Adaptive Few-shot Prompting for Machine Translation with Pre-trained Language Models

Lei Tang[1], **Jinghui Qin**[1*], **Wenxuan Ye**[2], **Hao Tan**[1], **Zhijing Yang**[1]
[1]Guangdong University of Technology
[2]The Chinese University of Hong Kong
tangtang302958@163.com, scape1989@gmail.com, nbvincentelite@gmail.com,
tanhao4869@gmail.com, yzhj@gdut.edu.cn

Abstract

Recently, Large Language Models (LLMs) with in-context learning have demonstrated remarkable potential in handling neural machine translation. However, existing evidence shows that LLMs are prompt-sensitive and it is sub-optimal to apply the fixed prompt to any input for downstream machine translation tasks. To address this issue, we propose an adaptive few-shot prompting (AFSP) framework to automatically select suitable translation demonstrations for various source input sentences to further elicit the translation capability of an LLM for better machine translation. First, we build a translation demonstration retrieval module based on LLM's embedding to retrieve top-k semantic-similar translation demonstrations from aligned parallel translation corpus. Rather than using other embedding models for semantic demonstration retrieval, we build a hybrid demonstration retrieval module based on the embedding layer of the deployed LLM to build better input representation for retrieving more semantic-related translation demonstrations. Then, to ensure better semantic consistency between source inputs and target outputs, we force the deployed LLM itself to generate multiple output candidates in the target language with the help of translation demonstrations and rerank these candidates. Besides, to better evaluate the effectiveness of our AFSP framework on the latest language and extend the research boundary of neural machine translation, we construct a high-quality diplomatic Chinese-English parallel dataset that consists of 5,528 parallel Chinese-English sentences. Finally, extensive experiments on the proposed diplomatic Chinese-English parallel dataset and the United Nations Parallel Corpus (Chinese-English part) show the effectiveness and superiority of our proposed AFSP.

Introduction

Neural Machine Translation (NMT) (Bahdanau, Cho, and Bengio 2015), the core of which lies in the encoder-decoder architecture, aims to translate texts in the source language into the target language automatically. NMT is a challenging task since it involves translating text among different languages and requires semantic alignment between languages (Fan et al. 2021; Costa-jussà et al. 2022; Yuan et al. 2023). Even so, it has made remarkable progress in recent years, especially with the emergence of large language models (LLMs) like ChatGPT & GPT-4 (Ouyang et al. 2022), GLM (Du et al. 2022), Llama (Touvron et al. 2023; Dubey et al. 2024), etc. Benefiting from the increasing scale of parameters and training corpus, these LLMs have gained a universal ability to handle various NLP tasks via in-context learning (ICL) (Brown et al. 2020) or prompt engineering (Chen et al. 2023), which is the process of structuring input text with exemplars and human-written instructions for LLMs, rather than conducting costly task-specific fine-tuning. Unsurprisingly, LLMs with ICL or prompting techniques have shown outstanding potential in machine translation (Zhang, Haddow, and Birch 2023a; Zhu et al. 2024; Zhang et al. 2023) by constructing elaborate instruction or

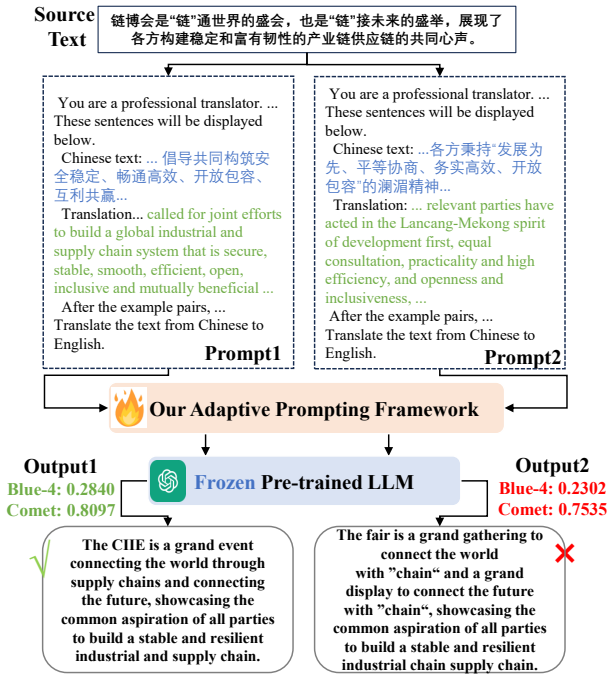

Figure 1: Illustration of translation results from different prompts using Llama3-8B. Our adaptive prompting framework can adaptively select the suitable prompt for an input text, yielding a better translation.

*Corresponding author

prompts with different prompting strategies.

Pioneering work (Zhang, Haddow, and Birch 2023a) conducted a systematic study on prompting strategies for machine translation with the testbed GLM-130B (Zeng et al. 2022), including zero-shot prompting and few-shot prompting. Coincidentally, another work (Zhang et al. 2023) evaluated 15 publicly available language models on machine translation tasks with zero-shot prompting and few-shot learning. (Zhu et al. 2024) explored the multilingual translation capabilities of eight popular LLMs, including ChatGPT and GPT-4. Although all these existing works have shown promising translation performance under the settings of both zero-shot prompting and few-shot ICL, they found that the prompt examples matter the translation performance, which means that LLMs are prompt-sensitive. Using suboptimal examples or instructions can degenerate translation. For example, as shown in Figure 1, the LLM with prompt 1 which is more related to the input text can generate a better translation result than the LLM with prompt 2 according to the BLEU and Comet. In terms of semantic consistency, we can also observe that the translation quality of the LLM with prompt 1 is higher than the LLM with prompt 2. Therefore, selecting suitable adaptive translation demonstrations to elicit the translation capability of an LLM is crucial for high-quality machine translation under in-context learning.

