs and DocMT-LLMs based on COMET across sentence-level MT benchmarks.

Comparing translation approaches: DocMT-LLMs vs. Agent-based methods. DocMT-LLMs consistently outperform (in quality and inference speed) agent-based methods like TRANSAGENTS (Wu et al., 2024b) and DELTA (Wang et al., 2024) in both Doc2Doc and contextual chunking translation tasks (Figure 6). Agent-based methods iteratively refine translations using memory-based mechanisms with dictionaries and using a hierarchical workflow. This multi-step pipeline and inter-agent communication can introduce significant latency, slowing inference. In contrast, training a DocMT-LLM requires 120 hours on 4 NVIDIA H100 80GB GPUs but runs efficiently on a single NVIDIA A6000 48GB GPU afterward. This initial cost is offset after processing just 42 documents of 400 sentences each. Therefore, both contextual chunking and Doc2Doc approaches with DocMT-LLMs achieve a superior balance among context preservation, speed, scalability, and translation quality.

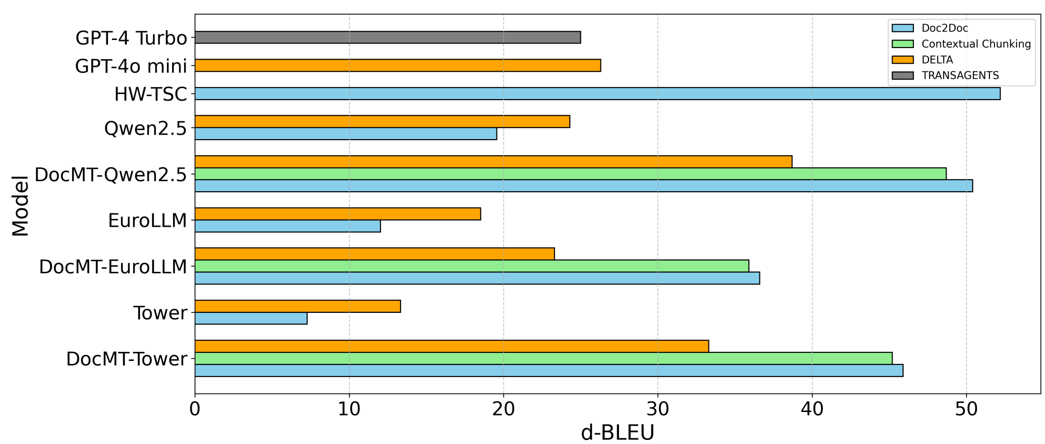

Figure 6: d-BLEU scores on the Guofeng test set, with chapter translations concatenated into a single document (Wu et al., 2024b). The figure compares document-level translation quality across different paradigms: blue for Doc2Doc, green for Contextual Chunking, orange for the DELTA agent-based approach, and gray for the TRANSAGENTS agent-based approach. In addition to evaluating our base models before and after DocMT training, we include strong external baselines representative of each paradigm – TRANSAGENTS with GPT-4 Turbo (Wu et al., 2024b), DELTA with GPT-4o mini (Wang et al., 2024), and HW-TSC (Xie et al., 2023) for Doc2Doc.

DocBlocks curation process impacts learning. As shown in Table 6, filtering improves model performance across most datasets. The Multiresolution Document-to-Document (MRD2D) technique provides significant benefits for both document-to-document translation and chunking by introducing additional subdocuments during training, especially for the IWSLT2017 dataset. Context-Aware Prompt Tuning (CAPT) is particularly effective for Contextual Chunking, as it provides context-driven examples that improve performance, particularly for the GuoFeng and IWSLT2017 (xx→en) tasks. Combining filtering with MRD2D and CAPT yields the best results overall, boosting translation quality and contextual understanding across document-level translation tasks. We further examine the impact of balancing the DocBlocks dataset, which is naturally skewed toward English-to-Chinese document pairs due to their greater availability. In this setting, we uniformly sample both from and into English across language directions to ensure an even distribution of translation pairs. Our results indicate that enforcing this balance yields no performance gains. As such, we choose to retain as much data as possible rather than discard samples to achieve balance, prioritizing overall model performance.

Models	Doc2Doc			Chunking			Contextual Chunking		
	GuoFeng	IWSLT2017		GuoFeng	IWSLT2017		GuoFeng	IWSLT2017	
	zh→en	en→xx	xx→en	zh→en	en→xx	xx→en	zh→en	en→xx	xx→en
TowerInstruct-Mistral-7B	58.23	29.68	31.79	68.81	67.61	52.26	70.21	70.62	69.53
SFT									
+ unfiltered DocBlocks	70.08	62.30	55.32	69.20	69.24	55.89	70.55	69.34	69.00
+ filtered DocBlocks	71.71	63.67	57.72	69.21	70.36	58.23	70.45	69.77	68.44
+ MRD2D	72.55	**74.60**	60.11	72.38	73.03	68.97	70.34	71.32	69.45
+ CAPT	71.70	66.33	58.31	68.80	68.55	65.37	**73.11**	73.05	**75.04**
+ MRD2D + CAPT	**72.84**	74.38	**60.29**	**72.86**	**73.51**	**69.21**	72.93	**73.18**	74.48
+ balanced DocBlocks	67.89	61.79	55.03	65.72	69.57	54.39	69.80	70.33	68.90

Table 6: Ablation results for the components of DocBlocks, reporting d-COMET scores.

