Towards Chapter-to-Chapter Literary Translation via Large Language Models

Anonymous ACL submission

Abstract

Discourse phenomena in existing document-level translation datasets are sparse, which has been a fundamental obstacle in the development of context-aware machine translation models. Moreover, most existing document-level corpora and context-aware machine translation methods rely on an unrealistic assumption on sentence-level alignments. To mitigate these issues, we first curate a novel dataset of Chinese-English literature, which consists of 132 books with intricate discourse structures. Then, we propose a more pragmatic and challenging setting for context-aware translation, termed chapter-to-chapter (CH2CH) translation, and investigate the performance of commonly used machine translation models under this setting. Furthermore, we introduce a potential approach to fine-tune large language models (LLMs) within the domain of CH2CH literary translation, yielding impressive improvements over baselines. Through our comprehensive analysis, we reveal that literary translation in the CH2CH setting is challenging in nature, with respect to both model learning methods and translation decoding algorithms.

1 Introduction

Despite the efforts on developing context-aware machine learning systems to meaningfully exploit inter-sentential information, recent work has investigated fundamental obstacles in existing document-level translation datasets and context-aware machine translation models (Jin et al., 2023). First, existing datasets lack the contextual information and/or discourse phenomena necessary for meaningful document-level translation (Lupo et al., 2022). Second, existing predominant context-aware translation methods assume sentence-level alignments available during training, which do not accurately represent real-world translation scenarios (Thai et al., 2022; Jin et al., 2023).

To remedy the issues, recent work has pivoted to literary translation and proposed a more realistic paragraph-to-paragraph setting, given that literary texts typically contain complex discourse structures that mandate a document-level frame of reference. Thai et al. (2022) released PAR3, a paragraph-level translation dataset sourced from recently-published 118 novels in 19 languages (about 6 novels per language on average). Jin et al. (2023) curated PARA2PARA, a small-scale dataset consisting of 10,545 parallel paragraphs across six novels. However, these datasets are either in small scale or the reference translations are automatically generated from machine translation systems (e.g. Google Translate (Wu et al., 2016) and fine-tuned GPT-3 (Brown et al., 2020)). In addition, there still exist some serious limitations in the paragraph-to-paragraph translation setting, including limited contextual information and equivocal paragraph splits in literary texts.

Large language models (LLMs) with decoder-only Transformer architectures have demonstrated outstanding performance as sentence-level translation systems (Vilar et al., 2023; Jiao et al., 2023; Kocmi and Federmann, 2023; Zhang et al., 2023; Yang et al., 2023). In the aspect of context-aware translation, recent studies have employed decoder-only LLMs to translate entire paragraphs using few-shot in-context learning methods, yielding impressive translation quality (Karpinska and Iyyer, 2023). However, how to finetune LLMs to process context-aware translation for literary texts in a more realistic and challenging scenario remains under-explored.

In this paper, we propose a more pragmatic and challenging setting for context-aware translation, named *chapter-to-chapter* (CH2CH), associated with a carefully curated dataset of Chinese-English literature. The dataset consists of 132 literary books, together with professional translations in Chinese. Then we investigate the performance of

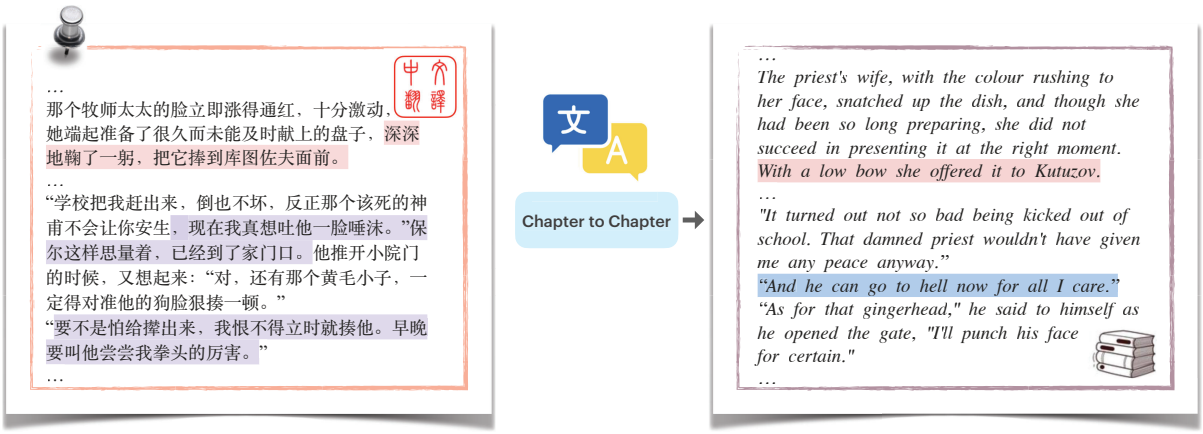

Figure 1: An example of CH2CH translation. Sentence Misalignment: Red parts are where a source sentence is separated into multiple sentences in the corresponding translation; blue parts are added by translators without a corresponding source segment; violet parts are deleted by translators.

commonly-used machine translation models under the proposed setting and dataset. In addition, we investigate the efficacy of applying LLMs in context-aware CH2CH literary translation and highlight several key challenges that impede the progress. Our main contributions are outlined as follows:

- We propose a more realistic setting for literary translation: chapter-to-chapter(CH2CH) translation, wherein a document is translated at the granularity of chapters. To support it, we release a chapter-aligned Chinese-English dataset (JAM), comprising 4,194 parallel chapters extracted from 132 novels, to catalyze future research endeavors.
- Through comprehensive analysis, we unveil the challenges in chapter-level translation, including long-context training and decoding strategies.
- With empirical experiments, we evaluate the performance of recent trending LLMs on the JAM dataset and propose an effective fine-tuning procedure tailored for LLMs to generate coherent translations of literary novels.

2 Preliminary Background

2.1 Context-aware NMT

Sentence-aligned Translation In the sentence-aligned setting of context-aware machine translation, we assume that the source and target sentences in a parallel document are well-aligned. Formally, given a document D comprising a set of source sentences $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_d\}$, there are the same number of sentences $\boldsymbol{Y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_d\}$ in the target side, which are aligned with sentences in $\boldsymbol{X}$ by the indices. The context-aware neural machine translation (NMT) model computes the probability of translating the source sentence $\boldsymbol{x}_i$ conditioned on the context C_i, wherein $0 \leq i \leq d$:

$$P_{\text{SentAlign}}(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{C}_i, \theta) = \prod_{j=1}^{N} P(y_i^j|y_i^{<j}, \boldsymbol{x}_i, C_i; \theta). \quad (1)$$

where C_i are contextual sentences surrounding $\boldsymbol{x}_i$ and/or $\boldsymbol{y}_i$. As illustrated in Figure 1, sentence-aligned translation does not accurately represent real-world translation scenarios.

Paragraph-to-Paragraph Translation To get rid of the assumption of sentence-level alignments and leverage richer contextual information, recent work (Thai et al., 2022; Jin et al., 2023) proposed a paradigm shift towards paragraph-to-paragraph (PARA2PARA) translation to relax the alignment assumption from sentence-level to paragraph-level. Concretely, a document D contains a set of aligned parallel paragraphs, $\boldsymbol{X} = \{\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_d\}$ and $\boldsymbol{Y} = \{\boldsymbol{Y}_1, \boldsymbol{Y}_2, ..., \boldsymbol{Y}_d\}$. Each pair of aligned paragraphs $\boldsymbol{X}_i$ and $\boldsymbol{Y}_i$ do not necessarily contain the same number of sentences:

$$P_{\text{Para2Para}}(\boldsymbol{Y}_i|\boldsymbol{X}_i, \theta) = \prod_{j=1}^{N} P(Y_i^j|Y_i^{<j}, \boldsymbol{X}_i; \theta) \quad (2)$$

where $Y_i^{<j}$ are all previously translated tokens in a paragraph. However, in literary texts the splits of paragraphs are equivocal, which limits the application of PARA2PARA to real-world scenario.

2.2 Datasets

Most commonly used corpora, including IWSLT-17 (Cettolo et al., 2012), NewsCom (Tiedemann, 2012), Europarl (Koehn, 2005), and OpenSubtitles (Lison et al., 2018) are sourced from news articles or parliamentary proceedings. Until recently,

Source	Target
"To think what we have been brought to!" Kutuzov cried suddenly, in a voice full of feeling, Prince Andrey's story evidently bringing vividly before him the position of Russia. "Wait a bit; wait a bit!" he added, with a vindictive look in his face, and apparently unwilling to continue a conversation that stirred him too deeply, he said: "I sent for you to keep you with me."	"弄到什么地步.....到什么地步！"库图佐夫突然说，他声音激动，显然，从安德烈公爵的叙述中，他清楚地想象到俄国目前的处境。"给我一段时间，给我一段时间！"他脸上带着愤怒的表情又说，很明显，他不愿继续这个使他激动的话题，他说："我叫你来，是想让你留在我身边。"
"We must, if everyone wants to; there is no help for it . . . But, mark my words, my dear boy! The strongest of all warriors are these two—time and patience. They do it all, and our wise counsellors n'entendent pas de cette oreille, voilà le mal. Some say ay, and some say no. What's one to do?" he asked, evidently expecting a reply. "Come, what would you have me do?" he repeated, and his eyes twinkled with a profound, shrewd expression. "I'll tell you what to do," he said, since Prince Andrey did not answer. "I'll tell you what to do. Dans le doute, mon cher"—he paused—"abstiens-toi." He articulated deliberately the French saying.	"打一仗是可以的，如果大家都愿意的话，没有什么可说的.....可是要知道，亲爱的朋友：没有比忍耐和时间这两个战士更强的了，这两位什么都能办成。可是顾问们不肯听这个，困难就在这里。一些人要这样，另一些又不这样。怎么办呢？"他问，显然在等着回答。 "你说说看，我怎么办？"他重复着，眼睛显得深沉、睿智。 "我告诉你怎么办。如果你犹豫不决，亲爱的，"他停了一下，"那你先干别的。"他慢条斯理地一字一句地说。

Table 1: Examples of paragraph misalignment. Each line represents an individual paragraph in the original text.

some document-level parallel corpora of literary texts have been released. Jiang et al. (2023) curated Bilingual Web Books (BWB), a sentence-aligned corpus that retains document-level information. BWB contains 9.6 million sentence pairs sourced from Chinese web novels and their corresponding English translations. However, BWB still follows the sentence-level alignment constrains. To support PARA2PARA translation, Thai et al. (2022) introduced PAR3, a paragraph-aligned corpus obtained through both human and automatic translators, containing multilingual non-English novels and their English translations. Al Ghussin et al. (2023) introduces a new paragraph-aligned corpus, with parallel paragraphs extracted from Paracrawl (Bañón et al., 2020) using automatic sentence alignments. This corpus includes data crawled from the Internet spanning various domains.

