Benchmarking and Improving Long-Text Translation with Large Language Models

Anonymous ACL submission

Abstract

Recent studies have illuminated the promising capabilities of large language models (LLMs) in handling long texts. However, their performance in machine translation (MT) of long documents remains underexplored. This paper aims to shed light on how LLMs navigate this complex task, offering a comprehensive evaluation of their capabilities and limitations in long-text MT. First, we collect and construct an instruction-based benchmark dataset, specifically designed for the finetuning and evaluation of LLMs, encompassing multilingual, multi-domain, and document-level parallel data. Second, we conduct a comprehensive comparison between MT and LLM models concerning document-level translation. Our analysis uncovers that LLMs exhibit shortcomings in long-text domains, and their performance diminishes as document size escalates. By exploiting various extrapolation strategies, we enhance the capacity of LLMs to translate longer texts. We will release data, code, and models, which we hope can promote research in this field.

1 Introduction

Recently, significant progress has been made in the field of large language models (LLMs) such as ChatGPT (Ouyang et al., 2022; OpenAI, 2023). These models have demonstrated remarkable capabilities in processing extensive texts, opening up new avenues for a variety of natural language processing (NLP) tasks (Li et al., 2023). Specifically, research on long-text LLMs primarily focuses on two aspects: 1) *Maximum Capacity*: how long text a model can handle (Chen et al., 2023; Ding et al., 2023); 2) *Efficient Utilization*: how efficiently that model can model long texts (Liu et al., 2023). Even though LLMs can process long texts, it does not necessarily imply they can effectively model such texts. The effectiveness and necessity of extrapolating LLMs for various NLP tasks remain ambiguous. In this paper, we predominantly delve into the use of LLMs for long-text machine translation (MT).

Diverging from other long-text tasks such as text summarization and completion, which involve either long input or output, long-text MT presents unique challenges. It necessitates not only comprehending extensive input but also preserving context and coherence over extended text sequences (Yang et al., 2019, 2020). This amalgamation of requirements intensifies the complexity of the task, positioning long-text MT as a rigorous benchmark for evaluating the potential and prowess of LLMs.

To facilitate this study, we first construct an instruction-based benchmark dataset designed for finetuning and evaluating the document-level translation capabilities LLMs. This dataset comprises a total of 410K documents with a document length 160∼54K tokens, spanning across seven domains (i.e. news, TED, subtitles, Parliamentary, social, Q&A, novels) and three language pairs (i.e. Chinese-English, German-English, Russian-English). Leveraging this dataset, we conduct a comprehensive comparison between MT and LLM models (e.g. DeepL, GPT-4, Llama). The experiments reveal that 1) all models experience a performance decline in the novels domain due to its exceedingly long nature; 2) the performance of LLMs drastically deteriorates as the document length increases from 2K to 8K. To further investigate the impact of enlarging document length on translation quality, we examine a variety of extrapolation strategies during finetuning stage. Our findings indicate that a mixture of document lengths with multi-task learning can effectively enhance LLMs' performance. The **main contributions** are:

- To facilitate future research on long-text translation with LLMs, we publicly release a large instruction-based dataset.This dataset is comprehensive, covering an array of domains and languages, and notably includes exceptionally long texts (Section 2).
- We establish a long-text MT benchmark by eval-

ID	Instruction	Instruction-Based Example
1	Give the TGT translation for these SRC sentences: **D**	
2	Please provide the TGT translation for these sentences: **D**	
3	Provide the TGT equivalent for the following SRC sentences: **D**	
4	Translate the following SRC sentences into TGT: **D**	
5	Translate the following sentences from SRC to TGT: **D**	
6	Convert the following sentences from SRC to TGT: **D**	
7	Render the following SRC sentences into TGT: **D**	
8	What is the TGT version of these SRC sentences? **D**	
9	What do the following sentences mean in TGT? **D**	
10	What are the translations of these SRC sentences in TGT? **D**	
11	How do these SRC sentences translate to TGT? **D**	
12	Give the TGT translation for this SRC **C**: **D**	
13	Please provide the TGT translation for the **C**: **D**	
14	Provide the TGT equivalent for the following SRC **C**: **D**	
15	Translate this **C** from SRC to TGT: **D**	
16	Translate the following SRC **C** into TGT: **D**	
17	Convert the following **C** from SRC to TGT: **D**	
18	Render the following SRC **C** into TGT: **D**	
19	What is the TGT version of this SRC **C**? **D**	
20	What do the following **C** mean in TGT? **D**	
21	What is the translation of this SRC **C** in TGT? **D**	
22	How do this SRC **C** translate to TGT? **D**	

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction:
Give the English translation for these Chinese sentences:

Input:
狂风吹过峡谷，在草原上掀起了阵阵波澜。正在俯首啃食着青嫩草叶的野牛抬起头来，望向天空。

Response:
As the wind howled through the valley, ripples spread across the grassland. The wild ox, with its head still buried in the green leaves of grass, raised its head and looked up at the sky.