Balancing sentence- and document-Level data in DocBlocks: impacts on translation performance. We conducted an ablation study to assess how varying the proportion of sentence-level data in DocBlocks affects translation performance, while keeping document-level data constant. The sentence-level portion is sourced from TowerBlocks (Alves et al., 2024) and uniformly sampled across languages. Results shown in Figure 7 demonstrate that just 10% sentence-level data is sufficient to maintain sentence-level quality, while higher proportions yield diminishing returns and reduce document-level performance. Allocating 90% of training data to document-level pairs provides the best balance, enhancing document-level performance without compromising sentence-level capabilities.

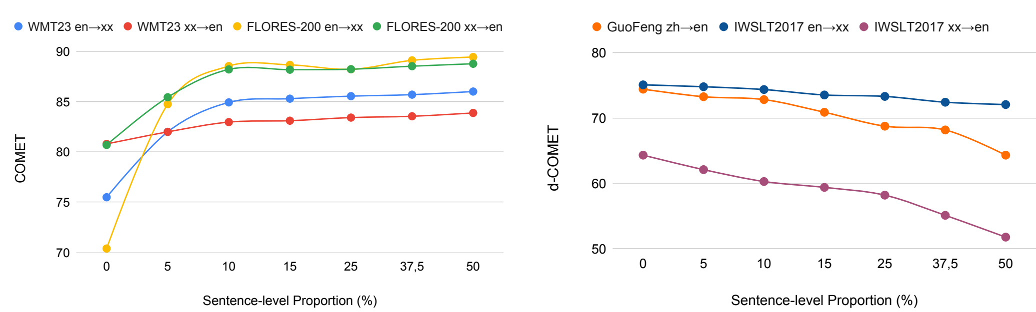

Figure 7: Effect of varying sentence-level data in DocBlocks training on sentence-level (left) and document-level (right) translation performance.

4 Related Work

Document-Level Machine Translation. Traditional sentence-level MT (Sutskever et al., 2014; Bahdanau et al., 2014; Vaswani et al., 2017) often leads to inconsistencies in terminology, tone, and context in longer texts (Xiong et al., 2018; Hu & Wan, 2023; Wang et al., 2023b). Document-level MT (Maruf et al., 2021) improves coherence by incorporating broader context across sentences and paragraphs. Recent approaches include document embeddings (Macé & Servan, 2019; Huo et al., 2020), multi-encoder architectures (Zhang et al., 2018; Voita et al., 2018; Bawden et al., 2018), and enhanced attention mechanisms for long-range dependencies (Zhang et al., 2020; Miculicich et al., 2018; Wu et al., 2023). While promising, progress is limited by the lack of high-quality document-level parallel data (Liu & Zhang, 2020; Wang et al., 2023b), which hinders training in complex domains such as literary translation (Ghazvininejad et al., 2018; Toral & Way, 2018). Recent efforts have drawn on web crawls (e.g., ParaCrawl (Bañón et al., 2020), BWB (Jiang et al., 2022)), news data (e.g., WMT (Kocmi et al., 2024)), parliamentary records (e.g., Europarl (Koehn, 2005)), public talks (e.g., IWSLT TED task (Cettolo et al., 2017)), literary texts (e.g., GuoFeng (Wang et al., 2023b)), and subtitles (e.g., OpenSubtitles (Lison & Tiedemann, 2016)), as well as large multilingual corpora (e.g., UNPC (Ziemski et al., 2016)). However, many of these suffer from domain constraints, noisy alignments, or unclear document boundaries. We address these issues by curating DOCBLOCKS, a high-quality document-level dataset built from public sources, emphasizing clean alignment, coherent document structure, and careful filtering.

LLMs for Machine Translation. LLMs have recently demonstrated strong performance on MT tasks (Alves et al., 2024). Their capacity to process long contexts opens new possibilities for modeling discourse structure and inter-sentence dependencies, which are essential for document-level translation. To leverage these capabilities, new strategies are being explored. These include context-aware prompting, which uses relevant context directly in the prompt to guide translation (Wang et al., 2023a; Wu et al., 2024), and document-level fine-tuning, which adapts instruction-tuned models to achieve stronger performance (Wu et al., 2024; Wu et al., 2024a; Li et al., 2024). In addition, agent-based approaches such as DELTA (Wang et al., 2024) and TRANSAGENTS (Wu et al., 2024b) further improve consistency by maintaining structured memory across sentences, though they introduce substantial inference overhead. Despite these advances, LLMs still face challenges in maintaining coherence over long documents due to limitations in training data (Wu et al., 2024a) and context modeling (Karpinska & Iyyer, 2023). To address these challenges, we fine-tune LLMs using DOCBLOCKS, enhancing their multilingual contextual understanding for document translation without incurring the high cost of full retraining.

5 Conclusion

This study advances the transition of state-of-the-art LLMs from sentence-level MT to document-level MT by addressing key challenges in contextual coherence and long-document consistency. Fine-tuning on the curated DOCBLOCKS dataset allows DocMT-LLMs to effectively manage long-range dependencies while preserving sentence-level translation quality. Moreover, context-aware prompt tuning and multi-resolution document-to-document training enhance adaptability across various translation paradigms, including document-to-document and several chunk-level methods. These findings provide a cost-effective and scalable approach to improving document-level translation quality.

