DragFT: Adapting Large Language Models with Dictionary and Retrieval Augmented Fine-tuning for Domain-specific Machine Translation

Anonymous ACL submission

Abstract

Large language models (LLMs) have shown great potential in domain-specific machine translation (MT). However, one major issue is that LLMs trained on general corpus might not generalize well to specific domains due to the lack of domain-specific knowledge. To address this issue, this paper focuses on enhancing the domain-specific MT capability of LLMs, by providing high-quality training datasets and proposing a novel fine-tuning framework denoted by ***DragFT***. DragFT augments LLMs via three techniques: (i) ***Dictionary-enhanced prompting*** improves domain-specific terminology translation; (ii) ***RAG-based few-shot example selection*** provides high-quality examples that simulate both the domain and style characteristics; (iii) ***Fine-tuning with few-shot examples*** further boosts fine-tuning with in-domain examples. We deploy DragFT on three well-known LLM backbones to validate its effectiveness. The results on three domain-specific datasets show that DragFT achieves a significant performance boost and shows superior performance compared to strong baselines such as GPT-3.5 and GPT-4o. The drastic performance improvement of DragFT over existing LLMs can be attributed to the incorporation of relevant knowledge while mitigating noise. Our three well-constructed datasets can accelerate future research in domain-specific MT: a benchmark dataset designed for MT within the IT domain, and two datasets constructed from publicly available datasets respectively in law and medicine.

1 Introduction

Although Large language models (LLMs) have demonstrated remarkable performance in MT, they often fall short of the performance achieved by domain-specific models. To improve the domain-specific machine translation (MT) capability of LLMs, existing works fall into two groups. The first group employs in-context learning (ICL) by feeding LLMs with in-domain translation examples as a demonstration without further fine-tuning (Aycock and Bawden, 2024; Vilar et al., 2023; Moslem et al., 2023a; Zhang et al., 2023a). ICL provides in-context examples that help the model quickly adapt to specific domains and styles. However, its performance depends heavily on the quality and relevance of examples. Another group fine-tunes LLMs with translation instructions to improve the domain-specific MT capability (Wei et al., 2022; Moslem et al., 2023b). However, it often requires high computational costs for extra training on specific domains and may weaken the general MT capabilities in LLMs due to over-specialization (Alves et al., 2023). Therefore, improving the domain-specific MT capability of general-purpose LLMs remains a challenge. First, current systems still struggle with terminology translation. Even domain-adapted models have difficulty with accurately translating domain-specific terminology (Sato et al., 2020). Second, high-quality in-domain parallel datasets are often required for fine-tuning LLMs.

This paper addresses the above challenges by boosting fine-tuning with few-shot examples to leverage both ICL and fine-tuning benefits. We propose a novel fine-tuning framework, denoted as ***DragFT*** (***D***ictionary and ***r***etrieval ***a***u***g***mented ***F***ine-***T***uning), to augment the performance of LLMs in domain-specific MT. DragFT contains three components: dictionary-enhanced prompting, RAG-based few-shot example selection, and fine-tuning with few-shot examples. We propose ***Dict-rephrasing***, a dictionary-enhanced algorithm, that rephrases the source sentence by replacing terminology with domain-specific terms in the target language. It can augment fine-tuning performance by improving domain-specific terminology translation. A RAG-based few-shot example selection mechanism is developed to boost fine-tuning with high-quality examples in instructions. We use extra corpora (self-constructed domain-specific corpora) to build a vector database and retrieve relevant examples to construct translation instructions, which are then fed into LLMs for fine-tuning. We construct three domain-specific translation instruction-following datasets and enhance the data quality by using LLM-based evaluation and human annotation to mitigate noise. In the component of fine-tuning with few-shot examples, we apply the Low-Rank Adaptation (LoRA) strategy to reduce the computational cost. Our main contributions are summarized as follows:

- We propose DragFT, a novel fine-tuning framework that enhances domain-specific MT by incorporating dictionary-enhanced prompting for improving terminology translation, and RAG-based selection mechanism for incorporating high-quality examples.
- We construct three bilingual translation corpora in specific domains and improve data quality through LLM-based evaluation and manual annotation, tackling the challenge of limited high-quality training data for fine-tuning in domain-specific MT.
- We conduct comprehensive experiments by adapting three well-known 13B backbone models over three datasets in different domains. The results show that DragFT can achieve significant improvements on existing LLMs in domain-specific MT. It also shows superior performance compared with strong baselines.

2 Related Works

2.1 ICL in Machine Translation

ICL feeds LLMs with extra translation examples within the prompts to improve the MT capabilities, without fine-tuning (Brown et al., 2020). Several works focused on improving the MT capabilities of LLMs via ICL. (Zhang et al., 2023a) revealed that prompt example effectiveness in MT depends on features like sequence length and semantic similarity, with back-translation being especially robust. (Agrawal et al., 2023) showed that optimizing in-context examples and prompts, especially using n-gram overlap and re-ranking, significantly improves the MT quality. Other works investigated prompting strategies for identifying appropriate examples. (Vilar et al., 2023) evaluated the MT performance of PaLM (Chowdhery et al., 2023) with different prompting strategies. (Garcia and Firat, 2022) used natural language-described prompts to control and improve multilingual MT, enabling translation into specific dialects and unseen languages. (Jiao et al., 2023b) demonstrated that effective prompts and example utilization can enhance ChatGPT [1] multilingual translation, with a pivot prompting strategy improving performance for distant languages.