Each instance consists of 4 parts: system prompt, instruction, input and response.

Table 1: The instructions used for document-level translation. SRC and TGT denote source and target languages, respectively. **D** represents the source sentences. In #12∼22 prompts, we randomly select one word from the candidates **C** ∈ ["document", "text", "paragraph", "passage"].

uating advanced MT and LLM models. This exploration not only highlights the limitations of current LLMs but also provides insights for improving their performance in translating long texts (Section 3).

- We investigate simple finetuning strategies, and identify effective methods to enhance the LLMs' proficiency in translating extensive texts (Section 4).

2 Instruction-Based Benchmark Dataset

According to Taori et al. (2023), we construct instruction-based datasets for long-text translation training and testing, spanning various domains and languages. In addition to leveraging existing parallel datasets, we create two challenging testsets.

2.1 Overview and Statistics

Following Wang et al. (2023), we ask GPT-4 to suggest a variety of instructions and manually refine them for the long-text translation task (as enumerated in Table 1). We then generate instruction-based instances by integrating the original document-level sentence pairs with these instructions. As seen, each instance consists of four parts: 1) *System Prompt*, a predefined text to set the context or frame for the model's response; 2) *Instruction*, describing the specific task; 3) *Input*, a sentence in the source language; and 4) *Response*, the corresponding translation in the target language. In the MT scenario, the *Input* and *Response* are parallel documents (e.g. Chinese-English).

Table 2 outlines the languages, domains, and statistics of our dataset. We choose three language pairs: Chinese-English (Zh-En), German-English (De-En) and Russian-English (Ru-En). We select texts from seven domains, of which five are consistent across both training and testing (i.e. News, TED, Subtitle, Parliamentar and Novels), while the remaining two are exclusive to the testing set (i.e. Social and Q&A):

- *News.* We used WMT2008∼2021 News Devsets and WMT2022 News Commentary for training, and WMT2022 News Testset for testing.[1]
- *TED.* We collected TED talks in the IWSLT2015, 2017 and 2014 corpora for training, and the standard testsets for testing.[2]

[1] https://www.statmt.org/wmt22.
[2] https://wit3.fbk.eu.

	Domain	Source	Language	|D|	|S|	|T|	|T|/|D|
Training	News	WMT08∼21 News Devsets	Zh-En	1.0K	12.8K	876.2K/47.6K	648
			De-En	1.5K	32.6K	1.4M/997.2K	795
			Ru-En	1.0K	21.8K	986.3K/645.0K	820
		WMT22 News Commentary	Zh-En	8.0K	313.7K	19.5M/10.0M	1,845
			De-En	9.8K	388.5K	17.9M/12.3M	1,544
			Ru-En	8.8K	331.5K	18.3M/11.0M	1,671
	TED	IWSLT2015	Zh-En	1.7K	213.4K	9.3M/5.1M	4,181
		IWSLT2017	De-En	1.7K	212.9K	6.4M/5.0M	3,348
		IWSLT2014	Ru-En	1.4K	181.0K	6.2M/4.3M	3,729
	Subtitle	OpenSubtitles 2018	Zh-En	12.0K	11.2M	189.2M/109.5M	12,456
			De-En	29.2K	22.5M	251.4M/218.5M	8,057
			Ru-En	34.5K	25.9M	338.3M/249.7M	8,511
	Parliamentary	United Nations v1.0	Zh-En	91.0K	15.9M	983.5M/564.0M	8,501
			Ru-En	133.0K	23.2M	1.3B/798.4M	7,990
		Europarl v7	De-En	11.3K	1.9M	87.7M/62.1M	6,633
	Novels	Par3	Zh-En	3.7K	23.5K	906.7K/534.8K	196
			De-En	8.9K	35.4K	1.2M/938.8K	120
			Ru-En	34.0K	104.7K	3.2M/2.4M	82
		Ours	Zh-En	22.6K	1.9M	79.6M/48.8M	2,759
Testing	News	WMT2022	Zh⇒En	38	505	41.2K/26.4K	889
	Social			25	478	32.6K/19.7K	1,046
	Novels	mZPRT	Zh⇒En	12	857	40.7K/23.6K	2,679
	Q&A			182	1,171	34.6K/24.7K	163
	TED	IWSLT2015	Zh⇒En	62	6,047	257.9K/136.6K	3,181
		IWSLT2017	En⇒De	23	2,271	51.9K/64.4K	2,527
	News	WMT2022	En⇒De	31	511	16.9K/23.4K	511
	Parliamentary	Europarl v7		360	5,134	167.9K/236.8K	562
	Subtitle	**Ours**	En⇒Ru	13	22,602	193.8K/232.2K	16.4K
	Novels-XL	**Ours**	Zh⇒En	12	16,742	824.4K/474.4K	54.1K

