Can Large Language Models Learn Translation Robustness from Noisy-Source In-context Demonstrations?

Leiyu Pan[†], **Yongqi Leng**[†], **Deyi Xiong**[*]
College of Intelligence and Computing, Tianjin University, Tianjin, China
{lypan, lengyq, dyxiong}@tju.edu.cn

Abstract

Large language models (LLMs) have been used for machine translation. When provided with prompts and source sentences, LLMs can achieve impressive translation results. However, the robustness of these LLMs remains a significant challenge, as they often struggle to accurately translate sentences in the presence of noise, even when using similarity-based in-context learning methods. This work proposes a research scheme for studying machine translation robustness on LLMs, investigating whether LLMs can learn translation robustness from noisy-source demonstration examples. Through experiments on different models, languages, and noise types, we empirically demonstrate that LLMs can learn how to handle noise and translation methods from noisy-source demonstration examples, thereby improving their translation performance on noisy sentences. Furthermore, we find that increasing the noise ratio appropriately for the noisy-source demonstration examples can enhance the translation robustness of LLMs. Additionally, we also attempt to investigate scenarios where LLMs are more likely to learn translation robustness for mixed and specific types of noise. We find that the model's performance varies across different noise settings.

Keywords: Large Language Model, Machine Translation, Robustness, In-context Learning

1. Introduction

The emergence of LLMs has posed a great impact on the field of natural language processing. They are not only capable of capturing knowledge of general domains, but also highly adaptable to the knowledge of various specialized domain. This multi-domain knowledge coverage makes LLMs a powerful tool for cross-domain natural language processing tasks, which can be applied to a variety of domains such as medicine (Thirunavukarasu et al., 2023), law (Cui et al., 2023), and finance (Wu et al., 2023). In addition, it can achieve superior capabilities on a wide range of tasks (Wei et al., 2022), achieving comparable results to task-specific SOTA models on tasks such as sentiment analysis (Wang et al., 2023b) and natural language inference (Qin et al., 2023). It has also shown excellent performance in the field of machine translation (Wang et al., 2023a).

Since fine-tuning LLMs is usually costly, it has become mainstream to design specific prompting strategies to enable LLMs to accomplish specific tasks. One effective prompting approach is to combine task descriptions and demonstration examples into prompts, i.e., in-context learning (Dong et al., 2022). It has been suggested that it can enhance the performance of a model by learning the mapping relationship between inputs and outputs (Pan, 2023). Similarly, employing in-context learning for translation can also enhance the translation performance of LLMs (Agrawal et al., 2022), selecting demonstration examples that are more similar to the test samples can bring more enhancement to the LLMs (Moslem et al., 2023).

In more complex robust machine translation scenarios, our preliminary experiments with LLMs reveal the same phenomenon with Moslem et al. (2023), i.e., selecting demonstration examples based on similarity leads to a greater improvement in the performance of the LLMs than choosing demonstration examples randomly. Nevertheless, it is important to note that our investigations also unveil the LLMs' limitations in effectively addressing noise within the target sentences, even when utilizing the most closely related sentence pair as a demonstration example.

In this work, we use a in-context learning approach to translate noisy sentences and attempt to explore whether LLMs can learn how to handle noise in test samples by incorporating noisy-source demonstration examples, thereby enhancing the robustness of LLMs for machine translation. To accomplish this goal, we propose a research scheme for the robustness of machine translation in the context of LLM. Our approach commences with the preparation of both synthetic and natural noise data. Subsequently, under the setting of considering the similarity between the sentence to be translated and the demonstration examples, we employ three distinct sampling methods to sample demonstration examples. These methods encompass sampling from clean data, sampling from a combination of different types of noisy data, and sampling exclusively from a single type of noisy data. Finally, we eval-

[†] Equal contribution.
[*] Corresponding author.

LREC-COLING 2024, pages 2798–2808
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

uate and analyze the translations generated by the model to determine whether it demonstrates an increased level of robustness in noisy contexts. In essence, a sufficient condition for enhancing a model's translation robustness entails maintaining or improving its performance on clean datasets while concurrently bolstering its performance on noisy datasets. Therefore, our primary focus lies in evaluating the model's accuracy on both clean and noisy datasets. We open source our code and data at https://github.com/tjunlp-lab/llm_translate_robust.

In summary, our contributions are as follows:

- We propose a research scheme for investigating the robustness of LLMs on machine translation.
- We construct a Chinese-English natural noise translation dataset based on the Multilingual Microblog Translation Corpus (McNamee and Duh, 2022), realizing fine-grained natural noise classification.
- Using the research scheme proposed in this paper, we find 1) LLMs are able to learn translation robustness from noisy-source demonstration examples with synthetic noise. 2) LLMs are more likely to learn robustness to character-level noise through type-specific synthesized noise, but less likely for robustness to word-level noise through mixed-type synthesized noise. 3) For specific and mixed types of natural noise, LLMs perform inconsistently in the learning of robustness over high and low resource languages.

