Efficiently Exploring Large Language Models for Document-Level Machine Translation with In-context Learning

Menglong Cui[*], **Jiangcun Du**[*], **Shaolin Zhu, Deyi Xiong** [†]

College of Intelligence and Computing, Tianjin University, Tianjin, China

{cuimenglongcs,d2000,zhushaolin,dyxiong}@tju.edu.cn

Abstract

Large language models (LLMs) exhibit outstanding performance in machine translation via in-context learning. In contrast to sentence-level translation, document-level translation (DOCMT) by LLMs based on in-context learning faces two major challenges: (1) document translations generated by LLMs are often incoherent; (2) the length of demonstration for in-context learning is usually limited. To address these issues, we propose a Context-Aware Prompting method (**CAP**), which enables LLMs to generate more accurate, cohesive, and coherent translations via in-context learning. CAP takes into account multi-level attention, selects the most relevant sentences to the current one as context, and then generates a summary from these collected sentences. Subsequently, sentences most similar to the summary are retrieved from the datastore as demonstrations, which effectively guide LLMs in generating cohesive and coherent translations. We conduct extensive experiments across various DOCMT tasks, and the results demonstrate the effectiveness of our approach, particularly in zero pronoun translation (ZPT) and literary translation tasks.

1 Introduction

With the increasing scale of parameters and training corpus, large language models (LLMs) have gained remarkable abilities to handle a variety of tasks via in-context learning (Li et al., 2023a; Wang et al., 2023b), which allows language models to perform tasks with a few given demonstrations and human-written prompts as context. One particular area where LLMs have shown outstanding potential is machine translation (MT) (Moslem et al., 2023; Zhang et al., 2023a,b; Gao et al., 2024). However, most existing studies (Zhu et al., 2024) have focused on sentence-level translation, which makes LLM-based document-level machine translation underexplored (Wang et al., 2023a; Li et al., 2024). To bridge this gap, we adapt LLMs to document-level machine translation (DOCMT) in this paper.

As a complex task different from sentence-level translation, one major challenge of DOCMT with LLMs is that **the length of demonstrations for in-context learning is limited**. For sentence-level MT, Agrawal et al. (2023) and Zhang et al. (2023a) show that using 32 or more randomly sampled bilingual parallel sentence pairs as prompts can effectively enhance the translation abilities of LLMs. However, for DOCMT, the length of the text segments to be translated or to be used as demonstrations inherently increase (Zeng et al., 2024). Consequently, the limited input window of large language models poses a significant obstacle (Pawar et al., 2024; Tang et al., 2023). Unlike sentence-level translation, the constraints of DOCMT hinder the seamless integration of such demonstrations. Moreover, the lengthier nature of document-level inputs exacerbates issues related to inference speed. As the length of the demonstrations increases, the computational resources required for processing also escalate, potentially impeding real-time translation capabilities.

Previous efforts to DOCMT with LLMs, e.g., the method presented in (Wu et al., 2024), involve fine-tuning LLMs using two strategies: Parameter-Efficient Fine-Tuning and Full Parameter Fine-Tuning. Such a framework may not be able to sufficiently overcome the limitation of the length of demonstration for in-context learning in LLMs. And it also reveals a significant issue: **context-independent translations**, which can result in translations that lack coherence and disambiguation (Zhang et al., 2022a; Macé and Servan, 2019).

To address these issues, we propose a novel method called **CAP**, which stands for **C**ontext-**A**ware **P**rompting. It consists of three essential steps. Firstly, by combining token- and sentence-

---

[*] Equal contribution
[†] Corresponding author

level attention scores, we obtain sentences with the highest sentence-level attention scores as context for each sentence. This step facilitates LLMs in generating coherent translations. Secondly, to address the limitation of the length of demonstration for in-context learning in LLMs, we summarize the context and retrieve parallel sentences from the datastore, which are most similar to the generated summary. Finally, we use the retrieved example pairs as demonstrations, prompting LLMs to produce document translations.

In summary, the contributions of this work are as follows:

- We investigate document-level translation via LLMs. Specially, we explore linguistic phenomena in document-level translation with LLMs, such as ZPT.

- We propose a novel method, CAP, to enable LLMs to perform accurate and coherent translation via in-context learning. Our method addresses the issue of translation coherence deficiency by employing a dynamic context window and resolves the limitation concerning the length of demonstration for in-context learning by summarizing the context.

- To demonstrate the effectiveness of our method, we conduct experiments on various DOCMT tasks. Experimental results show that our method significantly outperforms the baselines, especially for the ZPT task.

2 Related Work

Document-Level Machine Translation In recent years, numerous approaches have been proposed for DOCMT (Maruf et al., 2022; Lei et al., 2022; Zhang et al., 2022b; Tan et al., 2022; Zhang et al., 2020b). These studies could be further roughly categorized into two groups. Studies of the first group extend the source-side context from a sentence to the local context of few sentences (Tan et al., 2021; Lyu et al., 2021; Xu et al., 2021). Studies of the second group have focused on document-to-document (Doc2Doc) translation. The initial exploration of Doc2Doc NMT has involved concatenating multiple sentences into a single unit (Ma et al., 2020; Zhang et al., 2020a; Tan et al., 2019). Recent studies have successfully trained vanilla Transformer models for Doc2Doc translation by leveraging either large augmented datasets (Lupo et al., 2022; Sun et al., 2022) or pre-trained models (Yang et al.,

2021). Additionally, some latest work has achieved the truly Doc2Doc translation by enhancing the attention mechanism in the Transformer model (Bao et al., 2021; Li et al., 2023b).

