Towards Robust In-Context Learning for Machine Translation with Large Language Models

Shaolin Zhu‡, Menglong Cui‡, Deyi Xiong*
College of Intelligence and Computing, Tianjin University, Tianjin, China
{zhushaolin, cuimenglongcs, dyxiong}@tju.edu.cn

Abstract

Using large language models (LLMs) for machine translation via in-context learning (ICL) has become an interesting research direction of machine translation (MT) in recent years. Its main idea is to retrieve a few translation pairs as demonstrations from an additional datastore (parallel corpus) to guide translation without updating the LLMs. However, the underlying noise of retrieved demonstrations usually dramatically deteriorate the performance of LLMs. In this paper, we propose a robust method to enable LLMs to achieve robust translation with ICL. The method incorporates a multi-view approach, considering both sentence- and word-level information, to select demonstrations that effectively avoid noise. At the sentence level, a margin-based score is designed to avoid semantic noise. At the word level, word embeddings are utilized to evaluate the related tokens and change the weight of words in demonstrations. With both sentence- and word-level similarity, the proposed method provides fine-grained demonstrations that effectively prompt the translation of LLMs. Experimental results demonstrate the effectiveness of our method, particularly in domain adaptation.

Keywords: Machine translation, Large language model, In-context learning

1. Introduction

LLMs have recently exhibited fascinating abilities with ICL, mastering the skill of reproducing specific input-output text generation patterns without the need for additional fine-tuning (Gao et al., 2021; Lester et al., 2021; Chung et al., 2022). In particular, it has been demonstrated impressive machine translation capabilities through ICL. This is achieved by providing a few translation pairs (or prompt examples) as demonstrations and leveraging LLMs to perform machine translation tasks (Agrawal et al., 2023; Ghazvininejad et al., 2023; Moslem et al., 2023). These demonstrations can flexibly guide the LLMs to make better predictions for translation, especially for domain adaptation (Sia and Duh, 2023; Lyu et al., 2023). Compared with other methods that use LLMs to implement MT (Tan et al., 2022; Wang et al., 2021; Li et al., 2023), it has two advantages: 1) It is more scalable because we can directly boost the translation ability of LLMs by just manipulating the demonstrations. 2) It is more interpretable due to its observable retrieved demonstrations.

Generally, prompting LLMs for MT with ICL mainly involves two stages: (1) candidate demonstrations establishment and (2) taking the demonstrations as prefix inputs of LLMs to boost the translation ability. In the first stage, methods based on semantic similarity, using word- or sentence level embeddings, have been proposed to select similar translation pairs as demonstrations. Along this line, many efforts have been made to improve the translation ability of LLMs by selecting more suitable demonstrations. Ghazvininejad et al. (2023) propose that using prior knowledge from bilingual dictionaries to provide control hints in the prompts can provide an effective solution for translating rare words. Despite the success, prompting LLMs with ICL is with varying degrees of sensitivity to the choice of demonstrations for MT task. Agrawal et al. (2023) show that the domain of the in-context demonstrations matters and that unrelated demonstrations can have a catastrophic impact on output quality. Therefore, how to select demonstrations to enhance the translation ability of LLMs remains to be further explored.

In the second stage, the selected demonstrations are cascaded with test sentences as input of LLMs to implement MT (Zhang et al., 2023a). This is similar to *k*-Nearest-Neighbor NMT which retrieves useful translation pairs (demonstrations) from an additional parallel corpus to modify translations without updating the model (Wang et al., 2022; Zheng et al., 2021; Jiang et al., 2022). However, Jiang et al. (2022) point out that the performance will dramatically deteriorate due to the underlying noise in retrieved pairs for *k*NN NMT, as the retrieved pairs do not always contain ground-truth tokens. This suggests that the noises (e.g., words irrelevant to test sentences) of selected demonstrations may also affect the translation of LLMs. It is not clear whether prompting LLMs for MT with ICL is vulnerable to noisy perturbations in the demonstrations, for which we further conduct a preliminary study in Section 3 to validate this issue. Therefore, further exploration is needed to effectively handle noise when prompting LLMs with

‡ Equal contribution
* Corresponding author

20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

ICL for MT.

Putting these together, we propose a robust method to make LLMs able to overcome noise in demonstrations to implement robust translation with ICL. In particular, our method incorporates a multi-view approach, considering both sentence- and word-level information to select demonstrations to effectively avoid noise.

Min et al. (2022) and Zhang et al. (2023a) demonstrate that the key role of demonstrations lies in their support of the coherency in the input space. Therefore, we design a margin-based score to capture semantic similarity at the sentence level. This score is computed by integrating the cosine similarity between a given candidate and the average cosine similarity of its k nearest neighbors with the testing sentence. Unlike previous approaches that solely rely on nearest neighbor retrieval using cosine similarity, our proposed method takes into account the scale inconsistencies of this measure. It considers the margin between a given sentence pair and its closest candidates instead to avoid the semantic noise of sentence level. At the word level, we utilize word embeddings to evaluate the word similarity between the demonstrations and the testing sentences and add the weight of related words to avoid noisy perturbations.

Using both sentence- and word-level similarity, our method provides fine-grained demonstrations that effectively prompt the control of the LLMs. With carefully selected demonstrations using a multi-view approach with both sentence- and word-level information, our method helps LLMs maintain the quality and coherency of translations with ICL. This ultimately leads to more reliable and robust translation results.

The main contributions of this work are summarized as follows:

- To the best of our knowledge, we are the first to explore the robustness of in-context demonstrations to implement MT using LLMs.