Acknowledgments

We thank the members of the SARDINE lab for their useful and constructive comments. This work was supported by the Portuguese Recovery and Resilience Plan through project C645008882-00000055 (Center for Responsible AI), by the EU's Horizon Europe Research and Innovation Actions (UTTER, contract 101070631), by the project DECOLLAGE (ERC-2022-CoG 101088763), and by Fundação para a Ciência e Tecnologia through contract UIDB/50008/2020.

References

Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. Tower: An Open Multilingual Large Language Model for Translation-Related Tasks. *arXiv e-prints*, art. arXiv:2402.17733, February 2024. doi: 10.48550/arXiv.2402.17733.

Antonios Anastasopoulos, Alessandro Cattelan, Zi-Yi Dou, Marcello Federico, Christian Federmann, Dmitriy Genzel, Franscisco Guzmán, Junjie Hu, Macduff Hughes, Philipp Koehn, Rosie Lazar, Will Lewis, Graham Neubig, Mengmeng Niu, Alp Öktem, Eric Paquin, Grace Tang, and Sylwia Tur. TICO-19: the translation initiative for COvid-19. In Karin Verspoor, Kevin Bretonnel Cohen, Michael Conway, Berry de Bruijn, Mark Dredze, Rada Mihalcea, and Byron Wallace (eds.), *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online, December 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.nlpcovid19-2.5. URL https://aclanthology.org/2020.nlpcovid19-2.5.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv e-prints*, art. arXiv:1409.0473, September 2014. doi: 10.48550/arXiv.1409.0473.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen Technical Report. *arXiv e-prints*, art. arXiv:2309.16609, September 2023. doi: 10.48550/arXiv.2309.16609.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. ParaCrawl: Web-scale acquisition of parallel corpora. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4555–4567, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.417. URL https://aclanthology.org/2020.acl-main.417/.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. Evaluating discourse phenomena in neural machine translation. In Marilyn Walker, Heng Ji, and Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1304–1313, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1118. URL https://aclanthology.org/N18-1118.

Sheila Castilho, Maja Popović, and Andy Way. On context span needed for machine translation evaluation. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 3735–3742, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.461.

Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Jan Niehues, Sebastian Stüker, Katsuhito Sudoh, Koichiro Yoshino, and Christian Federmann. Overview of the IWSLT 2017 evaluation campaign. In *Proceedings of the 14th International Conference on Spoken Language Translation*, pp. 2–14, Tokyo, Japan, December 14-15 2017. International Workshop on Spoken Language Translation. URL https://aclanthology.org/2017.iwslt-1.1.

Daniel Deutsch, Eleftheria Briakou, Isaac Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, Shruti Rijhwani, Parker Riley, Elizabeth Salesky, Firas Trabelsi, Stephanie Winkler, Biao Zhang, and Markus Freitag. WMT24++: Expanding the Language Coverage of WMT24 to 55 Languages & Dialects. *arXiv e-prints*, art. arXiv:2502.12404, February 2025. doi: 10.48550/arXiv.2502.12404.

Bryan Eikema and Wilker Aziz. Sampling-based approximations to minimum Bayes risk decoding for neural machine translation. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 10978–10993, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.754. URL https://aclanthology.org/2022.emnlp-main.754.

Patrick Fernandes, Kayo Yin, Graham Neubig, and André F. T. Martins. Measuring and increasing context usage in context-aware machine translation. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 6467–6478, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.505. URL https://aclanthology.org/2021.acl-long.505.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474, 2021. doi: 10.1162/tacl_a_00437. URL https://aclanthology.org/2021.tacl-1.87.

Marjan Ghazvininejad, Yejin Choi, and Kevin Knight. Neural poetry translation. In Marilyn Walker, Heng Ji, and Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 67–71, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2011. URL https://aclanthology.org/N18-2011.

Zhengxian Gong, Min Zhang, and Guodong Zhou. Document-level machine translation evaluation with gist consistency and text cohesion. In Bonnie Webber, Marine Carpuat, Andrei Popescu-Belis, and Christian Hardmeier (eds.), *Proceedings of the Second Workshop on Discourse in Machine Translation*, pp. 33–40, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-2504. URL https://aclanthology.org/W15-2504.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz

Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. Exploring human-like translation strategy with large language models. *Transactions of the Association for Computational Linguistics*, 12:229–246, 2024. doi: 10.1162/tacl_a_00642. URL https://aclanthology.org/2024.tacl-1.13.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation. *arXiv e-prints*, art. arXiv:2302.09210, February 2023. doi: 10.48550/arXiv.2302.09210.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The Curious Case of Neural Text Degeneration. *arXiv e-prints*, art. arXiv:1904.09751, April 2019. doi: 10.48550/arXiv.1904.09751.

Xinyu Hu and Xiaojun Wan. Exploring discourse structure in document-level machine translation. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13889–13902, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.857. URL https://aclanthology.org/2023.emnlp-main.857.

Jingjing Huo, Christian Herold, Yingbo Gao, Leonard Dahlmann, Shahram Khadivi, and Hermann Ney. Diving deep into context-aware neural machine translation. In Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Yvette Graham, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, and Matteo Negri (eds.), *Proceedings of the Fifth Conference on Machine Translation*, pp. 604–616, Online, November 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.wmt-1.71.