Although moderate progress has been made, ICL is highly sensitive to the quality of provided examples. Poor examples may lead to sub-optimal LLM translation performance.

2.2 Instruction tuning in Machine Translation

Instruction tuning is a technique for fine-tuning language models to improve their abilities to follow specific instructions, enhancing their adaptability and performance across diverse downstream tasks. Given labeled domain-specific data, instruction tuning can be an alternative to improve the MT capabilities of LLMs. Instruction tuning is reported to outperform in-context learning in MT performance (Li et al., 2023). Several works enhanced the MT performance of LLMs by fine-tuning them with translation instructions on large amounts of parallel data (Wei et al., 2022; Yang et al., 2023b; Zhang et al., 2023c; Chen et al., 2023b). (Jiao et al., 2023a) incorporated hint fields and three instruction types to enhance chat translations. (Xu et al., 2024) revealed that large parallel datasets are unnecessary for high MT performance in LLMs, achieving significant improvements with a novel two-stage fine-tuning method involving monolingual fine-tuning and lightweight parallel fine-tuning.

2.3 Domain-specific Machine Translation

Even though trained on large amounts of data, these two groups of methods can struggle to translate inputs with rare words in domain transfer scenarios (Ghazvininejad et al., 2023). Therefore, several works focused on using the domain-specific vocabulary to supply translations in low-resource settings (Lu et al., 2023; Ghazvininejad et al., 2023; Moslem et al., 2023c). For instance, (Ghazvininejad et al., 2023) incorporated the additional dictionaries into zero-shot examples without training.

Our work takes full advantage of ICL and instruction tuning, incorporating high-quality and relevant translation examples during the fine-tuning

[1] https://chat.openai.com

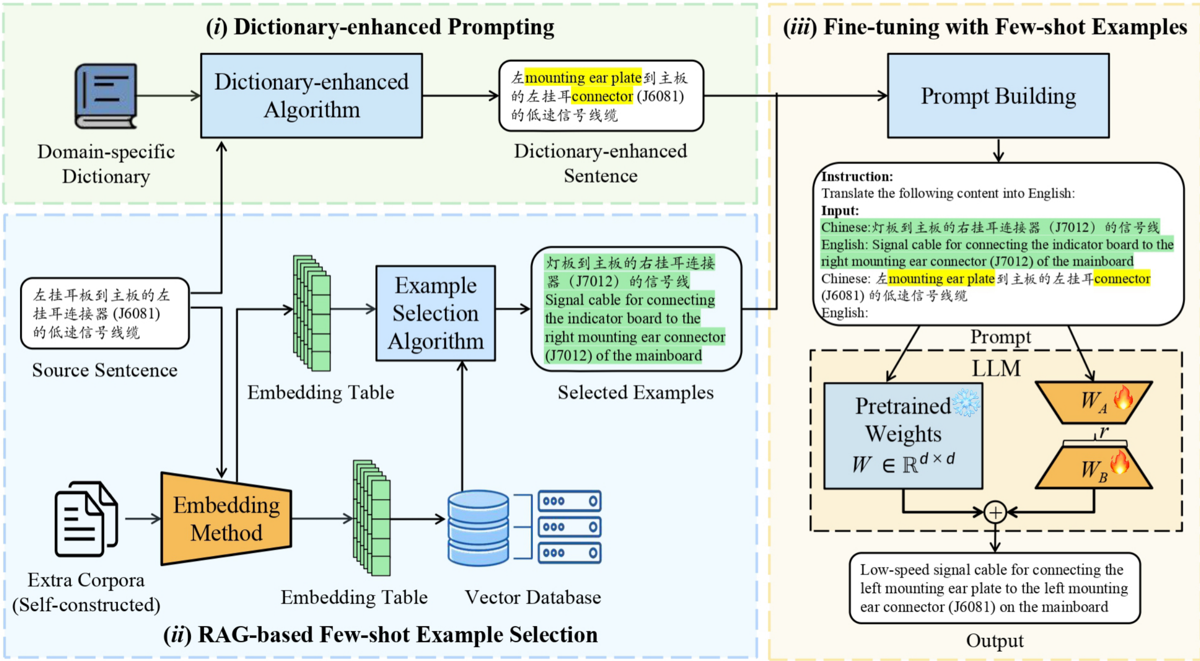

Figure 1: The framework of DragFT, including three techniques: (i) ***Dictionary-enhanced prompting***, (ii) ***RAG-based few-shot example selection***, and (iii) ***Fine-tuning with few-shot examples***.

stage. We introduce a RAG-based method for providing high-quality in-domain examples, ensuring the selected examples are semantically similar and contextually relevant to the training data. Additionally, we propose a novel dictionary augmentation method to address the challenge of translating terminology in specific domains.

3 DragFT

As shown in Figure 1, our DragFT enhances the domain-specific MT capabilities of LLMs through three techniques: (i) ***Dictionary-enhanced prompting*** is a dictionary augmented technique for improving domain-specific terminology translation; (ii) ***RAG-based few-shot example selection*** provides selected examples that closely match the source sentence in both translation style and vocabulary; (iii) ***Fine-tuning with few-shot examples*** incorporates in-domain examples into fine-tuning by taking advantages of both ICL and fine-tuning.