- We propose a robust method to enable LLMs to implement robust translation with ICL. The method incorporates a multi-view approach, considering both sentence- and word-level information, to select demonstrations that effectively avoid noise.

- To investigate the effectiveness and generality of our method, we conduct experiments on several commonly used benchmarks. Experimental results show that our model significantly outperforms the baselines, especially for domain adaptation.

2. Related Work

Recently, prompting LLMs for MT with ICL has shown effectiveness in improving the translation capabilities of LLMs (Puduppully et al., 2023; Zhu et al., 2023a; Lyu et al., 2023). Usually, they first retrieve relevant sentences as prompts with testing source sentences from the parallel corpus and then use them as a part of the input to boost translation capabilities of LLMs (Han et al., 2022; Vilar et al., 2023). For example, Garcia and Firat (2022) propose a MT approach based on natural language prompts, where the prompts are human-readable instructions that guide the model in generating accurate translations. By inputting the natural language prompts along with the source language sentences into the machine translation model, the model can generate more precise translations guided by the prompts. Garcia et al. (2023) introduce and demonstrate the effectiveness of using few-shot examples to control translation formality. It also supports the finding that the quality of the few-shot in-context examples plays a crucial role. Agrawal et al. (2023) address the challenge of selecting relevant and effective examples from a large parallel corpus to guide the machine translation process. It recognizes that the quality and relevance of the examples play a crucial role in the translation output. Our work provides both supporting and complementary pieces of evidence to these studies by 1) contributing a systematic analysis showing that although the demonstrations can improve the translation of LLMs, the noises of demonstrations deteriorate translation quality and 2) introducing a multi-view augmentation technique to enhance LLMs to improve the robustness of MT for noisy perturbations of demonstrations.

Prompting LLMs with ICL to implement MT task is similar to kNN NMT that retrieves useful translation pairs (demonstrations) from an additional parallel corpus to modify translations without updating the NMT model (Wang et al., 2022; Zhu et al., 2023b; Deguchi et al., 2023). Zheng et al. (2021) leverage contextual information and similarity measures to identify the most relevant and suitable nearest neighbors for each translation instance, leading to improved translation quality. Jiang et al. (2022) show that small perturbations in demonstrations can significantly impact the translation quality of kNN NMT. There have been numerous works (Pan et al., 2023; Qin et al., 2021; Zeng and Xiong, 2021; Cheng et al., 2020; Miao et al., 2022; Zhang et al., 2023b) that explore the robustness of NMT models. One area of these studies is to enhance the robustness of NMT models against small perturbations in training datasets. We believe that the presence of noise in demonstrations can indeed have a detrimental effect on the translation ability

Similarity score		> 0.9	0.75 ∼ 0.89	0.6 ∼ 0.74	<0.6
en-fr	1-shot	65	330	1,123	482
	3-shot	29	139	976	856
	5-shot	23	89	828	1,060
en-es	1-shot	108	447	1,007	438
	3-shot	42	226	969	763
	5-shot	31	150	888	931

Table 1: The numbers of sentences based on their similarity score interval to the source of demonstrations for English-French (en-fr) and English-Spanish (en-es) in 1-shot, 3-shot and 5-shot scenarios.

Language pairs & context		>0.9	0.75∼ 0.89	0.6∼ 0.74	<0.6
en-fr	1-shot	38.73	30.47	22.43	19.70
	3-shot	41.42	29.37	22.69	18.51
	5-shot	45.14	28.75	21.74	18.00
en-es	1-shot	44.82	29.59	23.90	20.25
	3-shot	61.85	28.48	23.20	19.88
	5-shot	67.86	27.94	21.12	18.82

Table 2: BLEU scores on each interval of OPUS test set for English-French (en-fr) and English-Spanish (en-es) in 1-shot, 3-shot and 5-shot scenarios. The demonstrations are selected via BM25.

of LLMs. In our study, we thoroughly investigate the impact of noise on LLMs in Section 3. Our findings confirm that the presence of noise in demonstrations does indeed deteriorate the translation ability of LLMs. To the best of our knowledge, our work is the first to address this problem and propose techniques to mitigate the impact of noise in demonstrations.

3. Preliminary Study

To investigate the impact of noise of demonstrations for translation of LLMs, we conduct a preliminary case experiment on BLOOM-7B with OPUS dataset[1] in this section. We use sentence-transformer[2] toolkit to extract semantically similar demonstrations with testing source sentences from the OPUS training dataset. Next, we partition the test set into multiple subsets based on the similarity scores.[3] Each subset represents a specific interval of similarity between the source sentences of the demonstrations and the testing sentences. This partitioning allows us to analyze the impact of noise in demonstrations on the translation ability of LLMs across different similarity intervals. Table 1 provides a detailed breakdown of the number of sentences in each partition for the 1-shot, 3-shot, and 5-shot scenarios. Table 2 shows the translation quality (BLEU) of LLMs in each partition for the 1-shot, 3-shot, and 5-shot scenarios. Table 3 presents the BLEU by using both zero-shot and few-shot prompting settings with different demonstration selection strategies on the OPUS test set. The template of prompting is designed as Moslem et al. (2023). This analysis enables us to examine how the translation performance varies as we move from highly similar demonstrations to less similar ones, shedding light on the influence of noise in demonstrations on the translation quality of LLMs.

Context	en-fr	en-es
0-shot	20.92	21.32
1-shot		
Random	14.84	15.81
BM25	21.42	22.80
3-shot		
Random	16.78	18.88
BM25	22.58	24.66
5-shot		
Random	17.69	20.23
BM25	23.15	25.58

Table 3: BLEU scores for zero-shot and few-shot prompting on OPUS test set.