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. Mistral 7B. *arXiv e-prints*, art. arXiv:2310.06825, October 2023. doi: 10.48550/arXiv.2310.06825.

Yuchen Jiang, Tianyu Liu, Shuming Ma, Dongdong Zhang, Jian Yang, Haoyang Huang, Rico Sennrich, Ryan Cotterell, Mrinmaya Sachan, and Ming Zhou. BlonDe: An automatic evaluation metric for document-level machine translation. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics:*

Human Language Technologies, pp. 1550–1565, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.111. URL https://aclanthology.org/2022.naacl-main.111.

Marzena Karpinska and Mohit Iyyer. Large language models effectively leverage document-level context for literary translation, but critical errors persist. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz (eds.), *Proceedings of the Eighth Conference on Machine Translation*, pp. 419–451, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.41. URL https://aclanthology.org/2023.wmt-1.41/.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv e-prints*, art. arXiv:1412.6980, December 2014. doi: 10.48550/arXiv.1412.6980.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, and Mariya Shmatova. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz (eds.), *Proceedings of the Eighth Conference on Machine Translation*, pp. 1–42, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.1. URL https://aclanthology.org/2023.wmt-1.1.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinthór Steingrímsson, and Vilém Zouhar. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz (eds.), *Proceedings of the Ninth Conference on Machine Translation*, pp. 1–46, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.wmt-1.1. URL https://aclanthology.org/2024.wmt-1.1/.

Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pp. 79–86, Phuket, Thailand, September 13-15 2005. URL https://aclanthology.org/2005.mtsummit-papers.11/.

Shankar Kumar and William Byrne. Minimum Bayes-risk decoding for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pp. 169–176, Boston, Massachusetts, USA, May 2 - May 7 2004. Association for Computational Linguistics. URL https://aclanthology.org/N04-1022.

Samuel Läubli, Rico Sennrich, and Martin Volk. Has machine translation achieved human parity? a case for document-level evaluation. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4791–4796, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1512. URL https://aclanthology.org/D18-1512.

Yachao Li, Junhui Li, Jing Jiang, and Min Zhang. Enhancing document-level translation of large language model via translation mixed-instructions, 2024. URL https://arxiv.org/abs/2401.08088.

Pierre Lison and Jörg Tiedemann. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 923–929, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL https://aclanthology.org/L16-1147.

Siyou Liu and Xiaojun Zhang. Corpora for document-level neural machine translation. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 3775–3781, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.466.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742, 2020. doi: 10.1162/tacl_a_00343. URL https://aclanthology.org/2020.tacl-1.47.

Marco Lui and Timothy Baldwin. langid.py: An off-the-shelf language identification tool. In Min Zhang (ed.), *Proceedings of the ACL 2012 System Demonstrations*, pp. 25–30, Jeju Island, Korea, July 2012. Association for Computational Linguistics. URL https://aclanthology.org/P12-3005.

Xinglin Lyu, Junhui Li, Zhengxian Gong, and Min Zhang. Encouraging lexical translation consistency for document-level neural machine translation. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3265–3277, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.262. URL https://aclanthology.org/2021.emnlp-main.262.

Valentin Macé and Christophe Servan. Using whole document context in neural machine translation. In Jan Niehues, Rolando Cattoni, Sebastian Stüker, Matteo Negri, Marco Turchi, Thanh-Le Ha, Elizabeth Salesky, Ramon Sanabria, Loic Barrault, Lucia Specia, and Marcello Federico (eds.), *Proceedings of the 16th International Conference on Spoken Language Translation*, Hong Kong, November 2-3 2019. Association for Computational Linguistics. URL https://aclanthology.org/2019.iwslt-1.21.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M. Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, Pierre Colombo, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. EuroLLM: Multilingual Language Models for Europe. *arXiv e-prints*, art. arXiv:2409.16235, September 2024. doi: 10.48550/arXiv.2409.16235.

Sameen Maruf, Fahimeh Saleh, and Gholamreza Haffari. A survey on document-level neural machine translation: Methods and evaluation. *ACM Comput. Surv.*, 54(2), March 2021. ISSN 0360-0300. doi: 10.1145/3441691. URL https://doi.org/10.1145/3441691.

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. Document-level neural machine translation with hierarchical attention networks. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2947–2954, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1325. URL https://aclanthology.org/D18-1325.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. No Language Left Behind: Scaling Human-Centered Machine Translation. *arXiv e-prints*, art. arXiv:2207.04672, July 2022. doi: 10.48550/arXiv.2207.04672.

OpenAI, 2023. URL https://github.com/openai/openai-python/blob/release-v0.28.1/chatml.md.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. GPT-4 Technical Report. *arXiv e-prints*, art. arXiv:2303.08774, March 2023. doi: 10.48550/arXiv.2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll 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. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin (eds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

Gema Ramírez-Sánchez, Jaume Zaragoza-Bernabeu, Marta Bañón, and Sergio Ortiz Rojas. Bifixer and bicleaner: two open-source tools to clean your parallel data. In André Martins, Helena Moniz, Sara Fumega, Bruno Martins, Fernando Batista, Luisa Coheur, Carla Parra, Isabel Trancoso, Marco Turchi, Arianna Bisazza, Joss Moorkens, Ana Guerberof, Mary Nurminen, Lena Marg, and Mikel L. Forcada (eds.), *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pp. 291–298, Lisboa, Portugal, November 2020. European Association for Machine Translation. URL https://aclanthology.org/2020.eamt-1.31/.