3.1 Machine Translation Task

Fine-tuning LLM for adaptation to domain-specific MT requires the guidance of translation instructions. Given a bilingual training dataset of $\mathbf{C}$, which contains pairs of parallel bilingual training data denoted as $(\mathbf{x}, \mathbf{y})$, the optimization function $\mathcal{L}$ for the MT task is defined as follows:

$$\mathcal{L} = \sum_{(\mathbf{x},\mathbf{y}) \in \mathbf{C}} -\log p(\mathbf{y}|\mathbf{x}, \mathcal{T}; \theta), \quad (1)$$

where $\mathbf{x} = \{x_1, ..., x_n\}$ is the source sentence, $\mathbf{y} = \{y_1, ..., y_m\}$ is its corresponding target translation, $\mathcal{T}$ is the translation instruction template, and θ represents the training parameters. The probability of a target sentence given the source sentence is:

$$p(\mathbf{y}|\mathbf{x}, \mathcal{T}; \theta) = \prod_{t=1}^{m} p\left(y_t | y_{<t}, \mathbf{x}, \mathcal{T}; \theta\right), \quad (2)$$

where y_t is the t-th generated token, $y_{<t}$ is the privious tokens.

3.2 Dictionary-enhanced Prompting

The main obstacle in domain-specific MT lies in the domain-specific terminology that is not commonly used in general domains, which results in inaccurate translations. To tackle this challenge, incorporating domain-specific terminology dictionaries into translation prompts is crucial. One straightforward method combines dictionary data along with the parallel corpus data to create a translation instruction format, called by ***Dict-instruction***. Inspired by (Zhang et al., 2023b), another approach appends the dictionary translation after the sentence translation in a chained manner, named

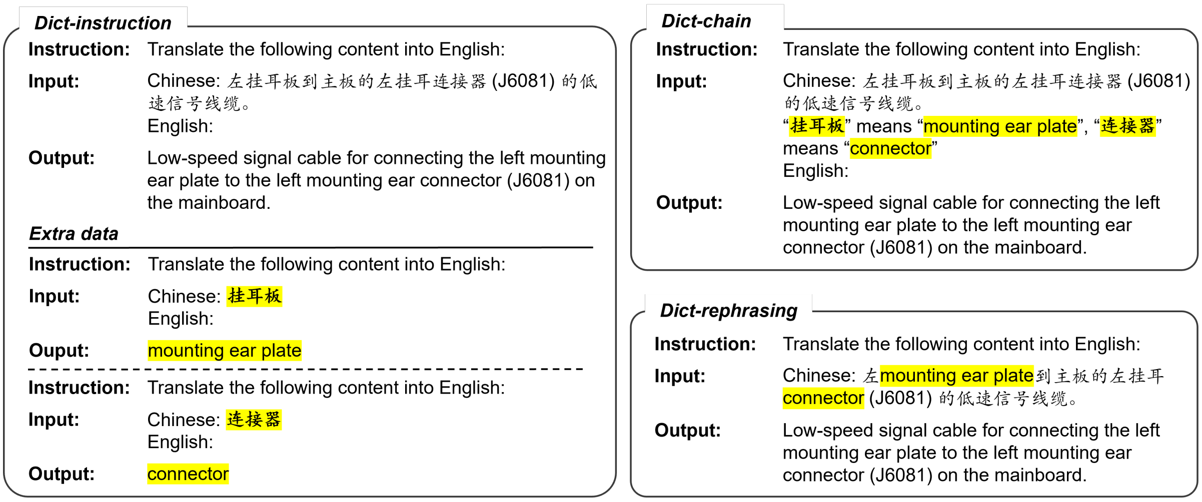

Figure 2: An illustration of three dictionary enhancement prompts, including Dict-instruction, Dict-chain, and Dict-rephrasing.

as ***Dict-chain***. However, the Dict-instruction increases the amount of fine-tuning data, while the Dict-chain extends the length of prompts, resulting in higher consumption of training resources and longer training time.

In this paper, we introduce a novel dictionary enhancement algorithm, denoted as ***Dict-rephrasing***. It directly replaces the domain-specific terminology in source sentences with their corresponding terms in the target language from the in-domain dictionary, as illustrated in Algorithm 1. Figure 2 shows examples of the three dictionary-enhanced prompting methods. Using the Dict-rephrasing, the terminology of "挂耳板" and "连接器" in the source sentence of "左挂耳板到主板的左挂耳连接器(J6081)的低速信号线缆" are directly rephrased to "mounting ear plate" and "connector", respectively. Therefore, the source sentence is rephrased as "左mounting ear plate到主板的左挂耳 connector(J6081)的低速信号线缆".

Dict-rephrasing helps LLMs better understand the terminology in context, effectively reducing the volume of training data compared to Dict-instruction and shortening the length of prompts compared to the Dict-chain. Our experiments in section 6.3 will further explore the effects of these methods.