2. Related Work

In-context learning is a commonly used prompting technique. It composes a prompt into the model by combining demonstrations with test inputs. Compared to zero-shot learning, utilizing in-context learning can enhance the performance of LLM on a variety of tasks (Brown et al., 2020). There have been a number of works that have investigated the reasons for the effectiveness of in-context learning. One of the studies points out that in-context learning can actually be decoupled into two mechanisms, task recognition and task learning (Pan, 2023). There are also studies that relate in-context learning to gradient descent and understand in-context learning as implicit fine-tuning (Dai et al., 2022). When using in-context learning methods, the demonstration example selection and ordering affects the performance of in-context learning (Zhao et al., 2021). Additionally, demonstrations embedded closer to the test input usually lead to better performance than those embedded further away (Liu et al., 2021).

On specific downstream tasks, there are studies that apply LLMs to machine translation and show their excellent capabilities. Just using simple prompts can make it possible to achieve results comparable to commercial MT systems in high-resource languages (Jiao et al., 2023). If task information and domain information are introduced in the prompts, it can stimulate the ability of LLMs for machine translation even further (Peng et al., 2023). In addition, there has been work exploring the use of in-context learning approaches to apply LLMs for machine translation, and selecting in-context examples that have a higher degree of overlap with the content of the sentence to be translated can significantly improve the translation effect of LLMs (Agrawal et al., 2022). However, the LLMs do not perform well enough on complex machine translation tasks, and the results obtained by evaluating the LLMs using a noisy translation test set show that the translation robustness of the LLMs still need to be strengthened (Jiao et al., 2023).

Translation robustness has been a pressing issue in machine translation. Recent studies have shed light on the impact of adversarial attacks at the source side vs. the target side, prompting a quest for more efficient methods of attack (Zeng and Xiong, 2021). Addressing these challenges, there have been endeavors to fortify machine translation systems against black-box (Wallace et al., 2020) and white-box attacks (Cheng et al., 2019). In particular, research has delved into mitigating specific types of noise, such as homophonic errors (Qin et al., 2021). Furthermore, Pan et al. (2023) find that robustness can also be transferable across languages to improve translation robustness in multilingual scenarios. Such robustness studies on neural machine translation could provide guidance for analyzing and enhancing translation robustness for LLMs.

While in general in-context learning can improve the performance of LLMs, injecting noise into demonstration examples has not been explored. In addition, previous work has only evaluated LLMs on translation test sets with noise, and has not gone further to explore how to make LLMs more robust to translation. In this work, we try to reveal whether LLMs can learn translation robustness from noisy-source demonstration examples. This is another aspect to study in-context learning and provides inspiration to improve translation robustness for LLMs.

3. Approach

In order to investigate whether LLMs can learn translation robustness from noisy-source in-context demonstrations, we have proposed a

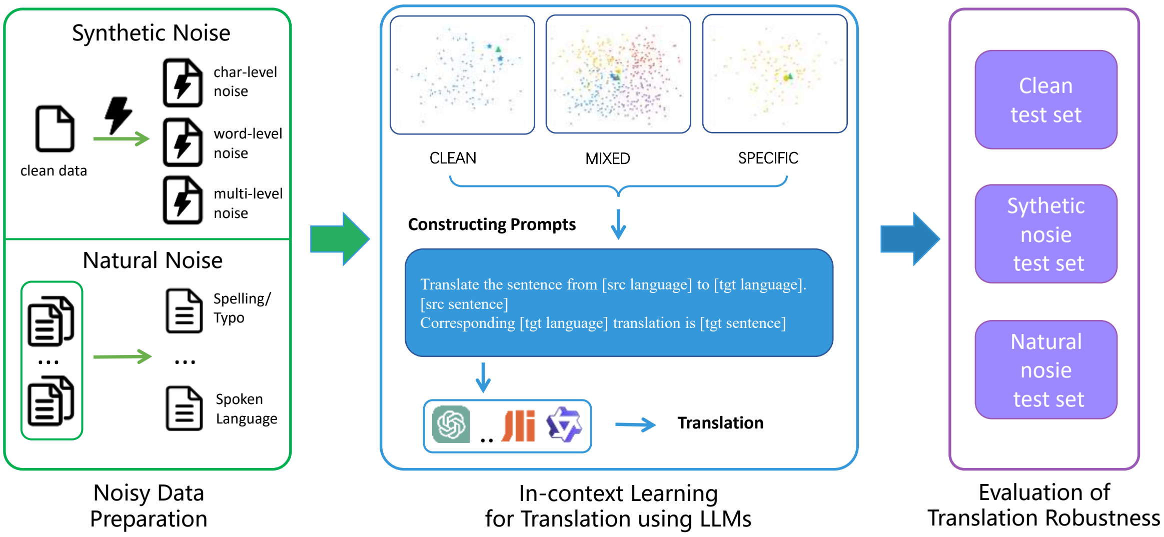

Figure 1: The proposed research scheme to explore whether LLMs can obtain translation robustness under noisy-source in-context demonstrations.

research scheme for robust machine translation in the context of LLMs. As shown in Figure 1, the scheme can primarily be divided into three stages: noise data preparation, in-context learning for translation using LLMs, and evaluation of translation robustness.