From Table 1, we can find that the similarity of most demonstrations is less than 0.7. It is reasonable because the retrieved demonstrations do not always contain ground-truth tokens. From Table 2, we observe that when the similarity is higher than 0.9, the translation quality of LLMs becomes very good. However, the quality significantly degrades when the similarity is less than 0.75. Comparing Table 2 and 3, we can conclude that noisy perturbations (unrelated tokens) in demonstrations significantly degrade the translation performance of LLMs, especially in cases of low-quality demonstrations. The impact of noisy demonstrations on LLMs is primarily manifested in two aspects. Firstly, noisy demonstrations introduce erroneous information, leading to incorrect translation outputs by LLMs. Secondly, noisy demonstrations interfere with the learning process of LLMs, making it difficult for them to capture the correct translation patterns accurately. The above experimental results indicate that the ICL for MT is sensitive to the quality of the demonstrations, which limits their applicability to real-world demonstrations. Therefore,

[1] https://opus.nlpl.eu/opus-100.php

[2] https://github.com/UKPLab/sentence-transformers

[3] The similarity score is 1.0 means the source sentence of the demonstration is identical to the testing sentence, and 0.0 means any words of two sentences are semantically unrelated.

it is of great significance to explore robust ICL of LLMs for MT.

4. Proposed Method

This section starts with the task formulation of prompting LLMs for MT with ICL and the multi-view integration to select demonstrations and cope with the noise, thus improving its robustness.

4.1. Task Formulation

The objective of generating translations with LLMs requires conditioning the decoder-only language model with in-context parallel demonstrations. The demonstrations provide valuable context and translation information about the source sentence, helping the model generate more accurate and contextually appropriate translations. Formally, for bilingual parallel pair (x, y), given k in-context examples $D = \{(\hat{\mathrm{x}}_1, \hat{\mathrm{y}}_1), (\hat{\mathrm{x}}_2, \hat{\mathrm{y}}_2), ..., (\hat{\mathrm{x}}_k, \hat{\mathrm{y}}_k)\}$, the prefix input of LLMs $\mathrm{x}_m = \sum_{i=1}^{k}(\hat{\mathrm{x}}_i + \hat{\mathrm{y}}_i) + \mathrm{x}$ is generated by concatenating the demonstration examples to the x, the $+$ is concatenation. The output is then generated via the LLMs with parameters θ via greedy decoding as follows:

$$\mathcal{L} = -\sum_{t=1}^{\mathrm{T}} \log \mathrm{p}(\mathrm{y}_t|\mathrm{y}_{<t}, \mathrm{x}_m; \theta) \tag{1}$$

where T is the number of tokens of target translation y.

This approach differs from standard sequence-to-sequence models, the conditioned input contains the demonstration D in addition to the source x. Previous works (Agrawal et al., 2023; Ghazvininejad et al., 2023; Zhang et al., 2023a) show good in-context examples can trigger LLMs to generate desired outputs and also elicit the information learned during the pre-training, suggesting that the in-context examples provide information about the MT task. This means those words that are related to x in demonstrations can assist the translation process. However, it is important to note that irrelevant bilingual words, which can be considered as noise, have the potential to mislead the translation generation (as discussed in Section 5.8). Most existing methods, such as (Agrawal et al., 2023; Vilar et al., 2023), directly treat the source sentence x_m as the input for LLMs. However, effectively dealing with the noise present in the demonstrations remains a significant challenge.

4.2. Prompt Selection

In Section 4.1, we have demonstrated that high-quality in-context examples can significantly enhance the translation ability of LLMs. Furthermore, Zhang et al. (2023a) emphasize the importance of maintaining the coherency of genuine source-target mapping in the demonstrations at sentence level. To avoid inconsistencies in the scale of cosine similarity across different sentence pairs, we propose a margin-based similarity method for selecting demonstrations using sentence embeddings. This method calculates the margin between the similarity of a candidate demonstration and the average similarity of its nearest neighbors (Zhu and Xiong, 2023; Zhu et al., 2024). The margin is defined as the ratio between the candidate similarity and the average cosine similarity of its nearest neighbors as follows:

$$\mathrm{Score}_l(\mathbf{x}, \hat{\mathbf{x}}) = \mathrm{margin}(\cos(\mathbf{x}, \hat{\mathbf{x}}), \mathrm{sNN}(\mathbf{x}, \hat{\mathbf{x}})) \tag{2}$$

$$\mathrm{sNN}(\mathbf{x}, \hat{\mathbf{x}}) = \sum_{\mathbf{z} \in nn_k(\mathbf{x})} \frac{\cos(\mathbf{x}, \mathbf{z})}{2k} + \frac{\cos(\hat{\mathbf{x}}, \mathbf{z})}{2k} \tag{3}$$

We utilize sentence-transformers toolkit to convert x and $\hat{\mathrm{x}} \in D$ into vector representations, denoted as $\mathbf{x}$ and $\hat{\mathbf{x}}$ respectively. The set $nn_k(\mathbf{x})$ represents the k nearest neighbors of $\mathbf{x}$, excluding any duplicates. $\mathrm{sNN}(\mathbf{x}, \hat{\mathbf{x}})$ is motivated to mitigate the hubness problem on Bilingual Lexicon Induction (BLI) over cross-lingual word embeddings. This approach aims to mitigate the issue of potentially different scales among candidate sentences, which can affect their relative ranking and exacerbate the hubness problem. Our proposed scoring method penalizes candidate sentences with high overall cosine similarities. We consider three different variants for the margin$(,)$ function, which computes the similarity score between $\mathbf{x}$ and $\hat{\mathbf{x}}$:

Absolute ($\mathrm{margin}(a, b) = a$): This variant ignores the average cosine similarity and is equivalent to the standard cosine similarity, serving as one of our baselines.