3.3 RAG-based Few-shot Example Selection

The main idea of Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) is integrating information from external data sources to supplement the input query or enhance the output. To ensure the quality of few-shot examples, we apply the idea

Algorithm 1 Dict-rephrasing

Input: domain-specific dictionary $\mathcal{D}$, domain-specific parallel corpus $\mathbf{C}$
Output: dictionary-enhanced parallel corpus $\mathbf{C}'$

Sort $\mathcal{D}$ by length $\downarrow$
for each translation pair (x, y) in $\mathbf{C}$ **do**
  Initialize $x' \leftarrow x$
  for each word pair (w_{src}, w_{tgt}) in $\mathcal{D}$ **do**
    if w_{src} in x **then**
      Replace w_{src} in x' with w_{tgt}
      $x' \leftarrow \text{Replace}(x', w_{src}, w_{tgt})$
    end if
  end for
  $\mathbf{C}' \leftarrow \mathbf{C}' \cup \{(x', y)\}$
end for

of RAG and design a few-shot example selection mechanism based on it. Specifically, we vectorize extra corpora using the BGE model (Xiao et al., 2023) and store these vectors to construct a domain-specific vector database of V. Given a source sentence, we convert it into a vector of s using the BGE model. To retrieve semantically similar and contextually relevant examples from V, we calculate the similarity score of c_i between s and the vector of $v_i \in V$.

$$c_i = \frac{s \cdot v_i}{\|s\| \|v_i\|} \tag{3}$$

where $\cdot$ represents the dot product function.

We set a similarity score threshold of k and a maximum number of examples n to refine the selection process. If the similarity score of c_i is greater than k, v_i is selected and added to the relevant examples set of R. When $|R|$ is equal to n, we stop retrieving to limit the volume of the fine-tuning dataset.

3.4 Fine-tuning with Few-shot Examples

We utilize the training dataset with few-shot translation examples to fine-tune LLMs. It is reported that fine-tuning with few-shot examples helps maintain the few-shot learning capabilities of LLMs while preserving the benefits of fine-tuning (Alves et al., 2023). The prompt example adopted in our study is shown in Figure 1. We use "*Translating the following content into <target-language>*" as the translation instruction with selected examples and sentences to be translated as inputs. To reduce training costs, we utilize the LoRA (Hu et al., 2021) fine-tuning strategy, which is designed for efficient fine-tuning of LLMs. As illustrated in Figure 1, the pre-trained weights of $W \in \mathbb{R}^{d \times d}$ are frozen, while two low-rank matrices of W_A and W_B with the rank of r are introduced to capture the parameter updates. This approach allows for efficient fine-tuning with reduced computational costs and GPU memory requirements.

4 Experimental Setups

4.1 Datasets & Evaluation Metrics

We conduct experiments on DragFT across three specific domains: IT, law, and medicine. Towards this end, we construct three bilingual instruction-following datasets in specific domains for fine-tuning LLMs.

We collect documents within the IT domain in both Chinese and English from well-known IT companies and segment them into sentences, which are aligned to form a parallel corpus. To improve data quality, we utilize the COMETKiwi (Rei et al., 2023), a model-based evaluation method that doesn't require extra translation references. Translation pairs with COMETKiwi scores below 80 are discarded and the remaining candidates are verified with manual annotations by domain experts.

We also conduct experiments on two datasets named Law and Medical, respectively belonging to the domain of law and medicine (Aharoni and Goldberg, 2020). As the original datasets are in English and German, we utilize Google Translate[2] to translate the contents into Chinese. We further improve the data quality by employing the same method with COMETKiwi and manual annotations and form two new datasets in both English and Chinese for the domains of law and medicine respectively.

We use two widely used evaluation metrics in MT, including the word-based metric of BLEU (Papineni et al., 2002), and the reference-based metric of COMET (Rei et al., 2022) for model evaluation.

To generate domain-specific dictionaries, we design prompts for GPT-3.5 to extract terminologies from the training sets. We then work with experts to manually filter out general words and annotate the translations. Detailed prompts are provided in the appendix.

4.2 Baselines

To investigate the effectiveness of DragFT, we adapt it to three 13B parameter-scale LLM backbones: ***Tigerbot-13B*** (Chen et al., 2023a), ***Baichuan2-13B*** (Yang et al., 2023a), and ***LLama2-13B*** (Touvron et al., 2023). Due to the baselines' poor adherence to translation instructions and frequent over-generation, we fine-tune these models using $20,000$ random samples from the WMT-19[3] dataset (in general domains) in the Zh⇔En directions. This fine-tuning process helps the baselines better follow translation instructions for model evaluation. We also consider three well-known strong baselines, including NLLB (Costa-jussà et al., 2022) from the NMT domain, GPT-3.5[4] and GPT-4o from the LLM domain.

4.3 Implementation Details

We fine-tune the backbone models using a learning rate of 3e-4, a training batch size of 2, a maximum sequence length of 512 tokens, a weight decay of 0.00001, and a warmup ratio of 0.01. For efficient training, we employ the Deepspeed[5] and Flash-Attention (Dao et al., 2022) acceleration frameworks for fine-tuning with LoRA, with the rank set to 16. In the inference stage, we adopt the vLLM (Kwon et al., 2023) framework to accelerate inference and reduce memory usage. We use the beam search algorithm with a beam width of 4, a temperature of 0 to minimize diversity in translation output and a length penalty of 1.0. In

[2]`http://translate.google.com`

[3]`https://www.statmt.org/wmt19`

[4]The GPT-3.5 version is gpt-3.5-turbo-1106.