2.3 Translation with Large Language Models

LLMs are not explicitly trained on parallel data for translation, yet they possess a profound understanding of languages and can produce coherent text, serving as a valuable foundation for translation tasks (Li et al., 2024). Particularly for resource-rich languages, colossal models with decoder-only architecture, such as GPT-4 (OpenAI et al., 2024), have approached or even exceeded traditional encoder-decoder models on sentence-level benchmarks and can generate more coherent and human-like translations drawing upon their extensive comprehension of both languages (Robinson et al., 2023; Hendy et al., 2023). Xu et al. (2023a) proposed a two-stage procedure to finetune Llama2-7b (Touvron et al., 2023) with a small amount of sentence-level parallel data and obtained impressive improvements over standard sentence-level NMT baselines without LLMs.

3 JAM: Chapter-Aligned Literary Translation Dataset

3.1 Chapter-to-Chapter Translation

In literary texts, the lengths of paragraphs vary and the splits of paragraphs are equivocal, particularly when dialogues are involved. For instance, in novels, dialogue lines are often presented as separate paragraphs, making it challenging to ensure accurate translations without access to the preceding context. As illustrated by the two examples shown in Table 1 , there are instances where multiple paragraphs from the source side are merged into one paragraph on the target side, and vice versa.

To address this issue, we propose *chapter-to-chapter* (CH2CH) translation, a pragmatic and challenging setting, by extending context-aware translation to chapter-level. Comparing to paragraph-level alignments, chapter-level alignments provide the model with more comprehensive context from both the source and target texts. This richer context theoretically offers greater potential for improvements and helps mitigate issues such as tense mismatches, particularly in languages like Chinese that lack explicit tense markers (Sun et al., 2020).

To conduct experiments and facilitate future research endeavours on CH2CH translation, we curate a chapter-aligned dataset of English-Chinese

	CHAP. #	SENTENCE # (EN/ZH)	WORD # (EN/ZH)
TRAIN	3546	334.8K / 445.0K	7.4M / 8.6M
VALID	391	36.5K / 47.9K	796.1K / 935.9K
TEST	257	29.5K / 40.6K	648.4K / 795.3K
TOTAL	4194	400.7K / 533.6K	8.8M / 10.4M

Table 2: JAM Corpus Statistics.

literature, named JAM, which comprises 132 English classic novels alongside professional Chinese translations. In professional literary translation, translators often leverage contexts to enhance the fluency and readability of the translation. To this end, translations may not strictly adhere to sentence alignment[1], and some typical sentence misalignment types are listed below, an example is shown in Figure 1 illustrates:

INSERT : new sentence(s) is added by translators without a corresponding source segment.

DELETE : a source sentence(s) is deleted by translators in translation.

SPLIT : a source sentence is separated into multiple ones in the corresponding translation.

As such, CH2CH translation is challenging in nature, given that chapters typically are lengthy and contain complex discourse structure.

3.2 Data Construction and Quality Control

We collect 132 bilingual literary books across different genres from the Internet, and format data by manually correcting chapter-level alignment [2]. Subsequently, we perform standard data cleaning steps (e.g. punctuation normalization) and filter the chapter pairs with a sequence length ratio > 3.0. The refined dataset contains a total of 4194 aligned chapters. The statistics of this dataset are shown in Table 2 and detailed corpus information is in Appendix A.1. We randomly select 16 books as the test set. The remaining corpus of 3937 chapters from 116 books was then split into an 90% training set and a 10% validation set.

4 Experimental Setup

4.1 Baselines

To examine the inherent capacity of the model in the translation task, we perform a benchmarking analysis against two baseline categories: **Encoder-decoder** and **Decoder-only**.

[1]In 50 sampled paragraphs from JAM there are 18 paragraphs with sentence mis-alignments.

[2]We select literary works with chapter breaks, then manually check the alignments of the first and last paragraphs.

In CH2CH translation, we train the decoder-only model by concatenating each source chapter with its corresponding target chapter, demarcated by a <SEP> token, and ended with an <EOS> token:

```
<SRC Chapter> <SEP> <TGT Chapter> <EOS>
```

Motivated by Zhang et al. (2018), we experiment with training a baseline model on the JAM dataset from scratch, as well as incorporating pre-trained baselines. In the pre-trained baselines, the model is first trained on the sentence-level WMT22 Zh→En dataset (Kocmi et al., 2022), before further fine-tuning on the JAM dataset.

Zero-shot Evaluation Recent work has showcased the proficiency of LLMs in sentence-level translation. To further probe the ability of LLMs in translating literary, we randomly sample 63 chapters from JAM test set and conduct a zero-shot evaluation on the sampled instances to compare with the following models:

NLLB-200-3.3B (Team et al., 2022): an encoder-decoder LLM, with 3.3b parameters.

LLAMA2-7B (Touvron et al., 2023): a generative text model with 7b parameters.

LLAMA3-8B (Touvron et al., 2023): a generative text model with 8b parameters.

ALMA-7B (Xu et al., 2023a): finetuned on 5 language pairs from Llama2-7b.

GPT-4 (OpenAI et al., 2024): a pre-trained large-scale multi-modal model.

The technical details are in Appendix B.3.

Finetuning We select ALMA-7B to finetune on JAM because of its impressive gains in translation tasks compared to other LLMs; its fine-tuning process is divided into two phrases: first, ALMA-7B-Stage1 finetuned LLAMA2-7B exclusively on monolingual data; then, the second stage ALMA-7B-Stage2 is subsequently finetuned on parallel data. Specifically, we finetune ALMA-7B-Stage1 on JAM to investigate whether pretraining with sentence-level parallel data is beneficial prior to fine-tuning on chapter-level data. We use causal language modeling (CLM) loss for finetuning and restrict loss computation only to the target tokens.

4.2 Handling Long Chapters

As some chapters exceed the maximal context length of some models, we equally segment those chapters into chunks, ensuring that each chunk contains less than 2048 tokens in both Zh and En sides.

Model	WMT	JAM	BLEU	d-BLEU	BlonDe					COMET
				all	pron.	entity	tense	d.m.		
Encoder-Decoder	✗	✓	1.87	2.12	8.70	49.23	19.22	42.30	17.21	0.4128
Decoder-only	✗	✓	1.09	1.45	7.23	47.46	20.77	40.40	16.54	0.4187
Encoder-Decoder	✓	✓	14.38	15.07	31.08	**89.78**	11.36	**86.88**	**81.96**	0.6617
Decoder-only	✓	✓	13.35	13.94	30.06	84.28	14.59	80.23	76.81	0.6377
ALMA-7B-Stage1	✗	✓	15.70	16.26	33.46	74.28	30.62	70.11	71.72	0.7806
ALMA-7B-Stage2	✓	✓	**18.80**	**19.28**	**36.90**	81.34	**32.72**	77.83	76.81	**0.8025**

Table 3: Automatic metric results on JAM test set. Note chapters are segmented by maximum 2048 tokens. ALMA-7B-Stage1 is only fine-tuned on monolingual data. ALMA-7B-Stage2 fine-tunes ALMA-7B-Stage1 on high-quality parallel data. (✗) denotes no fine-tuning on corresponding dataset; (✓) denotes fine-tuning.

During decoding, we also pack the maximum number of sentences into blocks within 2048 tokens. The model does not know how many sentences to generate in advance and decoding stops when <EOS> is predicted. As illustrated in Figure 8, <EOS> in our experiments is used to indicate the end of translation, not the end of a sentence.

4.3 Evaluation

For all tasks, we report both "sentence-level"[3] (e.g., BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005) and COMET (Rei et al., 2020)) and document-level automatic metrics in evaluation. In particular, we analyze the translation quality of LLMs related to specific discourse phenomena such as pronoun ellipsis, named entity coreference by BlonDe score (Jiang et al., 2022).

5 Result and Analysis

5.1 CH2CH is Challenging in Nature.

Motivated by Zhang et al. (2018), we experiment with training a baseline model on the JAM dataset from scratch, as well as incorporating a two-stage training procedure, in which the model is first trained on the sentence-level WMT22 Zh→En dataset (Kocmi et al., 2022), before further fine-tuning on the JAM dataset.