Translation Oriented LLMs LLMs have demonstrated remarkable proficiency across a wide range of NLP tasks (Scao et al., 2022; Touvron et al., 2023; Sun and Xiong, 2022; Min et al., 2024; Wei et al., 2022). Recent research has shown that in-context learning can significantly enhance their performance when following general language instructions. Specifically, there is a growing body of work exploring the translation capabilities of LLMs (Lu et al., 2023; Zhang et al., 2023a; Hendy et al., 2023). However, it is important to note that these efforts have primarily focused on sentence-level machine translation and have not delved into DOCMT (Wu et al., 2024). Latest noteworthy studies in DOCMT are conducted by Wu et al. (2024) and Wang et al. (2023a). Wang et al. (2023a) investigate the document-level translation capabilities of GPT-3.5-TURBO, making it the most closely related work to ours that explores the DOCMT ability of LLMs. Wu et al. (2024) explore fine-tuning LLMs using two strategies: Parameter-Efficient Fine-Tuning and Full Parameter Fine-Tuning, to enhance the DOCMT ability of LLMs. Such a framework may not be able to sufficiently overcome the limitation of the length of demonstration for in-context learning in LLMs. We observe that the existing translation-oriented LLMs tend to struggle with effectively translating sentences in the middle and ending of a document. Existing methods that explore DOCMT with LLMs are rarely considering the issues of (1) the length limitation of demonstrations for in-context learning and (2) maintaining coherence and resolving ambiguities. In this work, we mainly concentrate on mitigating the length limitation of in-context learning for DOCMT and addressing the issues of incoherence and ambiguity in DOCMT with LLMs.

3 CAP: Context-Aware Prompting

In this section, to address the challenges associated with utilizing LLMs for DOCMT via in-context learning, we propose a **C**ontext-**A**ware **P**rompting method.

As shown in Figure 1, our approach consists of three steps. Firstly, we take into consideration multi-level attention, subsequently opting for the sentences with the highest sentence-level attention
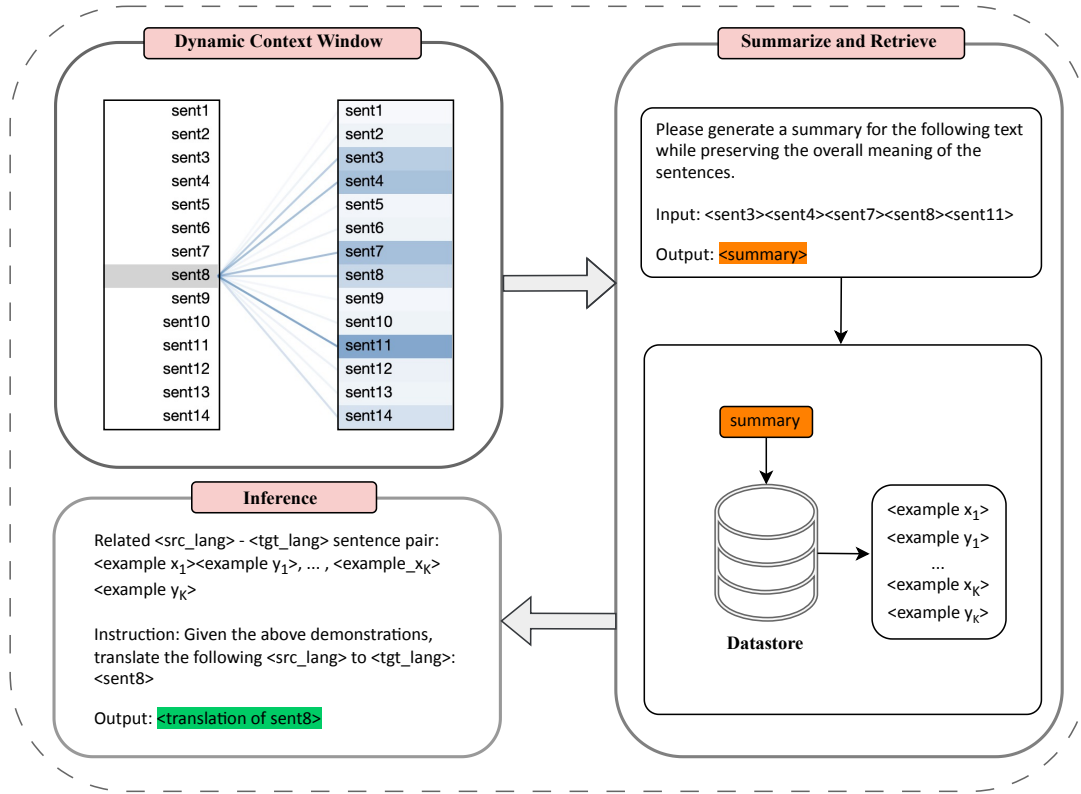
Figure 1: Diagram of the proposed CAP. It consists of three essential steps. (1) Dynamic Context Window: selecting context by sentence-level attention scores; (2) Summarize and Retrieve: summarizing the context and retrieving parallel sentences from the datastore, which are most similar to the summary; (3) Inference: with parallel sentences serving as demonstrations, the LLM translates the input text via in-context learning.

scores as the context. Secondly, we input the context into an LLM to obtain its summary, then retrieve K example pairs from the datastore, which are most similar to the summary. Thirdly, we use the retrieved example pairs as demonstrations in the prompt, guiding the model to generate accurate and coherent translations through a few-shot approach.