3.1. Noisy Data Preparation

In the noise data preparation stage, we prepare synthetic noise and natural noise data separately. This is due to the different data distributions of synthetic noise and natural noise, LLMs may produce different performances on data with different distributions of noise characteristics.

Synthetic Noise In terms of synthetic noise, we add three types of noise to the clean translation dataset. Specifically, it includes character-level, word-level, and multi-level noise. Three levels of noise can be added to the data by performing three levels of black-box attack methods respectively. Character-level black-box attack methods involve random character insertion, random character deletion, random character replacement, and random adjacent character swapping operations. Word-level black-box attack methods include insertion and replacement using words with similar word embeddings to the target word, random word deletion, and random word swapping operations. Character-level black-box attack methods and word-level black-box attack methods are combined to form a multi-level black-box attack methods. In addition, we only add noise to the part of the source language in the dataset, leaving the data in the target language unchanged.

Natural Noise In terms of natural noise, we collect user texts generated by social media as raw natural noise data. We refer to MTNT (Michel and Neubig, 2018) to classify natural noise into ten categories, namely spelling/typographical errors, grammar errors, spoken language, slang, proper nouns, dialects, code switching, jargon, emojis, and slurs. We then follow the steps of rule-based labeling, model-based labeling, and manual labeling to classify the natural noise categories of the data in order to reduce the classification bias.

3.2. In-context Learning for Translation using LLMs

After the synthetic noise data and natural noise data are prepared, we take an in-context learning approach to translate the noisy sentences using LLMs.

Formally, given k in-context examples $\{\boldsymbol{x_i}, \boldsymbol{y_i}\}_1^k$, each $\boldsymbol{x}$ and $\boldsymbol{y}$ is a pair of source and target sentences from the parallel corpus, the model input $\boldsymbol{x^P}$ can be constructed by concatenating the in-context examples to the test sentence to be translated. The model can be parameterized by $\boldsymbol{\theta}$. The model translation output $\hat{\boldsymbol{y}}$ can be generated as follows:

$$\hat{y}_t = \arg\max_{y'_t} P_{\text{LLM}}\left(y'_t \mid \boldsymbol{x^P}, \hat{\boldsymbol{y}}_{<t}; \boldsymbol{\theta}\right) \qquad (1)$$

In order to explore whether an LLM can learn to be robust against noise in in-context examples and then correctly translate noisy sentences, we take the following three approaches to sampling in-context examples:

- **CLEAN**: Sampling in-context examples from clean data. This sampling method is limited to synthetic noise. Because the raw data of synthetic noise is considered clean data, it can be randomly sampled as a sample set. However, the natural noise data collected are considered noisy and there is no clean data to sample corresponding to it.

- **MIXED**: Sampling in-context examples from mixed noise categories of data. In the case of synthetic noise, clean, character-level noise, word-level noise, and multi-level noise are combined as an in-context example sampling set for sampling. In the case of natural noise, the ten categories of noise data are combined as in-context example sampling sets for sampling.
- **SPECIFIC**: Sampling in-context examples from data with the same noise type as the sentence to be translated. We first determine the noise type of the sentence to be translated, and then select all sentences with the same type of noise as the sentence to be translated as the contextual example sampling set to be sampled.

All the above sampling methods select the most similar sentences to the test samples from the sampling set as demonstration examples. After completing the sampling of demonstration examples, we then construct the prompts for the demonstration examples according to the specified template, and add the noisy sentences to be translated into the prompts to form the final prompts. The constructed prompts are fed into the LLMs, and the model can return the translation results of the sentences to be translated.

3.3. Evaluation of Translation Robustness

In terms of synthetic noise, we mainly observe whether the performance of the LLMs on clean and noisy datasets improves at the same time, given the demonstrations with noise; in terms of natural noise, we only observe whether the performance of the LLMs on noisy datasets improves, since we only collect comment data with natural noise and lack clean test data in the corresponding domain.

4. Experiments

Using the proposed research scheme on robust machine translation in the background of LLMs, we conducted extensive experiments on different languages, different types of noise, and different models, to explore whether LLMs can learn translation robustness from noisy-source demonstrations.

4.1. Data

We prepared the synthetic and natural noise data according to the methodology of noise data preparation mentioned in the research scheme.

4.1.1. Synthetic Noise Data

We used the Chinese-English parallel corpus data from the publicly available WMT News Test Set dataset (Barrault et al., 2019) and the Indonesian-Chinese data from the TED TALKS 2020 dataset (Reimers and Gurevych, 2020), randomly selected a portion of these data as the raw data, and then added synthetic noise to them. For the Indonesian-Chinese translation, we used the nlpaug library (Ma, 2019) to implement character-level, word-level, and multi-level black-box attack operations to add noise. In terms of character-level black-box attacks, the proportion of attacked words to all words in each sentence is 0.3. Each attacked word has up to one character modified in it. The coverage of the four attack operations per sentence is 25%. For word-level black-box attacks, the proportion of attacked words to all words in each sentence is 0.3. Word insertion and substitution operations are language-dependent and require the use of language-specific pre-trained word embeddings to find words with similar semantics. We chose fastText pre-trained word embeddings[1] here and selected the most semantically similar words to replace. Similarly, the coverage of the four attack operations per sentence is 25%. In terms of multi-level black-box attacks, character-level and word-level attack operations are combined. The coverage of eight attack operations per sentence is 12.5%. For the Chinese-English language direction, the only difference is that in regards to the character-level black-box attack, the randomly replaced or inserted characters are those that are homophonic to the target character, which is more in line with realistic noise scenarios.