Vikas Raunak, Tom Kocmi, and Matt Post. SLIDE: Reference-free evaluation for machine translation using a sliding document window. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 205–211, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-short.18. URL https://aclanthology.org/2024.naacl-short.18.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In Philipp Koehn, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri (eds.), *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pp. 578–585, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.wmt-1.52.

Ricardo Rei, Nuno M. Guerreiro, José Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André F. T. Martins. Scaling up COMETKIWI: Unbabel-IST 2023 Submission for the Quality Estimation Shared Task. *arXiv e-prints*, art. arXiv:2309.11925, September 2023. doi: 10.48550/arXiv.2309.11925.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal 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 Fevry, Jason Alan Fries, Ryan Teehan, Tali Bers, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. Multitask Prompted Training Enables Zero-Shot Task Generalization. *arXiv e-prints*, art. arXiv:2110.08207, October 2021. doi: 10.48550/arXiv.2110.08207.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced Transformer with Rotary Position Embedding. *arXiv e-prints*, art. arXiv:2104.09864, April 2021. doi: 10.48550/arXiv.2104.09864.

Zewei Sun, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Shujian Huang, Jiajun Chen, and Lei Li. Rethinking document-level neural machine translation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 3537–3548, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.279. URL https://aclanthology.org/2022.findings-acl.279.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/file/a14ac55a4f27472c5d894ec1c3c743d2-Paper.pdf.

Antonio Toral. Reassessing claims of human parity and super-human performance in machine translation at WMT 2019. In André Martins, Helena Moniz, Sara Fumega, Bruno Martins, Fernando Batista, Luisa Coheur, Carla Parra, Isabel Trancoso, Marco Turchi, Arianna Bisazza, Joss Moorkens, Ana Guerberof, Mary Nurminen, Lena Marg, and Mikel L. Forcada (eds.), *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pp. 185–194, Lisboa, Portugal, November 2020. European Association for Machine Translation. URL https://aclanthology.org/2020.eamt-1.20.

Antonio Toral and Andy Way. What Level of Quality can Neural Machine Translation Attain on Literary Text? *arXiv e-prints*, art. arXiv:1801.04962, January 2018. doi: 10.48550/arXiv.1801.04962.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and Efficient Foundation Language Models. *arXiv e-prints*, art. arXiv:2302.13971, February 2023. doi: 10.48550/arXiv.2302.13971.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Giorgos Vernikos, Brian Thompson, Prashant Mathur, and Marcello Federico. Embarrassingly easy document-level MT metrics: How to convert any pretrained metric into a document-level metric. In Philipp Koehn, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri (eds.), *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pp. 118–128, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.wmt-1.6.

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. Context-aware neural machine translation learns anaphora resolution. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1264–1274, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1117. URL https://aclanthology.org/P18-1117.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. Document-level machine translation with large language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 16646–16661, Singapore, December 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.1036. URL https://aclanthology.org/2023.emnlp-main.1036.

Longyue Wang, Zhaopeng Tu, Yan Gu, Siyou Liu, Dian Yu, Qingsong Ma, Chenyang Lyu, Liting Zhou, Chao-Hong Liu, Yufeng Ma, Weiyu Chen, Yvette Graham, Bonnie Webber, Philipp Koehn, Andy Way, Yulin Yuan, and Shuming Shi. Findings of the WMT 2023 shared task on discourse-level literary translation: A fresh orb in the cosmos of LLMs. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz (eds.), *Proceedings of the Eighth Conference on Machine Translation*, pp. 55–67, Singapore, December 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.3. URL https://aclanthology.org/2023.wmt-1.3.

Yutong Wang, Jiali Zeng, Xuebo Liu, Derek F. Wong, Fandong Meng, Jie Zhou, and Min Zhang. DelTA: An Online Document-Level Translation Agent Based on Multi-Level Memory. *arXiv e-prints*, art. arXiv:2410.08143, October 2024. doi: 10.48550/arXiv.2410.08143.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned Language Models Are Zero-Shot Learners. *arXiv e-prints*, art. arXiv:2109.01652, September 2021. doi: 10.48550/arXiv.2109.01652.

Billy T. M. Wong and Chunyu Kit. Extending machine translation evaluation metrics with lexical cohesion to document level. In Jun'ichi Tsujii, James Henderson, and Marius Paşca (eds.), *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 1060–1068, Jeju Island, Korea, July 2012. Association for Computational Linguistics. URL https://aclanthology.org/D12-1097.

Minghao Wu, George Foster, Lizhen Qu, and Gholamreza Haffari. Document flattening: Beyond concatenating context for document-level neural machine translation. In Andreas Vlachos and Isabelle Augenstein (eds.), *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 448–462, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.33. URL https://aclanthology.org/2023.eacl-main.33.

Minghao Wu, Thuy-Trang Vu, Lizhen Qu, George Foster, and Gholamreza Haffari. Adapting Large Language Models for Document-Level Machine Translation. *arXiv e-prints*, art. arXiv:2401.06468, January 2024. doi: 10.48550/arXiv.2401.06468.