Table 2: Statistics of our benchmark dataset. We count the number of documents |D|, sentences |S|, tokens |T| in terms of source/target language, and the average length of a document |T|/|D| (K thousand and M million).

- *Subtitle.* We collected movie subtitles in the OpenSubtitles2018 corpus for training.[3] As the existing Ru-En testset contains limited length of context (Voita et al., 2019a), we employed professional translators to create a new one.
- *Parliamentary.* We employed the United Nations v1.0 and Europarl v7 training and testing sets.[4]
- *Novels.* The Par3 corpus is organized at the paragraph level.[5] In addition, we release a copyrighted corpus: 23K continuous chapters from 152 web fictions, covering 14 genres such as fantasy science and romance. The texts are originally written in Chinese and were translated into English by professional translators. Regarding testsets, we utilize both mZPRT and our specially curated long-text.
- *Q&A and Social.* We used subsets in BaiduKnows and website domains from WMT2022 and mZPRT[6] corpora, respectively.

The average document length (|T|/|D|) can be considered as an indicator of the complexity inherent in a document-level dataset. Extended documents introduce a two-fold challenge: they not only test the model's capacity to process extensive sequences but also necessitate the understanding and generation of nuanced discourse properties across extended contexts. As seen, the document lengths of our constructed testsets are 16K and 54K, currently making them the longest document-level

[3] https://opus.nlpl.eu/OpenSubtitles-v2018.php.

[4] https://conferences.unite.un.org/uncorpus and https://www.statmt.org/europarl.

[5] https://github.com/katherinethai/par3.

[6] https://github.com/longyuewangdcu/mZPRT.

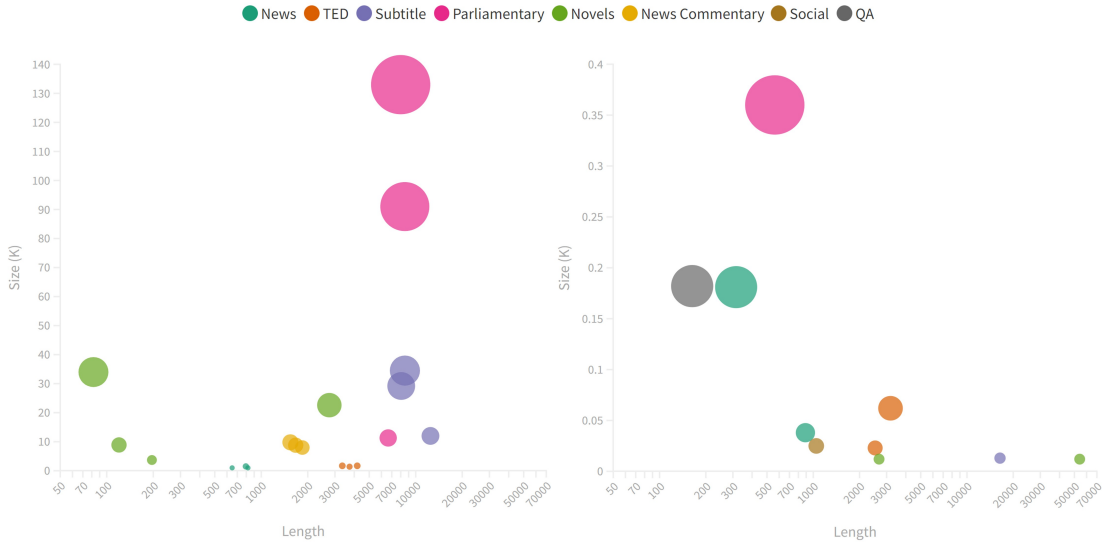

Figure 1: A visual representation of our dataset (left: training, right: testing). The x-axis denotes the average document length on a logarithmic scale, while the y-axis means the total number of documents in a dataset. Different colors correspond to distinct domains. Datasets from the same domain but in different languages are combined.

translation test sets available.