3.1 Dynamic Context Window

In the context of document-level translation for LLMs, issues of incoherence and disambiguation may arise, leading to context-independent translations. In order to address this problem, we propose a Dynamic Context Window as a solution. By extracting the most attended context surrounding the current sentence, it facilitates the model in generating accurate and coherent translation outputs.

Initially, we meticulously devise a method for computing sentence-level attention. Specifically, as demonstrated in Eq. (1), we extract the attention weights from the final layer of the LLM, followed by averaging across all attention heads to obtain the token-token attention scores. Next, as illustrated in

Eq. (2), the maximum value of token-token attention scores between tokens in the current sentence and tokens in all other sentences is computed as the token-sentence attention score. Ultimately, as shown in Eq. (3), the average of attention scores between all tokens in the current sentence and other sentences is calculated to derive the sentence-level attention score.

$$attn_{i,j} = \frac{1}{H} \sum_{h=1}^{H} attn_{i,j}^{h} \qquad (1)$$

$$\text{TS}_{i,S} = \max(attn_{i,k}), k \in \text{tokens}(S) \qquad (2)$$

$$\text{SS}_{I,S} = \text{avg}(\text{TS}_{i,S}), i \in \text{tokens}(I) \qquad (3)$$

where $\text{TS}_{i,S}$ is the attention score of **T**oken i and **S**entence S, $\text{SS}_{I,S}$ is the attention score between **S**entence I and **S**entence S. $attn_{i,j}^{h}$ denotes the attention score of the h-th head in the final transformer block, calculated between the i-th token and the j-th token in the current document. H is the number of attention heads in each transformer block.

After obtaining the sentence-level attention scores, we extract the top N sentences with the highest sentence-level attention scores as dynamic context, while maintaining their original sequential positions within the document.

3.2 Example-Specific In-Context Demonstrations Selection

Unlike sentence-level translation, the challenge in utilizing LLMs for DOCMT arises from the constrained context length, as the length of the text segments to be translated or used as demonstrations inherently increases. To effectively address these challenges, it is necessary to further process the context and extract as much relevant information as possible from the preceding sentences. We then input this extracted context into an LLM to generate a concise summary that encapsulates key topics and information.

Subsequently, as shown in Eq. (4), we utilize sentence-transformers (Reimers and Gurevych, 2020) to retrieve carefully designed example pairs from a specified datastore. These examples are selected based on their cosine similarity with the embedding of the summary sentences, ensuring that the retrieved examples exhibit semantic similarity to the context and content of the document.

$$\text{score}(S, X) = cos(\boldsymbol{S}, \boldsymbol{X}) \qquad (4)$$

where S, and X are the summary we obtained at the preceding step and the example sentence in datastore respectively. $\boldsymbol{S}$ and $\boldsymbol{X}$ are the embeddings of S and X respectively.

By employing these selected examples as few-shot demonstrations, we provide the LLM with a more precise understanding of the document. This approach not only enhances the model's accuracy in utilizing in-context learning for translation but also contributes to enhancing the robustness of document translation tasks.

3.3 Inference

We employ a prompt template, which is identical to that utilized in (He et al., 2023). In this approach, we integrate the retrieved example pairs from the preceding step seamlessly into the prompt template, which will then be input into the target LLM. This strategic integration serves to guide the LLMs in generating translations that exhibit enhanced cohesion and coherence.

Following the translation process, we employ a diverse set of evaluation metrics to thoroughly assess the quality of the translations generated by the LLM. This multifaceted evaluation allows us to gain comprehensive insights into the effectiveness and performance of the language models in producing accurate and appropriate translations.

4 Experiments

In this section, we conducted extensive experiments on various DOCMT tasks to examine the effectiveness of the proposed method.

4.1 Setup

Datasets: We used OPUS-100[1] and news-commentary-v15 as the datastore to retrieve the in-context demonstrations on four language pairs. To assess the effectiveness of our proposed method on LLMs, we employed WMT22 newstest as the test set. Following (Agrawal et al., 2023), we normalized punctuation using Moses[2] and removed sentence pairs with a source/target length ratio exceeding 1.5. To assess the accuracy of our method in ZPT, we utilized the GuoFeng (Xu et al., 2022) dataset, which covers five domains: movie subtitle, Q&A forum, government news, web fiction, and personal profile. To validate the capability of our approach in guiding the model to generate more cohesive and more coherent translations, we employed the WMT23 Literary Translation dataset,[3] which comprises two test sets: an in-domain test set and an out-of-domain test set. Additionally, we used its training set to retrieve in-context demonstrations.

Models: To compare the translation performance of LLMs using our method, we compared against the following three encoder-decoder based strong NMT models as baselines.

- MarianMT: A framework for translation models, using the same models as BART (Lewis et al., 2020). In our experiment, each language pair is associated with a distinct pre-trained model.

- M2M100 (Fan et al., 2021): It is a multilingual encoder-decoder model, which is pre-trained on 100 languages and is capable of translating 9,900 directions.