Minghao Wu, Yufei Wang, George Foster, Lizhen Qu, and Gholamreza Haffari. Importance-aware data augmentation for document-level neural machine translation. In Yvette Graham and Matthew Purver (eds.), *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 740–752, St. Julian's, Malta, March 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.eacl-long.44. URL https://aclanthology.org/2024.eacl-long.44/.

Minghao Wu, Yulin Yuan, Gholamreza Haffari, and Longyue Wang. (perhaps) beyond human translation: Harnessing multi-agent collaboration for translating ultra-long literary texts, 2024b. URL https://arxiv.org/abs/2405.11804.

Tong Xiao, Jingbo Zhu, Shujie Yao, and Hao Zhang. Document-level consistency verification in machine translation. In *Proceedings of Machine Translation Summit XIII: Papers*, Xiamen, China, September 19-23 2011. URL https://aclanthology.org/2011.mtsummit-papers.13.

Yuhao Xie, Zongyao Li, Zhanglin Wu, Daimeng Wei, Xiaoyu Chen, Zhiqiang Rao, Shaojun Li, Hengchao Shang, Jiaxin Guo, Lizhi Lei, Hao Yang, and Yanfei Jiang. HW-TSC's submissions to the WMT23 discourse-level literary translation shared task. In

Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz (eds.), *Proceedings of the Eighth Conference on Machine Translation*, pp. 302–306, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.32. URL https://aclanthology.org/2023.wmt-1.32/.

Hao Xiong, Zhongjun He, Hua Wu, and Haifeng Wang. Modeling Coherence for Discourse Neural Machine Translation. *arXiv e-prints*, art. arXiv:1811.05683, November 2018. doi: 10.48550/arXiv.1811.05683.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. A Paradigm Shift in Machine Translation: Boosting Translation Performance of Large Language Models. *arXiv e-prints*, art. arXiv:2309.11674, September 2023. doi: 10.48550/arXiv.2309.11674.

Biao Zhang, Barry Haddow, and Alexandra Birch. Prompting large language model for machine translation: A case study. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 41092–41110. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/zhang23m.html.

Jiacheng Zhang, Huanbo Luan, Maosong Sun, Feifei Zhai, Jingfang Xu, Min Zhang, and Yang Liu. Improving the transformer translation model with document-level context. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 533–542, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1049. URL https://aclanthology.org/D18-1049.

Pei Zhang, Boxing Chen, Niyu Ge, and Kai Fan. Long-short term masking transformer: A simple but effective baseline for document-level neural machine translation. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1081–1087, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.81. URL https://aclanthology.org/2020.emnlp-main.81.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. The United Nations parallel corpus v1.0. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 3530–3534, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL https://aclanthology.org/L16-1561.

A Document-level Training Hyperparameters

Table 7 contains the hyperparameter configuration for the training of DocMT-LLMs.

Batch size	32
Number of Epochs	2
Learning rate	7×10^{-6}
LR Scheduler	cosine
Warmup Steps	125
Weight Decay	0.01
Optimizer	Adam (Kingma & Ba, 2014)
Adam β_1	0.9
Adam β_2	0.999
Adam ϵ	1×10^{-8}
Maximum Sequence Length	32768

Table 7: Hyperparameter configuration to fine-tune DocMT-LLMs on DocBlocks.

B Full Translation Results

In this appendix, we provide detailed sentence-level and document-level evaluation results, broken down by individual language pairs. Tables 8, 9, and 10 report document-level evaluation results across IWSLT2017 language pairs using d-BLEU, d-COMET, and LTCR, respectively – similar to the results presented in Table 3. Additionally, Tables 11, 12, and 13 present the same comparisons as Table 5 in the main paper, illustrating the impact of document-level training on sentence-level translation performance across diverse datasets and language directions.

Models	IWSLT2017 (en→xx) de	es	fr	it	ko	nl	pt	ru	zh
Llama-3.3-70B-Instruct	30.51	35.17	34.45	28.96	19.59	**32.20**	25.08	15.12	22.64
Qwen2.5-72B-Instruct	24.86	34.34	34.63	28.65	20.54	30.73	24.74	13.50	27.32
GPT-4o	30.49	**50.21**	33.51	**30.29**	**22.71**	30.39	**42.21**	**35.21**	15.21
TowerInstruct-Mistral-7B	2.63	2.07	3.82	2.36	0.41	3.00	4.12	1.83	1.44
DocMT-TowerInstruct-Mistral-7B	28.58	23.92	**35.45**	29.34	18.59	26.51	27.72	13.15	27.13
EuroLLM-9B-Instruct	4.42	5.53	6.65	3.45	0.33	3.29	3.75	4.69	1.64
DocMT-EuroLLM-9B-Instruct	21.21	32.19	19.91	15.98	14.30	18.98	29.62	12.49	17.48
Qwen2.5-7B-Instruct	26.51	31.88	30.20	24.80	0.27	29.28	23.35	9.84	11.62
DocMT-Qwen2.5-7B-Instruct	**31.08**	33.59	34.85	23.19	19.43	27.83	20.52	8.09	**29.03**
Models	**IWSLT (xx→en) de**	**es**	**fr**	**it**	**ko**	**nl**	**pt**	**ru**	**zh**
Llama-3.3-70B-Instruct	46.02	**45.55**	45.56	**45.21**	16.63	**48.87**	**40.60**	28.75	24.01
Qwen2.5-72B-Instruct	**46.07**	44.11	45.66	44.86	24.86	48.52	40.06	28.20	26.96
GPT-4o	42.36	40.36	41.36	39.36	13.36	43.36	36.36	23.36	21.36
TowerInstruct-Mistral-7B	3.15	2.49	3.57	2.94	0.78	4.08	2.97	0.95	0.32
DocMT-TowerInstruct-Mistral-7B	18.38	31.98	23.01	38.61	30.79	19.31	27.99	17.84	**33.20**
EuroLLM-9B-Instruct	6.31	0.55	6.86	4.29	3.74	4.46	1.91	4.66	2.95
DocMT-EuroLLM-9B-Instruct	31.92	33.87	21.32	22.43	28.35	37.94	30.41	18.37	31.54
Qwen2.5-7B-Instruct	40.78	40.28	41.39	41.38	20.83	41.15	36.61	24.70	24.18
DocMT-Qwen2.5-7B-Instruct	46.00	44.05	**47.33**	44.93	**36.94**	46.45	37.88	**30.35**	31.11