As illustrates in Table 3, Encoder-Decoder and Decoder-only models trained from scratch on JAM significantly under-perform the models trained with the 2-stage procedure. The significant performance gap demonstrates the challenging nature of CH2CH (e.g., 1.87 and 1.09 on BLEU), i.e., the inherent difficulty of training on chapter-level, long-sequence data. Translation models that trained with the 2-stage procedure to leverage the sentence-level WMT22 exhibit a notable improvement, attesting the difficulty of the CH2CH translation task.

[3]Sentence-level metrics are performed on chunks, as our dataset has no sentence alignment, see details in B.2

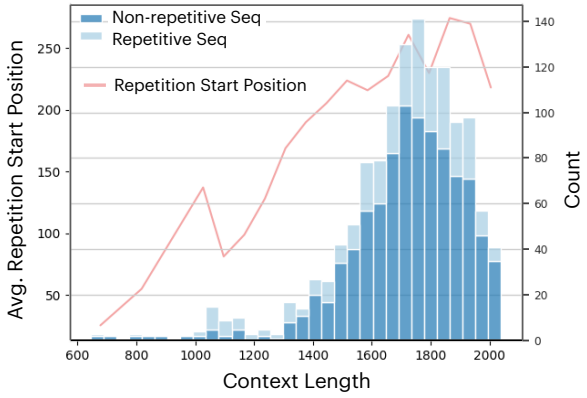

Figure 2: Repetition distribution.

5.2 Fine-tuning and Decoding Strategy

Does sentence-level fine-tuning help? We next investigate the prerequisite of sentence-level fine-tuning prior to the training on JAM dataset by comparing ALMA-7B-Stage1 and ALMA-7B-Stage2 respectively, with the latter has been fine-tuned on sentence-level parallel datasets. Table 3 indicates that such sentence-level fine-tuning improves BLEU from 15.7 to 18.80 and BlonDe from 33.46 to 36.95, suggesting that fine-tuning at sentence-level contributes positively to the accuracy of chapter-level literary translation.

In contrast, the improvement on COMET is marginal, possibly attributable to COMET's focus on assessing the coherence and fluency of the generated translations. These qualities might already be sufficiently robust in an LLM.

Repetition Problem in Decoding. Deutsch et al. (2023) founds that translation does not degrade as the sequence becomes longer. However, according to our results, this is not universally the case; the effectiveness of translation diminishes as the context becomes really lengthy. To investi-

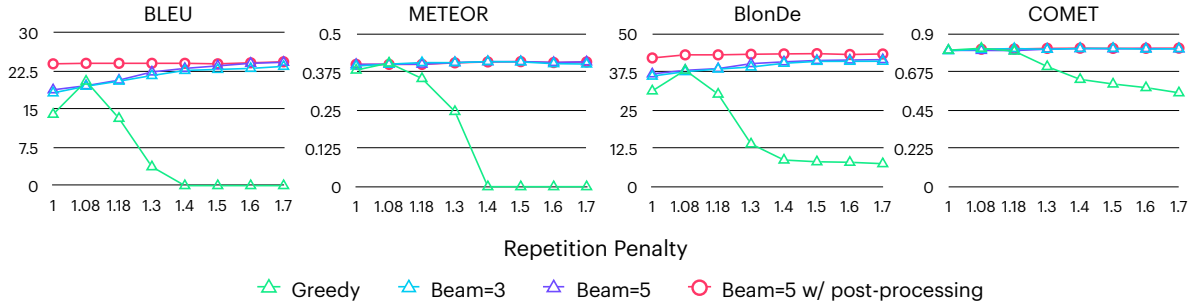

Figure 3: Automatic metric results across different decoding strategies. Repetition penalty $\gamma = 1$ represents pure greedy or beam search w/o penalty; $\gamma > 1$ denotes near-greedy decoding.

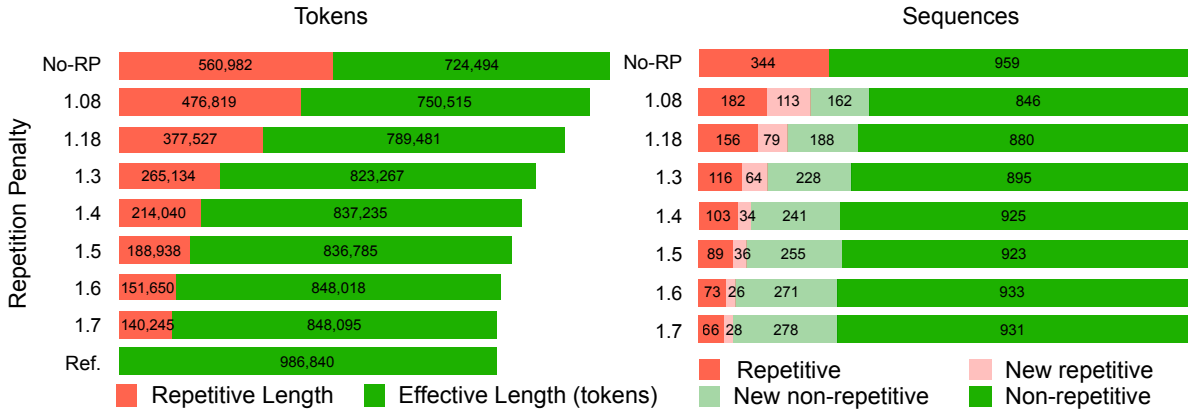

Figure 4: Left: **Effective token counts**; Right: **Sequence repetition analysis**. *(Non-)Repetitive* means sequences that staying (non-)repetitive w/ different γ; *New (non-)repetitive* indicates sequences that newly become (non-)repetitive. *No-RP* denotes no repetition penalty ($\gamma = 1$). *Ref.* means total number of tokens in the reference.

gate the insights, we examine the translations of JAM test set on the fine-tuned ALMA-7B-Stage2 model and observe a notable pattern of undesirable repetitions—either phrases or entire sentences—emerges within the translations.

Specifically, 26.4% of the translations within our test set exhibit some form of repetition. As illustrates in Figure 2, repetition occurs predominantly located within the first half of the translations (Shown as the red curve). Furthermore, sentences exceeding 1300 tokens are more likely to generate repetitive words, phrases or sentences[4]. This observation is consistent with earlier studies indicating text generation with LLMs often results in consecutive sentence-level repetitions, attributed to the use of maximization-based decoding algorithms (Holtzman et al., 2020; Xu et al., 2023b). The detailed analysis by Xu et al. (2022) sheds light on the underlying causes: these models have an inherent tendency to repeat previous sentences due to overestimating their probability. This problem is particularly evident in long-context translation, where increasing the chunk length amplifies the risk of the model falling into repetitive loops.

Comparison of Decoding Strategies By default, beam search is employed for all models, with beam size 5. However, upon training certain LLMs on the CH2CH task, we observe sub-optimal performance with beam search. We investigate the performance of three decoding strategy: *greedy*, *beam search* and *near greedy* decoding, which introduces repetition penalty γ to discount the scores of previously generated tokens (Keskar et al., 2019).

Figure 3 presents the effect of applying the penalty γ to both greedy and beam search decoding with different beam sizes. For beam search (with beam size = 3 or 5), both BLEU and BlonDe scores improve significantly. Concretely, with beam size = 5, BLEU and BlonDe increase from 18.80 to 24.20 and from 36.90 to 41.42, respectively. In contrast, the improvements in METEOR and COMET scores are comparatively smaller, suggesting that the overall translation quality may not be improving as expected. In addition, for beam search decoding, increasing γ keeps improving translation performance and there are marginal variances once $\gamma \geq 1.5$. For greedy decoding, however, translation quality rapidly declines when $\gamma > 1.2$.

[4]Repetition analysis for all zero-shot generations across various architectures are in Appendix B.5

γ	Rep.	New rep.	New non-rep.	Non-rep.
1.0	7.5 → 9.4	**18.2 → 18.2**	8.3 → 17.1	**22.5**
1.1	8.4 → 9.7	10.0 → 15.7	12.5 → 18.7	**22.5**
1.2	8.9 → 11.9	11.0 → 14.6	13.6 → 19.0	22.3
1.3	**11.1 → 13.0**	11.8 → 16.7	16.0 → 19.7	22.3
1.4	9.7 → 12.5	13.2 → 16.4	17.5 → 20.2	22.3
1.5	10.0 → 10.5	13.5 → 16.8	18.9 → 20.4	22.2
1.6	10.6 → 11.4	11.3 → 15.0	19.5 → 20.4	22.1
1.7	7.8 → 9.4	5.9 → 12.0	**20.7 → 20.7**	21.2

Table 4: BLEU scores across different groups. → denotes after post-processing.

We then explore the number of effective (i.e., non-repetitive) tokens generated as γ increases (Figure 4 (left)). We further analyze repetition sentence by sentence by separating test sequences into four categories: *repetitive*, *non-repetitive*, *new repetitive*, and *new non-repetitive* to illustrate how different repetition penalties would fare on the occurrence of repetition (Figure 4 (right)). In general, less sequences become repetitive as the penalty becomes stronger.

Post-processing To further evaluate the model's translation ability, we implement post-processing to eliminate repetitions in the generations. Before evaluation, we employ a sliding window with a length of 10 words, calculating the hash value of the substring within the window. As we slide the window, if the hash value of the current substring matches any previously seen hash value, we compare the actual substrings to confirm the repetition and then trim accordingly[5]. After cleaning, the blocks belonging to the same chapter are merged back together for evaluation at the chapter level.

