Towards Demonstration-Aware Large Language Models for Machine Translation

Chen Li[1] **Meishan Zhang**[1] **Xuebo Liu**[1*] **Zhaocong Li**[2] **Derek F. Wong**[2] **Min Zhang**[1]
[1]Institute of Computing and Intelligence, Harbin Institute of Technology, Shenzhen, China
[2]NLP[2]CT Lab, Department of Computer and Information Science, University of Macau
lichen@stu.hit.edu.cn, {zhangmeishan,liuxuebo,zhangmin2021}@hit.edu.cn,
nlp2ct.zhaocong@gmail.com, derekfw@um.edu.mo

Abstract

Tuning-based large language models for machine translation (aka large translation model, LTM) have demonstrated significant performance in the field of machine translation. Despite their success, these models often face difficulties in leveraging demonstrations to further improve their performance. To tackle this challenge, we introduce a novel approach that integrates demonstration-aware training and inference strategies within the framework of tuning-based LTMs, hereby referred to as demonstration-aware LTMs. During training, we enrich the model's learning process by incorporating both sentence- and document-level demonstrations derived from its original training dataset. During inference, the model synergizes its own contextual translations with retrieved high-quality demonstrations, leading to more precise and contextually appropriate outputs. Empirical results reveal that our demonstration-aware LTM not only mitigates the negative impacts traditionally associated with demonstrations but also secures substantial improvements in translation accuracy, particularly in domain-specific and document-level translation tasks. Source code and scripts are freely available at https://github.com/ChenLi0620/Demo-Aware-LLM-MT.

1 Introduction

Large language models (LLMs) (Robinson et al., 2023; Touvron et al., 2023) have demonstrated significant advancements in the field of machine translation (MT). The current state-of-the-art involves LLMs for machine translation (aka large translation model, LTM) that leverage LLMs, categorized into prompting-based and tuning-based approaches. Prompting-based LTMs (Zhang et al., 2023a; Zhu et al., 2023; Lu et al., 2023; Peng et al., 2023; He et al., 2024) initiate translations by appending examples to a base LLM, offering a training-free method that allows for swift customization to specific translation needs, such as domain-specific translations. However, this approach is constrained by static parameters, limiting its adaptability and translation depth. In contrast, tuning-based LTMs (Jiao et al., 2023; Zhang et al., 2023b; Zeng et al., 2023; Xu et al., 2024; Mao and Yu, 2024; Wang et al., 2024) enhance translation capabilities through supervised fine-tuning on MT-specific datasets, resulting in improved performance due to parameter updates. Yet, this method tends to overlook the model's ability to leverage in-context learning, particularly in scenarios beyond zero-shot sentence-level translation.

Acknowledging the advantages and limitations of both LTM types, we propose the demonstration-aware LTM, which is developed by lightweight LoRA (Hu et al., 2022) fine-tuning on the samples with demonstrations. To cover possible translation scenarios during training, we propose a mixture of demonstration types, which determines the demonstration type of a training sample by randomly choosing either the sentence pairs from the training set as sentence-level demonstrations or the contiguous contextual text as document-level demonstrations. We also design different strategies during the inference phase. For the domain translation, we use only sentence-level demonstrations and retrieve high-quality samples from the training data as demonstrations. When switching to the document-level translation, to ensure the specific information (e.g., style), we propose using hybrid demonstrations that concatenate the document-level and retrieval-based demonstrations.

Our experimental findings highlight that vanilla tuning-based LTMs may not effectively learn to utilize demonstrations. Conversely, our demonstration-aware LTM consistently outperforms a strong tuning-based LTM in demonstration-utilizing scenarios (e.g., domain-specific translation and document-level translation), underscoring

*Corresponding Author

Findings of the Association for Computational Linguistics: ACL 2024, pages 13868–13881

the synergistic potential of combining prompting and tuning-based approaches. Additionally, our analysis corroborates the utility of demonstrations in facilitating the translation of rare words. However, it also brings to light a critical challenge: mitigating the detrimental effects of noisy demonstrations, which can compromise the translation quality of medium- or high-frequency words.

The main contributions are as follows:

- We identify a key limitation in tuning-based LTMs, specifically their inadequate utilization of demonstration knowledge. To overcome this, we introduce a demonstration-aware LTM that incorporates lightweight LoRA fine-tuning on samples with demonstrations, enhancing the model's learning efficiency.
- We innovate by integrating both sentence-level and document-level demonstrations during training. This dual-level approach provides a comprehensive coverage of potential translation scenarios, significantly enriching the model's contextual understanding.
- We design a demonstration-aware model inference strategy by amalgamating retrieval-based examples with model-generated contextual data as demonstrations. This hybrid approach significantly increases translation accuracy in document-level translation.