Choosing suitable translation demonstrations for different input text is challenging and nontrivial. To address this issue, we propose an **A**daptive **F**ew-**S**hot **P**rompting (AFSP) framework to automatically select suitable translation demonstrations for various source input sentences to further elicit the translation capability of an LLM for better machine translation. First, we build a translation demonstration retrieval module based on LLM's embedding to retrieve top-k semantic-similar translation demonstrations from aligned parallel translation corpus. The retrieval top-k translation demonstrations will be filled into the handcrafted instruction prompt template which is used for various source sentences uniformly. These translation demonstrations are crucial in eliciting the translation capability of an LLM to generate more semantic-consistent target sentences with current input source sentences. M3-Embedding (Chen et al. 2024) shows that conducting semantic retrieval with a combination of different retrieval functionalities can achieve better retrieval performance by improving the discrimination of embeddings. Inspired by this, we construct a demonstration retrieval module based on dense embedding, sparse embedding, and multi-vector embedding to build better input representation for retrieving more semantic-related translation demonstrations. The dense embedding, sparse embedding, and multi-vector embedding of a sentence are generated from deployed LLM which is also used for machine translation. Then, we use a constructed adaptive few-show prompt to obtain the translation result in the target language. There is output diversity in an LLM (Kirk et al. 2023) due to the probabilistic sampling. Different outputs can lead to different translation quality. To mitigate semantic bias caused by LLMs' probabilistic sampling and ensure semantic-consistent translation, we force the deployed LLM to generate multiple output candidates in the target language and rerank these candidates by a rerank model based on a small language model (SLM). Since there is no available large-scale annotated corpus about the translation quality of different translation outputs and annotating such a corpus is costly, we train the rerank model at a lower cost with a self-supervision way by negative sampling with different text perturbation. With the rerank model, we can choose better translation results, ensuring better semantic consistency between source inputs and target outputs.

Besides, Language evolves throughout time. To better evaluate the effectiveness of our AFSP framework on the latest language and extend the research boundary of neural machine translation, we construct a high-quality diplomatic Chinese-English parallel dataset that consists of 5,528 parallel Chinese-English sentences about the question answers with Chinese foreign-ministry spokesman and foreign journalists. These parallel sentences have very high semantic consistency since they are diplomatically oriented and have been rigorously vetted and proofread. Extensive experiments on our proposed diplomatic Chinese-English parallel dataset and United Nations Parallel Corpus (Chinese-English part) show that the effectiveness and superiority of our AFSP.

Related Work

The emergence of LLMs has shown outstanding potential in the field of machine translation. Unlike traditional neural machine translation methods (Bahdanau, Cho, and Bengio 2014; Sennrich, Haddow, and Birch 2015; Wang et al. 2022) which need to be trained with a large-scale machine translation dataset, LLMs were trained on general large-scale corpus and could effectively finish downstream machine translation tasks via prompt engineering or in-context learning without extra model tuning. Current research evaluating and improving the machine translation capabilities of LLMs can be included in two lines. The first line focuses on comprehensive evaluations of LLMs under various translation scenarios, including multilingual translation (Jiao et al. 2023; Hendy et al. 2023), document-level translation (Wang et al. 2023; Hendy et al. 2023), low-resource translation (Jiao et al. 2023; Bawden and Yvon 2023), etc. Another line focuses on designing novel mechanisms to improve the machine translation capabilities of LLMs, including the design of prompt templates (Zhang, Haddow, and Birch 2023b; Jiao et al. 2023), demonstration selection for in-context learning (Zhang, Haddow, and Birch 2023b; Vilar et al. 2022; García et al. 2023; Yao et al. 2023; Merx et al. 2024; Jiang and Zhang 2024a), self-refinement (Feng et al. 2024b,a), agentic workflow (Wu et al. 2024; Guo et al. 2024), etc.

Among these research lines, the most relevant to our work is the demonstration selection. (Vilar et al. 2022) investigated various strategies for choosing translation examples for few-shot prompting. (García et al. 2023) outperformed the best-performing system on the WMT'21 English-Chinese news translation task by only using five random examples of English-Chinese parallel data at inference. Both these two works found that example quality is the most important factor, but random sampling will influence their performances. (Yao et al. 2023) proposed a low-resource LLM prompting technique In-Context Sampling

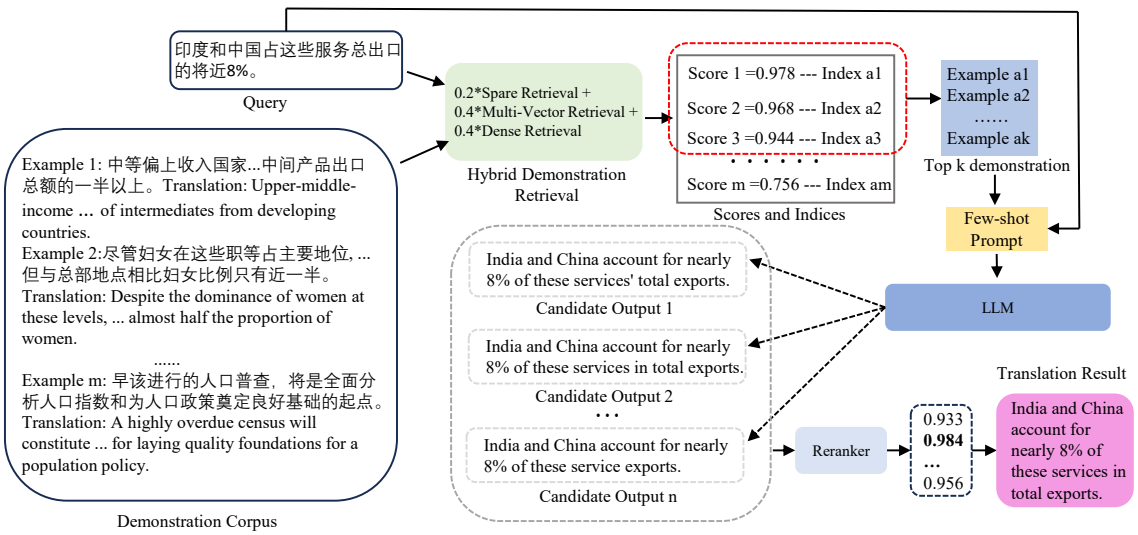

Figure 2: The overview of our proposed Adaptive Few-shot Prompting (AFSP) framework.

(ICS) to produce confident predictions by optimizing the construction of multiple ICL prompt inputs. Leveraging a novel corpus derived from a Mambai language manual and additional sentences translated by a native speaker, (Merx et al. 2024) examine the efficacy of few-shot prompting for machine translation (MT) in the low-resource context by prompting with the strategic selection of parallel sentences and dictionary entries, enhancing translation accuracy.

Different from them, our AFSP automatically selects suitable translation demonstrations for various source input sentences to elicit the translation capability of an LLM for better machine translation. Rather than using other embedding models for semantic demonstration retrieval, our AFSP first deploys a translation demonstration retrieval module based on the deployed LLM's embedding to retrieve top-k semantic-similar translation demonstrations from aligned parallel translation corpus. Then, to ensure better semantic consistency between source inputs and target outputs, we force the deployed LLM to generate multiple output candidates in the target language with the help of translation demonstrations and rerank these candidates with an SLM which is trained in a self-supervised way.