According to Figure 3, although applying repetition penalty in decoding procedure shows significant improvements in BLEU and BlonDe scores, the METEOR and COMET scores do not reflect similar gains. To determine whether repetition penalty genuinely improves translation quality rather than simply reducing repetition, we carefully examine the BLEU scores across the four categories before and after post-processing (→). The division of the four groups is based on the results of $\gamma = 1.7$ compared to the case with no repetition penalty applied ($\gamma = 1$).

As Table 4 shows, the repetition penalty affects the four groups differently: for sequences that cease to be repetitive after the penalty is applied (*New Non-repetitive*), increasing γ consistently improves translation quality. In contrast, for *Non-repetitive* sequences which stay non-repetitive before and after applying the penalty, increasing γ slightly diminishes performance. It demonstrates that repetition penalty did not produce more meaningful translations for this group. On the other hand, applying an appropriate repetition penalty can slightly improve translation effectiveness for sequences that stay repetitive before and after applying the penalty (*Repetitive*). It should be noted that an excessively high penalty may negatively impact performance for sequences that are prone to repeat. Unsurprisingly, for sequences in *New Repetitive* which start to be repetitive after applying the penalty, the translation quality declines rapidly. This leads to a potential direction of future work to develop advanced decoding algorithms to avoid repetitions in translation.

[5]Most repetitions exhibit a self-reinforcement effect, continuously repeating the same sentences or phrases. Therefore, once a repetition is detected, we remove all subsequent words.

5.3 How Do LLMs Perform on JAM?

In order to evaluate the capacity of LLMs on CH2CH translation, we perform zero-shot evaluation on the JAM dataset across different models. To further analyze performance variations across different context lengths, we segment chapters into at most 512, 1024, and 2048 tokens, respectively. The results are presented in Figure 5.

GPT-4 outperforms all other models across both sentence-level and document-level metrics. Rather, translation-oriented models, such as NLLB-3.3B and ALMA-7B-Stage2, struggle in the CH2CH task, i.e., performance drop dramatically especially when the sequence become longer than 1024 tokens. One reason as to why ALMA-7B-Stage2 faces challenges in translating long sentences is that it has been finetuned exclusively on short parallel sequences. This may impair its capability to handle long-sequence translation and fully exploit the advantages of chapter-level contextual information to improve translation quality. However, we observe notable improvements after fine-tuning ALMA-7B on our chapter-level dataset JAM even in the most challenging setting where the context extends up to 2048 tokens, as shown in Table 3.

Despite LLMs such as LLAMA2 being theoretically capable of handling contexts of up to 4096 tokens, their performance in translation tasks over extensive contexts remains subpar. Before delving into more nuanced improvements in discourse-level translation, it is crucial to enhance the model's capacity for high-quality long-context translation.

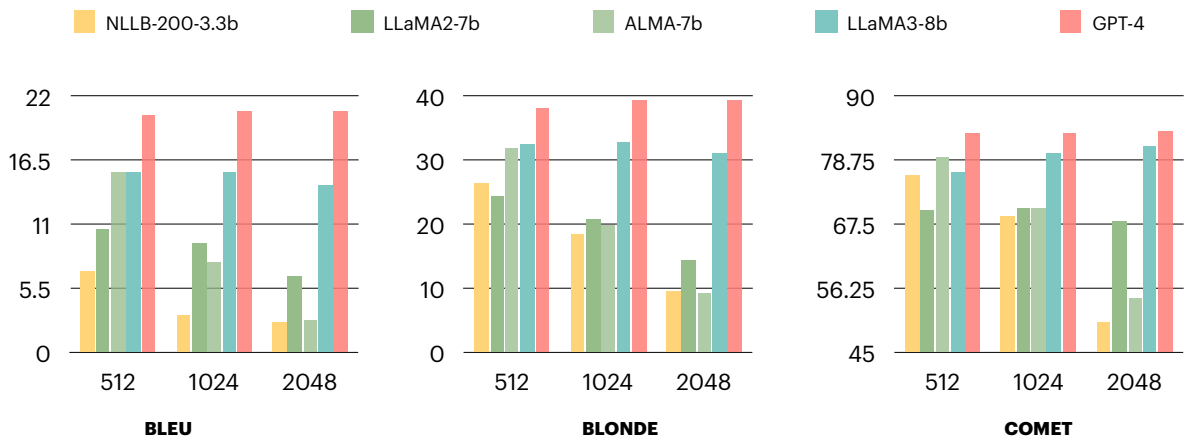

Figure 5: Zero-shot translation performance on JAM data across LLMs. The chapter-level data are segmented into chunks containing at most 512, 1024, 2048 tokens.

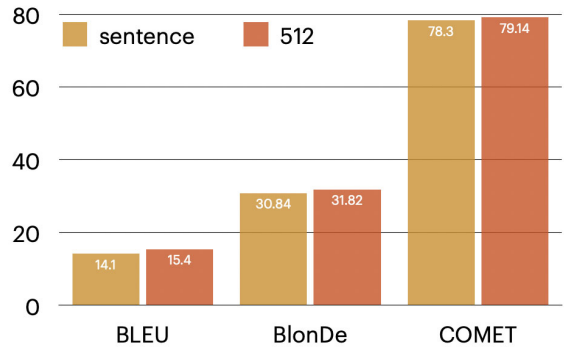

Figure 6: Zero-shot translation performance.

CH2CH vs. Sentence Translation The high-level objective of CH2CH translation is to leverage more training signals from chapter-level dataset. To test the effectiveness of this setting, we conduct an experiment to segment chapters into sentences for comparison. Concretely, we first split each chapter into separated sentences using the NLTK [6] package, then execute translation individually on each sentence with ALMA-7B. The translated sentences are concatenated back to calculate document-level evaluation metrics. Figure 6 indicates that ALMA-7B under the 512-tokens setting outperforms the sentence-segmented setting across all metrics, attesting the significance of CH2CH translation.

Decoder-only vs. Encoder-Decoder Architecture Under the zero-shot setting (Figure 5), ALMA-7B-Stage2 continues to surpass encoder-decoder translation model NLLB-200-3.3B on BLEU scores. In terms of document-level evaluation metrics, ALMA-7B-Stage2 performs on par with, or even better than NLLB-200-3.3B on the most BlonDe metrics, e.g., pronnoun and discourse marker (d.m.). One potential explanation is that the backbone LLM LLAMA2-7B has a better context understanding and text generating ability. For example, discourse markers, e.g., *however, on the other hand*, are crucial for maintaining the coherence and cohesion of text, areas in which LLMs are trained. NLLB-200-3.3B tends to generate shorter text compared to other models. One hypothesis is that it is primarily trained on a sentence-aligned dataset, where the source and target sentences do not differ significantly in length.

After finetuning on JAM, though Encoder-Decoder perform slightly better than Decoder-only model, yet still under-perform ALMA models on most of the evaluation metrics (Table 3). The above results demonstrates the effectiveness of decoder-only models in handling complex literary translation. Particularly noteworthy is the fact that LLMs do not rely heavily on large amounts of parallel data and are inherently capable of translating long context sequences after finetuning.

6 Conclusion

In our paper, we show that CH2CH translation is a viable approach for *context-aware* NMT, exemplified by our novel dataset, JAM. Chapter-level data, derived from professional translations, offers richer context signals and presents a more realistic scenario. Through detailed empirical experiments, we discover that LLMs are aptly suited for CH2CH translation following a two-step fine-tuning process: first at the sentence level, then at the chapter level. Nevertheless, challenges of repetition inheriting from long-context generation signals the need for improved long-sequence decoding strategies in future research.

[6] https://github.com/nltk/nltk

7 Limitations

Because JAM dataset is constructed from professional literary works, we acknowledge that literary books may introduce stylistic and cultural biases into the dataset. This might limit the model's generalization to other domains like technical or conversational language and potentially amplify cultural or temporal biases inherent in the source materials. While literary texts enrich the model's ability to handle complex and expressive language, their focus may underrepresent diverse voices and non-literary contexts. Another limitation of our work is that the current dataset is restricted to the Chinese-English language pair, which limits the model's applicability to other language combinations. In future work, we plan to extend our dataset to include a wider range of languages, enabling more comprehensive evaluations and applications. Additionally, we aim to incorporate texts from diverse domains to enhance the model's generalizability and adaptability to varied contexts.

References

Yusser Al Ghussin, Jingyi Zhang, and Josef van Genabith. 2023. Exploring paracrawl for document-level neural machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1304–1310, Dubrovnik, Croatia. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

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. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

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

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. WIT3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation*, pages 261–268, Trento, Italy. European Association for Machine Translation.

Daniel Deutsch, Juraj Juraska, Mara Finkelstein, and Markus Freitag. 2023. Training and meta-evaluating machine translation evaluation metrics at the paragraph level. *Preprint*, arXiv:2308.13506.

Patrick Fernandes, Kayo Yin, Graham Neubig, and André F. T. Martins. 2021. Measuring and increasing context usage in context-aware machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6467–6478, Online. Association for Computational Linguistics.

Zihao Fu, Wai Lam, Qian Yu, Anthony Man-Cho So, Shengding Hu, Zhiyuan Liu, and Nigel Collier. 2023. Decoder-only or encoder-decoder? interpreting language model as a regularized encoder-decoder. *Preprint*, arXiv:2304.04052.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. *Preprint*, arXiv:1904.09751.

Yuchen Eleanor Jiang, Tianyu Liu, Shuming Ma, Dongdong Zhang, Mrinmaya Sachan, and Ryan Cotterell. 2023. Discourse-centric evaluation of document-level machine translation with a new densely annotated parallel corpus of novels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7853–7872, Toronto, Canada. Association for Computational Linguistics.