Relative ($\mathrm{margin}(a, b) = a - b$): This variant subtracts the average cosine similarity from the similarity of the given candidate, capturing the difference between them.

Ratio ($\mathrm{margin}(a, b) = \frac{a}{b}$): This variant calculates the ratio between the similarity of the candidate and the average cosine similarity of its nearest neighbors.

Previous studies (Agrawal et al., 2023; Moslem et al., 2023) have shown that word overlapping between the source and retrieved sentences can effectively improve the translation quality of LLMs. Therefore, we also consider the consistency of word level information while ensuring consistent sentence level similarity. For selected demonstrations $(\hat{\mathrm{x}}, \hat{\mathrm{y}})$, we evaluate the consistency of word level by calculating the cosine similarity between the words from (x̂, ŷ) and x as follows:

Algorithm 1 Multi-view prompt selection

```
Given source sentence x
Using Eq.(2) to get a demonstration $(\hat{x}, \hat{y})$ by retrieving bilingual corpus
Using Eq.(4) and (5) to get related word pairs $(w_m^{\hat{x}}, w_j^{\hat{y}})$
Using $(\hat{x}, \hat{y})$, $(w_m^{\hat{x}}, w_j^{\hat{y}})$ and x to construct template as Eq.(6)
Using the template as input to implement translation with LLMs
```

$$\text{Score}_w^{st}(\mathbf{w}_i^{\mathbf{x}}.\mathbf{w}_j^{\hat{\mathbf{y}}}) = \cos(\mathbf{w}_i^{\mathbf{x}}.\mathbf{w}_j^{\hat{\mathbf{y}}}) \tag{4}$$

$$\text{Score}_w^{ts}(\mathbf{w}_j^{\hat{\mathbf{y}}}.\mathbf{w}_m^{\hat{\mathbf{x}}}) = \cos(\mathbf{w}_j^{\hat{\mathbf{y}}}.\mathbf{w}_m^{\hat{\mathbf{x}}}) \tag{5}$$

where $\mathbf{w}_i^{\mathbf{x}} \in$ x, $\mathbf{w}_m^{\hat{\mathbf{x}}} \in \hat{x}$ and $\mathbf{w}_j^{\hat{\mathbf{y}}} \in \hat{y}$. If the $\text{Score}_w^{st}(.) > \alpha$ and $\text{Score}_w^{ts}(.) > \beta$[4], we add the word pair $(w_m^{\hat{x}} : w_j^{\hat{y}})$ into the prompt template as follows:

$$\text{[psrc]} : \hat{x}_1 \quad \text{[ptgt]} : \hat{y}_1 \quad \text{[psrc]} : w_m^{\hat{x}} \quad \text{[ptgt]} : w_j^{\hat{y}} \quad \ldots \quad \text{[psrc]} : \text{x} \quad \text{[ptgt]} : \tag{6}$$

where [psrc] and [ptgt] denote prompt language(s), i.e., the source and target language name of the prompt example, respectively. The main idea behind our method is to ensure consistency at both the sentence and word levels. To achieve this, we introduce a weighting mechanism that adds related word pairs to the prompt template. It can assign higher weights to related word pairs in the demonstrations and reduces the weights of unrelated word pairs, while preserving the overall semantic meaning of the sentence. This approach allows the model to pay less attention to unrelated word pairs, effectively mitigating the problem of noisy perturbations in demonstrations. The detailed process of our method is outlined in **Algorithm** 1. This algorithm provides a step-by-step guide on how to implement our approach and ensure robust translation with improved consistency at both the sentence and word levels.

5. Experiments

In this section, we conducted extensive experiments with multiple language pairs to examine the effectiveness of the proposed method.

5.1. Dataset

We used OPUS-100[5] as the datastore to retrieve the in-context demonstrations on bidirectional three language pairs. To assess the effectiveness of our proposed method on LLMs, we employed Flores-200 and OPUS testing set as the test set. Following (Agrawal et al., 2023), we normalized punctuation using Moses[6] and removed sentences longer than 250 tokens and sentence pairs with a source/target length ratio exceeding 1.5. For evaluation across different domains, we used the multi-domain dataset from Aharoni and Goldberg (2020), covering domains: Medical, Law, IT, and Koran in German-English.

5.2. Setting and Baselines

We adopted BLOOMZ-7b1-mt (Yong et al., 2023) as our foundational model which involves 46 natural languages and 13 programming languages. This model consists of 30 transformer-decoder layers, featuring an embedding size of 4096, a hidden size of 16384, and 32 attention heads. All experiments were performed on 2 NVIDIA A100 GPUs. We measured translation quality with case-sensitive detokenized BLEU by SacreBLEU (Post, 2018).

Baselines We compared our method against three baselines:

- Random: Selecting several sentence pairs from the corpus as demonstrations randomly.
- BM25: It is a bag-of-words unsupervised retrieval function that ranks a set of documents based on the query terms appearing in the documents.
- Fuzzy (Moslem et al., 2023): They use sentence embedding similarity-based retrieval to select demonstration. We use sentence-transformers to reimplement this method against our method on our used public test sets.

5.3. Main Results

Table 4 presents the results of our experiments. In the comparison between the "Random" baseline and the 0-shot setting, we observe a clear improvement in the translation ability of the LLM when the number of demonstrations is increased for the "Random" setting. However, when there is only one demonstration and its words are significantly different from the testing sentences, it confuses the LLM and hinders its translation performance.