Adaptive Few-shot Prompting (AFSP)

We introduce our **A**daptive **F**ew-**S**hot **P**rompting framework, which first adaptively retrieves suitable demonstrations to fill into the placeholder in the prompt template from the demonstration corpus and then sorts the multiple candidate sampled outputs generated by the deployed LLM to obtain final translation result. In this work, except for the demonstration placeholder, we use fixed prompt words in the prompt template for general task description. The overview of the inference phase in our AFSP framework is shown in Figure 2. AFSP relies on three key components: a translation demonstration corpus, a hybrid demonstration retrieval module based on the deployed LLM-driven embedding, and a re-ranker module. The translation demonstration corpus aims to provide high-quality parallel translation pairs. The retrieval module takes charge of selecting suitable demonstrations to fill into the prompt template for each input source text. The hybrid demonstration retrieval module is train-free and produces a relevance score based on multiple types of embedding ways for each input source and the text in the demonstration corpus by the deployed LLM for machine translation, rather than using third-party embedding models. With the retrieval demonstrations, we fill the demonstrations into a predefined few-shot prompt and enter it into an LLM for multiple candidate output generation. Finally, we deploy a re-ranker module, which is a small language model (SLM) trained in a self-supervised manner, to sort the generated candidate outputs and obtain the final translation result.

Prompt Template and Demonstration Corpus

In AFSP, as shown in Table 1, we only use a fixed prompt template with variable placeholders inspired from prior works (Jiang and Zhang 2024b; Agarwal et al. 2024). We do not focus on the diversified design of prompt templates in this work and achieve adaptive prompts for different source text by filling suitable demonstrations according to the retrieved results from the demonstration corpus. Demonstration corpus can be any high-quality parallel translation corpus. In practice, it can be expanded with extra parallel translation corpus. For the sake of simplicity, we simply use the training set from specific translation tasks as the demonstration sources to build the demonstration corpus. For example, we use the training part in the UN Open Corpus v1.0 (Chinese and English versions) as the demonstration corpus when conducting Chinese-English bilingual translation. Similarly, we also use the training subset of our newly constructed Diplomatic corpus as the demonstration corpus when conducting machine translation on its test set.

You are a professional translator. I will give you one or more examples of text fragments, where the first one is in $\{src\ lang\}$ and the second one is the translation of the first fragment into $\{tgt\ lang\}$. These sentences will be displayed below. 1. $\{src\ lang\}$ text: $\{src\ demo\ 1\}$ $\{tgt\ lang\}$ translation: $\{tgt\ demo\ 1\}$ 2. $\{src\ lang\}$ text: $\{src\ demo\ 2\}$ $\{tgt\ lang\}$ translation: $\{tgt\ demo\ 2\}$ ... k. $\{src\ lang\}$ text: $\{src\ demo\ k\}$ $\{tgt\ lang\}$ translation: $\{tgt\ demo\ k\}$ After the example pairs, I will provide a/an $\{src\ lang\}$ sentence and I would like you to translate it into $\{tgt\ lang\}$. Please provide only the translation result without any additional comments, formatting, or chat content. Translate the text from $\{src\ lang\}$ to $\{tgt\ lang\}$.

Table 1: Illustration of the few-shot prompt used in our work. The placeholders $\{src\ lang\}$ and $\{tgt\ lang\}$ will be replaced with specific source and target language names in practice, such as Chinese, English, etc. Similarly, the placeholders $\{src\ demo\ i\}$ and $\{tgt\ demo\ i\}$ will also be replaced with retrieved parallel translation demonstrations where $i \in [1, ..., k]$.

Hybrid Demonstration Retrieval

As claimed in the pioneering works (Zhang, Haddow, and Birch 2023a; Zhang et al. 2023; Zhu et al. 2024), the number, quality, and semantic similarity of prompt examples matter the translation performance. Therefore, it is crucial to adaptively retrieve high-quality and highly semantic similar demonstrations for different input texts to achieve better LLM prompting for better eliciting the translation capability of an LLM. To achieve this goal, superior embedding-based semantic representation is essential. (Chen et al. 2024) shows that a combination of different embedding-based retrieval functionalities can improve the discrimination of embedding-based semantic representation. Inspired by them, we build an embedding-based hybrid demonstration retrieval module for demonstration retrieval in a training-free way by utilizing the embedding matrix of the deployed LLM that conducts machine translation. The retrieval results are sorted by a weighted combination of relevance scores based on dense embedding, sparse embedding, and multi-vector embedding. The reason we use the deployed LLM as the embedding generator rather than other embedding models is that LLM is pre-trained with large-scale general corpus and can represent text accurately.

Formally, given a query text q in the source language, the demonstration retrieval module can retrieve translation demonstration $\langle d^{src}, d^{tgt} \rangle$ from the corpus $\mathcal{D}$ based on the hybrid relevance score s_rank of q and d^{src}: $\langle d^{src}, d^{tgt} \rangle = f_h(q, \mathcal{D})$. Here, $f_h(\cdot)$ denotes the retrieval function based on the hybrid relevance score. For the text q, dense embedding, sparse embedding, and multi-vector embedding can be formalized separately as follows: 1) dense embedding e_q^{dense}: the text q is first transformed into the embedding vectors $\mathbf{E}_q$ based on the embedding layer in the LLM. Then, we obtain e_q^{dense} by conducting max pooling on $\mathbf{E}_q$ and normalization: $e_q^{dense} = norm(MaxPooling(\mathbf{E}_q))$. 2) sparse embedding e_q^{sparse}: the embedding vector $\mathbf{E}_q$ is also used to estimate the importance of each token to facilitate lexical representation. For each token t within the text q, the token weight is calculated as $w_{q_t} = ReLU(\mathbf{W}_{sparse}^T \mathbf{E}_q[t])$, where $ReLU$ is rectified linear unit and $\mathbf{W}_{sparse} \in \mathbb{R}^{H \times 1}$. H is the dimension size of the embedding and $\mathbf{W}_{sparse}$ is a projection matrix mapping token embedding into a float number as its importance. It is only initialized by Gaussian initialization since we found Gaussian initialization is enough to make the model work fine without any model training. 3) multi-vector embedding e_q^{multi}: it is an extension of dense embedding by utilizing the entire output embeddings for text representation: $e_q^{multi} = norm(\mathbf{W}_{multi}^T \mathbf{E}_q)$, where $\mathbf{W}_{multi} \in \mathbb{R}^{H \times H}$ is a projection matrix initialized by Gaussian initialization.