Yuchen Eleanor Jiang, Tianyu Liu, Shuming Ma, Dongdong Zhang, Jian Yang, Haoyang Huang, Rico Sennrich, Ryan Cotterell, Mrinmaya Sachan, and Ming Zhou. 2022. Blonde: An automatic evaluation metric for document-level machine translation. *Preprint*, arXiv:2103.11878.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine. *arXiv preprint arXiv:2301.08745*.

Linghao Jin, Jacqueline He, Jonathan May, and Xuezhe Ma. 2023. Challenges in context-aware neural machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15246–15263, Singapore. Association for Computational Linguistics.

Marzena Karpinska and Mohit Iyyer. 2023. Large language models effectively leverage document-level context for literary translation, but critical errors persist. In *Proceedings of the Eighth Conference on Machine Translation*, pages 419–451, Singapore. Association for Computational Linguistics.

Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *Preprint*, arXiv:1909.05858.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. *Preprint*, arXiv:2302.14520.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *Preprint*, arXiv:1808.06226.

Jiahuan Li, Hao Zhou, Shujian Huang, Shanbo Cheng, and Jiajun Chen. 2024. Eliciting the translation ability of large language models via multilingual finetuning with translation instructions. *Preprint*, arXiv:2305.15083.

Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. 2018. OpenSubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Lorenzo Lupo, Marco Dinarelli, and Laurent Besacier. 2022. Divide and rule: Effective pre-training for context-aware multi-encoder translation models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4557–4572, Dublin, Ireland. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, and et al. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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

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

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Zewei Sun, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Shujian Huang, Jiajun Chen, and Lei Li. 2020. Rethinking document-level neural machine translation. *arXiv preprint arXiv:2010.08961*.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, and et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

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. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Katherine Thai, Marzena Karpinska, Kalpesh Krishna, Bill Ray, Moira Inghilleri, John Wieting, and Mohit Iyyer. 2022. Exploring document-level literary machine translation with parallel paragraphs from world literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9882–9902, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. Prompting palm for translation: Assessing strategies and performance. *Preprint*, arXiv:2211.09102.

Yonghui Wu, Mike Schuster, Z. Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason R. Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Gregory S. Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *ArXiv*, abs/1609.08144.

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

Jin Xu, Xiaojiang Liu, Jianhao Yan, Deng Cai, Huayang Li, and Jian Li. 2022. Learning to break the loop: Analyzing and mitigating repetitions for neural text generation. *Preprint*, arXiv:2206.02369.

Nan Xu, Chunting Zhou, Asli Celikyilmaz, and Xuezhe Ma. 2023b. Look-back decoding for open-ended text generation. *Preprint*, arXiv:2305.13477.

Wen Yang, Chong Li, Jiajun Zhang, and Chengqing Zong. 2023. Bigtranslate: Augmenting large language models with multilingual translation capability over 100 languages. *Preprint*, arXiv:2305.18098.

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

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

Appendix: Towards Chapter-to-Chapter Context-Aware Literary Translation via Large Language Models

A JAM Dataset

A.1 Corpus Information

The whole JAM corpus contains world literatures; for a source text to be included in JAM, it must be (1) a literary work that has a published electronic version with chapter breaks along with (2) its corresponding human-written, Chinese translations from professional translators available on the Internet. Books genres include both fiction (e.g., romance, science, adventure, etc) and non-fiction literature (e.g., biography and self-help).

All books in JAM have entered the public domain with cleared copyright, from the earliest published in 1817 to the latest in 1949. Table 6 shows 20 sample books from the JAM dataset, in which the ACL column is obtained by using LlamaTokenizerFast. The data statistics are shown in Table 2

A.2 Paragraph Misalignment Example

subsection A.2 shows examples of paragraph misalignment, which make paragraph-level translation not feasible.

B Implementation Details

B.1 Data

Data for baseline models is encoded and vectorized with byte-pair encoding (Sennrich et al., 2016) using the `SentencePiece` (Kudo and Richardson, 2018) framework. We use a 32K joint vocabulary size for Zh→En. Full corpus statistics of WMT22 are in Table 5.

Dataset	Lg. Pair	Train	Valid	Test
WMT22	Zh→En	25134743	2002	2001

Table 5: Sentence counts across WMT22 datasets.

Due to the limitation to certain context length of backbone LLM, we have to segment chapters into chunks within 2048 token length. Our data assumes chapter-alignment without sentence-alignment, chunking might potentially result in misalignment. However, the loss caused by misalignment is marginal considering the whole length of the context. In addition, the main advantages in Ch2Ch is brought by sentence-misalignment, which is not diminished by chunking.

To segment JAM chapter-level dataset into chunks, we first decide the number of chunks to split in a chapter by ensuring that each chunk includes no more than 2048 English and Chinese tokens, then equally segment the chapter into the computed number of chunks. There is no overlap between chunks, and we keep a sentence a complete unit when we split chapters.

B.2 Evaluation

Because there is no sentence-level alignment in our dataset JAM, we are calculating BLEU scores for each chunk, which is "segment-level" BLEU, considered pretty similar to d-BLEU. For comparison, we also report d-BLEU on the whole chapters in Table 3. The differences between segment-level BLEU and d-BLEU is small, as there are often less than 5 chunks in each chapter.

```
Prompt
Translate this from [src lang] to [tgt lang]:
[src lang]: <src chapter>
[tgt lang]:
```

Figure 7: Prompt template for LLMs.

B.3 Baseline Traning

We use the following two architectures for baseline:

Encoder-Decoder Architecture We use the Transformer (Vaswani et al., 2017) `base` version, which consists of 6 encoder layers, 6 decoder layers, a model dimension of 512, and an FFN hidden dimension of 2048.

Decoder-only Architecture Compared to the prevalent encoder-decoder architecture, the decoder-only framework is often simpler in architecture and computationally efficient (Fu et al., 2023).

We train baseline models (Encoder-decoder and Decoder-only) on the `fairseq` framework. Following Vaswani et al. (2017); Fernandes et al. (2021), we use the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.98$, dropout set to 0.3, an inverse square root learning rate scheduler with an initial value of 10^{-4}, and the warm-up step set to 4000. Here, we only train the Transformer `base` version, and the decoder-only model is also derived from the base Transformer `base` architecture. We keep the parameter size of both Encoder-decoder and Decoder-only architecture similar for fair comparison.

Title	Author	Year	#Chapts	ACL (en/zh)
1984	George Orwell	1949	24	5.8K/10.2K
A Tale of Two Cities	Charles Dickens	1859	44	4.3K/8.0K
Alice's Adventures in Wonderland	Lewis Carroll	1865	9	3.1K/5.7K
Ancient Greek Myths	/	/	58	488.2/862.1
Around the World In Eighty Days	Jules Verne	1872	36	2.6K/5.5K
Black Beauty	Anna Sewell	1877	13	1.9K/3.0K
Don Quixote	Miguel de Cervantes	1605	125	4.4K/6.9K
Five Weeks in a Balloon	Jules Verne	1863	44	3.1K/5.9K
How The Steel Was Tempered	Nikolai Ostrovsky	1934	18	11.7K/24.8K
Little Prince	Antoine de Saint-Exupéry	1943	28	822.3/1.4K
Little Women	Louisa May Alcott	1868	47	5.8K/10.7K
Oliver Twist	Charles Dickens	1838	53	4.4K/8.7K
Robinson Crusoe	Daniel Defoe	1719	8	20.9K/35.4K
Tess of the d'Urbervilles	Thomas Hardy	1891	59	3.7K/7.8K
The Adventures of Tom Sawyer	Mark Twain	1876	35	3.1K/5.7K
The Moon and Sixpence	William Somerset Maugham	1919	58	1.8K/3.9K
The Mysterious Island	Jules Verne	1875	62	4.5K/8.2K
The Time Machine	H. G. Wells	1895	13	3.4K/6.2K
Women in Love	D. H. Lawrence	1920	27	10.3K/9.5K
Wuthering Heights	Emily Brontë	1847	34	5.1K/9.3K

Table 6: Corpus information for 20 sample books. ACL = average chapter length in tokens.

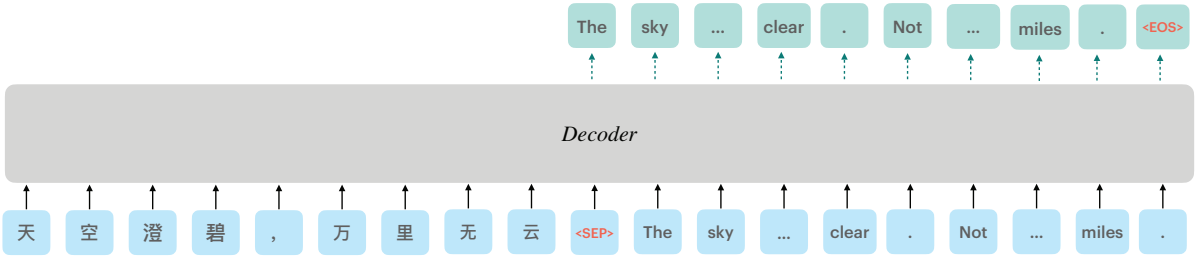

Figure 8: Decoder-only architecture.

B.4 LLM Training

All models are trained with 8xA40 GPUs and DeepSpeed+ZeRO3. Following Xu et al. (2023a), we use Adam optimizer, weight decay set to 0.01, and the warm-upratio set to 0.01, an inverse square root learning rate scheduler with an initial value of 2×10^{-5}.