2 Related Work

Prompting-based LTM Prompting-based LTM has demonstrated substantial significant ICL capabilities, where carefully crafted prompts can yield remarkable outcomes in MT tasks (Bang et al., 2023). Optimal selection of context examples is pivotal, as it can activate the intrinsic mechanisms of prompting-based LTMs to produce the anticipated outputs, as evidenced by Brown et al. (2020). Consequently, there has been considerable research focused on optimizing prompting strategies for LTMs in MT, encompassing the development and evaluation of prompt templates (Zhang et al., 2023a; Hendy et al., 2023), the curation of demonstration sets (Agrawal et al., 2022), and the in-depth exploration of the models' capacity to learn from such demonstrations (Tan et al., 2023; Peng et al., 2024). Further investigations have explored the method of using a pre-trained neural retriever to retrieve knowledge from databases and integrate external knowledge sources into LTMs to elevate translation accuracy (Lewis et al., 2020; Lu et al., 2023; He et al., 2024).

Despite the demonstrated efficiency in domain adaptation and ease of utilization, prompting-based LTMs have not yet achieved the general translation efficacy of their tuning-based counterparts, as highlighted by comparative analyses (Zhu et al., 2023). This limitation somewhat constrains their broader applicability in MT contexts.

Tuning-based LTM Tuning-based LTMs fine-tune foundational LLM models on machine translation datasets, adjusting parameters to better perform on specific translation tasks. With the advent of open-source large models, research into tuning-based LTM (Mishra et al., 2022; Wei et al., 2022) increasingly gains attention. To rapidly adapt models to the translation domain, researchers have constructed numerous instruction-tuning datasets for MT (Xu et al., 2023; Taori et al., 2023; Zheng et al., 2023). To enhance the translation capabilities of LTM, Zeng et al. (2023) incorporate additional training data to improve tuning-based LTM performance. Given that most current large models are primarily based on English as the basic language, there has been research aimed at optimizing translations from English to other languages (Zhang et al., 2023b). Additionally, novel tuning methods have been designed. Xu et al. (2024) focus on models primarily on tuning from extensive multilingual non-parallel corpora. Their method involves two stages: initial fine-tuning on monolingual data followed by fine-tuning on a small set of high-quality parallel data. Mao and Yu (2024) propose a tuning-based LTM for low-resource languages and introduce contrastive alignment instructions to enhance cross-lingual supervision. Koneru et al. (2023) try adapting LTM as automatic post-editors for document-level translation, showing that fine-tuning for automatic post-editors significantly improves translation metrics. Wang et al. (2024) propose a two-stage generation LTM through self-reflection, where the LTM first generates the initial translation, then conducts self-assessments, and refines the translation in the next stage based on the evaluation results. While these advancements have significantly elevated translation capabilities, they often overlook the potential of integrating prompting strategies during the inference stage.

This paper endeavors to amalgamate the strengths of prompting-based LTMs with those of tuning-based LTMs, aiming to endow the result-

ing model with superior demonstration learning capabilities while ensuring exemplary translation performance.

3 Demonstration-Aware LTM

3.1 Motivation

Current LTMs are typically trained with supervised fine-tuning methods, employing simple instructions along with source and target sentence pairs for training, achieving commendable results. Numerous studies have demonstrated these large language models' robust ICL capabilities, which can enhance translation performance through the provision of relevant examples. However, after fine-tuning for machine translation, there's a noticeable decline in the models' ICL ability. During demonstration-aware inference, even when provided with high-quality demonstrations, the improvement in translation performance is not significant and may even diminish. This paper explores how to better integrate both approaches, ensuring that the model fully acquires translation knowledge during the training phase while also possessing superior demonstration learning capabilities during the inference stage, thereby achieving improved translation outcomes. Table 1 presents our proposed four types of demonstrations. Appendix A.1 details the specific prompt forms for these four types of demonstrations for the model training and inference processes.