With the above three embeddings with different granularities, we can calculate three relevance scores for multi-granularity retrieval. For dense retrieval, given a text q and source demonstration p, we can compute the relevance score s_{dense} by the inner product between the two embeddings e_q^{dense} and e_p^{dense} as follows: $s_{dense} = e_q^{dense} \cdot e_p^{dense}$. For sparse retrieval, we can compute s_{sparse} by the joint importance of the co-existed tokens (denoted as $q \cap p$) as follows: $s_{sparse} = \sum_{t \in q \cap p} (w_{q_t} \times w_{q_t})$. For multi-vector retrieval, we can compute s_{multi} by late interaction as follows: $s_{sparse} = \frac{1}{l_q} \sum_{i=1}^{l_q} max_{j=1}^{l_p} e_q^{multi}[i] \cdot e_p^{multi}[j]^T$, where l_q and l_p are the lengths of text q and source demonstration p.

Based on the above three relevance scores, we conduct the demonstration retrieval in a hybrid process according to s_{rank} which can be defined as follows:

$$s_{rank} = \alpha_1 \times s_{dense} + \alpha_2 \times s_{sparse} + \alpha_3 \times s_{multi} \quad (1)$$

where α_1, α_2, and α_3 are three hyper-parameters to adjust the weights of three retrieval functionality.

Result Re-ranking

There is output diversity in an LLM (Kirk et al. 2023) due to the probabilistic sampling. The different outputs may have different semantic biases, which will influence the final translation performance. To mitigate this issue, we force the deployed LLM to generate multiple output candidates in the target language and rerank these candidates by a re-ranker model based on a small language model (SLM). The re-ranker takes charge of scoring the output candidates. However, training this re-ranker is challenging since there is available large-scale annotated corpus about the translation quality of different translation outputs and annotating such a corpus is costly. Therefore, we design a self-supervised training method to train such a re-ranker at a low cost by conducting negative sampling with different text perturbations.

Negative Sampling Formally, given the parallel translation corpus $\mathcal{D} = \left\{< d_1^{src}, d_1^{tgt} >, < d_2^{src}, d_2^{tgt} >, ..., < d_N^{src}, d_N^{tgt} >\right\}$ where d_i^{src} and d_i^{tgt} are the texts in the source language and the target language respectively. To construct a dataset $\mathcal{D}' = \left\{< d_1^{tgt'}, s_1' >, < d_2^{tgt'}, s_2' >, ..., < d_M^{tgt'}, s_M' >\right\}$

to train the re-ranker, we can disturb the d_i^{tgt} by multiple degeneration operation set A including converting to the parallel text ($Parallel$), back translation ($Back$), inserting source text ($Insert$), spelling mistake (Se), repeated translation (Ret), synonym replacement ($Replace$). $d_i^{tgt'}$ and s_i' are the degenerated text and corresponding quality score, respectively. We define the quality score of the original target text d_i^{tgt} as 1. Assuming that B contains a null operation that means we just copy the original text into $\mathcal{D}'$ and all possible combinations of the degeneration operation in A, for each possible combination $b_i \in B$, we can obtain the degenerated text $d_i^{tgt'}$ and calculate its score s' as follows:

$$\begin{aligned} d_i^{tgt'} &= f_{b_i}(d_i^{tgt}), b_i \in B \\ s_i' &= 1 - 0.2 \cdot |b_i| \end{aligned} \tag{2}$$

where $f_{b_i}(\cdot)$ represents the degeneration function with the degenration operation combination b_i and $|b_i|$ is the number of degeration operations in b_i. In this way, we can generate a large-scale dataset $\mathcal{D}'$ from the parallel translation corpus $\mathcal{D}$ to train the re-ranker in a self-supervised manner.

Re-ranker and Learning Objectives We deploy BERT (Devlin et al. 2019) as the backbone of the SLM in the Re-ranker. For Chinese-English translation, we use Bert-large-cased[1] while we use Bert-based-Chinese[2] as the SLM for English-Chinese translation. Given a degenerated text $d_i^{tgt'}$ and its quality score s_i', the re-ranker takes the degenerated text $d_i^{tgt'}$ as input and predicts a quality assessment score s_i^{rerank} as close to the annotated score s_i' as possible. The re-ranker calculates the quality assessment score by applying a Linear layer to map the output encoding of [CLS] token into 1-D float number followed by the $Sigmoid$ function to normalize the output to between 0 and 1. To optimize the re-ranker, we adopted Mean Squared Error as its objective function to enable the re-ranker to predict quality assessment scores. Therefore, the re-ranker and its learning objective can be modeled as follows:

$$\begin{aligned} \mathbf{E} &= BERT(d_i^{tgt'}) \\ s_i^{rerank} &= Sigmoid(Linear(\mathbf{E}[0])) \\ \mathcal{L} &= \left\| s_i^{rerank} - s_i' \right\|_2 \end{aligned} \tag{3}$$

Experiments

Experiment Settings

Datasets To validate the effectiveness of the proposed AFSP, we first crawled a high-quality parallel Chinese-English dataset named Diplomatic corpus from the China Diplomatic website[3]. The Diplomatic corpus consists of speeches made by spokespersons during routine press conferences, including questions posed by journalists and responses from Chinese spokespersons on a range of diplomatic issues. There are some advantages in the Diplomatic corpus. The first is accessibility. All data is publicly available on the China Diplomatic website and can be easily found online. The second is high quality. All bilingual materials are translated by professional translators from specialized institutions, ensuring superior quality. The third is language complexity. The Diplomatic corpus contains certain political terminology and specialized terms, which may pose challenges for the LLM in the context of China's political landscape. The final one is recency. All texts are recorded from 2022 to 2023 which can reflect the latest advances in language. Besides, we also use a Chinese-English subset from the UN Open Corpus v1.0 as the second testbed, which can show the universality of the proposed AFSP. We use UN to denote this subset. The UN Parallel Corpus is a parallel corpus that includes official UN documents and statements from meetings. The content covers various fields such as politics, economics, culture, and technology. This corpus records texts written and manually translated from 1990 to 2014, aligned at the sentence level. We randomly selected 120,000 entries as the data for evaluation. The statistics for both two datasets are shown in Table 2. For both two datasets Diplomatic and UN, we randomly selected 500 parallel translation pairs to serve as the test set for evaluating AFSP. The remaining pairs are used as the demonstration corpus for adaptive demonstration retrieval.