---

[1]https://opus.nlpl.eu/opus-100.php
[2]http://www2.statmt.org/moses/
[3]https://www2.statmt.org/wmt23/literary-translation-task.html

methods	Qwen-1.8B-Chat		Qwen-7B-Chat		Qwen-14B-Chat		Qwen-72B-Chat	
	de-en	zh-en	de-en	zh-en	de-en	zh-en	de-en	zh-en
Zero-shot	**19.84**	**17.28**	34.21	24.02	35.99	26.37	37.42	29.11
Random	17.58	15.12	32.91	23.58	35.28	26.42	37.55	28.40
BM25	17.70	14.92	32.90	23.54	34.86	25.90	37.49	28.59
Similar	17.81	15.34	33.43	23.85	35.29	26.42	37.72	28.50
Precedent	15.80	15.23	34.49	23.65	35.75	26.49	38.01	29.15
Ours	18.56	15.51	**34.86**	**24.34**	**36.57**	**26.77**	**38.14**	**29.32**

Table 1: Document-level BLEU scores of variants of the Qwen chat version LLM, employing different prompt strategies, on the WMT22 newstest dataset.

- NLLB (Costa-jussà et al., 2022): It is a neural multilingual machine translation model, which supports translation between any pair of 200 languages. There are numerous variants of it, and the one utilized in our experiments is NLLB-3.3B.

To evaluate the translation abilities of LLMs, we utilized LLMs that are tuned with SFT, as they can follow instructions to generate translations. Hence, we chose the following two LLMs, which support both Chinese and English and have undergone instruction tuning:

- Baichuan (Yang et al., 2023): It is trained on a high-quality corpus with 2.6 trillion tokens and has achieved the best performance in authoritative Chinese and English benchmarks of the same size.

- Qwen (Bai et al., 2023): It is pretrained on over 3 trillion tokens, including Chinese, English, multilingual texts, code, and mathematics, covering general and professional fields.

Prompt Selection Strategies: To assess the effectiveness of the proposed method, we compared our method against five prompt selection strategies:

- Zero-shot: Zero-shot is a powerful baseline. Previous studies (Radford et al., 2019) show that LLMs have excellent zero-shot abilities in many NLP tasks, including reading comprehension, translation, and summarization.

- Random: This strategy randomly selects K sentence pairs from the datastore as demonstrations.

- BM25: It calculates the similarity between translated sentences and sentences in the corpus with BM25, and then selects top K sentences with the highest similarity as demonstrations.

- Similar: This method uses sentence embedding similarity-based retrieval to select K demonstrations (Moslem et al., 2023). In our work, we used sentence-transformers (Reimers and Gurevych, 2020) to implement this method.

- Precedent: This approach guides large language models in generating translations by utilizing the preceding K sentences of the current sentence within a document, along with those translations generated in previous steps as demonstrations.

For prompt selection experiments, we employed models from four distinct sizes of the Qwen LLMs. Meanwhile, for a trade-off between performance and cost, we set K to 3.

Metrics: We employed a variety of metrics to assess the quality of document-level translation generated by LLMs, including document-level BLEU(d-BLEU), and ChrF2 using SacreBLEU for document-level evaluation. For the ZPT task, we employed accuracy to evaluate the precision of zero pronoun translation. Additionally, we have also evaluated document-level translation quality with Blonde (Jiang et al., 2022), which is an automatic evaluation metric for document-level machine translation. More experimental results can be found in Appendix A.

All experiments were performed on 4 NVIDIA A100 GPUs.

4.2 Comparing Different Prompt Selection Strategies

Experimental results are presented in Table 1. We find that the prompt guidance constructed using our approach yields the highest translation quality across LLMs of different sizes, except for Qwen-1.8B-Chat. We attribute this to the fact that small

Methods	de-en		zh-en		en-de		en-zh	
	d-BLEU	chrF2	d-BLEU	chrF2	d-BLEU	chrF2	d-BLEU	chrF2
Encoder-Decoder								
MarianMT	36.25	68.28	22.15	58.60	30.99	58.29	27.08	24.89
M2M100-1.2B	35.11	66.86	17.32	49.15	31.13	67.51	28.45	26.13
NLLB-3.3B	36.35	67.74	25.48	60.03	**34.08**	**69.23**	24.03	23.02
LLMs								
Baichuan-7B-Chat	33.59	68.32	25.06	60.70	20.87	63.08	33.34	33.57
Baichuan-13B-Chat	33.62	68.83	26.02	64.91	22.32	64.83	36.30	33.94
Qwen-7B-Chat	34.86	68.84	24.34	62.57	24.01	64.72	36.59	33.67
Qwen-14B-Chat	36.57	70.00	26.77	64.24	29.36	67.88	42.28	37.85
Qwen-72B-Chat	**38.14**	**70.44**	**29.32**	**65.43**	31.48	68.11	**46.80**	**40.90**

Table 2: Document-level BLEU and chrF2 scores of models with different architectures on four language pairs from the WMT22 newstest dataset.

methods	1.8B	7B	14B	72B
Zero-shot	29.29	36.00	38.30	36.53
Random	27.81	38.59	37.55	38.13
BM25	28.13	37.45	37.94	37.95
Similar	29.14	37.49	37.93	37.39
Precedent	29.91	42.02	40.45	39.07
Ours	**32.58**	**42.13**	**41.87**	**41.65**

Table 3: ZPT accuracy of various variants of the Qwen chat version model on the GuoFeng dataset. 1.8B, 7B, 14B, and 72B represent the model sizes. We computed the average accuracy across five domains for the model.

models often struggle to extract meaningful context.