2.2 Analysis and Discussion

Data Distribution Figure 1 provides a visualization of the benchmark datasets, illustrating the natural distribution of data size and document length across various domains. As seen, 1) the data scale across different domains is imbalanced. For instance, there is an abundance of data in the Parliamentary, while spoken language domains such as TED are relatively scarce; 2) the natural length of documents varies across different domains. For example, Novels and Subtitles tend to be quite lengthy, whereas News are typically shorter; 3) the distribution between the training and testing sets is not strictly identical, reflecting a challenge that also exists in real-world applications.

Data Contamination There exists a risk of data contamination as publicly available data are often leveraged during different stages of LLM training (e.g., pre-training, SFT, or RLHF). To mitigate this risk, we have compiled three new datasets for training and testing models:

- *Novels Zh-En training set*. We crawl monolingual web novels and manually conduct cross-lingual alignment. The collected Par3 dataset is organized at the paragraph level, while ours is more extensive, organized at the chapter level.
- *Novels Zh-En testing set*. We collect latest Chinese web novels and then employ professional translators to translate them into English. We combine 20 continuous chapters into one document, resulting in a super-long document.
- *Subtitle Ru-En testing set*. We collect latest movie subtitles in English and engage professional translators to translate them into Russian.

3 Benchmark Results

We provide a comprehensive view of how MT and LLM systems handle the complexities of long-text translation across various languages and domains.

3.1 Experimental Setup

Models We utilize *testing sets* in the benchmark dataset. We conduct a systematic comparison of:

- *Commercial MT Systems.* The commercial translation products always perform best in translation performance. However, they typically preprocess documents into individual sentences and then perform translation on a sentence-by-sentence basis. We mainly compare Google Translate,[7] DeepL Translate,[8] and Tencent Translate,[9] as adopted in (Jiao et al., 2023b).
- *Commercial LLM Systems.* ChatGPT are known for their extensive context modeling capabilities and can achieve state-of-the-art results across various languages and tasks. Between August 1st and 30th, 2023, we used GPT-3.5 (4K) and

[7] https://translate.google.com.
[8] https://www.deepl.com.
[9] https://transmart.qq.com.

Model	#	Zh⇒En						En⇒De			En⇒Ru
		News	Soc.	Nov.	Q&A	TED	Nov._{XL}	TED	News	Par.	Sub.
Google	*sent*	27.7	35.4	16.0	12.0	28.1	26.6	32.6	51.9	32.2	29.1
DeepL	*sent*	30.3	33.4	16.1	11.9	29.1	23.3	33.5	48.8	34.7	27.7
Tencent	*sent*	29.3	38.8	20.7	15.0	30.4	31.4	32.3	48.8	31.9	24.3
GPT-3.5	*4K*	31.6	36.1	19.4	19.5	27.3	28.2	33.5	49.2	30.9	29.1
GPT-4	*8K*	32.6	41.9	19.8	21.0	29.0	23.9	33.8	50.8	31.9	28.9
Llama-2-chat	*4K*	9.3	7.4	4.4	9.7	10.8	10.3	8.0	18.7	11.0	4.4
Llama-mt	*2K*	29.1	30.2	18.4	7.3	28.7	26.3	32.5	46.9	31.8	27.8
Llama-mt	*4K*	29.0	26.2	16.8	5.6	27.9	23.0	32.7	46.6	31.9	25.8

Table 3: The comparison of MT and LLM systems on our benchmark *testing* set, spanning various domains and languages. We report performances using the case-sensitive d-BLEU metric. We evaluated the GPT-3.5-4K and GPT-4-8K in August 2023. The Llama-2-chat is a 7B model, which is pre-trained and fine-tuned with 4,096 max length. The Llama-mt are Llama 7B models, which are pre-trained with 2,048 max length, and finetuned on *training* set with pseudo-document length of 2,048 and 4,096. The numbers in color represent the best results within each category, while the highlighted background indicates the best performance across all models.

GPT-4 (8K) APIs to translate documents.[10]

- *Open-sourced LLM Models.* Llama and Llama-2-chat are widely-used, open-source LLMs, pre-trained and fine-tuned with a maximum length of 4,096 tokens (Touvron et al., 2023a,b). This enables to handle a relatively longer context window size, making it suitable for document translation. About Llama-mt, we enhance Llama on document-level MT by using the training set of our benchmark for SFT. We randomly select 5K instances from each subset in Table 2, resulting in approximately 200K instances.