Table 8: Evaluation based on d-BLEU across IWSLT2017 language pairs.

Models	IWSLT2017 (en→xx)								
	de	**es**	**fr**	**it**	**ko**	**nl**	**pt**	**ru**	**zh**
Llama-3.3-70B-Instruct	67.90	71.31	66.91	69.81	61.11	72.49	70.33	74.00	81.18
Qwen2.5-72B-Instruct	71.26	**73.16**	69.50	72.49	**81.45**	74.55	72.03	**77.82**	84.08
GPT-4o	74.90	72.38	64.67	75.39	80.69	76.49	69.38	74.38	**84.30**
TowerInstruct-Mistral-7B	27.78	28.63	30.04	27.18	30.57	27.29	30.41	30.72	34.50
DocMT-TowerInstruct-Mistral-7B	**76.39**	63.74	**72.47**	**77.51**	74.62	**79.13**	**74.04**	75.88	75.64
EuroLLM-9B-Instruct	20.70	30.50	20.49	24.24	20.72	23.78	29.04	33.26	21.99
DocMT-EuroLLM-9B-Instruct	64.01	66.37	58.64	63.91	57.12	67.94	68.34	69.65	39.68
Qwen2.5-7B-Instruct	68.92	71.42	66.51	68.39	76.40	69.91	69.49	71.45	83.08
DocMT-Qwen2.5-7B-Instruct	71.13	72.75	68.55	72.59	77.39	75.15	64.40	71.84	83.21
Models	**IWSLT2017 (xx→en)**								
	de	**es**	**fr**	**it**	**ko**	**nl**	**pt**	**ru**	**zh**
Llama-3.3-70B-Instruct	71.14	72.82	72.24	69.97	78.11	67.30	69.29	73.97	73.24
Qwen2.5-72B-Instruct	68.95	70.49	65.72	62.37	70.45	64.92	71.12	70.39	71.55
GPT-4o	**75.74**	**73.45**	66.79	**76.45**	78.95	**77.25**	67.89	**75.60**	**83.25**
TowerInstruct-Mistral-7B	34.08	32.01	30.77	33.88	30.93	33.34	31.87	24.63	34.59
DocMT-TowerInstruct-Mistral-7B	55.78	68.36	51.34	64.15	54.80	54.30	69.13	68.47	56.30
EuroLLM-9B-Instruct	29.38	30.10	28.25	29.48	29.29	27.55	32.17	30.90	31.15
DocMT-EuroLLM-9B-Instruct	63.62	70.64	53.36	68.65	44.61	67.55	72.65	71.65	51.67
Qwen2.5-7B-Instruct	72.54	71.90	71.66	73.81	77.42	74.23	73.40	73.47	71.17
DocMT-Qwen2.5-7B-Instruct	74.63	63.36	**73.28**	74.80	**79.14**	75.39	**74.33**	72.64	69.60

Table 9: Evaluation based on d-COMET across IWSLT2017 language pairs.