Additionally, in the same few-shot experimental setup, the translation quality of selecting demonstrations based on sentence embedding similarity surpasses that of the random approach. This is also corroborated in (Agrawal et al., 2023). Additionally, we observe experimental results consistent with (Radford et al., 2019), indicating that the quality of generative outputs guided in a "Zero-shot" method is generally superior to those generated through "Random" selection or "Similarity-based" selection in a few-shot method.

We observe that "Precedent" serves as a strong baseline and exhibits a similar performance trend to our method: when the model size is too small, it struggles to produce high-quality translations. As noted by Agrawal et al. (2023), the quality of demonstrations in prompt is crucial for translation performance. Therefore, we attribute this to the cumulative effect of low-quality translations produced by small models, which leads to progressively worse translations.

Experiment results suggest that as the model size increases, our method continues to effectively guide LLMs in generating high-quality translations. This also demonstrates the universality of our approach.

4.3 Comparing with Traditional NMT Models

We compared our method against traditional NMT models. As shown in Table 2, the results show that when the model size is large enough, the translation abilities of LLMs comprehensively exceed those of dedicated translation models. This is observed except in the en-de language pair, where we attribute the discrepancy to the insufficient understanding of the German language by the Baichuan and Qwen models, primarily trained on Chinese and English corpora. Specifically, in the en-zh translation direction where Baichuan and Qwen excel, the translation results of LLMs significantly surpass those of specialized models.

Moreover, our experiments reveal interesting insights into the impact of model size on translation performance. As expected, larger LLMs, such as Qwen-72B-Chat, consistently outperforms smaller counterparts across various language pairs and evaluation metrics. This trend demonstrates the importance of model scale in harnessing the full potential of LLMs for translation tasks.

However, the most striking finding is the competitive performance of small LLMs, particularly those with parameters in the 7B to 14B range. Despite their smaller size compared to their larger counterparts, these models exhibit translation capabilities on par with or even outperforming dedicated translation models. This suggests that with an appropriate prompt strategy, small LLMs can

methods	Qwen-1.8B-Chat		Qwen-7B-Chat		Qwen-14B-Chat		Qwen-72B-Chat	
	test01	test02	test01	test02	test01	test02	test01	test02
Zero-shot	13.77	8.93	19.33	14.09	21.43	15.52	22.96	16.61
Random	12.90	8.24	18.82	14.21	21.03	15.46	21.94	16.68
Similar	12.48	7.96	19.48	14.13	21.02	15.24	21.67	17.55
Ours	**14.81**	**9.95**	**19.65**	**14.73**	**21.56**	**15.87**	**23.10**	**17.85**

Table 4: Document-level BLEU scores on the WMT literary translation dataset, which includes two test sets, namely "test01" and "test02". "test01" and the datastores are within the same domain, while "test02" is in a different domain from the datastores.

offer a cost-effective solution for translation tasks.

Our method has improved document translation performance on different LLMs, further verifying the robustness of our method. This underscores the generality and efficiency of our approach in leveraging LLMs for translation tasks across diverse domains and languages.

4.4 Analyzing the Effectiveness on ZPT

The increasing interest among researchers in ZPT can be attributed to its increasing complexity and significance within the field of language translation. ZPT involves the intricate task of proficiently addressing the omission of zero pronouns in the source language text when it undergoes translation into the target language. The primary aim is to ensure the meticulous restoration or expression of these omitted pronouns in the translated output, thereby upholding linguistic accuracy, fluency, and contextual consistency.

To assess the effectiveness of our proposed method in the context of the ZPT task, a series of experiments were conducted on the comprehensive GuoFeng dataset. Experimental results are displayed in Table 3. Remarkably, across models of varying sizes, a discernible enhancement in ZPT accuracy is consistently observed when compared to the baseline performance. It is noteworthy, however, that the relationship between the model size and the accuracy of ZPT is not characterized by continuous improvement. This phenomenon can be ascribed to the realization that a high overall translation quality of the model does not necessarily translate into a proportionate increase in accuracy specifically related to ZPT. Additionally, we found that the "Precedent" method outperforms similarity-based prompt selection strategies, indicating the importance of contextual information in the ZPT task. Our approach, which considers dynamic contextual information, effectively captures relevant context while filtering out irrelevant

information. As a result, our method demonstrates superior performance in the ZPT task.

Hence, it becomes evident that a mere augmentation of the model size does not suffice for the improvement of ZPT accuracy. Consequently, the exploration and adoption of more effective methods become imperative in the pursuit of advancing the state-of-the-art in this domain.

4.5 Analyzing the Effectiveness on Literary Translation

The significant challenges for translating literary works include entity consistency, anaphora resolution, and lexical choice (Matusov, 2019). Literary works exhibit longer contextual spans compared to texts in other domains, such as news articles. Therefore, LLMs must possess the ability to model distant contexts in order to learn translation consistency and appropriate lexical choices. Our approach, by focusing on the context relevant to the current sentence and subsequently summarizing the context, is able to encapsulate more contextual information within a limited-length prompt. To validate whether our method can guide the model to generate more consistent translations, we conducted experiments on the WMT23 Literary Translation dataset.

The experimental results are presented in Table 4. We observe consistent improvements of our approach over the baseline method, both on in-domain and out-of-domain test sets. Consistent with the above experimental results, the strategy of simply using sentence embedding similarity to select demonstrations often performs even worse than random selection. This implies that a straightforward prompt selection strategy based on similarity is not suitable for challenging tasks like literary translation. However, our method is able to effectively extract contextual information, thereby guiding LLMs to generate higher-quality translations, specifically achieving better entity consistency.