Dataset	Language	#Sent.	#Word	#Average.
Diplomatic	English	5.528K	415K	75.20
	Chinese	5.528K	316K	57.29
UN	English	120K	3,500K	29.11
	Chinese	120K	3,220K	26.75

Table 2: Statistics of the Diplomatic Corpus and UN. The statistics include the number of sentences (#Sent.), the number of words (#Word), and the average number of words per sentence (#Average.).

Metrics To conduct a comprehensive assessment of translation quality, we employed the most commonly used BLEU-4 (B-4) (Papineni et al. 2002), METEOR (ME) (Banerjee and Lavie 2005), ROUGE-1 (R-1), ROUGE-2 (R-2), ROUGE-L (R-L) (Lin 2004), CHRF (CH) (Popović 2015), and COMET-Kiwi (CK)(Rei et al. 2022) as evaluation metrics. We also evaluate our AFSP by conducting a human evaluation of fluency (FLU), accuracy (ACC), and consistency of style (STY).

Implementation Details We evaluate AFSP on 3 open-source LLMs, ChatGLM3-6B, InternLM2-7B, and Llama3-8B, as well as a closed-source LLM ChatGPT-3.5-turbo-0125 by comparing with three baselines, Zero-shot (Jiang and Zhang 2024a), Few-shot (Jiang and Zhang 2024a), kNN-based few-shot (Nori et al. 2023) in both Chinese-to-English and English-to-Chinese translation directions. The α_1, α_2, and α_3 are set to 0.4, 0.4, 0.2 for computing the final relevance score s_{rank}. The k for few-shot prompts is set to 3 due to the limited memory of NVIDIA RTX 3090. For the closed-source ChatGPT-3.5-turbo-0125, we deploy ChatGLM3-6B as the embedding model for hybrid

[1]https://huggingface.co/google-bert/bert-large-cased

[2]https://huggingface.co/google-bert/bert-base-chinese

[3]https://www.fmprc.gov.cn/

Methods	English-to-Chinese							Chinese-to-English						
	B-4	ME	R-1	R-2	R-L	CH	CK	B-4	ME	R-1	R-2	R-L	CH	CK
ChatGLM3-6B														
Zero-shot	18.9	50.6	21.2	10.3	21.1	42.8	88.1	22.7	54.6	59.9	32.0	50.9	66.5	81.4
Few-shot	20.3	51.7	20.5	10.3	20.5	43.8	88.2	23.6	55.7	60.8	33.1	51.6	67.1	81.4
KNN Few-shot	20.2	51.8	20.8	10.4	20.7	43.9	88.3	22.9	55.1	60.4	32.5	51.1	66.7	81.3
AFSP w/o rerank (Ours)	25.9	56.8	22.5	11.4	22.4	48.6	89.0	28.1	59.3	63.8	37.3	54.9	69.3	82.3
AFSP (Ours)	**27.4**	**58.2**	**22.9**	**11.5**	**22.9**	**49.8**	**89.1**	**29.2**	**60.0**	**64.0**	**37.6**	**55.2**	**69.7**	**82.7**
InternLM2-7B														
Zero-shot	20.6	52.3	22.4	11.1	22.3	44.5	88.4	23.4	56.1	60.4	32.6	51.2	67.7	81.9
Few-shot	24.5	55.7	22.7	11.6	22.5	47.2	89.0	24.2	57.1	61.4	33.8	52.2	68.5	81.7
KNN Few-shot	24.9	56.3	22.8	11.7	22.6	47.6	89.0	24.7	57.4	61.7	34.3	52.6	68.7	81.9
AFSP w/o rerank (Ours)	30.7	60.8	22.9	11.9	22.7	52.3	89.5	30.1	61.2	65.0	39.3	56.3	71.0	82.5
AFSP (Ours)	**31.8**	**61.3**	**22.9**	**11.9**	**22.7**	**52.9**	**89.6**	**31.3**	**61.9**	**64.9**	**39.7**	**56.1**	**70.9**	**82.7**
Llama3-8B														
Zero-shot	10.7	37.2	17.4	8.1	17.2	32.0	84.6	24.0	56.7	60.9	33.6	51.7	68.2	82.3
Few-shot	15.5	45.7	18.5	8.8	18.3	39.0	86.2	25.7	58.0	62.4	35.2	53.1	68.8	82.8
KNN Few-shot	16.5	46.8	18.7	9.0	18.5	39.9	86.5	25.3	57.6	62.3	35.1	53.0	68.4	82.8
AFSP w/o rerank (Ours)	26.3	55.6	19.8	9.6	19.6	48.0	88.0	30.7	61.5	65.7	40.2	57.1	71.1	83.3
AFSP (Ours)	**27.7**	**56.7**	**19.9**	**9.9**	**19.7**	**49.1**	**88.4**	**31.0**	**61.5**	**65.7**	**40.3**	**57.0**	**71.0**	**83.3**
Chatgpt-3.5-turbo-0125														
Zero-shot	23.0	54.9	22.3	11.1	22.2	46.5	89.1	27.7	59.9	64.1	37.3	55.5	70.7	83.2
Few-shot	24.6	56.7	22.3	11.2	22.2	48.2	89.5	28.1	60.5	64.6	37.7	56.0	71.0	83.4
KNN Few-shot	25.6	57.8	22.4	11.4	22.3	49.1	89.5	28.3	60.5	64.5	37.9	56.00	70.9	83.5
AFSP w/o rerank (Ours)	30.3	62.1	23.5	11.7	23.4	52.9	89.9	31.3	62.5	66.4	40.8	58.00	72.1	83.9
AFSP (Ours)	**32.3**	**63.5**	**23.3**	**11.7**	**23.3**	**54.3**	**90.3**	**32.3**	**63.2**	**66.9**	**41.4**0	**58.7**	**72.5**	**84.0**

Table 3: Performance Comparison on Diplomatic Corpus. The best result is highlighted in **bold**.