4.1.2. Natural Noise Data

For Chinese-English translation, we used the Chinese-English part of the MMTC dataset (McNamee and Duh, 2022). We performed secondary processing of the data, including data filtering, modification and labeling of noise categories. We started by filtering for repeated sentences. Also we filtered the data for discriminatory statements, as the LLMs may reject translations of these discriminatory statements. We also observed that there is a mismatch between username mentions and URLs in the parallel corpus. That is, when the username mention exists in the source sentence, the reference translation may or may not translate

[1] https://fasttext.cc/docs/en/crawl-vectors.html

the username mention, which has an impact on the quality of the data. Therefore, we uniformly remove username mentions and URLs from the sentences. Finally we follow the following three steps for labeling the natural noise categories.

1) **Rule-based labeling**: The emojis noise in sentences is labeled by using regular expressions. In particular, for Chinese, dialect noise in sentences can be labeled by using the opencc library[2], and code switching noise in sentences can be labeled by using regular expressions.

2) **Model-based labeling**: We use the GPT-3.5-turbo api[3] for labeling. The manually labeled examples are used as context samples, which are then entered into the model as prompts along with the samples to be labeled. Including context samples in the prompts effectively specifies the format of the model output noise categories and facilitates the extraction of noise categories from the model output.

3) **Manual labeling**: The above steps of automatic labeling may produce errors, we need to manually confirm and modify the labeling results to ensure the correctness of the labeling results.

After labeling, the raw natural noise data is divided into ten categories. Since a sentence may contain multiple natural noises, it is possible that the sentences in each category may overlap.

In Indonesian-Chinese translation, we used the publicly available Indonesian-Chinese noise translation dataset. Its data is also derived from social media, which is more consistent with the distribution of natural noise data in the Chinese-English language direction. In addition, it has the same noise classification as that set forth herein.

Having been labeled and organized, the statistical information of the natural noise dataset in the Chinese-English language direction and the Indonesian-Chinese language direction is shown in Figure 2. As can be seen from the figure, both language directions have more grammatical errors, spoken language, slang, and proper nouns natural noise errors, and fewer spelling/typo and slurs natural noise errors.

4.2. Model

We used two families of LLMs, Baichuan2 (Yang et al., 2023), Qwen (Bai et al., 2023). For Baichuan2, we used the chat version of 7B and 13B. For Qwen, we used the chat version of 7B and 14B.

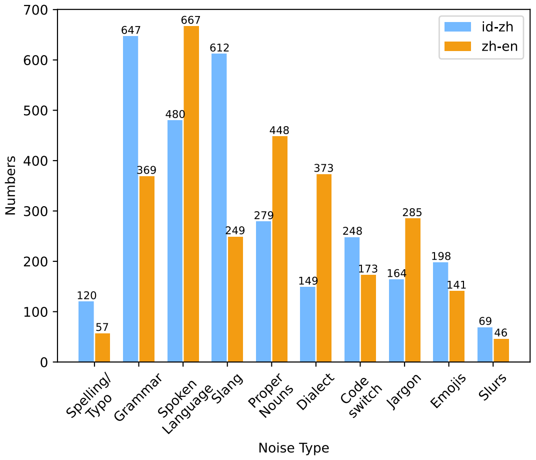

Figure 2: Statistical information on natural noise categories in the Chinese-English and Indonesian-Chinese translation.

4.3. Sampling Details

In selecting demonstration examples, we first determined the sampling set based on different sampling methods. Then we calculated the sentence embeddings of the sampling set and the sentence to be translated separately using the sentence-transformer (Reimers and Gurevych, 2020). After that, we computed the cosine similarity between the sentence embeddings to be translated and the sentence embeddings of the sampling set. Finally, a specified number of demonstration examples that are most similar to the sentence to be translated are selected.

4.4. Results

LLMs can learn translation robustness from demonstration examples with synthetic noise. From Table 1, it is evident that LLM translation results with MIXED sampling method yield slightly higher BLEU scores than the CLEAN sampling method on the clean test set. Since the CLEAN sampling setup samples clean demonstration examples, while in the MIXED sampling setup samples basically demonstration examples with noise, this suggests that introducing moderate noise to demonstration examples has no adverse impact on LLM translation of clean sentences. This may be due to the fact that noise in the demonstration example serves as augmented data, enhancing model performance. BLEU scores achieved with MIXED and SPECIFIC sampling methods are consistently higher than those obtained with CLEAN sampling method on character-level noise, word-level noise, and multi-level noise test sets. This implies that for translating noisy sentences, demonstration examples with noise are more valuable than clean ones. The model can learn from