[5]`https://github.com/microsoft/DeepSpeed`

Model	Zh ⇒ En						En ⇒ Zh					
	IT		Law		Medical		IT		Law		Medical	
	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
Advanced Models												
NLLB-3.3B	26.37	82.76	46.27	83.87	37.52	81.32	26.96	83.37	42.99	84.46	38.15	80.00
GPT-3.5	29.33	84.58	34.44	84.12	41.51	86.56	34.44	85.58	47.71	86.18	53.77	86.05
GPT-4o	31.23	85.43	38.71	85.60	**45.55**	**87.96**	37.16	86.44	54.22	88.31	61.19	88.44
Base Model: Tigerbot-13B												
Tigerbot-13B	25.79	82.47	32.30	83.22	37.04	85.11	27.79	82.22	39.85	83.56	44.61	84.66
DragFT	45.49	**85.64**	45.65	85.32	44.93	86.02	**45.31**	86.92	58.95	89.26	64.44	89.10
Base Model: Baichuan2-13B												
Baichuan2-13B	26.81	82.67	34.56	82.69	40.41	85.96	30.02	82.87	45.69	81.39	55.81	86.26
DragFT	43.24	84.65	44.89	**85.73**	44.78	86.67	44.56	87.05	**60.18**	**89.28**	64.48	89.31
Base Model: Llama2-13B												
Llama2-13B	22.21	80.36	31.32	82.28	34.16	83.07	23.31	79.56	24.21	74.53	28.17	75.78
DragFT	**45.64**	85.55	**47.35**	85.11	44.85	86.50	45.16	**87.07**	57.08	88.19	**65.43**	**89.83**

Table 1: Translation performance of advanced models and applying DragFT method on three backbone models (TigerBot-13B, Baichuan2-13B, and Llama2-13B) on IT, Law, and Medical datasets (Zh ⇔ En).

the RAG-based few-shot example selection mechanism, we set the similarity score threshold k to 0.7, and the maximum number of examples n to 2. All experiments were conducted on one NVIDIA A100 GPU.

5 Results

We show the main results of the domain-specific translation for Zh⇔En in Table 1. To ensure consistency between training and testing, we apply the corresponding dictionary-enhanced methods to construct the test set during the inference stage. Overall, our DragFT significantly improves the translation quality of existing LLMs and shows superior performance compared with strong baselines. We have the following observations:

(i) DragFT achieves a significant performance boost in three LLM backbones over three domain-specific test sets of IT, Law, and Medical. This can be attributed to the incorporation of relevant knowledge while mitigating noise, which also indicates the effectiveness of three techniques in DragFT.

(ii) Among three strong baselines of GPT-3.5, GPT-4o, and NLLB-3.3B, GPT-4o achieves the best performance. Compared to GPT-4o, DragFT significantly outperforms it in most datasets and shows comparative performance over the dataset in the medical domain (Zh ⇒ En).

(iii) DragFT demonstrates drastic improvement in the BLEU metric compared to the COMET metric. Since BLEU evaluates translation quality at word and phrase levels, our dictionary-enhanced prompting can augment LLMs by translating domain-specific terminologies. This also indicates the effectiveness of Dict-rephrasing.

6 Analysis

6.1 Effect of Instruction Tuning on MT

To evaluate the effect of instruction tuning on MT tasks, we conduct a comparative experiment using the Tigerbot-13B. We use the WMT22 test set (Zh⇔En) [6] as the test set, which is formatted into translation instructions. Additionally, we extract 20,000 samples from the WMT19 parallel corpus (Zh⇔En) to form the training set.

The experiment includes the following settings:

Pre-trained: The test set is directly fed into the original model without fine-tuning.

Fine-tuned: The model is fine-tuned using training data without translation instruction tuning.

Instruction-tuned: The model is fine-tuned using training data formatted with translation instructions.

Reference: The referenced translations of the test set.

We show the length distribution result of tokenized outputs when translating the WMT22 test set (Zh⇒En) on different training setups as shown in Figure 3. We observe that the outputs of the pre-trained model are generally too short, indicating a failure to accurately understand the MT task without fine-tuning. On the other hand, the fine-tuned

[6] https://www.statmt.org/wmt22

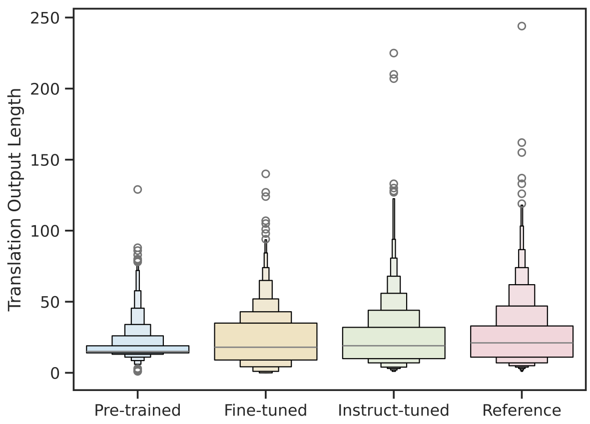

Figure 3: The length distribution of tokenized outputs on the WMT22 test set (Zh⇒En).

model produces excessively long outputs, demonstrating the over-generation problem. In contrast, the instruction-tuned model generates outputs with length distribution closer to the reference. This indicates that instruction tuning effectively guides the model to complete the MT task without generating redundant information.

6.2 Effect of Dictionary-enhanced Prompting

To investigate whether our proposed dictionary-enhanced algorithm can improve the performance of LLMs in domain-specific MT, we conduct comparative experiments on Tigerbot-13B. We employ three different dictionary-enhanced methods introduced in section 3.3 to construct training data for fine-tuning and then evaluate the translation quality on a domain-specific test set. We also conduct an experiment on fine-tuning without dictionary augmentation, denoted as ***Dict-none***. The experimental results are shown in Figure 4.