3.2 Demonstration-Aware Model Training

Before introducing the types of demonstrations in training, we formulate the general ICL for machine translation scenarios. With K-shot demonstrations $\boldsymbol{C}^K$, the probability of target sentence $\boldsymbol{y}$ is:

$$p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{C}^K) = \prod_{t=1}^{T} p(y_t|\boldsymbol{x}, \boldsymbol{y}_{<t}, \boldsymbol{C}^K) \quad (1)$$

Sentence-level Demonstrations The sentence-to-sentence translation is the basic setting of MT. Therefore, we illustrate the formulation of sentence-level demonstrations for translation tasks. Specifically, the K-shot sentence-level demonstrations $\boldsymbol{C}_S^K$ is concatenated by K sentence-pairs as:

$$\boldsymbol{C}_S^K = |t(\boldsymbol{x}^1, \boldsymbol{y}^1); ...; t(\boldsymbol{x}^K, \boldsymbol{y}^K)| \quad (2)$$

where t denotes the format of one written demonstration and ";" denotes the separator between demonstrations. Accordingly, the probability of translation is $p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{C}_S^K)$. During training, the sentence-level demonstrations are randomly sampled from the whole training data.

Document-level Demonstrations Sentence-level demonstrations may limit the use of the demonstration-aware MT model on document-level translation since sentence-level demonstrations lack contextual information. Therefore, we introduce document-level demonstrations, which are the continuous sentence pairs in a document. The organization format of document-level demonstrations follows the sentence-level demonstrations.

However, different from the sentence-level demonstrations, the composition of the i-th instance $(\boldsymbol{x}^i, \boldsymbol{y}^i)$ in document-level demonstrations should be the continuous span within a specific document:

$$\boldsymbol{C}_D^{i,K} = |t(\boldsymbol{x}^{i-K}, \boldsymbol{y}^{i-K}); ...; t(\boldsymbol{x}^{i-1}, \boldsymbol{y}^{i-1})| \quad (3)$$

Mixture of Two Types Since we propose to consider both the sentence-level and document-level demonstrations for different settings of translation, the training data should include two types of demonstrations. We consider randomly choosing one type of demonstration for the training sample $(\boldsymbol{x}, \boldsymbol{y})$. Specifically, the choosing process of demonstrations $\boldsymbol{C}^K$ follow Bernoulli distribution:

$$\boldsymbol{C}^K = \begin{cases} \boldsymbol{C}_S^K, & q \\ \boldsymbol{C}_D^K, & 1-q \end{cases} \quad (4)$$

where q is the probability of choosing sentence-level demonstrations $\boldsymbol{C}_S^K$ during the training process. The probability of choosing document-level demonstrations $\boldsymbol{C}_D^K$ is $1-q$.

Lightweight LoRA Fine-tuning In this paper, we introduce extra modules for continually lightweight finetuning. Given the parameters θ of the base model, we add extra tunable parameters θ_m into the model. Consequently, the training loss of training sample $(\boldsymbol{x}, \boldsymbol{y})$ is:

$$L_{ce} = -\log p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{C}^J, \theta^*, \theta_m) \quad (5)$$

where $J \sim \text{Uniform}\{1, 2, ..., K\}$, and $*$ denote that the parameters of a well-trained LTM θ is frozen during training. As we use the light-weight LoRA finetuning, the comparison of parameter size between θ and θ_m should satisfy:

$$\theta_m \ll \theta \quad (6)$$

More importantly, the vanilla sentence-level translation performance will not be affected by the newly introduced demonstration-aware translation knowledge by parameter deactivation.

Role	Example
Demonstrations $\boldsymbol{C}^K$	EN: It's been a long day without you, my friend ZH: 没有老友你的陪伴 日子真是漫长 EN: And I'll tell you all about it when I see you again ZH: 与你重逢之时 我会敞开心扉倾诉所有
Source Sentence $\boldsymbol{x}^i$	EN: We've come a long way from where we began
Target Sentence $\boldsymbol{y}^i$	ZH: 回头凝望 我们携手走过漫长的旅程

Demonstration Type	Example
Sentence-level $\boldsymbol{C}_S^K$	EN: The moon shines on my bed brightly ZH: 床前明月光 EN: East or west, home is the best ZH: 东也好，西也好，还是家最好
Document-level $\boldsymbol{C}_D^{i,K}$	EN: It's been a long day without you, my friend ZH: 没有老友你的陪伴 日子真是漫长 EN: And I'll tell you all about it when I see you again ZH: 与你重逢之时 我会敞开心扉倾诉所有
Retrieval-based $\boldsymbol{C}_R^K$	EN: We've come a long way from where we started ZH: 我们已经从最初的起点走了很长的路 EN: Reflecting on how far from where we began ZH: 反思我们从起点来了多远
Hybrid $\boldsymbol{C}^{i,K}$	EN: We've come a long way from where we started ZH: 我们已经从最初的起点走了很长的路 EN: And I'll tell you all about it when I see you again ZH: 与你重逢之时 我会敞开心扉倾诉所有

Table 1: Our proposed four types of demonstrations for the given example $(\boldsymbol{x}^i, \boldsymbol{y}^i)$. **Model Training**: $\boldsymbol{C}_S^K$ are demonstrations we randomly select from training dataset, and $\boldsymbol{C}_D^{i,K}$ are demonstrations that precede $(\boldsymbol{x}^i, \boldsymbol{y}^i)$ according to the sentence order. **Model Inference**: $\boldsymbol{C}_R^K$ refers to the demonstrations selected through retrieval methods for their high similarity to the $(\boldsymbol{x}^i, \boldsymbol{y}^i)$. $\boldsymbol{C}^{i,K}$ is a hybrid of demonstrations based on similarity retrieval and self-generated contextual demonstrations at inference stage.