[2] https://github.com/BYVoid/OpenCC
[3] https://openai.com/product

	shots	Baichuan2-7B-Chat				Baichuan2-13B-Chat			
		Clean	Character Noise	Word Noise	Multi Noise	Clean	Character Noise	Word Noise	Multi Noise
	0 shot	15.34	10.98	8.66	10.15	14.77	11.97	9.53	10.64
CLEAN	1 shot	24.16	18.19	15.42	17.28	26.32	20.99	16.76	19.19
	3 shot	23.75	17.60	15.70	16.90	**26.68**	21.59	17.82	19.72
	5 shot	24.32	18.77	15.99	17.85	25.86	20.64	18.10	19.39
MIXED	1 shot	24.04	18.94	15.55	17.76	26.39	21.04	17.60	19.67
	3 shot	24.42	17.87	15.47	17.34	26.32	21.68	18.02	20.12
	5 shot	**24.49**	19.54	**16.65**	17.00	26.45	21.98	**18.57**	20.44
SPECIFIC	1 shot	24.16	18.59	15.70	17.50	26.32	21.12	17.65	19.19
	3 shot	23.75	18.55	16.05	18.04	**26.68**	22.06	18.14	20.14
	5 shot	24.32	**20.03**	15.27	**19.06**	25.86	**22.35**	17.59	**21.20**

Table 1: Results of Baichuan2-7B-Chat model and Baichuan2-13B-Chat on Chinese-English dataset of sythetic noise. (underline: the maximum value of the data in this column for the current sampling method; **bold**: the maximum value of data in this column for all sampling methods).

	shots	Qwen-7B-Chat				Qwen-14B-Chat			
		Clean	Character Noise	Word Noise	Multi Noise	Clean	Character Noise	Word Noise	Multi Noise
	0 shot	21.41	7.79	12.52	11.48	24.25	13.51	14.51	14.34
CLEAN	1 shot	22.33	9.22	13.43	11.26	25.89	15.47	14.54	16.05
	3 shot	22.89	9.03	13.43	12.62	26.04	17.20	16.09	17.70
	5 shot	22.67	9.68	13.84	12.81	26.01	15.98	16.20	17.57
MIXED	1 shot	22.25	10.17	13.92	12.35	25.96	16.87	16.18	16.95
	3 shot	**22.96**	10.38	13.78	12.25	**26.26**	17.97	17.34	18.55
	5 shot	22.90	9.98	**14.37**	**13.30**	26.15	18.16	**17.57**	18.21
SPECIFIC	1 shot	22.33	**10.51**	13.90	12.22	25.89	17.28	15.37	17.12
	3 shot	22.89	9.87	12.86	12.91	26.04	**18.26**	16.08	18.39
	5 shot	22.67	9.28	12.97	12.99	26.01	18.12	17.45	**18.77**

Table 2: Results of Qwen-7B-Chat model and Qwen-14B-Chat on Indonesian-Chinese dataset of sythetic noise.

demonstration examples with synthesized noise how to handle the noise and translate it so that it can better cope with the sentence to be translated. This scenario remains consistent even when selecting a different number of demonstration examples. Similarly, Table 2 demonstrates consistent findings in low-resource languages. Therefore, we can empirically conclude that Language Models (LLMs) can acquire translation robustness from examples with synthetic noise. Additionally, when employing the same sampling method configuration, the greater the number of demonstration examples used, the more likely LLMs are to exhibit improved translation performance. However, in certain cases, there may also be situations where 3-shots perform better than 5-shots. Our case studies reveal that an increased number of demonstration examples may lead to a higher likelihood of the model exhibiting hallucination, potentially explaining why 5-shots do not perform as well as 3-shots in certain instances.

LLMs are more likely to learn robustness to character-level noise through type-specific synthetic noise and robustness to word-level noise through mixed-type synthetic noise. By referring to Table 1 and Table 2, we can observe that, for the character-level noise test set, using the SPECIFIC sampling method yields better results for LLMs compared to the MIXED sampling method. For instance, the Baichuan2-7B-Chat model achieved the highest BLEU score of 20.03 on the Chinese-English character-level noise test set using SPECIFIC sampling. However, for word-level noise test sets, LLMs perform better using the MIXED sampling method compared to the SPECIFIC sampling method. These findings hold true for both high- and low-resource languages. This is to some extent due to the fact that coarse-grained word-level noise includes fine-grained character-level noise, so the model can learn robustness to word-level noise through character-level noise, but not vice versa.

The robustness of LLMs in learning from various types of natural noises varies across high and low-resource languages. Through Tables 3 and 4, we can observe that in high-resource

	shots	Code Switch	Dialect	Emojis	Grammar	Jargon	Proper Nouns	Slang	Slurs	Spelling/ Typo	Spoken Language	Average
	0 shot	15.77	15.19	**15.17**	14.40	17.11	17.31	14.40	12.83	16.77	15.09	15.40
MIXED	1 shot	15.60	**16.73**	14.63	**16.79**	**17.74**	**17.92**	**16.90**	**14.97**	19.42	**15.41**	**16.61**
	3 shot	15.47	14.85	12.67	15.28	16.98	16.05	15.14	12.37	**20.66**	15.09	15.45
	5 shot	14.55	13.50	11.92	15.05	15.90	15.55	14.45	10.18	16.09	13.58	14.08
SPECIFIC	1 shot	15.28	15.33	14.13	14.35	17.13	16.41	14.69	11.37	18.44	13.56	15.07
	3 shot	**16.52**	15.75	12.45	14.25	16.96	17.08	14.35	9.73	16.85	13.96	14.79
	5 shot	16.16	15.54	13.12	15.20	17.34	16.99	14.98	11.35	17.17	14.25	15.21