Figure 2: Average attention scores of the final layer of the Qwen-14B-Chat model. The darker the color, the higher the attention score assigned to that token. It is noteworthy that the model attends to "他" (meaning "he" in English) from the examples we extract when generating "he". Our examples successfully help the model in translating the zero pronoun "他" (meaning "he" in English).

methods	de-en	zh-en
w/o DCW	35.84	26.03
max ⇒ avg	36.05	26.38
Ours	**36.57**	**26.77**

Table 5: Ablation study on the de-en and zh-en WMT22 newstest dataset.

5 Analysis

In this section, we conducted experiments to further analyze the effectiveness of our approach.

5.1 Ablation Study

To verify the effectiveness of various factors on our method, we further compared our method with the following variants and present the results in Table 5:

(1) *w/o DCW*. In this variant, we directly used a Fixed Context Window to select the context. Specifically, we utilized the preceding and subsequent two sentences of the current sentence as the context. As reported in Table 5, this variant dramatically decreases performance on two language pair test sets. It reveals the significance of dynamically selecting context through attention scores.

(2) *max ⇒ avg*. In this variant, we replaced the max function with the average function in Eq. (2). From Table 5, we can observe that this variant performs worse than our method. It demonstrates the effectiveness of our approach in computing the sentence-level attention score.

5.2 Visualization

In order to explain why our approach successfully translates omitted zero pronouns in Chinese, we extracted the attention weights from the final layer of the Qwen-14B-Chat model and visualized it using

appropriate tools.[4] As illustrated in Figure 2, we observe that when the model attempts to translate the zero pronoun "He", it exhibits greater attention to the subject in the context and the associated information about the subject. The visualized results reflect the effectiveness of our approach, showcasing its ability to capture relevant contextual information and guide the model in generating accurate and coherent translations. This indicates that our approach involves more than merely translating words at the surface level; rather, it entails translating based on a profound comprehension of sentence structure and context, thereby offering guidance and support for the model to generate more precise translations, thereby demonstrating the effectiveness of our method.

5.3 Case Study

In order to assess the precision of ZPT across various methods in the ZPT task, a comprehensive case study was undertaken. As illustrated in Table 6, it is observed that the "Zero-shot", "Random", and "Similar" methods exhibit inaccuracies in translating omitted zero pronouns in Chinese. Conversely, our approach demonstrates proficiency by accurately translating the omitted zero pronouns as "His". This success can be attributed to the method's adeptness at extracting relevant contextual information, enabling the model to deduce the omitted zero pronouns effectively.

In summary, our method stands out in its capacity to extract valuable contextual information, guiding the model towards generating translations that are not only more accurate but also more cohesive and coherent. Further examples and elaborate experimental results are available in Appendix B for reference.

---

[4]https://github.com/alan-cooney/CircuitsVis

Source	主要从事媒体大数据计算、网络多媒体与跨媒体智能等研究工作。
Reference	His research fields primarily concentrate on Media Big Data Computing, Network Multimedia and Cross-Media Intelligence, etc.
Zero-shot	Our research focuses on media big data computing, network multimedia, and cross-media intelligence.
Random	My main research works involve media big data computation, network multimedia and cross-media intelligence.
Similar	The main research areas include media big data computing, network multimedia and cross-media intelligence.
Ours	His main research focuses on media big data computation, network multimedia, and cross-media intelligence.

Table 6: A Case Study on the translation of Qwen-14B-Chat with various prompt selection strategies. Red text denotes incorrect translations of Zero Pronoun while blue text indicates accurate translations.

6 Conclusion

In this paper, we have presented a novel approach that guides the model to generate more accurate, cohesive, and coherent document-level translations via in-context learning. By selecting relevant contexts using our carefully designed method, we can extract topic information from the context summaries. These summaries serve as a guide for selecting the most beneficial examples from the datastore to enhance the in-context learning procedure of LLMs, hence mitigating the length limitation issue of in-context learning for DOCMT. Experimental results on various benchmarks demonstrate the effectiveness of our approach, particularly in ZPT and literary translation tasks, where our method outperforms the baseline significantly.

Limitations

By selecting better contexts and demonstrations, we have effectively increased the ZPT accuracy and BLEU score of DOCMT with LLMs via in-context learning. We have validated the effectiveness of our method using ZPT accuracy and BLEU score. However, beyond the general BLEU and Blonde metric, DOCMT encompasses challenges such as coreference consistency. We also need to make improvements to address these challenges. Moreover, we should also investigate the performance on more language pairs, which we consider as future work.

Yet another limitation in our work is the increased computational demand. Since our approach necessitates the computation of dynamic context and summarization of context, the computational load is slightly higher compared to the baseline. We intend to explore more efficient document-level machine translation methods in future work.

Acknowledgements

References

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2023. In-context examples selection for machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 8857–8873. Association for Computational Linguistics.

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

Guangsheng Bao, Yue Zhang, Zhiyang Teng, Boxing Chen, and Weihua Luo. 2021. G-transformer for document-level 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, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 3442–3455. Association for Computational Linguistics.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *CoRR*, abs/2207.04672.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22:107:1–107:48.

Pengzhi Gao, Zhongjun He, Hua Wu, and Haifeng Wang. 2024. Towards boosting many-to-many multilingual machine translation with large language models. *CoRR*, abs/2401.05861.

Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2023. Exploring humanlike translation strategy with large language models. *CoRR*, abs/2305.04118.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are GPT models at machine translation? A comprehensive evaluation. *CoRR*, abs/2302.09210.

Yuchen 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. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 1550–1565. Association for Computational Linguistics.

Yikun Lei, Yuqi Ren, and Deyi Xiong. 2022. Codonmt: Modeling cohesion devices for document-level neural machine translation. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 5205–5216. International Committee on Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Jia Li, Ge Li, Chongyang Tao, Jia Li, Huangzhao Zhang, Fang Liu, and Zhi Jin. 2023a. Large language model-aware in-context learning for code generation. *CoRR*, abs/2310.09748.

Yachao Li, Junhui Li, Jing Jiang, Shimin Tao, Hao Yang, and Min Zhang. 2023b. P-transformer: Towards better document-to-document neural machine translation. *IEEE ACM Trans. Audio Speech Lang. Process.*, 31:3859–3870.

Yachao Li, Junhui Li, Jing Jiang, and Min Zhang. 2024. Enhancing document-level translation of large language model via translation mixed-instructions. *CoRR*, abs/2401.08088.

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

Lorenzo Lupo, Marco Dinarelli, and Laurent Besacier. 2022. Focused concatenation for context-aware neural machine translation. In *Proceedings of the Seventh Conference on Machine Translation, WMT 2022, Abu Dhabi, United Arab Emirates (Hybrid), December 7-8, 2022*, pages 830–842. Association for Computational Linguistics.

Xinglin Lyu, Junhui Li, Zhengxian Gong, and Min Zhang. 2021. Encouraging lexical translation consistency for document-level neural machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 3265–3277. Association for Computational Linguistics.

Shuming Ma, Dongdong Zhang, and Ming Zhou. 2020. A simple and effective unified encoder for document-level machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 3505–3511. Association for Computational Linguistics.

Valentin Macé and Christophe Servan. 2019. Using whole document context in neural machine translation. In *Proceedings of the 16th International Conference on Spoken Language Translation, IWSLT 2019, Hong Kong, November 2-3, 2019*. Association for Computational Linguistics.

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

Evgeny Matusov. 2019. The challenges of using neural machine translation for literature. In *Proceedings of the Qualities of Literary Machine Translation*, pages 10–19, Dublin, Ireland. European Association for Machine Translation.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2024. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Comput. Surv.*, 56(2):30:1–30:40.

Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023. Adaptive machine translation with large language models. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation, EAMT 2023, Tampere, Finland, 12-15 June 2023*, pages 227–237. European Association for Machine Translation.

Saurav Pawar, S. M. Towhidul Islam Tonmoy, S. M. Mehedi Zaman, Vinija Jain, Aman Chadha, and Amitava Das. 2024. The what, why, and how of context length extension techniques in large language models - A detailed survey. *CoRR*, abs/2401.07872.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. *CoRR*, abs/2004.09813.

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

Haoran Sun and Deyi Xiong. 2022. Language branch gated multilingual neural machine translation. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 5046–5053. International Committee on Computational Linguistics.

Zewei Sun, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Shujian Huang, Jiajun Chen, and Lei Li. 2022. Rethinking document-level neural machine translation.

In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3537–3548. Association for Computational Linguistics.

Xin Tan, Longyin Zhang, Fang Kong, and Guodong Zhou. 2022. Towards discourse-aware document-level neural machine translation. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 4383–4389. ijcai.org.

Xin Tan, Longyin Zhang, Deyi Xiong, and Guodong Zhou. 2019. Hierarchical modeling of global context for document-level neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 1576–1585. Association for Computational Linguistics.

Xin Tan, Longyin Zhang, and Guodong Zhou. 2021. Coupling context modeling with zero pronoun recovering for document-level natural language generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 2530–2540. Association for Computational Linguistics.

Ruixiang Tang, Dehan Kong, Longtao Huang, and Hui Xue. 2023. Large language models can be lazy learners: Analyze shortcuts in in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 4645–4657. Association for Computational Linguistics.

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

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023a. Document-level machine translation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 16646–16661. Association for Computational Linguistics.

Qinyong Wang, Zhenxiang Gao, and Rong Xu. 2023b. Exploring the in-context learning ability of large language model for biomedical concept linking. *CoRR*, abs/2307.01137.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy

Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Trans. Mach. Learn. Res.*, 2022.

Minghao Wu, Thuy-Trang Vu, Lizhen Qu, George F. Foster, and Gholamreza Haffari. 2024. Adapting large language models for document-level machine translation. *CoRR*, abs/2401.06468.

Mingzhou Xu, Liangyou Li, Derek F. Wong, Qun Liu, and Lidia S. Chao. 2021. Document graph for neural machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 8435–8448. Association for Computational Linguistics.