[4] We set $\alpha = 0.8$ and $\beta = 0.8$. we also analyse the effect of α and β in experimental section.

[5] https://opus.nlpl.eu/opus-100.php

[6] http://www2.statmt.org/moses/

Methods	en-fr		fr-en		en-es		es-en		en-pt		pt-en	
	O	F	O	F	O	F	O	F	O	F	O	F
0-shot	20.9	32.9	17.0	34.7	21.3	21.7	24.2	36.9	13.4	27.0	19.7	38.6
1-shot												
Random	14.8	25.0	15.7	31.6	15.8	15.0	16.6	20.1	12.2	25.7	16.2	24.1
BM25	21.4	32.5	21.3	38.4	22.8	20.0	21.4	24.4	17.2	31.3	18.3	31.4
Fuzzy	21.6	38.7	21.8	43.2	23.0	21.8	22.3	28.1	17.6	36.1	21.6	35.2
Absolute	22.4	40.7	22.3	45.2	23.5	20.7	26.4	31.0	19.2	37.4	23.9	41.5
Relative	22.8	41.1	21.9	**45.8**	**23.7**	21.1	**27.1**	31.3	19.9	37.9	**24.5**	41.8
Ratio	**23.2**	**41.7**	**22.5**	45.6	23.1	**21.5**	26.8	**31.8**	**20.1**	**38.1**	24.2	**42.1**
3-shot												
Random	16.8	35.4	15.2	33.0	18.9	21.2	18.6	23.0	16.6	35.4	17.2	27.8
BM25	22.6	41.7	22.4	38.2	24.7	23.8	23.9	25.4	19.5	38.7	21.3	33.7
Fuzzy	22.3	43.7	23.3	42.6	23.7	24.8	22.9	29.0	19.4	39.5	21.3	39.5
Absolute	23.3	44.3	25.7	49.6	27.0	25.7	30.2	39.3	22.1	45.1	26.2	50.9
Relative	23.9	**45.3**	25.9	**51.1**	**27.6**	26.2	30.9	39.8	23.1	45.6	26.5	**51.8**
Ratio	**24.1**	45.2	**26.0**	50.8	27.3	**26.5**	**31.1**	**40.2**	**23.5**	**46.1**	**26.8**	51.3
5-shot												
Random	17.7	39.4	16.4	34.9	20.2	23.4	19.1	25.0	17.2	39.6	17.9	27.8
BM25	23.2	42.5	21.8	37.8	25.6	24.2	22.7	27.0	20.3	40.0	21.8	34.7
Fuzzy	22.3	43.4	22.8	42.9	22.5	25.2	24.1	28.9	18.7	40.4	22.5	39.4
Absolute	22.9	48.1	27.1	52.4	27.7	27.8	31.0	40.7	22.1	46.7	27.6	52.8
Relative	**23.3**	48.9	**27.7**	**53.0**	**28.3**	**28.6**	**31.9**	41.3	22.5	**47.4**	28.1	53.3
Ratio	23.2	**49.2**	27.5	52.8	28.1	28.4	31.5	**41.4**	**22.6**	47.3	**28.5**	**53.6**

Table 4: BLEU scores for zero-shot and few-shot prompting on different language pairs with different demonstration selection strategies. "O" denotes the OPUS test set. "F" is the Flores-200 test set.

Comparing the "BM25" and "Fuzzy", which use different similarity computation approaches to select semantically similar demonstrations, we find that both methods outperform the 0-shot and "Random" baselines. This indicates that selecting demonstrations that are semantically similar to the testing sentences can enhance the translation ability of the LLM. From Table 4, an interesting observation is that adding more demonstrations for the "BM25" and "Fuzzy" methods does not consistently improve the translation ability across most language pairs. In some cases, there is even a decline in performance (e.g., es-en). It can be attributed to the fact that the retrieved demonstrations may not always contain the ground-truth words. The presence of unrelated words in the demonstrations negatively impacts the LLM's translation performance.

Regarding the variants of our method, the "Absolute" variant is similar to the "Fuzzy" method in terms of computing the similarity score at the sentence level (both using cosine similarity). The difference lies in the inclusion of weights for related words in the demonstrations. This demonstrates that incorporating the weight of related words can mitigate the impact of unrelated words. The "Relative" and "Ratio" variants show that the proposed sentence level similarity computation can further improve the translation ability of LLMs. On all translation tasks, our proposed method significantly surpasses all baselines on both the OPUS and Flores-200 test sets.

5.4. Robustness of Proposed Method

To verify the robustness of our model, we investigated the performance of LLMs with retrieved demonstrations of varying qualities. We partitioned the test set into subsets based on the similarity scores between the demonstrations and the testing sentences, as described in Section 3. Firstly, we present the performance of our model and the baselines in Figure 1 for different few-shot scenarios. Overall, our model performs well in all situations and even surpasses all baselines. Notably, as we increase the number of demonstrations from 1-shot to 5-shot, all three variants consistently achieve more stable and improved results. This observation suggests that the performance of our method becomes more reliable and robust with the inclusion of additional demonstrations. Secondly, we analyze the performance of our model and the baselines across different similarity intervals in Figure 1. Intuitively, demonstrations with lower similarity contain more unrelated words and can potentially confuse the LLMs. Compared to baselines, our model exhibits significantly less performance decline. Particularly, when the