Table 3: Results of Baichuan2-7B-Chat model on Chinese-English dataset of natural noise.

	shots	Code Switch	Dialect	Emojis	Grammar	Jargon	Proper Nouns	Slang	Slurs	Spelling/ Typo	Spoken Language	Average
	0 shot	17.83	17.77	17.41	19.40	20.43	23.19	18.73	12.35	18.07	17.18	18.24
MIXED	1 shot	18.17	17.59	**21.72**	19.60	21.32	22.21	18.45	12.72	20.46	**19.41**	19.16
	3 shot	16.55	**18.84**	17.92	19.09	18.95	23.22	20.26	**15.48**	16.20	19.13	18.56
	5 shot	18.91	17.51	15.08	20.67	21.50	20.94	19.86	15.03	14.87	19.09	18.35
SPECIFIC	1 shot	**19.23**	17.49	20.66	20.02	21.30	23.38	18.53	13.80	20.16	19.36	19.39
	3 shot	18.36	17.45	21.62	17.14	21.72	23.77	**20.41**	14.38	**21.57**	19.00	**19.54**
	5 shot	15.43	15.04	16.04	**21.09**	**23.13**	**24.81**	15.85	13.57	21.37	17.77	18.41

Table 4: Results of Qwen-7B-Chat model on Indonesian-Chinese dataset of natural noise.

languages, LLMs achieve higher BLEU scores for translation results under the MIXED sampling setting compared to the SPECIFIC setting. However, the opposite trend emerges in low-resource languages. This holds with different numbers of demonstration examples. This may be attributed to the fact that LLMs have more comprehensive knowledge in high-resource languages, enabling them to use mixed type noise for translation robustness. Conversely, in low-resource languages, where knowledge is lacking, specific types of natural noise are needed to learn translation robustness. Therefore, it can be empirically concluded that the performance of LLMs in acquiring translation robustness varies inconsistently in high-resource and low-resource languages when exposed to various types of natural noise.

5. Analysis

We explore the effect of the noise ratio in the demonstrations on the learning effect of the LLMs' translation robustness, and investigate the translation robustness of the LLM learned from noisy-source demonstrations for specific examples.

5.1. Effect of Noise Ratio in Demonstrations

We explore the effect of the noise ratio in the demonstration example on model robustness learning on the Chinese-English dataset for three synthetic noise types. Specifically, we use the BLEU value of the MIXED sampling method subtracted from the BLEU value of the CLEAN sampling method under the same shot setting as a measure of the robustness of the LLMs, and conduct the experiments under the settings of noise ratio of 0.1, 0.3, and 0.5, respectively, and the results are shown in Fig 3.

We find that as the noise proportion increases, the average BLEU improvement of LLMs for three types of noise in all shot scenarios also increases. The trends in BLEU improvement values for the three types of noise under specific shot settings exhibit a similar pattern for the most part. Therefore, we empirically conclude that the robustness of LLMs to noise they learn improves within a certain noise proportion range as the noise proportion increases. However, this increasing trend will slow down as the noise level increases. Therefore, we predict that the improvement of LLMs' robustness to noise will not continuously increase with the increase in noise levels. Only when the noise level is within a certain range, more noise imparts greater knowledge to LLMs, and the performance improvement of LLMs' robustness to noise will increase accordingly. Nevertheless, when the proportion of noise exceeds this interval, the difficulty of the task exceeds the upper limit of the LLMs' capability, and thus the learning effect of the LLMs' robustness to noise shows a decreasing trend.

5.2. Case Study

Through Table 5, we can find that when the LLMs are provided with noisy-source demonstration examples, the LLMs can better solve the noise problem in the test input and output high-quality translated sentences. We argue that there are noisy source sentences to be translated and corre-