Methods	English-to-Chinese							Chinese-to-English						
	B-4	ME	R-1	R-2	R-L	CH	CK	B-4	ME	R-1	R-2	R-L	CH	CK
ChatGLM3-6B														
Zero Few-shot	18.9	52.9	42.2	17.9	41.7	45.2	86.9	21.5	58.0	60.0	34.2	53.2	64.1	83.6
Few-shot	19.1	52.6	42.1	17.6	41.6	45.1	86.3	21.8	58.2	59.4	34.6	52.8	65.3	82.5
KNN Few-shot	18.8	52.3	41.6	17.5	41.2	44.9	86.6	22.7	59.0	60.1	35.3	53.5	65.1	83.1
AFSP w/o rerank (Ours)	**24.8**	57.3	**42.8**	18.3	**42.5**	50.2	88.0	**29.4**	63.4	64.6	41.7	58.2	69.0	84.6
AFSP (Ours)	24.6	**57.6**	42.8	**18.3**	42.4	**50.5**	**88.2**	29.1	**64.5**	**65.1**	**41.8**	**58.4**	**69.6**	**84.9**
InternLM2-7B														
Zero Few-shot	17.3	50.3	37.4	17.8	36.8	43.5	86.3	25.5	61.9	63.5	38.7	57.0	67.2	84.7
Few-shot	17.9	51.5	41.1	18.4	40.7	44.4	86.7	25.8	62.6	63.5	38.8	57.0	68.4	83.9
KNN Few-shot	17.8	51.3	39.5	18.2	39.0	44.3	86.6	26.3	62.6	64.0	39.3	57.3	68.0	84.5
AFSP w/o rerank (Ours)	**26.2**	58.5	40.5	19.2	40.1	51.6	88.3	31.3	61.9	64.9	39.7	56.1	**72.8**	86.1
AFSP (Ours)	26.1	**58.8**	**40.5**	**19.3**	**40.1**	**51.8**	**88.9**	**34.0**	**67.4**	**68.6**	**46.3**	**62.5**	71.9	**86.5**
Llama3-8B														
Zero Few-shot	8.9	35.4	34.2	15.4	33.7	30.3	81.1	18.6	55.4	58.2	32.1	51.1	63.9	83.4
Few-shot	13.1	44.5	34.1	15.3	33.4	38.4	83.4	21.9	58.1	60.8	35.1	53.7	65.9	84.3
KNN Few-shot	12.6	43.8	29.4	15.0	28.9	37.9	83.5	21.7	57.6	60.7	35.1	53.6	65.7	84.3
AFSP w/o rerank (Ours)	22.7	54.0	33.1	16.5	32.5	47.8	86.5	30.1	63.9	66.8	44.2	60.7	71.3	**86.3**
AFSP (Ours)	**23.6**	**54.5**	**34.3**	**16.9**	**33.8**	**48.5**	**87.2**	**31.0**	**64.9**	**67.1**	**44.9**	**61.0**	**71.7**	86.2
Chatgpt-3.5-turbo-0125														
Zero Few-shot	20.0	52.9	32.0	17.5	31.5	46.3	87.1	24.5	61.3	63.3	38.0	56.9	67.8	85.4
Few-shot	21.3	55.4	43.0	18.6	42.5	47.9	87.9	27.0	63.9	65.3	40.6	58.9	69.7	85.9
KNN Few-shot	20.3	53.8	35.1	18.0	34.7	47.1	87.5	27.1	63.1	64.9	40.6	58.7	69.3	85.9
AFSP w/o rerank (Ours)	28.4	61.1	40.7	19.0	40.3	53.9	89.1	32.7	66.9	68.9	46.2	63.1	73.0	86.9
AFSP (Ours)	**29.1**	**62.0**	**42.4**	**19.1**	**42.0**	**54.6**	**89.5**	**34.1**	**67.5**	**69.3**	**47.0**	**63.7**	**73.3**	**87.3**

Table 4: Performance Comparison on UN. The best result is highlighted in **bold**.

demonstration retrieval. We conduct top-30 sampling for ChatGLM3-6B, InternLM2-7B, and Llama3-8B and top-5 sampling for ChatGPT-3.5-turbo-0125.

Experiment Result

Main Results Table 3 and Table 4 show the performance of our AFSP and baselines on different translation

Method	FLU	ACC	STY	FLU	ACC	STY
	Chinese-to-English			English-to-Chinese		
Zero-shot	14.3	20.7	20.0	14.3	12.1	14.3
Few-shot	25.7	25.7	27.1	14.3	25.0	17.1
KNN Few-shot	27.1	22.9	19.3	17.8	14.3	15.0
ASFP (Ours)	**32.9**	**30.7**	**30.7**	**53.6**	**48.6**	**53.6**

Table 5: Human evaluation results on the translation performance for the Diplomatic corpus.

Emb	BGE-M3	E5-Large	BGE-large	BCE	ChatGLM3
B-4	23.6	24.1	20.3	24.7	**27.4**
ME	54.9	55.1	51.8	55.6	**58.2**
R-1	21.5	22.0	20.6	21.8	**22.9**
R-2	10.8	10.9	10.4	10.8	**11.5**
R-L	21.4	21.9	20.5	21.7	**22.9**
CH	46.7	47.0	43.9	47.5	**49.8**
CK	88.7	88.7	88.3	88.9	**89.1**

Table 6: The translation performance of ASFP with different embedding models for ChatGLM3-6B on the English-to-Chinese part of Diplomatic Corpus.

datasets and different LLMs. Compared to baselines, our AFSP demonstrates superior performance by always generating higher-quality translation according to various metrics. For instance, when Llama-3-8B translates from Chinese to English on the UN, our ASFP achieves significant improvements over KNN Few-shot with 9.3 improvement in BLEU-4, 7.3 improvement in METEOR, 7.4 improvement in ROUGE-L, and 1.9 improvement in COMET-Kiwi. Other models also show significant metric improvements in translation across different datasets and LLMs, highlighting the effectiveness of our AFSP method.

Human Evaluation To further validate the effectiveness of the AFSP, we also conducted a human evaluation of both two datasets and two translation directions to compare AFSP with baselines. For each translation direction, we randomly selected 5 examples from each dataset. Participants judged the options based on fluency, accuracy, and style retention by selecting the sentence they deemed best. We tested the translation results generated by Llama3-8B and invited 14 teachers or students fluent in English or Chinese to participate in the evaluation for each translation direction. To avoid bias, the output order was randomized. The evaluation results in Table 5 demonstrate that our AFSP outperforms all baselines in fluency, semantic accuracy, and style consistency.

The Choice of Embedding Model In our hybrid demonstration retrieval, we use the embedding model in the deployed LLM to compute relevance scores. To show the effectiveness of using the embedding model of the deployed LLM model, we conduct an ablation study on various embedding models including BGE-M3, E5-Large, BGE-large, and BCE. The results in Table 6 show using the embedding model of the deployed LLM model is a better choice than using third-party embedding models.