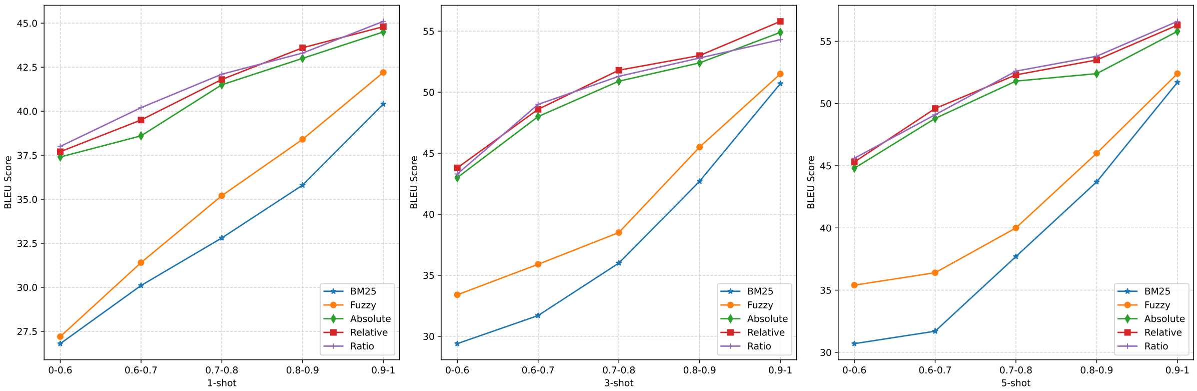

Figure 1: The robustness of different methods on pt-en language pair of Flores-200 test set.

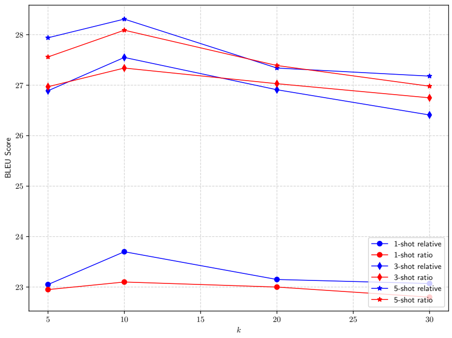

Figure 2: BLEU scores of our method on the OPUS test set for English-Spanish (en-es) with different k.

similarity drops below 0.6, our model consistently outperforms the baselines by a large margin (over 30%). These results demonstrate that our method enhances the translation robustness of LLMs.

5.5. Analysis on the Effect of Domain Adaptation

To investigate the effect of our method when dealing with texts from different domains, we conducted experiments on the multi-domain dataset. The experimental results are presented in Table 5. Our method has a more substantial impact on the model's translation capabilities compared to baselines. In some instances, our method even leads to improvements exceeding 10 BLEU points. These findings highlight the effectiveness of our proposed method in enhancing the translation performance of LLMs, particularly when dealing with texts from different domains. The ability to achieve significant improvements in translation quality demonstrates the potential of our method in real-world applications where domain adaptation is crucial.

Methods	IT	Medical	Koran	Law
0-shot	14.1	12.8	3.8	14.4
1-shot				
Random	10.2	10.7	2.9	7.9
BM25	14.7	19.6	7.4	17.2
Fuzzy	16.0	20.2	6.7	16.3
Absolute	21.2	25.2	11.2	24.2
Relative	21.2	25.2	11.2	24.2
Ratio	21.2	25.2	11.2	24.2
3-shot				
Random	10.2	12.3	2.7	8.4
BM25	16.4	22.2	7.6	18.5
Fuzzy	16.9	22.9	7.5	16.6
Absolute	20.8	28.9	11.3	26.3
Relative	20.8	28.8	11.2	26.2
Ratio	20.9	28.9	11.2	26.0
5-shot				
Random	12.8	13.0	3.0	8.9
BM25	16.1	22.1	6.5	17.0
Fuzzy	16.5	23.8	7.3	14.8
Absolute	21.4	29.4	11.3	25.9
Relative	21.2	29.4	11.4	25.8
Ratio	21.2	29.2	11.4	25.9

Table 5: BLEU scores on the test sets of different domains.

5.6. Analysis on the Effect of k

In our framework, the value of k in Eq.(3), which represents the number of nearest neighbors to select for prompt examples, plays a crucial role. In this subsection, we will explore the effects of different values of k on the model's performance. Regarding the variants of our method, the "Absolute" variant is independent of k, our focus is solely on examining the impact of k on the "Relative" and "Ratio" variants. As depicted in Figure 2, we observe that the model achieves optimal performance when k is set to 10. When k is too small

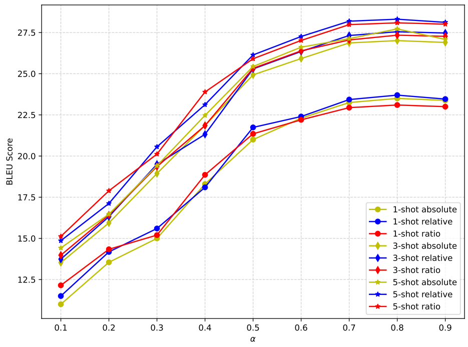

Figure 3: BLEU scores of our method on the OPUS test set for en-es with different α.

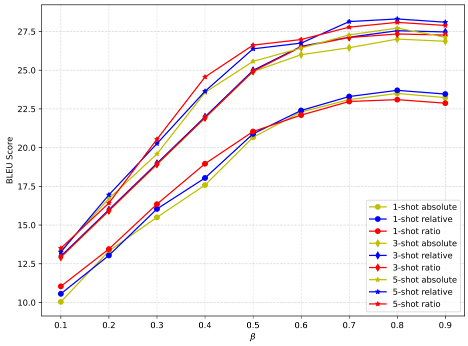

Figure 4: BLEU scores of our method on the OPUS test set for en-es with different β.