Pre-processing Pseudo-Document In testing commercial and open-source LLM systems, we concatenate N consecutive sentences into a pseudo-document (Wang et al., 2017; Voita et al., 2019b; Wu et al., 2023) until it matches either the *maximum LLM input* or the *actual document length*. Assuming a 1:1.5 input to output ratio, we can calculate the maximum size of the *maximum LLM input* based on the total length $\in$ {2K, 4K, 8K}. We transform these pseudo-document parallel data into instruction formats as Jiao et al. (2023a).

Finetuning LLMs for Document Translation We finetune Llama-7B for 3 epochs with a batch size of 128. Our experiments are conducted on 16 NVIDIA A100 GPUs with DeepSpeed ZeRO stage 3 facilitating model parallelism.[11] To conserve GPU memory, we integrate Flash Attention mechanism (Dao et al., 2022). The SFT task is akin to the casual language modeling task, where the loss is computed solely based on the *output* portion of the input sequence.

Evaluation According to Wang et al. (2023), the document-level sacreBLEU (d-BLEU) (Liu et al., 2020) are consistent with performances of discourse-specific phenomena (e.g. consistency of terminology translation and accuracy zero pronoun translation). We employ case-sensitive d-BLEU as our metric, which is computed by matching n-grams in the whole document.

3.2 Main Results

Table 3 presents a comprehensive comparison of various MT and LLM models on *benchmark testing set* across different domains and languages. When evaluating individual testsets, we notice discernible differences between models. Taking Zh-En testsets for example, Tencent Translate excels in Novels, while GPT-4 is more adept with web-based content like Social and Q&A. Concerning language pairs, Google and DeepL models show proficiency in tasks targeting non-English languages, whereas GPT-4 predominates in tasks with English as the target. These variances can largely be attributed to differences in training data. Consequently, we will draw our conclusions holistically rather than relying solely on individual test sets.

On average across all testsets, GPT-4 outperforms all other models (i.e. 31.36 BLEU), particularly excelling in the Zh-En News, Social, Q&A, and En-De TED. The performance of GPT-3.5 is

[10] https://platform.openai.com.
[11] https://github.com/microsoft/DeepSpeed.

comparable to that of sentence-level MT systems, such as Tencent Translate (e.g. GPT-3.5 30.29 vs. Tencent Translate 30.48 BLEU). In contrast, Llama-2-chat registers a significantly lower score of 9.4 BLEU, attributable to its limited MT data during SFT and coverage of other languages (e.g. Russian). Addressing the need for long-text MT, we developed Llama-mt models by finetuning Llama using our benchmark training data. This adaptation enhances the translation quality of Llama by almost 200% (Llama-mt-2K 27.9 vs. Llama-2-chat 9.4 BLEU), placing it on par with commercial sentence-level MT systems (Llama-mt-2K 27.9 vs. DeepL 28.9 BLEU). Upon closer comparison between Llama-mt 2K and 4K, we observed a significant drop in performance, particularly on the Novel-XL testset, with a decrease of -3.3 BLEU. We cannot enhance the long-text translation capability of Llama by simply finetuning and decoding it with longer texts. This highlights the complexities and challenges inherent in attempting to bolster LLMs' proficiency in long-text translation.

When comparing the performance of ChatGPT reported in Wang et al. (2023), we notice a general upward trend across the overlapping testsets. While this might indicate performance enhancements over time with model evolution, it's crucial to acknowledge the different experimental parameters: 1) Their evaluations were conducted in March 2023, whereas ours took place in August 2023. 2) They accessed ChatGPT via its web interface, whereas we used its API. 3) Their methodology involved pseudo-documents of predetermined lengths (4 or 10 sentences), taking into account multi-turn contexts. Conversely, we employed the maximum lengths allowed by the API (2K or 4K) and disregarded multi-turn contexts.

3.3 Effects of Document Lengths

We further investigate the impacts of directly extending document length at decoding phase on the Nov.$_{XL}$ dataset (more analysis are discussed in Appendix §A.1). For comparison, we select two representative systems: DeepL and Llama-2-chat with 4K length of texts in pre-trained and SFT. Figure 2 shows the translation performance as the document length increases $\in$ {256, 1K, 2K, 4K, 8K} during the decoding phase. While the performance of sentence-level DeepL remains unaffected by decoding length, Llama-2-chat paints a different picture. It initially achieves its peak d-BLEU score at a document length of 1K but sees a sharp decline from 2K to 8K. This underscores that even though LLMs can process long texts, it does not necessarily imply they can effectively model such texts. This motivates us to exploit extrapolation methods to enhance the ability of LLMs to translate texts that extend beyond their pretraining length. Another interesting observation is that the MT performance of Llama-2-chat does not reach its zenith at 4K, diverging from the lengths used in pre-training and SFT. We suspect that this may be cased by various training signals at SFT phase. To focus specifically on the MT task, without being influenced by signals from other tasks, we primarily employ the Llama-mt settings for further experiments (in Section 4).