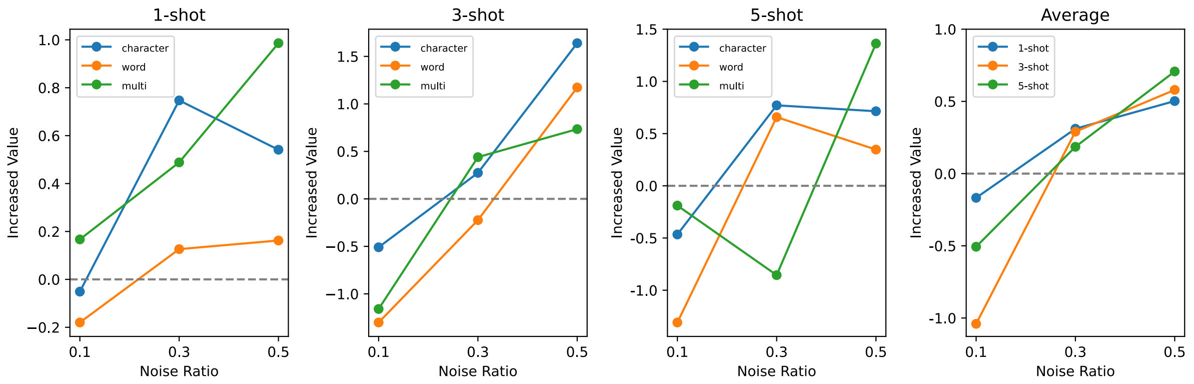

Figure 3: The relationship between the BLEU score change and noise proportion in demonstration examples when comparing the MIXED sampling method with three types of synthetic noise on a Chinese-English dataset to the CLEAN sampling method. The first three subfigures respectively demonstrate the relationship between the noise ratio and the change in BLEU score for 1-shot, 3-shot, and 5-shot settings. The fourth subfigure averages the BLEU score changes for all types of noise under the same shot and noise ratio settings, and displays the relationship between the noise ratio and this average value.

Source	这位 17 岁的攻击型前卫在上个赛季总共出场五次，他是英超联赛中出现的首个 1999 年出生的球员，在英国青少年球队中也受到高度评价。
Reference	The 17-year-old attacking midfielder made five appearances in total last season, becoming the first player born in 1999 to appear in the Premier League, and is also highly rated in the England underage system.
Demonstration	普利斯在职业生涯中起用年轻球员的效率之慢谓可臭名昭~~著~~，在上个季赛以 1-1 战平利物浦队的最后一场篦赛中入进其乞视野的三名青少年球员（乔纳森 · 莱科、姆山 · 菲尔德和~~泰~~勒 · 罗伯茨）怖不太可能成限常态。 Pulis has been notoriously slow to promote young players throughout his career and the sight of three teenagers - Jonathan Leko, Sam Field and Tyler Roberts - in the final game of last season, a 1-1 draw with Liverpool, is unlikely to become the norm.
Noisy input	这位 17 隋得攻击型歉前卫在上铬赛季总共出场次五，他是超英~~联~~赛中出~~现~~的首个 1999念出生的球员，在哉英国少青年球队中揶受到高度评价。
LLM output	The 17-year-old attacking midfielder made five appearances in total during the last season, becoming the first player born in 1999 to appear in the Premier League, and receiving high praise in the English youth teams.

Table 5: LLM for generating cases in the Chinese-English dataset, where demonstration examples are selected for 1 shot. (Orange: Insertion noise; Red: Swapping noise; Magenta: Replacement noise; ~~Blue~~: Deletion noise.)

sponding clean target-translated sentences in the demonstration examples, from which the LLMs are able to learn how to transform against the noise, and that when confronted with noisy test input sentences, the LLMs are able to apply this ability to deal with the noise.

Furthermore, when there is the same type of noise as the test input in the noise demonstration example, it might be more beneficial for the LLMs to learn about the way to deal with this type of noise. For example, as shown in Table 5, there are Insertion, Swapping, Replacement and Deletion noises in the demonstration example, and these types of noises in the test input are also well resolved by the LLM.

5.3. Results for Other Models and Language Pairs

To ensure the broad applicability of our findings and mitigate potential biases stemming from specific models and language pairs, we expanded our experiments to cover more models and language pairs. For the Baichuan2 model, we augmented our experiments to include en-zh, fr-en and id-zh translation directions. We employed four distinct experimental settings: 0 shot, CLEAN, MIEXD, and SPECIFIC. In each case, three contextual examples were sampled for the respective sampling methods. Results are presented in Table 6. Likewise, for the Qwen model, we broadened the experiments to encompass three translation direc-

Settings	**en-zh**			**fr-en**			**id-zh**		
	clean	character	word	clean	character	word	clean	character	word
0 shot	35.27	22.64	19.02	21.61	12.98	10.55	16.01	5.47	9.19
CLEAN	38.28	24.52	21.84	23.00	13.44	10.67	**19.56**	6.19	10.87
MIXED	**38.62**	27.64	**23.16**	**23.82**	13.71	**11.79**	19.05	**6.53**	**10.94**
SPECIFIC	38.28	**27.91**	22.06	23.00	**14.06**	11.38	**19.56**	6.26	9.56

Table 6: Results of Baichuan2-7B-Chat on the synthetic noise dataset under various settings.

Settings	**en-zh**			**fr-en**			**zh-en**		
	clean	character	word	clean	character	word	clean	character	word
0 shot	33.62	22.34	16.36	17.67	10.27	7.61	23.65	17.83	14.95
CLEAN	36.46	26.41	19.64	26.82	13.11	10.07	28.53	22.05	19.92
MIXED	**36.86**	**28.32**	**20.54**	**28.72**	14.74	**11.47**	**28.94**	22.22	**20.20**
SPECIFIC	36.46	28.16	20.32	26.82	**15.40**	10.46	28.53	**22.80**	20.08

Table 7: Results of Qwen-7B-Chat on the synthetic noise dataset under various settings.