Compared to Dict-none, all three dictionary-enhanced methods demonstrate translation performance improvements, indicating that they can effectively improve domain-specific terminology translation. Among them, our proposed Dict-rephrasing algorithm shows the most significant improvement, although it performs slightly worse than the Dict-chain in the Medical dataset. This strongly validates the effectiveness of our proposed Dict-rephrasing, which directly embeds terminology information into the source sentences. This approach neither requires additional dictionary data for training nor increases the prompt length, allowing the LLMs to better understand the context of terminology during training, and therefore improving the translation quality.

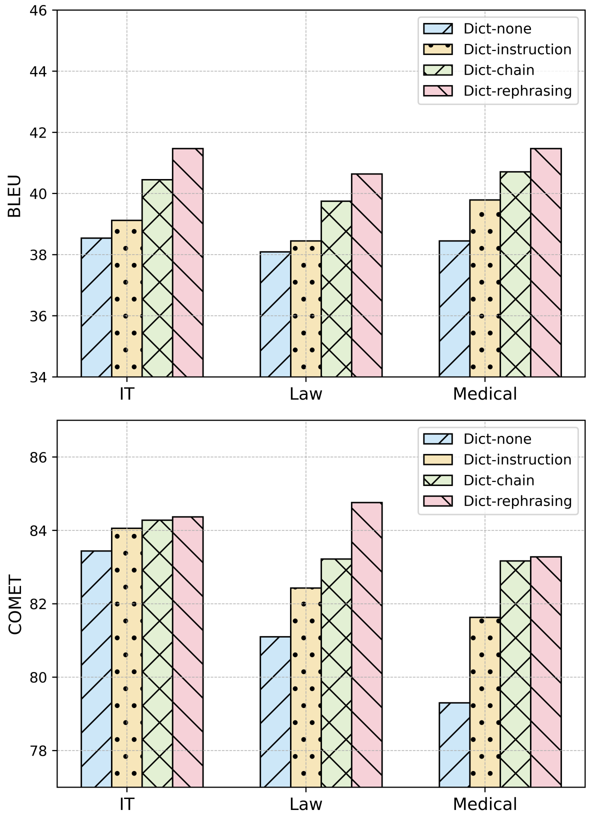

Figure 4: Performance comparison of different dictionary-enhanced prompting methods on domain-specific test sets.

6.3 Ablation Study

We conduct an ablation study to analyze the effects of different components of DragFT. Table 2 shows the results on Tigerbot-13B, which highlights the importance of each component in DragFT.

Without (w/o) Dict-rephrasing. We remove Dict-rephrasing and use the source sentence. From result IDs of 0 and 1 in Table 2, we observe a significant drop in translation quality without dictionary-enhanced prompting. This indicates its essential role in domain-specific MT. The results of 0, 4, and 5 show that the Dict-rephrasing algorithm achieves superior performance compared to the Dict-instruction and Dict-chain methods, which also validates our findings in section 6.2, indicating the effectiveness of the Dict-rephrasing algorithm for domain-specific MT.

Without (w/o) RAG-based selection. We replace the RAG-based example selection mechanism with a strategy that randomly selects two examples for each training data from extra corpora. The results of 0 and 2 in Table 2 reveal a remarkable performance decline in the LLM without RAG selection, which also indicates the quality and relevance of

ID	Method	IT		Law		Medical	
		BLEU	COMET	BLEU	COMET	BLEU	COMET
0	**DragFT [Dict-rephrasing]**	**45.49**	**85.64**	**45.65**	85.32	**44.93**	**86.02**
1	w/o Dict-rephrasing	42.25	84.02	42.59	84.84	42.47	83.74
2	w/o RAG-based selection	39.42	80.41	40.25	83.51	40.77	75.48
3	w/o few-shot example	41.47	84.37	40.64	84.76	41.47	83.28
4	DragFT [Dict-instruction]	43.89	84.34	43.27	85.11	43.32	84.98
5	DragFT [Dict-chain]	44.47	84.87	43.44	**85.42**	44.15	84.78

Table 2: Ablation study. We report the BLEU and COMET scores in Zh⇒En direction with Tigerbot-13B.

examples can affect the performance.

Without (w/o) few-shot example. We directly conduct instruction tuning on the LLM without providing any translation examples. From the results of 0 and 3, we find a drastic decline in translation quality when performing instruction tuning without few-shot examples. This suggests that simple instruction tuning is insufficient to fully leverage the ICL capabilities of LLMs.

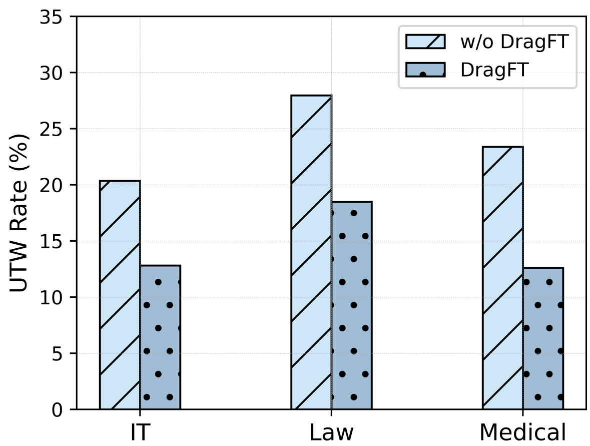

Figure 5: Comparison between the UTW before and after applying DragFT.