(e.g., 5), the prompts lack diversity, leading to limited coverage of different translation patterns. On the other hand, when k is too large (e.g., 30), the prompts introduce excessive semantic bias, potentially leading to overfitting and reduced generalization ability. Considering both model performance and computational efficiency, we have carefully chosen to set k to 10. This value strikes a balance between capturing diverse translation patterns and avoiding excessive semantic bias, resulting in improved translation quality.

5.7. Analysis on the Effect of α and β

We conducted a grid search over the values of α and β, ranging from 0.1 to 0.9, to determine their optimal settings. The experimental results are illustrated in Figure 3 and Figure 4. It can be observed that as the values of α and β increase, the model's performance also improves. We found that excessively large values of α and β tend to result in the selection of words with repetitive semantic meanings, leading to a lack of semantic diversity. Conversely, when α and β are too small, the chosen words do not adequately capture the semantic information present in the source sentence, compromising the model's translation capability. Based on our analysis, we set α and β to 0.8.

Source	Brazil: Was the shooting of Ricardo Gama politically motivated?
Reference	Brasil: ¿ Fue el tiroteo de Ricardo Gama por motivos políticos?
Random	巴西: ¿El asesinato de Ricardo Gama fue motivado políticamente?
BM25	巴西: ¿Fue el asesinato de Ricardo Gama un acto de motivación política?
Fuzzy	巴西: ¿Motivó la muerte de Ricardo Gama la política?
Absolute	Brasil: ¿Fue el asesinato de Ricardo Gama un acto de motivación política?
Relative	Brasil: ¿Fue el asesinato de Ricardo Gama un acto político?
Ratio	Brasil: ¿Fue el asesinato de Ricardo Gama un acto político?

Table 6: Case Study

5.8. Case Study

To examine the translation effectiveness of various methods on the translation task, we conducted a case study. As shown in Table 6, we observe that random-based methods, "BM25" and "Fuzzy" are all prone to producing translations in other languages, which is not observed with our approach. We attribute this phenomenon to the similarity of word embeddings for words with the same semantics in different languages within large models. While the model may have learned the semantic information of sentences, it has not acquired the correspondence between the translation languages in the prompt. In contrast, our method introduces word pairs with the same semantics into the prompt, enabling the model to learn the language correspondence within the prompt's instruction. As a result, our approach effectively captures the topic and semantic information of the target sentence, leading to improved translation results.

6. Conclusion

In this paper, we have presented a robust method for improving the translation capabilities of LLMs using ICL. We address the issue of vulnerability to noise in demonstrations, which can significantly deteriorate the performance of LLMs. To overcome this, we propose a multi-view approach that considers both sentence- and word-level information. At the sentence level, we introduce a margin-based similarity that takes into account the scale inconsistencies of cosine similarity. At the word level, we introduce a weighting method to avoid the the issue of noise in demonstrations. Therefore, our method provides fine-grained demonstrations that effectively prompt the control of LLMs. Experimental results on various benchmarks demonstrate the effectiveness of our method, particularly in domain adaptation, where our model significantly outperforms the baselines.

Acknowledgments

The present research was supported by the National Natural Science Foundation of China Youth Foud (Grant No. 62306210) and Natural Science Foundation of Xinjiang Uygur Autonomous Region (Grant No. 2022D01D43). We would like to thank the anonymous reviewers for their insightful comments.

7. Bibliographical 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.

Roee Aharoni and Yoav Goldberg. 2020. Unsupervised domain clusters in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7747–7763. Association for Computational Linguistics.

Yong Cheng, Lu Jiang, Wolfgang Macherey, and Jacob Eisenstein. 2020. Advaug: Robust adversarial augmentation for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5961–5970. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416.

Hiroyuki Deguchi, Taro Watanabe, Yusuke Matsui, Masao Utiyama, Hideki Tanaka, and Eiichiro Sumita. 2023. Subset retrieval nearest neighbor machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 174–189. Association for Computational Linguistics.

Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André F. T. Martins, Graham Neubig, Ankush Garg, Jonathan H. Clark, Markus Freitag, and Orhan Firat. 2023. The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. *CoRR*, abs/2308.07286.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. 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 3816–3830. Association for Computational Linguistics.

Xavier Garcia, Yamini Bansal, Colin Cherry, George F. Foster, Maxim Krikun, Melvin Johnson, and Orhan Firat. 2023. The unreasonable effectiveness of few-shot learning for machine translation. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 10867–10878. PMLR.

Xavier Garcia and Orhan Firat. 2022. Using natural language prompts for machine translation. *CoRR*, abs/2202.11822.

Marjan Ghazvininejad, Hila Gonen, and Luke Zettlemoyer. 2023. Dictionary-based phrase-level prompting of large language models for machine translation. *CoRR*, abs/2302.07856.

Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Supryadi, Linhao Yu, Yan Liu, Jiaxuan Li, Bojian Xiong, and Deyi Xiong. 2023. Evaluating large language models: A comprehensive survey. *CoRR*, abs/2310.19736.

Lifeng Han, Gleb Erofeev, Irina Sorokina, Serge Gladkoff, and Goran Nenadic. 2022. Examining large pre-trained language models for machine translation: What you don't know about it. In *Proceedings of the Seventh Conference on Machine Translation, WMT 2022, Abu Dhabi, United Arab Emirates (Hybrid), December 7-8, 2022*, pages 908–919. Association for Computational Linguistics.