The zero-shot evaluation on JAM dataset across different chunk sizes is shown in Table 7. Note that we also evaluated the JAM dataset without chunking using a long-context model, such as GEMINI-2.0-PRO (Team et al., 2024). We include the results in Table 7. We do not see the performance of long-context models surpass GPT-4, which indicates the impact of chunking of our dataset is marginal. However, GEMINI-2.0-PRO performs much better than other LLaMA-based models.

Building upon Xu et al. (2023a), we prepend a fixed prompt (Figure 7) to each chapter.

B.5 Repetition Analysis on Zero-shot Translations

As illustrated in Figure 9, repetition is not an issue for sentence-level translation. However, the repetition ratio significantly increases as the input context length increases from 512 to 1024. Furthermore, Figure 10 shows that as the input length increases, the repetition start position also occurs earlier.

B.6 Post-processing on Fine-tune Translations

Post-processing eliminate repeated words and phrases in generated translations. Table 8 shows a comprehensive automatic metric comparison between translations with post-processing versus.

Model	BLEU	d-BLEU	BlonDe					COMET	ACL
			all	pron.	entity	tense	d.m.		
				512 tokens					
NLLB-200-3.3b	6.90	7.31	26.37	63.26	23.96	63.53	61.59	0.7592	870
LLaMA2-7b	10.60	11.21	24.49	73.89	17.51	72.70	66.85	0.6990	1551
ALMA-7b	15.40	15.90	31.82	88.35	19.69	88.22	82.30	0.7914	1608
GPT-4	**20.40**	**20.63**	**38.24**	**91.03**	**39.43**	**90.34**	**82.35**	**0.8324**	**1863**
				1024 tokens					
NLLB-200-3.3b	3.20	3.82	18.32	47.37	17.17	46.15	44.29	0.6888	709
LLaMA2-7b	9.30	10.21	20.57	64.09	11.60	66.44	59.74	0.7025	1648
ALMA-7b	7.70	8.15	19.82	68.49	13.30	71.00	62.49	0.7017	2223
GPT-4	**20.60**	**21.05**	**39.20**	**91.12**	**40.87**	**90.32**	**82.87**	**0.8347**	**1821**
				2048 tokens					
NLLB-200-3.3b	2.50	2.91	9.48	41.62	7.37	50.66	25.98	0.5009	1254
LLaMA2-7b	6.40	7.40	14.40	49.45	8.63	53.66	39.69	0.6778	1780
ALMA-7b	2.70	3.59	9.09	42.27	6.35	47.98	27.77	0.5433	2382
GPT-4	**20.70**	**20.70**	**21.44**	**91.39**	**41.81**	**91.39**	**83.67**	**0.8359**	**1765**
				Whole Chapter w/o chunking					
Gemini-2.0	18.87	19.34	37.39	87.42	36.51	89.29	82.97	0.7983	1859

Table 7: Zero-shot performance on JAM data across LLMs. The chapter-level data are segmented into chunks containing at most 512, 1024, 2048 tokens. ACL = average chapter length in tokens; The ACL of sampled instances=1850.

Model	WMT	JAM	Post-processing	BLEU	BlonDe					COMET
					all	pron.	entity	tense	d.m.	
ALMA-7B-Stage1	✗	✓	✗	15.70	33.46	74.28	30.62	70.11	71.72	0.7806
ALMA-7B-Stage2	✓	✓	✗	18.80	36.90	81.34	32.72	77.83	76.81	0.8025
ALMA-7B-Stage1	✗	✓	✓	21.6	39.54	86.43	35.43	84.52	82.98	0.7986
ALMA-7B-Stage2	✓	✓	✓	**23.9**	**42.73**	**90.69**	**38.41**	**89.02**	**84.95**	**0.8106**

Table 8: Automatic metric result of ALMA-7B translations on JAM, with versus without post repetition removal processing. **Bold** denotes best performance.

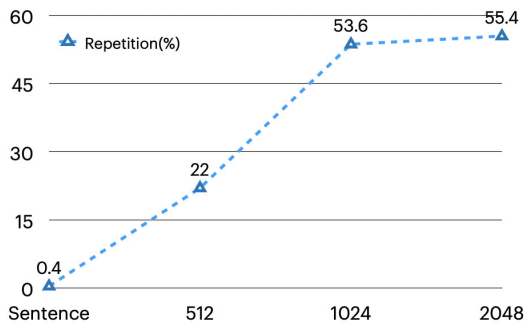

Figure 9: Repetition ratio in the generation results for different input context length

without post-processing.

C Qualitative Results

We present the qualitative examples translating a sample from JAM test dataset across different repetition penalty in Table 9. Without repetition penalty during decoding, the generation of long chapters will fall into repetition of phrases/sentences.

In Table 10, we present a comparison of zero-shot translations produced by various state-of-the-art LLMs, including LLaMA2-7B, LLaMA3-8B, GPT-4, and NLLB-54B. GPT-4 demonstrates a strong performance by avoiding repetitions, whereas LLaMA3-8B begins generating unrelated text instead of translating the source, ultimately ending with repetitive output. LLaMA2-7B translates approximately half of the source text before repeating the same sentence and also fails to generate the correct character name (Oliver). NLLB-54B translates only a small portion of the text before falling into repetitive patterns.

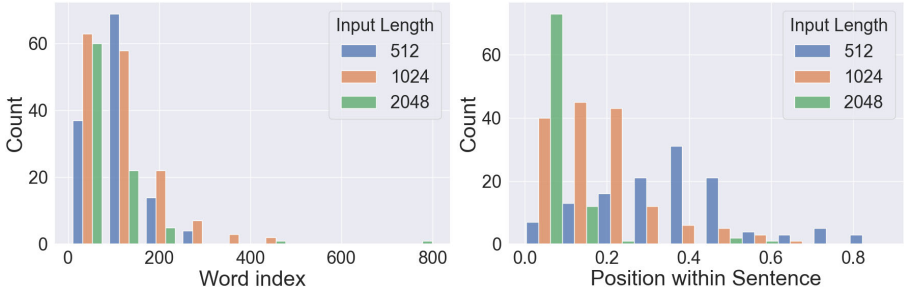

Figure 10: Repetition start position across different input lengths. Left: The word index of repetition, Right: The relative position of repetition.

Model	Translation
Source	他们来到街上. 这是一个令人扫兴的早晨,风疾雨猛,漫天阴云,像是要来一场暴风雨. 夜里雨下得很猛,路上积起了无数的大水洼,水沟也都满了. 天空透出一道隐隐可见的微光,预示着新的一天即将来临,而这一道亮光非但没有减轻反倒加重了景物的幽暗,使街灯射出的光芒变得一片苍白,没有在湿漉漉的屋顶和凄凉的街道上洒下一丝温暖,明亮的色彩. 这一带街区似乎还没有人起床,房屋的窗户全都关得紧紧的,他们经过的街道也是一片沉寂,空无一人. 直到他们拐进贝丝勒尔草地大道,天色才总算亮起来了. 灯光大多已经熄灭,几辆乡间的大车朝伦敦缓缓驶去,时而有一辆糊满泥污的公共马车咔哒咔哒地飞驰而过,车把式在赶到前边去的时候,总要惩戒性地照着呆头呆脑的大车老板来一鞭子,他们占错了车道,很可能会害得他比规定时间迟十几秒钟到站. 点着煤气灯的酒馆已经开堂,别的商号也一家接一家开始营业,路上有了零零星星的行人. 接着,络绎不绝地涌来了一群群上班的工人,头上顶着鱼筐的男男女女,装有各种蔬菜的驴车,满载活畜或是宰好的全猪全羊的双轮马车,手提牛奶桶的妇人——一股源源不断的人流携带着各种食品,艰难地向东郊移动着. 到了商业中心区附近,喧闹声与车辆行人的往来更是有增无已. 当赛克斯拉着奥立弗挤过肖狄奇区和伦敦肉市场之间的街道时,这种车水马龙的景象终于汇成一片喧嚣与奔忙. 天已经完全亮了,同往日没什么两样,大概一直要持续到黑夜重新来临. 伦敦城一半的市民迎来了他们繁忙的早晨. 赛克斯先生带着奥立弗拐进太阳街,克朗街,穿过芬斯伯雷广场,沿着契士韦尔路急步闪人望楼街,又溜进长巷否证科学理论系统的任何一个部分,主张用"有用","有,来到伦敦肉市场,这个地方传出一片纷乱的喧闹,使奥立弗退斯特大为惊讶. 这天早晨正逢赶集. 地面覆盖着几乎漫过脚踝的污泥浊水,浓浊的水气不断地从刚刚宰杀的牲畜身上腾起,与仿佛是驻留在烟囱顶上的雾混合起来,沉甸甸地垂挂在市场上空. 在这一大片平地的中心,所有的畜栏,连同许许多多还可以往这片空地里挤一挤的临时棚圈,都关满了羊,水沟边的木桩上拴着三四排菜牛和枯牛. 乡下人,屠户,家畜经纪人,沿街叫买的小贩,顽童小偷,看热闹的,以及各个社会底层中的流氓无赖,密密麻麻挤成一团. 家畜经纪人打着日哨,狗狂吠乱叫,公牛边蹬蹄子边吼,羊咩咩地叫,猪嗯叽嗯叽地哼哼;小贩的叫卖声,四面八方的呼喊,咒骂,争吵;一家家酒馆里钟鸣铃响,人声喧哗;
rp=1.3	They came out into the street. It was a dismal morning, with a strong wind and heavy clouds, as if there were going to be a storm. It had rained hard during the night, and there were great pools in the roads, and the gutters were full. There was a glimmer of light in the sky, promising the approach of a new day, but that glimmer only deepened the gloom, and made the gas-lamps look white and colourless, and shed no warmth or brightness on the wet roofs and dreary streets. The neighbourhood did not seem to be astir yet, and the windows of the houses were all shut up. The streets through which they passed were silent and deserted. It was not until they turned into Bishopsgate Road that the day began to break. Most of the lamps were out, and country waggons were slowly making their way to London, with here and there a public conveyance splashing through the mud with a clattering of whips at the heads of the horses, who took the wrong side of the road, and were in danger of being ten minutes late. The public-houses with their gas were open, and one by one the other shops began to be astir. There was a sprinkling of people in the streets. Then there poured in a steady stream of workpeople, men and women with fish-baskets on their heads, carts of all sorts of vegetables, loaded waggons of live-stock or slaughtered sheep and pigs, milk-women with pails, and a ceaseless procession of food of all sorts, struggling towards the East End. As they drew nearer to the heart of the commercial world, the noise and traffic increased. When Sikes and Oliver threaded their way through the streets between Shoreditch and the London meat-market, it became a roar and a rush. It was broad daylight, and it was likely to continue so until night should come again.Half the population of London was astir on this busy morning. Mr Sikes took Oliver into Sun-street, then into Cruncher-street, through Finsbury Square, along Chiswell-street, and up Long-lane, with a brisk trot. Then he slipped into any one of the alleys that deny a portion of their science to any part of the System of Pure Logic.Oliver was astonished at the din and bustle of the London meat-market. It was market-day, and the ground was covered with nearly ankle-deep mud, and the heavy atmosphere was thick with the fumes of newly-slaughtered animals, which mingled with the fogs that seemed to linger in the chimney-tops, and hung heavily over the market. In the centre of this great open space, all the pens, with many temporary sheds crowded in among them, were full of sheep, and three or four pens of cattle, and three or four pens of calves, were tied to stakes at the edge of the gutters. Country people, butchers, cattle-dealers, street hawkers, boys and thieves, idlers, and all the scum of the lower classes, were huddled together. Cattle-dealers were whistling, dogs were barking, bullocks were bellowing, sheep were bleating, pigs were grunting; the cries of the hawkers, the shouts, oaths, and quarrels of the crowd, the ringing of bells from the public-houses, and the general uproar.