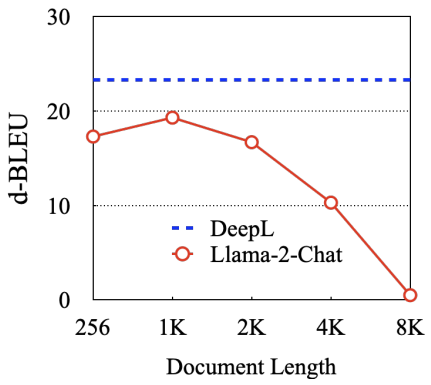

Figure 2: The effects of translation quality on document length. We draw *line chart* by extending the pseudo-document length during decoding on Novels-XL.

4 Context Extrapolation on Translation

In practice, it is unrealistic to keep increasing the context window of LLMs in the training process due to resource limits and unsatisfied performance (Chen et al., 2023). Therefore, it is interesting to investigate the potential of LLMs in translating documents that are noticeably longer than the predefined context window. In this section, we focus on curating the training data of long texts and explore various strategies for context extrapolation.

We employ Llama (7B) as testbed, which is pre-trained and fine-tuned with a maximum length of 2,048 token (Touvron et al., 2023a). Following the Llama-mt configurations detailed in Section 3.1, we pre-process each training subset into pseudo-documents with defined length. Then we randomly select 5K instances from each training subset, culminating ∼200K instances for SFT phase. About evaluation, we utilize Chinese⇒English

Finetune	Decode	*News*	*Novels*
SENT	sent	24.7	16.6
	doc (256)	23.8	15.9
	doc (512)	24.4	14.1
DOC (2048)	sent	27.5	16.1
	doc (1024)	**29.1**	18.3
	doc (2048)	**29.1**	**18.4**
	doc (4096)	22.6	9.2
DOC (4096)	sent	27.5	17.2
	doc (2048)	29.0	17.7
	doc (4096)	29.0	16.8
	doc (8192)	23.4	6.6

Table 4: The performance of document-level translation using different supervised finetuning (Finetune) and decoding (Decode) strategies in terms of d-BLEU scores. We conducte experimens on Llama 7b model on Chinese⇒English News and Novels datasets. The "DOC/doc (n)" denotes the documents with the size of context windows n used in finetuning/decoding while "sent" means the sentence level strategies.

Scaling		Setting		*News*	*Novels*
FT	DEC	CO	MDS		
2048	2048	✗	✗	29.1	18.4
		✓	✗	26.9	16.8
4096	4096	✗	✗	29.0	16.8
		✓	✗	24.3	17.8
		✗	✓	**29.5**	18.1
		✓	✓	26.8	17.7
4096	4608	✗	✗	29.3	17.3
		✗	✓	29.0	18.3
		✓	✓	26.6	17.1
	5120	✗	✗	29.1	16.0
		✗	✓	28.9	**18.7**
		✓	✓	26.2	17.0

Table 5: Effects of content overlapping (CO) and mixture of document lengths (MDL) on longer-text translation (2048→4096+) in terms of d-BLEU scores. The scaling document sizes is listed in Table 4, which is represented as finetuning (FT) and decoding (DEC).

WMT2022 News and mZPRT Novels as representative datasets and report d-BLEU score. We have examined different methods for utilizing data to improve the effectiveness and efficiency of modeling text at the document level. By leveraging these findings, we have improved the model's performance under data-constrained conditions. We will maintain the overall bilingual token count used during instruction finetuning.

4.1 Effects of Scaling Document Length

First of all, we present how LLMs perform when directly scaling the document size for training and decoding. For comparison, we also include the performance of the sentence-to-sentence approach like typical MT systems. At the SFT phase (FT), we employ two context windows (2048 and 4096), which are used to split the full documents with sliding block techniques. In the decoding phase (DEC), we assess LLMs on a variety of document lengths: shorter than the context window, matching the context window, and exceeding it. For example, when finetuning LLMs with the context window of 2048, we test on documents with context windows of 1024, 2048, and 4096, respectively. Here, 2048 corresponds to the context window employed during Llama's pre-training and finetuning, while 4096 represents an extrapolated context window.