3.3 Demonstration-Aware Model Inference

After demonstration-aware training, the model has enhanced its ability to understand context. Through ICL and the use of appropriate samples, we can make the model applicable to a wider range of scenarios. For this purpose, we have considered two scenarios.

Domain Translation Typically, translation models are trained on limited data, often restricted to a specific domain, leading to sub-optimal performance in other domains. Even with ICL, it might not yield satisfactory results and could even have negative effects. We use the R-BM25 (Agrawal et al., 2022) method to select translation pairs with high domain similarity to the sentence to be translated as our distractions. R-BM25, a method based on n-gram matching for linguistic similarity, enables the selection of superior data as few-shot examples. Subsequently, utilizing our enhanced model for inference, we could further enhance domain translation. Superficially, we use *retrieval-based demonstrations* for domain translation, satisfying the equation $\boldsymbol{C}^K = \boldsymbol{C}_R^K$. With retrieving scoring function R and the retrieving corpus $\boldsymbol{Z}$, the retrieved R-BM25 demonstrations is:

$$\boldsymbol{C}_R^K = \underset{(\boldsymbol{x}_z, \boldsymbol{y}_z) \in \boldsymbol{Z}}{\text{top-}K} \{R(\boldsymbol{x}, \boldsymbol{x}_z)\} \tag{7}$$

Document-level Translation Document-level translation often involves significant associations between sentences, yet leveraging their contextual relationships for inference can be challenging. To address this, we introduce an online adaptation approach where we use the sentences immediately preceding the source sentence and their translations generated by LTM within the same document as document-level demonstrations for ICL when translating a new sentence. However, the effectiveness of using self-generated sentences as few-

shot prompts can be significantly influenced by the model's translation capabilities, which might not effectively grasp the document's translation style. To mitigate this issue, we further augment the model inference with additional retrieval-based demonstrations. This assists the model in achieving better translations. By combining these two types of demonstrations, which we call it *hybrid demonstrations*, LTM can leverage the document's contextual information and ensure translation quality with specific information. With hybrid demonstrations $\boldsymbol{C}^{i,K}$, the translation $\widehat{\boldsymbol{y}}^i$ of i-th sentence $\boldsymbol{x}^i$ is generated as:

$$\widehat{\boldsymbol{y}}^i = \arg\max_{\boldsymbol{y}} p(\boldsymbol{y}|\boldsymbol{x}^i, \boldsymbol{C}^{i,K}) \tag{8}$$

The hybrid demonstrations are concatenated by the retrieval-based demonstrations $\boldsymbol{C}_R^U$ and document-level demonstrations $\boldsymbol{C}_D^V$. Therefore, the final decoding criterion with hybrid demonstration is:

$$\boldsymbol{C}^{i,K} = \boldsymbol{C}_R^U; \boldsymbol{C}_D^{i,V} \tag{9}$$

where u and v represent the number of sentence-level and document-level demonstrations respectively. Given the number of total demonstrations K, we have $K = U + V$. Finally, by utilizing our enhanced model with hybrid demonstrations for inference, we achieve better document-level translation.

3.4 Discussion

Our approach employs lightweight LoRA training, facilitating easy deployment on current LTM. For training data, we exclusively reconstruct the original dataset used during the model's initial training phase, without introducing any additional data. This method not only optimizes the use of existing training data but also simplifies implementation.

4 Experiments

4.1 Data

Training and Development Data In our experiments, there is no need to introduce additional data during training and we simply utilize the training dataset used for instruction fine-tuning of the LTM models. Following the approach of ALMA (Xu et al., 2024), during the instruction fine-tuning phase, we employed training data from the WMT'17 to WMT'20 test datasets, along with the development and test sets from Flores-200. This includes 58,702 training examples across 10 translation directions: cs-en, de-en, is-en, zh-en, ru-en. Test data from WMT'21 are used for the development dataset.