Model	Translation
rp=1.18	They came out into the street. It was a dismal morning, with a strong wind and heavy clouds, as if it were going to be a storm. The night had been very wet, and there were great puddles in the road, and the gutters were full. There was a glimmer of light in the sky, promising a new day, but that glimmer only deepened the gloom, and made the gas-lamps look white and colourless, and shed no warmth or brightness on the wet roofs and dreary streets. The neighbourhood did not seem to have waked up yet, and the windows of the houses were all shut, and the streets were silent and deserted. It was not until they turned into Belle-Saxe Road that the day began to break. Most of the lamps were out, and country waggons were slowly making their way to London, with here and there an omnibus splashing through the mud, and jerking its whip at the stupid carter who had taken the wrong side of the road, and was liable to be late by ten or fifteen seconds. The public-houses with their gas were open, and one by one the shops began to be astir, and there were stray pedestrians in the streets. Then came a steady stream of working people, men and women with fish-baskets on their heads, donkey-carts full of vegetables, waggons laden with live-stock or slaughtered sheep and pigs, milk-women with pails, and a ceaseless procession of food of all kinds, moving slowly towards the East End. As they drew nearer to the heart of the commercial world, the noise and traffic increased. When Mr. Sownds and Oliver had scrambled through the streets between Shoreditch and the London Meat Market, it became a roar and bustle. It was broad daylight, and it might have been expected to last till night came round again.Half the population of London was astir on this busy morning. Mr. Sownds took Oliver down Sun-street, along Cruncher-street, across Finsbury Square, and up Long-Bow Lane, with a brisk trot, and slipped into some alley or court which disproved any part of the system of Science. He brought Oliver to the London Meat Market, where there was a confused uproar, which astonished Oliver very much.It was market-day. The ground was covered with mud nearly up to the ankles, and the heavy atmosphere was thick with the fumes of newly-slaughtered animals, which mingled with the fogs that seemed to be stationed at the top of every chimney, and hung heavily over the market. In the centre of this great flat space, all the cattle-trucks, with many temporary stalls crowded in among them, were full of sheep, and pens at the edge of the gutters were three and four deep with cattle and calves. Country people, butchers, cattle-dealers, street hawkers, boys and thieves, lookers-on, and all sorts of ruffians from the lowest depths of society, were huddled together. Cattle-dealers were whistling, dogs were barking, bullocks were bellowing, sheep were bleating, pigs were grunting; the cries of the hawkers, the shouts, oaths, and quarrels; the ringing of bells from the public-houses, and the noise of voices and laughter within them;
rp=1 (no rp)	They came out into the street. It was a dismal morning, with a strong wind and heavy clouds, as if there were going to be a storm. It had rained hard during the night, and there were great pools in the roads, and the gutters were full. There was a faint glimmer of light in the sky, promising a new day, but this glimmer only deepened the gloom, and made the street lamps look white and colourless, and shed no warmth or cheerfulness on the wet roofs and dismal streets. It seemed as if no one had got up yet in this part of the town, and all the windows were shut, and the streets were silent and empty. It was not until they turned into Belle-Saxe Road that the day began to break. Most of the lamps were out, and a few country waggons were slowly making their way into London. Now and then a mud-bespattered omnibus would rattle by, with the conductor lashing it on in front, as it took a wrong turning, and was liable to be ten or fifteen seconds late in reaching its destination. The gas-lighted public-houses were opening, and one by one the other shops were beginning to do business. There was a sprinkling of people in the streets. Then came a steady stream of workers, men and women with fish-baskets on their heads, carts piled with vegetables, waggons laden with live-stock or slaughtered carcases of pigs and sheep, milk-women with their pails, and a ceaseless procession of food-bearing humanity creeping slowly eastward. As they neared the heart of the commercial world, the noise and traffic increased. When Sikes and Oliver had threaded their way through the streets between Shoreditch and the metropolitan butchers' shops, it became a roar and a rush. It was broad daylight, and it was likely to continue so until night should come again.Half the population of London was astir on this busy morning. Mr Sikes took Oliver into Sun-street, and down Cruncher-street, through Finsbury Square, and along Chiswell-street, with a brisk trot, and slipped into the by-streets to evade any part of the system of science that might be inconsistent with the theory of 'usefulness.'The place of rendezvous was the metropolitan butchers' market, where a confused noise and bustle astonished Oliver. It was market-day. The ground was covered with nearly ankle-deep mud, and the heavy atmosphere was thick with the fumes of newly-killed animals, which mingled with the fogs that seemed to linger in the chimney-tops, and hung heavily over the market. In the centre of this great flat space, all the pens, with many temporary sheds crowded in among them, were full of sheep, and three or four tiers of cattle and calves were tied to the posts at the edge of the gutters. Country people, butchers, cattle-dealers, street hawkers, boys and thieves, lookers-on, and all the scum of the lowest classes, were huddled together. The cattle-dealers were whistling, dogs were barking, bullocks were bellowing, sheep were bleating, pigs were grunting; the cries of the hawkers, the shouts, oaths, and quarrels of the people, the ringing of bells from the public-houses, the uproar of voices and laughter within them, the tramp of horses, the rattle of wheels, the clatter of hoofs, the creaking of waggon-poles, the clanking of chains, the grinding of carts, the jingling of money, the clashing of knives and forks, the smell of stale beer, the smell of new beer, the smell of roast meat, the smell of raw meat, the smoky smell of cooking meat, the smell of fish, the smell of cheese, the smell of onions, the smell of spiced ale, the smell of spiced beer, the smell of spiced gin, the smell of spiced rum, the smell of spiced brandy, the smell of spiced tobacco, the smell of spiced pipes, the smell of spiced tobacco-pipes, the smell of spiced tobacco-snuff, the smell of spiced tobacco-snuff-pipes, the smell of spiced tobacco-snuff-pipes, the smell of spiced tobacco-snuff-pipes, the smell of spiced tobacco-snuff-pipes, the smell of spiced tobacco-snuff-pipes, the smell of spiced tobacco-snuff-pipes, the smell of spiced tobacco-snuff-pipes, ...