Ablation on Weights of Hybrid Demonstration Retrieval The hybrid demonstration retrieval uses multiple retrieval functions to compute relevance scores. To investigate the influence of different weights α_1, α_2, and α_3, we conduct experiments with ChatGLM3-6B on the Diplomatic Corpus by setting different weights. The results in Table 7 show that α_1, α_2, and α_3 are set to 0.4, 0.4, and 0.2 can achieve the best performance on most of the metrics.

α_1	0.2	0.3	0.4	0.25	0.25	0.35	0.4
α_2	0.4	0.3	0.3	0.25	0.35	0.35	0.4
α_3	0.4	0.4	0.3	0.5	0.4	0.3	0.2
B-4	25.8	25.7	25.9	25.7	25.6	25.7	**25.9**
ME	56.7	56.5	56.6	56.6	56.5	56.6	**56.8**
R-1	22.2	22.2	22.4	22.4	22.2	22.2	**22.5**
R-2	11.3	11.2	11.4	11.2	11.1	11.2	**11.4**
R-L	22.1	22.1	22.4	22.3	22.2	22.1	**22.4**
CH	48.5	48.3	48.4	48.4	48.3	48.4	**49.8**
CK	**89.0**	88.9	**89.0**	**89.0**	**89.0**	**89.0**	**89.0**

Table 7: The translation performance of ChatGLM3-6B on the English-to-Chinese part of Diplomatic Corpus with different weights.

Demos	1	2	3	1	2	3
	Chinese-to-English			English-to-Chinese		
B-4	26.7	27.7	**29.2**	24.2	26.0	**27.4**
ME	58.1	58.9	**60.0**	55.2	56.9	**58.2**
R-1	62.0	63.6	**64.0**	22.2	22.0	**22.9**
R-L	54.1	54.7	**55.2**	22.1	22.0	**22.9**
CH	68.7	69.1	**69.7**	47.2	47.8	**49.8**
CK	81.8	82.3	**82.7**	88.8	88.9	**89.1**

Table 8: The performance of different numbers of demonstrations on the Diplomatic Corpus with ChatGLM3-6B.

The Number of Translation Demonstrations We verify the effects of different numbers of demonstrations for few-shot prompting by using the Diplomatic Corpus dataset on ChatGLM3-6B. The results in Table 8 show the best translation performance can be achieved when the number of demonstrations is 3.

Conclusion

In this work, we propose an adaptive few-shot prompting (AFSP) framework to automatically select suitable translation demonstrations for various source input sentences to further elicit the translation capability of an LLM for better machine translation. First, we retrieve top-k semantic-similar translation demonstrations from aligned parallel translation corpus based on hybrid demonstration retrieval. Then, to ensure better semantic consistency between source inputs and target outputs, we force the deployed LLM itself to generate multiple output candidates in the target language with the help of translation demonstrations and rerank these candidates. Besides, to better evaluate the effectiveness of our AFSP framework on the latest language and extend the research boundary of neural machine translation, we construct a high-quality diplomatic Chinese-English parallel dataset that consists of 5,528 parallel sentences. Extensive experiments on the proposed Diplomatic dataset and UN show the effectiveness and superiority of our AFSP.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant No. 62206314, GuangDong Basic and Applied Basic Research Foundation under Grant No. 2022A1515011835 and 2023A1515012561, Science and Technology Projects in Guangzhou under Grant No. 2024A04J4388 and 2024A04J4387.

References

Agarwal, R.; Singh, A.; Zhang, L. M.; Bohnet, B.; Chan, S.; Anand, A.; Abbas, Z.; Nova, A.; Co-Reyes, J. D.; Chu, E.; Behbahani, F. M. P.; Faust, A.; and Larochelle, H. 2024. Many-Shot In-Context Learning. *ArXiv*, abs/2404.11018.

Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. *CoRR*, abs/1409.0473.

Bahdanau, D.; Cho, K. H.; and Bengio, Y. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*.

Banerjee, S.; and Lavie, A. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *IEEvaluation@ACL*.

Bawden, R.; and Yvon, F. 2023. Investigating the Translation Performance of a Large Multilingual Language Model: the Case of BLOOM. In *European Association for Machine Translation Conferences/Workshops*.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Chen, B.; Zhang, Z.; Langrené, N.; and Zhu, S. 2023. Unleashing the potential of prompt engineering in Large Language Models: a comprehensive review. *arXiv preprint arXiv:2310.14735*.

Chen, J.; Xiao, S.; Zhang, P.; Luo, K.; Lian, D.; and Liu, Z. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

Costa-jussà, M. R.; Cross, J.; Çelebi, O.; Elbayad, M.; Heafield, K.; Heffernan, K.; Kalbassi, E.; Lam, J.; Licht, D.; Maillard, J.; et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein, J.; Doran, C.; and Solorio, T., eds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Du, Z.; Qian, Y.; Liu, X.; Ding, M.; Qiu, J.; Yang, Z.; and Tang, J. 2022. GLM: General Language Model Pretraining with Autoregressive Blank Infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 320–335.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*.

Fan, A.; Bhosale, S.; Schwenk, H.; Ma, Z.; El-Kishky, A.; Goyal, S.; Baines, M.; Celebi, O.; Wenzek, G.; Chaudhary, V.; et al. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107): 1–48.

Feng, Z.; Chen, R.; Zhang, Y.; Meng, Z.; and Liu, Z. 2024a. Ladder: A Model-Agnostic Framework Boosting LLM-based Machine Translation to the Next Level. *arXiv preprint arXiv:2406.15741*.

Feng, Z.; Zhang, Y.; Li, H.; Wu, B.; Liao, J.; Liu, W.; Lang, J.; Feng, Y.; Wu, J.; and Liu, Z. 2024b. TEaR: Improving LLM-based Machine Translation with Systematic Self-Refinement.

García, X.; Bansal, Y.; Cherry, C.; Foster, G. F.; Krikun, M.; Feng, F.; Johnson, M.; and Firat, O. 2023. The unreasonable effectiveness of few-shot learning for machine translation. *ArXiv*, abs/2302.01398.

Guo, S.; Zhang, S.; Ma, Z.; Zhang, M.; and Feng, Y. 2024. SiLLM: Large Language Models for Simultaneous Machine Translation. *ArXiv*, abs/2402.13036.

Hendy, A.; Abdelrehim, M. G.; Sharaf, A.; Raunak, V.; Gabr, M.; Matsushita, H.; Kim, Y. J.; Afify, M.; and Awadalla, H. H. 2023. How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation. *ArXiv*, abs/2302.09210.