Settings	**en-zh**			**fr-en**			**zh-en**		
	clean	character	word	clean	character	word	clean	character	word
0 shot	31.47	19.22	15.05	26.24	10.76	10.22	20.15	12.98	9.52
CLEAN	32.78	19.59	16.01	26.68	11.47	11.27	22.22	14.10	11.75
MIXED	**33.31**	22.78	**16.28**	**27.78**	12.52	**11.63**	**23.08**	15.55	11.36
SPECIFIC	32.78	**23.01**	14.19	26.68	**12.79**	11.55	22.56	**16.28**	**12.43**

Table 8: Results of InternLM-Chat-7B on the synthetic noise dataset under various settings.

tions: en-zh, fr-en, and zh-en. The corresponding experimental results can be found in Table 7. Furthermore, we introduced a new model InternLM and conducted experiments across three translation directions en-zh, fr-en, and zh-en, with results provided in Table 8.

According to these experimental results, it is evident that the BLEU scores achieved through the MIXED and SPECIFIC sampling methods, across various models and translation directions, consistently surpass the BLEU scores obtained via the CLEAN sampling method. This trend holds true across different test sets, including the clean test set, character-level noise test set, and word-level noise test set. These results further corroborate the assertion that LLMs exhibit enhanced translation robustness by learning from context enriched with synthetic noise. Notably, on the character-level noise test set, LLMs with the SPECIFIC sampling method consistently outperform those with the MIXED sampling method in the majority of experimental setups. Conversely, on the word-level noise test set, LLMs that use the MIXED sampling method demonstrate superior performance compared to those with the SPECIFIC sampling method. These observations align with our research findings, highlighting the general validity of our conclusions across diverse models and language pairs.

6. Conclusion

In this paper, we have presented a research scheme on the robustness of machine translation via LLMs, aiming at investigating whether LLMs can learn to deal with noise and translation methods in contextual environments with noise. We find empirically through experiments that LLMs can indeed learn machine translation robustness from demonstration examples with synthetic noise, both on high- and low-resource languages. And we find that within the appropriate range, increasing the noise level in the demonstration examples can enhance the translation robustness of LLMs. Moreover, we observe that LLMs are more likely to learn robustness to character-level noise through type-specific synthesized noise as well as robustness to word-level noise through mixed-type synthesized noise. Finally, we conduct an exploration on both publicly available and self-annotated natural noise translation noise test sets and find that for various types of natural noise, the robustness learning performance of LLMs varies between high and low-resource languages. High-resource languages tend to learn robustness in demonstration examples with mixed-type noise, while low-resource languages tend to learn robustness in demonstration examples with specific types of noise.

Acknowledgments

The present research was supported by the Key Research and Development Program of Yunnan Province (Grant No.202203AA080004) and Natural Science Foundation of Xinjiang Uygur Autonomous Region (Grant No.2022D01D43). We would like to thank the anonymous reviewers for their insightful comments.

Bibliographical References

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2022. In-context examples selection for machine translation. *arXiv preprint arXiv:2212.02437*.

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

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Yong Cheng, Lu Jiang, and Wolfgang Macherey. 2019. Robust neural machine translation with doubly adversarial inputs. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4324–4333. Association for Computational Linguistics.

Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *arXiv preprint arXiv:2306.16092*.

Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Zhifang Sui, and Furu Wei. 2022. Why can gpt learn in-context? language models secretly perform gradient descent as meta optimizers. *arXiv preprint arXiv:2212.10559*.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*.

Wenxiang Jiao, Wenxuan Wang, JT Huang, Xing Wang, and ZP Tu. 2023. Is ChatGPT a good translator? yes with gpt-4 as the engine. *arXiv preprint arXiv:2301.08745*.

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*.

Edward Ma. 2019. NLP augmentation. https://github.com/makcedward/nlpaug.

Paul McNamee and Kevin Duh. 2022. The multilingual microblog translation corpus: Improving and evaluating translation of user-generated text. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 910–918.

Paul Michel and Graham Neubig. 2018. Mtnt: A testbed for machine translation of noisy text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 543–553.

Yasmin Moslem, Rejwanul Haque, and Andy Way. 2023. Adaptive machine translation with large language models. *arXiv preprint arXiv:2301.13294*.

Jane Pan. 2023. *What In-Context Learning "Learns" In-Context: Disentangling Task Recognition and Task Learning*. Ph.D. thesis, Princeton University.

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.

Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023. Towards making the most of ChatGPT for machine translation. *arXiv preprint arXiv:2303.13780*.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is ChatGPT a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*.

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.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, pages 1–11.

Eric Wallace, Mitchell Stern, and Dawn Song. 2020. Imitation attacks and defenses for black-box machine translation systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 5531–5546. Association for Computational Linguistics.

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

Zengzhi Wang, Qiming Xie, Zixiang Ding, Yi Feng, and Rui Xia. 2023b. Is ChatGPT a good sentiment analyzer? a preliminary study. *arXiv preprint arXiv:2304.04339*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

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. 2023. Prompting large language model for machine translation: A case study. *arXiv preprint arXiv:2301.07069*.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pages 12697–12706. PMLR.

Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Neil Zhenqiang Gong, Yue Zhang, et al. 2023. Promptbench: Towards evaluating the robustness of large language models on adversarial prompts. *arXiv preprint arXiv:2306.04528*.