Table 9: Qualitative Results

Model	Translation
Source	“我丝毫也不怀疑,朋友. "老先生回答. 他把鼻梁上的眼镜扶扶正,四下里找起墨水壶来. 奥立弗的命运到了一个关键时刻. 倘若墨水壶是在老绅士想像中的地方,他就会把鹅毛笔插下去,然后签署证书,奥立弗也就一径被人匆匆带走了. 可墨水壶偏偏是在老绅士的鼻子底下,接下来他照例满桌子都找遍了,还是没有找到. 就在他一个劲地往前找的时候,目光落在了奥立弗退斯特那张苍白而惊恐的脸上. 虽说邦布尔在一旁递眼色警告他,掐他,奥立弗全然不顾,目不转睛地望着未来的主人的丑恶嘴脸,那种厌恶与恐慌交融在一起的神情任何人也不会看错,哪怕是一位眼神不济的治安推事. 老先生停了下来,放下鹅毛笔,看看奥立弗,又看了看利姆金斯先生,这位先生装出在吸鼻烟,一副愉快而又若无其事的样子. "孩子. "老先生从写字台上俯下身来,说道. 这声音吓了奥立弗一跳,他这种反应倒也情有可原,听听这话有多温和就是了,然而没有听熟的声音总是叫人害怕的,他不住地打着哆嗦,眼泪夺眶而出. "孩子,"老绅士说,"瞧你,脸都吓白了. 出什么事了?""干事,离他远一点儿,"另一位推事说着,放下报纸,饶有兴致地向前探出身子. "行了,孩子,告诉我们是怎么回事,别害怕. "奥立弗扑地跪下来,双手紧紧地握在一起,哀求他们把自己送回那间黑屋子去——饿死他——揍他——高兴宰掉也行——就是不要打发他跟那个可怕的人走. "呃,"邦布尔先生说道,他抬起双手,眼珠朝上翻了翻,神情庄重得非常令人感动. "呃,奥立弗,阴险狡猾,心术不正的孤儿我见得多了,你是其中最无耻的一个. ""闭嘴,干事. "邦布尔先生刚把带"最"字的形容词说出来,第二位老绅士便说道. "对不起,大人,"邦布尔先生说道,他怀疑自己是不是听错了. "您指的是我吗?""不错,闭上你的嘴巴. "邦布尔先生惊得目瞪日呆. 竟然喝令一位教区干事闭嘴. 真是改天换地了. 戴了一副玳瑁眼镜的老绅士看了自己的同事一眼,那一位意味深长地点点头. "这些契约我们不予批准. "老绅士将那张羊皮纸往旁边一扔,说道. "我希望,"利姆金斯先生结结巴巴地说,"我希望两位大人不要单凭一个孩子毫无理由的抗议,就认为院方有管理不善的责任. ""治安推事不是专管排难解纷的,"第二位老绅士厉声说道,"把孩子带回济贫院去,好好对待他,看来他有这方面的需要. "这天傍晚,白背心绅士非常自信,非常明确地断言,奥立弗不光要受绞刑,而且还会被开肠剖肚,剁成几块. 邦布尔先生闷闷不乐,有些神秘地直摇脑袋,宣称自己希望奥立弗终得善报. 对于这一点,甘菲尔先生回答说,他希望那小子还是归自己,尽管他大体上同意干事的话,但表达出来的愿望似乎完全相反. 第二天清晨,公众再次获悉:重新转让奥立弗,任何人只要愿意把他领走,可获得酬金五镑.
GPT-4	"I have absolutely no doubt, my friend," the old gentleman replied. He adjusted the glasses on his nose, and started looking for the ink bottle. A pivotal moment had arrived in Oliver’s fate. If the ink bottle was where the old gentleman imagined, he would dip the quill in and sign the certificate, and Oliver would be hastily taken away. But the ink bottle happened to be right under the old gentleman’s nose, and despite a thorough search of the desk, he still couldn’t find it. As he continued to search forward, his eyes landed on Oliver Twist’s pale and terrified face. Despite Bumble’s warnings and nudges, Oliver stared unflinching at his future master’s ugly face, filled with a blend of disgust and terror that couldn’t be misread by anyone, even by a poorly sighted law officer. The old gentleman stopped, put down the quill, studied Oliver, then Mr. Limkins, who made a show of leisurely snuffing tobacco, looking cheerful and unconcerned. "Child," the old gentleman leaned over from his writing desk and said in a voice that startled Oliver; a reaction that was entirely understandable considering the gentleness of the tone and the fact that unfamiliar voices can be frightening. Shivering uncontrollably, tears streaming down his face, Oliver knelt down and begged to be sent back to the dark room—to starve, to be beaten, even slaughtered—just not to be sent away with this dreaded man. "Ah," sighed Mr. Bumble, raising his hands and turning his eyes upward in a deeply touching expression of gravity. "Ah, Oliver, I have seen many sneaky and evil orphans, and you are the most shameless one.” "Shut up, Officer," the second gentleman interjected, as soon as Mr. Bumble described Oliver’s shamelessness. "Excuse me, Sir,” Mr. Bumble hesitated, wondering if he had misheard. "Do you mean me?" "Yes, close your mouth," the second gentleman instructed, leaving Mr. Bumble agape. It was extremely unusual for a parish officer to be ordered to keep quiet. The old gentleman with the turtle shell glasses glanced over at his colleague, who nodded in agreement. "We will not approve these agreements," the old gentleman casually discarded the parchment. "I hope," Mr. Limkins stuttered, "I hope, gentlemen, you won’t assume irresponsibility on the part of the management just because of a child’s unjustified protest." "A magistrate is not responsible for resolving disputes," said the second gentleman sharply, "Take the child back to the workhouse, and treat him properly, it seems he needs just that." In the evening of the same day, the ‘white-waistcoat’ gentleman confidently and definitively stated that not only would Oliver be hanged, but he would also be eviscerated and chopped up into pieces. Mr. Bumble, disheartened and mysteriously shaking his head in disagreement, hoped for a better fate for Oliver. In response, Mr. Gamfield said he hoped Oliver would come back to him, and despite generally agreeing with the officer, he expressed a contrary wish. The next morning, the public was again informed: Oliver was being transferred again, anyone willing to take him would receive a reward of five pounds.

Model	Translation
LLaMA3-8B	“I have no doubt, friend. "The old gentleman answered. He adjusted his spectacles on the bridge of his nose, and began to hunt for the inkstand. Oliver’s fate was at a crisis. If the inkstand were where the old gentleman supposed it to be, he would dip his pen into it, and sign the certificate, and Oliver would be hurried away. But the inkstand was not where the old gentleman supposed it to be, and the old gentleman, after looking for it in all sorts of unexpected places, and under all sorts of unlikely names, at last gave it up. Just as he gave it up, his eyes happened to fall on Oliver’s face, and he looked at him very hard. Mr. Bumble, who was standing by, cleared his throat very loud, and looked at Oliver very hard, too. The old gentleman looked at Mr. Bumble, and Mr. Bumble looked at the old gentleman. The old gentleman looked at Oliver again, and Oliver looked at the old gentleman again. The old gentleman looked at Mr. Bumble a third time, and Mr. Bumble looked at the old gentleman a third time, and then they both looked at Oliver a third time, and then Oliver looked at them both a third time, and then they all three looked down at the inkstand a third time, as if they expected the inkstand would do something new, on the strength of what had passed. But the inkstand took no notice, and looked straight on in front of it, as if it were determined to know nothing. The old gentleman looked at Oliver a fourth time, and Mr. Bumble looked at Oliver a fourth time, and Oliver looked at Mr. Bumble a fourth time, and Mr. Bumble looked at the old gentleman a fourth time, and the old gentleman looked at the key of the inkstand a fourth time, and then they all three looked at the inkstand a fourth time, as if they expected the inkstand would do something new on the strength of what had last passed. But the inkstand took no notice, and looked straight on in front of it, as before. The old gentleman looked towards Mr. Bumble, and nodded slightly. Mr. Bumble looked towards Oliver, and nodded slightly. Oliver looked towards Mr. Bumble, and nodded slightly. Mr. Bumble looked towards the old gentleman, and nodded slightly. The old gentleman looked towards Oliver, and nodded slightly. Oliver looked towards Mr. Bumble, and nodded slightly. Mr. Bumble looked towards the old gentleman, and nodded slightly. The old gentleman looked towards Oliver, and nodded slightly. Oliver looked towards Mr. Bumble, and nodded slightly. Mr. Bumble looked towards the old gentleman, and nodded slightly. . . (exclude repetition)
LLaMA2-7B	“I have no doubt, my friend. "The old gentleman answered. He took off his pince-nez and wiped it with his handkerchief. "Mr. Oliphant’s fate has come to a crucial moment. If the ink bottle were in the old gentleman’s imagination, he would put on his spectacles, sign the document, and take Oliphant away. But the ink bottle happened to be under the old gentleman’s nose, and he had to hunt all over the table before he found it. Just as he was about to go forward, his eyes fell on Oliphant’s pale and terrified face. The look of hatred and terror on that face would have frightened anyone, even a policeman who had never seen it before. The old gentleman stopped, put down his pen, and looked at Oliphant. "My boy," he said, "look at you. Your face is as white as a sheet. What’s the matter?" "I’m frightened," said Oliphant. "Don’t be frightened, my boy," said the old gentleman. "Look at me. I’m smoking a cigar, and I’m as happy as can be. You needn’t be afraid of me. I’m not going to send you back to that horrible house. I’m not going to starve you to death. I’m not going to beat you. I’m not going to kill you. I’m not going to do anything to you. I’m only going to send you back to that horrible house." "Well," said the second gentleman, "Oliphant, you’re a sneak and a scoundrel. You’re a sneak and a scoundrel. You’re a sneak and a scoundrel. You’re a sneak and a scoundrel. You’re a sneak and a scoundrel. You’re a sneak and a scoundrel. You’re a sneak and a scoundrel. You’re a sneak and a scoundrel. You’re a sneak and a scoundrel. You’re a sneak and a scoundrel. You’re a sneak and a scoundrel. . . (exclude repetition)
NLLB-54B	"The inkpot was right under the old gentleman’s nose, and he looked all over the table for the inkpot, and found none. Just as he was going forward with his eyes fixed on Oliver’s pale and terrified face, he would have inserted the marker and signed the certificate, and Oliver would have been carried away with it. But the inkpot was under the old gentleman’s nose, and he looked all over the table for the inkpot, and found none. Just as he was going forward with his eyes fixed on Oliver’s pale and terrified face, he would have inserted the marker and signed the certificate, and Oliver would have been carried away with it.

Table 10: Zero-shot Qualitative Results across different LLMs