Table 4 lists the results. Compared with the sentence-by-sentence approach (SENT-sent), increasing the context window during training can lead to much better translation results, which corroborates the findings of paragraph-level translation demonstrated in Karpinska and Iyyer (2023). Both approaches highlight the importance of context-aware translation techniques in achieving better results for long texts. Besides, we have two specific observations concerning the context windows: (1) With a context window of 2048 for SFT, the test performance on documents with the context window of 4096 degrades significantly, indicating the poor ability of LLMs in context extrapolation. (2) When conducting SFT with a context window of 4096, which is larger than that used in llama pre-training, we have not observed further improvements over the context window of 2048. Specifically, both DOC(4096)-doc(2048) and DOC(4096)-doc(4096) underperform DOC(2048)-doc(2048), especially on *Novels*, which indicates the ineffectiveness of merely increasing context window. Therefore, it becomes crucial to explore other methods for context extrapolation.

4.2 Effects of Content Overlapping and Mixture of Document Lengths

As an attempt, we explore the compositional generalization ability of LLMs in learning from short documents and generalizing to long documents. The default method for constructing training data has two problems that may limit the generalization

Scaling		+TE	*News*	*Novels*
DOC	doc			
4096	4096	✗	29.5	18.1
		✓	**29.9**	**19.2**
4096	4608	✗	29.0	18.3
		✓	29.7	19.1
	5120	✗	28.9	18.7
		✓	26.9	17.4

Table 6: Effects of multi-task finetuning (TE, text completion) on translating longer texts (2048→4096+) in terms of d-BLEU scores.

ability of LLMs: (1) The full documents are split into blocks with no overlap; (2) The length of input documents is fixed rather than following the length distribution in practice. For the first problems, we propose shift block to allow the content overlap between adjacent document blocks (CO, Content Overlapping). For the second, we propose to mix document blocks with varied sizes (MDS, Mixture of Document Sizes).

We present the results in Table 5. Unexpectedly, content overlapping leads to worse performance which may result from the less unique tokens during finetuning. In contrast, the model trained with the mixture of document sizes achieves noticeable improvement, compared to that with a fixed document size. The improvement is especially enlarged when we adopt a context window exceeding the one used in training, e.g., 5120 vs. 4096. It demonstrates that learning on various sizes of documents indeed improves the generalization of LLMs to longer documents.

4.3 Effects of Multi-task Finetuning

Unlike sentence-level translation, document-level translation tasks suffer from the scarcity issue of bilingual parallel corpora severely, which may limit the performance of long-text finetuning. We exploit the advantages of monolingual novel data and combine it with our document-level translation corpora in the form of text completion (TE). Similar to Section 2, we construct an instruction-based dataset for TE by combining instructions (in Table 7) with Chinese and English monolingual Novels (different from parallel data in Table 2). We employ the data split by MDS (in Section 4.2) and the scale of the data is equivalent to that of the Chinese-English document translation corpus.

As shown in Table 6, incorporating the monolingual text completion tasks can improvement the performance considerably. However, it can also degrade the translation performance slightly when extrapolating the context, making more hallucinations due to the relatively free generation style of text completion. Nonetheless, monolingual text completion tasks can improve the modeling ability of Llama models in non-English languages, and also ameliorate the language imbalance issue of English-centric parallel data. We will explore this direction more extensively in the future.

ID	Instruction
1	Given the story in the input, please provide a continuation in the output. **D**
2	Based on the narrative provided, kindly offer a follow-up in the resulting text. **D**
3	Taking into account the story presented, please supply a subsequent part in the response. **D**
4	Considering the tale in the input, please deliver an extension in the output. **D**
5	With the story mentioned, please generate a progression in the resulting content. **D**
6	In light of the story shared, kindly produce a sequel in the output. **D**
7	Given the plot in the input, please create a further development in the response. **D**
8	Keeping the story in mind, please contribute a continuation in the output. **D**
9	Reflecting on the narrative given, please furnish an additional segment in the response. **D**
10	Acknowledging the story in the input, please extend it in the output. **D**
11	Upon reviewing the story provided, kindly add a follow-up in the resulting text. **D**

Table 7: The prompts suggested by ChatGPT for story continuation. Chinese continuation and English continuation tasks use the same instruction. **D** represents the story that needs to be continued.

5 Conclusion

We provided valuable insights into the performance of LLMs in long-text translation. By creating a comprehensive, multi-domain, and multilingual document-level corpus for training and testing, we have been able to analyze the impact of various extrapolation methods on translation performance and propose a simple yet effective approach. By releasing data, code, and models, we aim to significantly promote further research in this field, ultimately improving the performance of LLMs and MT systems in handling long translation tasks.