Vivek Iyer, Pinzhen Chen, and Alexandra Birch. 2023. Towards effective disambiguation for machine translation with large language models. *CoRR*, abs/2309.11668.

Hui Jiang, Ziyao Lu, Fandong Meng, Chulun Zhou, Jie Zhou, Degen Huang, and Jinsong Su. 2022. Towards robust k-nearest-neighbor machine 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 5468–5477. Association for Computational Linguistics.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. 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 3045–3059. Association for Computational Linguistics.

Shangjie Li, Xiangpeng Wei, Shaolin Zhu, Jun Xie, Baosong Yang, and Deyi Xiong. 2023. MMNMT: modularizing multilingual neural machine translation with flexibly assembled moe and dense blocks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 4978–4990. Association for Computational Linguistics.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*.

Chenyang Lyu, Jitao Xu, and Longyue Wang. 2023. New trends in machine translation using large language models: Case examples with chatgpt. *CoRR*, abs/2305.01181.

Zhongjian Miao, Xiang Li, Liyan Kang, Wen Zhang, Chulun Zhou, Yidong Chen, Bin Wang, Min Zhang, and Jinsong Su. 2022. Towards robust neural machine translation with iterative scheduled data-switch training. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 5266–5277. International Committee on Computational Linguistics.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? 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 11048–11064. Association for Computational Linguistics.

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.

Leiyu Pan, Supryadi, and Deyi Xiong. 2023. Is robustness transferable across languages in multilingual neural machine translation? In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 14114–14125. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Ratish Puduppully, Raj Dabre, Ai Ti Aw, and Nancy F. Chen. 2023. Decomposed prompting for machine translation between related languages using large language models. *CoRR*, abs/2305.13085.

Wenjie Qin, Xiang Li, Yuhui Sun, Deyi Xiong, Jianwei Cui, and Bin Wang. 2021. Modeling homophone noise for robust neural machine translation. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021*, pages 7533–7537. IEEE.

Eduardo Sánchez, Pierre Andrews, Pontus Stenetorp, Mikel Artetxe, and Marta R. Costa-jussà. 2023. Gender-specific machine translation with large language models. *CoRR*, abs/2309.03175.

Suzanna Sia and Kevin Duh. 2023. In-context learning as maintaining coherency: A study of on-the-fly machine translation using large language models. *CoRR*, abs/2305.03573.

Zhixing Tan, Xiangwen Zhang, Shuo Wang, and Yang Liu. 2022. MSP: multi-stage prompting for making pre-trained language models better translators. 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 6131–6142. Association for Computational Linguistics.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George F. Foster. 2023. Prompting palm for translation: Assessing strategies and performance. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 15406–15427. Association for Computational Linguistics.

Dexin Wang, Kai Fan, Boxing Chen, and Deyi Xiong. 2022. Efficient cluster-based k-nearest-neighbor machine translation. 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 2175–2187. Association for Computational Linguistics.

Shuo Wang, Zhaopeng Tu, Zhixing Tan, Wenxuan Wang, Maosong Sun, and Yang Liu. 2021. Language models are good translators. *CoRR*, abs/2106.13627.

Jitao Xu, Josep Maria Crego, and Jean Senellart. 2020. Boosting neural machine translation with similar translations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1580–1590. Association for Computational Linguistics.

Zheng Xin Yong, Hailey Schoelkopf, Niklas Muennighoff, Alham Fikri Aji, David Ifeoluwa Adelani, Khalid Almubarak, M. Saiful Bari, Lintang Sutawika, Jungo Kasai, Ahmed Baruwa, Genta Indra Winata, Stella Biderman, Edward Raff, Dragomir Radev, and Vassilina Nikoulina. 2023. BLOOM+1: adding language support to BLOOM for zero-shot prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 11682–11703. Association for Computational Linguistics.

Zhiyuan Zeng and Deyi Xiong. 2021. An empirical study on adversarial attack on NMT: languages and positions matter. 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 2: Short Papers), Virtual Event, August 1-6, 2021*, pages 454–460. 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.

Huaao Zhang, Qiang Wang, Bo Qin, Zelin Shi, Haibo Wang, and Ming Chen. 2023b. Understanding and improving the robustness of terminology constraints in neural machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 6029–6042. Association for Computational Linguistics.

Xin Zheng, Zhirui Zhang, Junliang Guo, Shujian Huang, Boxing Chen, Weihua Luo, and Jiajun Chen. 2021. Adaptive nearest neighbor 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 2: Short Papers), Virtual Event, August 1-6, 2021*, pages 368–374. Association for Computational Linguistics.

Shaolin Zhu, Shiwei Gu, Shangjie Li, Lin Xu, and Deyi Xiong. 2024. Mining parallel sentences from internet with multi-view knowledge distillation for low-resource language pairs. *Knowl. Inf. Syst.*, 66(1):187–209.

Shaolin Zhu and Deyi Xiong. 2023. TJUNLP: system description for the WMT23 literary task in chinese to english translation direction. In *Proceedings of the Eighth Conference on Machine Translation, WMT 2023, Singapore, December 6-7, 2023*, pages 307–311. Association for Computational Linguistics.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Lingpeng Kong, Jiajun Chen, Lei Li, and Shujian Huang. 2023a. Multilingual machine translation with large language models: Empirical results and analysis. *CoRR*, abs/2304.04675.

Wenhao Zhu, Jingjing Xu, Shujian Huang, Lingpeng Kong, and Jiajun Chen. 2023b. INK: injecting knn knowledge in nearest neighbor machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 15948–15959. Association for Computational Linguistics.