6.4 Effects of DragFT

To analyze the impact of the DragFT method, we compare the **Unaligned Translation Words (UTW)** rate between before and after applying DragFT on Tigerbot-13B. The alignment is measured using the method from (Dou and Neubig, 2021), also used by (Hendy et al., 2023). The results are shown in Figure 5, we can observe that after domain adaptation with DragFT, the UTW significantly decreased, indicating improved word translation precision and overall translation performance. This validates DragFT's advantage in handling domain-specific terms.

7 Conclusion

To enhance the domain-specific MT capabilities of LLMs, this paper proposes a novel fine-tuning framework denoted as DragFT. DragFT employs dictionary-enhanced prompting to improve domain-specific terminology translation and RAG-based few-shot example selection to provide high-quality few-shot examples to boost fine-tuning with in-domain examples. We deploy DragFT on three well-known LLM backbones, and the results on three domain-specific datasets show that DragFT can achieve a remarkable performance boost in three backbones and surpass strong baselines. The performance improvement of DragFT over existing LLMs can be attributed to the incorporation of relevant knowledge while mitigating noise. We also construct three domain-specific translation instruction-following datasets to accelerate future research in domain-specific MT. Our current proposed framework fine-tunes all instances, irrespective of whether a test instance requires fine-tuning or not, which may lead to the deterioration of translation quality for some sentences. In the future, we plan to identify those sentences that require fine-tuning and adapt only to them. Meanwhile, we perform dictionary-enhanced prompting for all instances, irrespective of whether a terminology requires enhancement or not, which may lead to the deterioration of translation quality for some sentences. Moving forward, we will focus on identifying domain-specific terms that require rephrasing or dictionary chaining and adopt only those.

Limitation

We focus on the Zh⇔En translation directions, which may limit the generalizability of our findings. Due to time and resource constraints, we rely on machine translation metrics rather than human evaluation to assess translation quality.

Ethics Statement

This work relies on large language models which, as detailed in (Brown et al., 2020) and (Chowdhery et al., 2023), can carry inherent risks. Potential issues include the presence of toxic content due to training on extensive web corpora (Gehman et al., 2020), and high energy consumption during training (Strubell et al., 2019). In constructing the domain-specific dataset, the data were collected with respect to individual privacy, and proper consent was obtained where applicable. Personal or sensitive information was anonymized to ensure protection. Furthermore, to enhance the quality of the dataset, we engage annotators who are duly compensated for their time and expertise, ensuring fair practices by established standards.

References

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2023. In-context examples selection for machine translation. In *ACL Findings*, pages 8857–8873.

Roee Aharoni and Yoav Goldberg. 2020. Unsupervised domain clusters in pretrained language models. *arXiv preprint arXiv:2004.02105*.

Duarte Alves, Nuno Guerreiro, João Alves, José Pombal, Ricardo Rei, José de Souza, Pierre Colombo, and Andre Martins. 2023. Steering large language models for machine translation with finetuning and in-context learning. In *EMNLP*, pages 11127–11148.

Seth Aycock and Rachel Bawden. 2024. Topic-guided example selection for domain adaptation in llm-based machine translation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 175–195.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *NIPS*, pages 1877–1901.

Ye Chen, Wei Cai, Liangmin Wu, Xiaowei Li, Zhanxuan Xin, and Cong Fu. 2023a. Tigerbot: An open multilingual multitask llm. *arXiv preprint arXiv:2312.08688*.

Yijie Chen, Yijin Liu, Fandong Meng, Yufeng Chen, Jinan Xu, and Jie Zhou. 2023b. Improving translation faithfulness of large language models via augmenting instructions. *arXiv preprint arXiv:2308.12674*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 35:16344–16359.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128.

Xavier Garcia and Orhan Firat. 2022. Using natural language prompts for machine translation. *Preprint*, arXiv:2202.11822.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369.

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

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

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Wenxiang Jiao, Jen-tse Huang, Wenxuan Wang, Zhiwei He, Tian Liang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023a. ParroT: Translating during chat using large language models tuned with human translation and feedback. In *EMNLP*, pages 15009–15020.

Wenxiang Jiao, Wenxuan Wang, Jen tse Huang, Xing Wang, and Zhaopeng Tu. 2023b. Is chatgpt a good translator? a preliminary study. *ArXiv*, abs/2301.08745.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Jiahuan Li, Hao Zhou, Shujian Huang, Shanbo Cheng, and Jiajun Chen. 2023. Eliciting the translation ability of large language models via multilingual finetuning with translation instructions. *Preprint*, arXiv:2305.15083.

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

Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023a. Adaptive machine translation with large language models. In *EAMT Annual Conference*, pages 227–237.

Yasmin Moslem, Rejwanul Haque, and Andy Way. 2023b. Fine-tuning large language models for adaptive machine translation. *Preprint*, arXiv:2312.12740.

Yasmin Moslem, Gianfranco Romani, Mahdi Molaei, John D. Kelleher, Rejwanul Haque, and Andy Way. 2023c. Domain terminology integration into machine translation: Leveraging large language models. In *WMT*, pages 902–911.