Limitations

We list the main limitations of this work as follows:

1. *Extrapolation Methods*: This study focuses on enhancing long-text translation through data manipulation and training strategies. Nonetheless, a wide array of alternative approaches merit exploration for boosting model performance. Delving into these methodologies may reveal more efficacious means of bolstering the robustness and precision of predictive models.
2. *Evaluation Methods*: In this paper, we mainly use the commonly-used automatic metric. The d-BLEU allows for a more nuanced assessment of translation quality, reflecting the unique challenges and requirements of long-text translation. However, it is still an open question whether human and automatic evaluation metrics are complementary or mutually exclusive in measuring the document-level translation quality.

Ethics Statement

We take ethical considerations very seriously, and strictly adhere to the ACL Ethics Policy. Most datasets used in this paper are publicly available. Besides, we are the copyright owners of the newly proposed dataset. We ensure that the findings and conclusions of this paper are reported accurately and objectively.

References

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. 2023. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems*.

Jiayu Ding, Shuming Ma, Li Dong, Xingxing Zhang, Shaohan Huang, Wenhui Wang, and Furu Wei. 2023. Longnet: Scaling transformers to 1,000,000,000 tokens. *arXiv preprint arXiv:2307.02486*.

Wenxiang Jiao, Jen-tse Huang, Wenxuan Wang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023a. ParroT: Translating during chat using large language models. *arXiv preprint arXiv:2304.02426*.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023b. Is chatgpt a good translator? a preliminary study. *arXiv preprint arXiv:2301.08745*.

Marzena Karpinska and Mohit Iyyer. 2023. Large language models effectively leverage document-level context for literary translation, but critical errors persist. *arXiv preprint arXiv:2304.03245*.

Dacheng Li, Rulin Shao, Anze Xie, Ying Sheng, Lianmin Zheng, Joseph E. Gonzalez, Ion Stoica, Xuezhe Ma, and Hao Zhang. 2023. How long can open-source llms truly promise on context length?

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*.

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

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

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

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

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

Elena Voita, Rico Sennrich, and Ivan Titov. 2019a. 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.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019b. 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*.

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

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.

Minghao Wu, George Foster, Lizhen Qu, and Gholamreza Haffari. 2023. Document flattening: Beyond concatenating context for document-level neural machine translation. *arXiv preprint arXiv:2302.08079*.

Baosong Yang, Longyue Wang, Derek F Wong, Lidia S Chao, and Zhaopeng Tu. 2019. Convolutional self-attention networks. 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 4040–4045.

Yilin Yang, Longyue Wang, Shuming Shi, Prasad Tadepalli, Stefan Lee, and Zhaopeng Tu. 2020. On the sub-layer functionalities of transformer decoder. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4799–4811.

A Appendix

A.1 Effects of Different Domains and Languages

As shown in Figure 3, the Social and Q&A domains exhibit a more informal structure and length, leading to varied performance across models. Besides, most models consistently excel in the News and Parliamentary domains. However, they underperform in the Novel and Subtitle domains, which are characterized by its extended context. This underscores the challenges LLMs face in long-text translation and motivates our pursuit to further explore translations of longer texts. Concerning languages, most models exhibit consistent performance: they shine in En-De translation while showing relatively subdued results in Zh-En and En-Ru. Therefore, we will delve deeper into the impact of document lengths using the Zh-En Novels dataset.

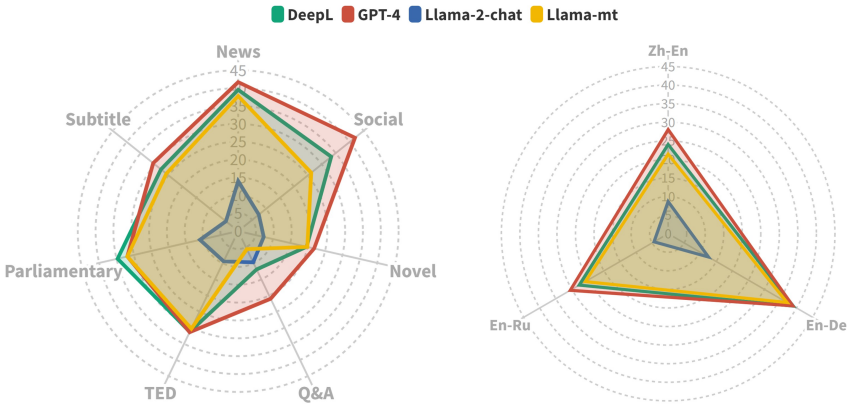

Figure 3: The effects of translation performance on domain and language. We draw *radar chart* by combining results in Table 3.