Models	IWSLT2017 (en→xx)								
	de	**es**	**fr**	**it**	**ko**	**nl**	**pt**	**ru**	**zh**
Llama-3.3-70B-Instruct	60.37	76.33	61.01	64.43	24.09	58.58	62.65	**56.77**	54.52
Qwen2.5-72B-Instruct	62.41	**79.40**	61.25	66.03	41.19	59.35	65.65	56.23	**55.24**
GPT-4o	**71.83**	61.97	60.32	65.66	**56.01**	61.75	57.97	48.97	46.97
TowerInstruct-Mistral-7B	25.96	26.55	24.68	44.73	26.17	27.10	29.06	19.27	24.24
DocMT-TowerInstruct-Mistral-7B	59.93	72.67	**64.31**	63.59	41.62	66.95	**68.64**	55.77	50.66
EuroLLM-9B-Instruct	45.05	33.63	48.45	48.25	26.78	37.82	26.75	17.58	32.48
DocMT-EuroLLM-9B-Instruct	56.79	73.23	60.23	62.60	31.65	47.98	68.51	48.13	50.11
Qwen2.5-7B-Instruct	57.74	70.77	59.10	61.28	37.16	54.79	65.87	47.87	52.66
DocMT-Qwen2.5-7B-Instruct	61.21	76.78	62.86	**71.58**	45.22	**67.71**	68.25	50.32	54.51
Models	**IWSLT (xx→en)**								
	de	**es**	**fr**	**it**	**ko**	**nl**	**pt**	**ru**	**zh**
Llama-3.3-70B-Instruct	84.69	88.38	88.54	88.60	69.32	**88.49**	81.22	77.56	73.02
Qwen2.5-72B-Instruct	83.00	87.61	87.90	87.80	71.46	85.79	**81.35**	77.60	71.63
GPT-4o	82.07	83.98	82.79	85.40	65.81	79.92	77.23	74.94	68.75
TowerInstruct-Mistral-7B	30.12	32.42	27.00	38.22	27.69	29.69	21.60	24.88	28.29
DocMT-TowerInstruct-Mistral-7B	84.28	**90.10**	87.97	**91.87**	60.02	82.90	80.16	**79.26**	63.16
EuroLLM-9B-Instruct	39.17	34.80	40.49	35.17	26.13	56.22	28.75	20.54	22.83
DocMT-EuroLLM-9B-Instruct	77.73	83.92	81.75	86.86	69.93	79.49	80.07	75.73	68.33
Qwen2.5-7B-Instruct	82.55	83.48	89.01	88.71	68.41	86.08	79.33	75.44	71.49
DocMT-Qwen2.5-7B-Instruct	**85.65**	84.49	**91.16**	91.72	**74.74**	86.82	80.88	76.26	**85.09**

Table 10: Evaluation based on LTCR across IWSLT2017 language pairs.

Models	WMT23					
	en→de	en→ru	en→zh	de→en	ru→en	zh→en
TowerInstruct-Mistral-7B	83.91	85.77	86.10	85.48	83.15	80.46
DocMT-TowerInstruct-Mistral-7B	83.70	85.47	85.62	85.39	82.93	80.59
EuroLLM-9B-Instruct	84.83	85.64	86.15	85.48	83.34	80.90
DocMT-EuroLLM-9B-Instruct	84.55	85.62	85.61	85.31	82.72	80.61
Qwen2.5-7B-Instruct	78.45	76.78	84.45	81.92	78.92	78.59
DocMT-Qwen2.5-7B-Instruct	81.86	84.21	85.66	83.44	81.73	80.23

Table 11: Evaluation based on COMET across WMT23 language pairs.

Models	TICO-19				
	en→es	en→fr	en→pt	en→ru	en→zh
TowerInstruct-Mistral-7B	88.60	82.03	89.51	88.31	88.85
DocMT-TowerInstruct-Mistral-7B	88.73	81.72	89.55	88.16	87.70
EuroLLM-9B-Instruct	88.61	81.82	90.01	88.48	88.74
DocMT-EuroLLM-9B-Instruct	88.96	81.94	89.49	88.82	86.79
Qwen2.5-7B-Instruct	85.51	78.17	87.56	78.77	86.79
DocMT-Qwen2.5-7B-Instruct	88.04	81.03	88.99	87.12	87.37

Table 12: Evaluation based on COMET across TICO-19 language pairs.

Models	FLORES-200 (en→xx)								
	de	es	fr	it	ko	nl	pt	ru	zh
TowerInstruct-Mistral-7B	88.25	87.05	88.89	89.20	90.11	88.67	89.77	90.13	88.56
DocMT-TowerInstruct-Mistral-7B	88.52	86.79	88.55	88.75	89.20	88.59	89.68	89.76	87.01
EuroLLM-9B-Instruct	88.92	87.38	89.10	89.29	89.96	88.68	90.39	90.38	88.79
DocMT-EuroLLM-9B-Instruct	88.71	86.94	88.74	88.94	89.81	88.59	89.39	90.26	87.73
Qwen2.5-7B-Instruct	83.21	84.10	84.92	83.77	77.81	80.24	87.04	79.40	87.21
DocMT-Qwen2.5-7B-Instruct	86.97	86.37	87.81	87.71	88.39	86.53	88.93	88.94	88.28
Models	**FLORES-200 (xx→en)**								
	de	**es**	**fr**	**it**	**ko**	**nl**	**pt**	**ru**	**zh**
TowerInstruct-Mistral-7B	89.50	87.62	89.55	88.50	88.50	87.95	90.07	87.09	87.25
DocMT-TowerInstruct-Mistral-7B	89.47	87.24	89.44	88.20	88.30	87.67	89.78	86.84	86.95
EuroLLM-9B-Instruct	89.41	87.25	89.59	88.29	88.54	87.59	89.75	86.93	87.44
DocMT-EuroLLM-9B-Instruct	89.65	87.77	89.60	88.39	88.35	87.57	90.02	87.17	87.17
Qwen2.5-7B-Instruct	87.98	86.24	88.37	86.82	85.73	85.59	87.92	85.37	86.28
DocMT-Qwen2.5-7B-Instruct	89.25	87.38	89.20	88.04	87.74	87.41	89.58	86.77	87.16

Table 13: Evaluation based on COMET across FLORES-200 language pairs.