Mingzhou Xu, Longyue Wang, Derek F. Wong, Hongye Liu, Linfeng Song, Lidia S. Chao, Shuming Shi, and Zhaopeng Tu. 2022. Guofeng: A benchmark for zero pronoun recovery and translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 11266–11278. Association for Computational Linguistics.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, Fei Deng, Feng Wang, Feng Liu, Guangwei Ai, Guosheng Dong, Haizhou Zhao, Hang Xu, Haoze Sun, Hongda Zhang, Hui Liu, Jiaming Ji, Jian Xie, Juntao Dai, Kun Fang, Lei Su, Liang Song, Lifeng Liu, Liyun Ru, Luyao Ma, Mang Wang, Mickel Liu, MingAn Lin, Nuolan Nie, Peidong Guo, Ruiyang Sun, Tao Zhang, Tianpeng Li, Tianyu Li, Wei Cheng, Weipeng Chen, Xiangrong Zeng, Xiaochuan Wang, Xiaoxi Chen, Xin Men, Xin Yu, Xuehai Pan, Yanjun Shen, Yiding Wang, Yiyu Li, Youxin Jiang, Yuchen Gao, Yupeng Zhang, Zenan Zhou, and Zhiying Wu. 2023. Baichuan 2: Open large-scale language models. *CoRR*, abs/2309.10305.

Pengcheng Yang, Pei Zhang, Boxing Chen, Jun Xie, and Weihua Luo. 2021. Context-interactive pre-training for document machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 3589–3595. Association for Computational Linguistics.

Shulin Zeng, Jun Liu, Guohao Dai, Xinhao Yang, Tianyu Fu, Hongyi Wang, Wenheng Ma, Hanbo Sun, Shiyao Li, Zixiao Huang, Yadong Dai, Jintao Li, Zehao Wang, Ruoyu Zhang, Kairui Wen, Xuefei Ning, and Yu Wang. 2024. Flightllm: Efficient large language model inference with a complete mapping flow on fpgas. *CoRR*, abs/2401.03868.

Biao Zhang, Ankur Bapna, Melvin Johnson, Ali Dabirmoghaddam, Naveen Arivazhagan, and Orhan Firat. 2022a. Multilingual document-level translation enables zero-shot transfer from sentences to documents. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 4176–4192. Association for Computational Linguistics.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023a. Prompting large language model for machine translation: A case study. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 41092–41110. PMLR.

Linlin Zhang, Zhirui Zhang, Boxing Chen, Weihua Luo, and Luo Si. 2022b. Context-adaptive document-level neural machine translation. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, pages 6232–6236. IEEE.

Pei Zhang, Boxing Chen, Niyu Ge, and Kai Fan. 2020a. Long-short term masking transformer: A simple but effective baseline for document-level neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 1081–1087. Association for Computational Linguistics.

Pei Zhang, Xu Zhang, Wei Chen, Jian Yu, Yanfeng Wang, and Deyi Xiong. 2020b. Learning contextualized sentence representations for document-level neural machine translation. In *ECAI 2020 - 24th European Conference on Artificial Intelligence, 29 August-8 September 2020, Santiago de Compostela, Spain, August 29 - September 8, 2020 - Including 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020)*, volume 325 of *Frontiers in Artificial Intelligence and Applications*, pages 2298–2305. IOS Press.

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

Shaolin Zhu, Menglong Cui, and Deyi Xiong. 2024. Towards robust in-context learning for machine translation with large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 16619–16629. ELRA and ICCL.

source	手动还是自动，确定离合掌握好了？
reference	Is your automatic or manual? Are you sure that you deal with your clutch well?
zero-shot	Should I use manual or automatic, and have I got a good grip on it?
random	Should we use manual or automatic methods to solve the problem? Have we grasped the situation well?
similar	Should I use manual or automatic control? Have I got a good grip on the distance and angle?
ours	Manual or automatic, have you determined the gear shift handle is in the correct position?

Table 7: A Case Study on the translation of Qwen-14B-Chat with various prompt selection strategies. Red text denotes incorrect translations of Zero Pronoun while blue text indicates accurate translations.

Figure 3: Average attention scores of the final layer of the Qwen-14B-Chat model. The darker the color, the higher the attention score assigned to that token.

A Additional Experimental Results

methods	Qwen-7B-Chat		Qwen-14B-Chat	
	de-en	zh-en	de-en	zh-en
Zero-shot	46.88	35.15	50.01	38.27
Random	46.47	34.74	49.35	38.27
BM25	45.94	34.63	49.37	38.00
Similar	46.36	34.77	49.72	38.82
Precedent	46.13	35.04	49.50	38.43
Ours	**47.00**	**35.20**	**50.23**	**38.90**

Table 8: Blonde scores of Qwen chat LLMs, with different prompt strategies, on the WMT22 newstest dataset.

In Section 4.2, we have employed only d-BLEU and ChrF2 to evaluate translation quality. We use Blonde scores here to further measure translation quality.

Specifically, we evaluated the translation performance of Qwen-7B-Chat and Qwen-14B-Chat on the WMT22 newtest test set using the Blonde metric under different prompt strategies. Experimental results, shown in Table 8, indicate that our proposed method significantly outperforms the baselines, demonstrating its effectiveness.

B Case Study and Visualization Analysis

In this section, we present more examples to further validate our approach with case study and attention visualization.

B.1 Case Study

As shown in Table 7, we provide another example where in the original Chinese text, the subject "你" (you) is omitted due to language convention. In English, it should be correctly restored and translated as "you". We find that the "Zero-shot", "Random", and "Similar" methods all incorrectly recover zero pronouns, whereas our method successfully recovers the correct zero pronoun "you". This demonstrates that our method outperforms better than baselines in ZPT.

B.2 Visualization

As illustrated in Figure 3, we observe that when the model is tasked with translating "He", it predominantly focuses on the subject of the context and its associated information, encompassing entities such as "白崇恩" and the designation "教授". This experimental result illustrates how our approach guides LLMs to generate more accurate translations via in-context learning, thus proving the effectiveness of our method.