Jiang, Z.; and Zhang, Z. 2024a. Can ChatGPT Rival Neural Machine Translation? A Comparative Study. *arXiv preprint arXiv:2401.05176*.

Jiang, Z.; and Zhang, Z. 2024b. Convergences and Divergences between Automatic Assessment and Human Evaluation: Insights from Comparing ChatGPT-Generated Translation and Neural Machine Translation.

Jiao, W.; Wang, W.; tse Huang, J.; Wang, X.; and Tu, Z. 2023. Is ChatGPT A Good Translator? A Preliminary Study. *ArXiv*, abs/2301.08745.

Kirk, R.; Mediratta, I.; Nalmpantis, C.; Luketina, J.; Hambro, E.; Grefenstette, E.; and Raileanu, R. 2023. Understanding the effects of rlhf on llm generalisation and diversity. *arXiv preprint arXiv:2310.06452*.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Annual Meeting of the Association for Computational Linguistics*.

Merx, R.; Mahmudi, A.; Langford, K.; de Araujo, L. A.; and Vylomova, E. 2024. Low-Resource Machine Translation through Retrieval-Augmented LLM Prompting: A Study on the Mambai Language. *ArXiv*, abs/2404.04809.

Nori, H.; Lee, Y. T.; Zhang, S.; Carignan, D.; Edgar, R.; Fusi, N.; King, N.; Larson, J.; Li, Y.; Liu, W.; Luo, R.; McKinney, S. M.; Ness, R. O.; Poon, H.; Qin, T.; Usuyama, N.; White, C.; and Horvitz, E. 2023. Can Generalist Foundation Models Outcompete Special-Purpose Tuning? Case Study in Medicine. arXiv:2311.16452.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Annual Meeting of the Association for Computational Linguistics*.

Popović, M. 2015. chrF: character n-gram F-score for automatic MT evaluation. In Bojar, O.; Chatterjee, R.; Federmann, C.; Haddow, B.; Hokamp, C.; Huck, M.; Logacheva, V.; and Pecina, P., eds., *Proceedings of the Tenth Workshop on Statistical Machine Translation*, 392–395. Lisbon, Portugal: Association for Computational Linguistics.

Rei, R.; Treviso, M.; Guerreiro, N. M.; Zerva, C.; Farinha, A. C.; Maroti, C.; C. de Souza, J. G.; Glushkova, T.; Alves, D.; Coheur, L.; Lavie, A.; and Martins, A. F. T. 2022. CometKiwi: IST-Unbabel 2022 Submission for the Quality Estimation Shared Task. In Koehn, P.; Barrault, L.; Bojar, O.; Bougares, F.; Chatterjee, R.; Costa-jussà, M. R.; Federmann, C.; Fishel, M.; Fraser, A.; Freitag, M.; Graham, Y.; Grundkiewicz, R.; Guzman, P.; Haddow, B.; Huck, M.; Jimeno Yepes, A.; Kocmi, T.; Martins, A.; Morishita, M.; Monz, C.; Nagata, M.; Nakazawa, T.; Negri, M.; Névéol, A.; Neves, M.; Popel, M.; Turchi, M.; and Zampieri, M., eds., *Proceedings of the Seventh Conference on Machine Translation (WMT)*, 634–645. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics.

Sennrich, R.; Haddow, B.; and Birch, A. 2015. Neural Machine Translation of Rare Words with Subword Units. *ArXiv*, abs/1508.07909.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Vilar, D.; Freitag, M.; Cherry, C.; Luo, J.; Ratnakar, V.; and Foster, G. F. 2022. Prompting PaLM for Translation: Assessing Strategies and Performance. *ArXiv*, abs/2211.09102.

Wang, L.; Lyu, C.; Ji, T.; Zhang, Z.; Yu, D.; Shi, S.; and Tu, Z. 2023. Document-Level Machine Translation with Large Language Models. In *Conference on Empirical Methods in Natural Language Processing*.

Wang, W.; Jiao, W.; Hao, Y.; Wang, X.; Shi, S.; Tu, Z.; and Lyu, M. R. 2022. Understanding and Improving Sequence-to-Sequence Pretraining for Neural Machine Translation. In *Annual Meeting of the Association for Computational Linguistics*.

Wu, M.; Yuan, Y.; Haffari, G.; and Wang, L. 2024. (Perhaps) Beyond Human Translation: Harnessing Multi-Agent Collaboration for Translating Ultra-Long Literary Texts. *ArXiv*, abs/2405.11804.

Yao, B.; Chen, G.; Zou, R.; Lu, Y.; Li, J.; Zhang, S.; Liu, S.; Hendler, J.; and Wang, D. 2023. More Samples or More Prompt Inputs? Exploring Effective In-Context Sampling for LLM Few-Shot Prompt Engineering. *arXiv preprint arXiv:2311.09782*.

Yuan, F.; Lu, Y.; Zhu, W.; Kong, L.; Li, L.; Qiao, Y.; and Xu, J. 2023. Lego-MT: Learning Detachable Models for Massively Multilingual Machine Translation. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Findings of the Association for Computational Linguistics: ACL 2023*, 11518–11533. Toronto, Canada: Association for Computational Linguistics.

Zeng, A.; Liu, X.; Du, Z.; Wang, Z.; Lai, H.; Ding, M.; Yang, Z.; Xu, Y.; Zheng, W.; Xia, X.; et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

Zhang, B.; Haddow, B.; and Birch, A. 2023a. Prompting large language model for machine translation: A case study. In *International Conference on Machine Learning*, 41092–41110. PMLR.

Zhang, B.; Haddow, B.; and Birch, A. 2023b. Prompting Large Language Model for Machine Translation: A Case Study. *ArXiv*, abs/2301.07069.

Zhang, X.; Rajabi, N.; Duh, K.; and Koehn, P. 2023. Machine Translation with Large Language Models: Prompting, Few-shot Learning, and Fine-tuning with QLoRA. In Koehn, P.; Haddow, B.; Kocmi, T.; and Monz, C., eds., *Proceedings of the Eighth Conference on Machine Translation*, 468–481. Singapore: Association for Computational Linguistics.

Zhu, W.; Liu, H.; Dong, Q.; Xu, J.; Huang, S.; Kong, L.; Chen, J.; and Li, L. 2024. Multilingual Machine Translation with Large Language Models: Empirical Results and Analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, 2765–2781.