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

Ricardo Rei, José GC De Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André FT Martins. 2022. Comet-22: Unbabel-ist 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585.

Ricardo Rei, Nuno M Guerreiro, José Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José GC de Souza, and André FT Martins. 2023. Scaling up cometkiwi: Unbabel-ist 2023 submission for the quality estimation shared task. *arXiv preprint arXiv:2309.11925*.

Shoetsu Sato, Jin Sakuma, Naoki Yoshinaga, Masashi Toyoda, and Masaru Kitsuregawa. 2020. Vocabulary adaptation for domain adaptation in neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4269–4279, Online. Association for Computational Linguistics.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in nlp. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. Prompting palm for translation: Assessing strategies and performance. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *ICLR*.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighof. 2023. C-pack: Packaged resources to advance general chinese embedding. *arXiv preprint arXiv:2309.07597*.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024. A paradigm shift in machine translation: Boosting translation performance of large language models. In *The Twelfth International Conference on Learning Representations*.

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

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

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023a. Prompting large language model for machine translation: A case study. In *ICML*, pages 41092–41110.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023b. Prompting large language model for machine translation: A case study. In *International Conference on Machine Learning*, pages 41092–41110. PMLR.

Shaolei Zhang, Qingkai Fang, Zhuocheng Zhang, Zhengrui Ma, Yan Zhou, Langlin Huang, Mengyu Bu, Shangtong Gui, Yunji Chen, Xilin Chen, et al. 2023c. Bayling: Bridging cross-lingual alignment and instruction following through interactive translation for large language models. *arXiv e-prints*, pages arXiv–2306.

Prompt:
You are a seasoned translator specializing in the medical field. Please review the provided Chinese-English translation pairs and identify the most specialized medical terms from each pair. Skip any pairs that do not contain specialized medical terms.

Input:
Chinese: 对于局部肿瘤扩张可能损害重要解剖结构的患者，建议考虑使用皮质类固醇进行预处理。
English: It is recommended that pretreatment with corticosteroids be considered for patients in whom local tumour expansion may compromise vital anatomic structures.

Output:
局部肿瘤扩张 local tumour expansion
解剖结构 anatomic structures
皮质类固醇 corticosteroids

Figure 1: An example of extracting specialized medical bilingual dictionaries.

Appendix A

A1 Domain-specific Dictionary Generation

We employ a method combining LLM models and manual annotation to build domain-specific dictionary data. The process is outlined as follows:

1. For three domain-specific datasets (IT, Law, Medical), we initially input data into ChatGLM [7] using predefined prompts, as shown in Figure 1.

2. The LLM model extracts domain-specific words from the data guided by the prompts.

3. Domain experts perform manual annotations to enhance the accuracy of translating specialized terms.

This approach integrates automated text processing capabilities with domain expertise from human professionals, enabling the efficient generation of high-quality and precise domain-specific dictionary data.

Domain	Train	Test	Vector Database
IT	60000	4920	74699
Law	6000	3950	100000
Medical	6000	3770	87288

Table 1: The data statistics of the datasets we construct on three domain-specific datasets.

[7] https://open.bigmodel.cn/

Method	IT	Law	Medical
Dict-chain	1.54M	4.87M	4.10M
Dict-rephrasing	1.24M	2.66M	2.02M

Table 2: Length of token using different dictionary enhancement methods.

Method	IT	Law	Medical
Dict-instruction	64k	88k	75k
Dict-rephrasing	60k	60k	60k

Table 3: The number of training data using different dictionary enhancement methods.

A2 Dataset Statistics

After separating the test set, we select 60,000 manually screened, high-quality bilingual parallel data for fine-tuning in each of the three domains (IT, Law, and Medical). The remaining data is used to build the vector database. Table 1 shows the statics of the datasets we construct on three specific domains.

A3 Benefits of Dict-rephrasing

We apply three dictionary enhancement methods and conduct data statistics on three training sets. Table 2 shows the total token length of instructions and inputs, while Table 3 displays the number of training data. It can be observed that compared to the Dict-chain method, the training set enhanced by the Dict-rephrasing has a reduced total token length. In comparison to the Dict-instruction method, Dict-rephrasing significantly reduces the volume of train-

Method	Zh ⇒ En				En ⇒ Zh			
	WMT22		Flores-200		WMT22		Flores-200	
	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
WMT22 Winners	**33.50**	**81.0**	**54.3**	**86.8**	-	-	-	-
NLLB-3.3B	21.07	76.92	32.52	81.56				
Tigerbot-13B	15.72	76.62	27.20	86.64	36.34	85.35	**39.89**	86.94
DragFT	23.23	79.93	27.43	86.64	**40.31**	**86.38**	38.91	86.59

Table 4: Translation performance of our DragMT on WMT22 test set and Flores-200 test set with Tigetbot-13B model.

ing data. Overall, the Dict-rephrasing method effectively shortens training time by reducing prompt length and data scale, saving time and computational resources.

A4 Translation performance in general domain

To validate the performance of the model fine-tuned with DragFT in the general domain, we evaluate translation metrics on the WMT22 and Flores-200 test sets and compare them with advanced models. The backbone model is Tigerbot-13B. Table 4 shows the results in the general domain. It is evident that DragFT maintains robust domain-specific translation capabilities while demonstrating excellent translation performance on general domain datasets WMT22 and Flores-200 (Costa-jussà et al., 2022).