Test Data of Domain Translation For evaluating domain translation, we utilized the multi-domain German-English corpus introduced by Koehn and Knowles (2017), encompassing textual data across five distinct domains: subtitles, medical, law, Koran, and IT.

Test Data of Document-Level Translation For the evaluation of document-level translation, we employed two datasets. The Fiction dataset, derived from mZPRT (Xu et al., 2022), consists of 24 chapters across five genres of books, sourced from Chinese and English Webnovel websites. These chapters have been manually aligned to create parallel Chinese-English sentence pairs. The GuoFeng (Wang et al., 2023) dataset is a Webnovel Corpus, comprising works originally written in Chinese by novel writers and subsequently translated into English by professional translators. This dataset encompasses 22,567 continuous chapters from 179 web novels, spanning 14 genres, including fantasy science and romance.

4.2 Model Configuration

We selected the prompting-based LTMs `ParroT-7B`[1] and `BayLing-13B`[2], and the tuning-based LTMs `LLaMA-2-7B`[3] and `ChatGLM3-6B`[4] (Du et al., 2022) to evaluate the performance of LTM in domain and document-level translation, as well as their ability to perceive demonstrations. We used the representative tuning-based LTM `ALMA`[5] as our backbone models due to its effectiveness. ALMA is an LTM based on the MT-centered instruction tuning of LLaMA-2-7B and 13B, undergoes an initial fine-tuning process on monolingual data, followed by further fine-tuning on a small set of high-quality parallel data and has good multilingual translation capabilities.

4.3 Training

During our training process, we implemented the methodology described in §3.2, constructing the

[1] https://huggingface.co/wxjiao/ParroT-7b
[2] https://huggingface.co/ICTNLP/bayling-13b-v1.1
[3] https://huggingface.co/meta-llama/Llama-2-7b
[4] https://huggingface.co/THUDM/chatglm3-6b
[5] https://huggingface.co/haoranxu/ALMA-7B

Model	Domain						Document-level		
	Medical	Law	Koran	Subtitles	IT	Avg.	Fiction	GuoFeng	Avg.
Prompting-based LTM									
LLaMA-2-7B	82.58	82.73	70.58	78.49	81.36	79.15	14.56	19.55	17.06
w/ Demonstrations	84.91	85.27	71.68	78.74	85.15	81.15	14.95	24.40	19.68
ChatGLM3-6B	81.00	80.15	67.88	75.94	80.45	77.08	14.96	22.55	18.76
w/ Demonstrations	83.11	83.44	71.21	75.62	82.96	79.27	15.67	23.89	19.78
Tuning-based LTM									
ParroT-7B	82.39	82.32	70.28	77.79	79.75	78.51	12.55	15.19	13.87
w/ Demonstrations	80.16	79.46	64.03	72.73	81.16	75.50	9.94	10.90	10.42
BayLing-13B	81.67	81.99	69.90	77.77	79.23	78.11	12.66	18.55	15.61
w/ Demonstrations	84.26	85.78	72.52	77.50	84.55	80.92	13.44	17.85	15.65
Our Recipe with Base Model: ALMA-7B									
Vanilla LTM-7B	83.30	83.40	71.92	79.52	82.09	80.05	17.08	22.51	19.80
w/ Demonstrations	84.49	85.69	71.44	79.15	85.45	81.24	15.50	20.96	18.23
Demonstration-Aware LTM	**85.10**	**86.48**	**72.79**	**79.97**	**85.96**	**82.06**	**18.86**	**23.02**	**20.94**
Our Recipe with Base Model: ALMA-13B									
Vanilla LTM-13B	83.58	84.09	72.59	79.67	82.29	80.44	16.68	23.62	20.15
w/ Demonstrations	84.95	85.93	71.02	79.19	85.36	81.29	19.84	22.79	21.32
Demonstration-Aware LTM	**85.06**	**86.66**	**72.73**	**80.00**	**86.22**	**82.13**	**20.15**	**24.15**	**22.15**

Table 2: COMET scores for domain translation and d-BLEU scores for document-level translation of prompting-based LTM, tuning-based LTM and our demonstration-aware LTM. We use 5-shot retrieval-based demonstrations for domain translation and 5-shot hybrid demonstrations for document-level translation.

sentence-level demonstrations and the document-level demonstrations for the original training data. We set the parameter q to 0.5, opting for an equal probability in selecting demonstrations, before training with the newly constructed data as our final training dataset. We employed lightweight LoRA technique and set the hyperparameters of LoRA (r, α) to $(16, 32)$, updating 0.1% of the parameters. The batch size is 256, the warm-up ratio is 0.01, and the max-tokens are 1,280. The model underwent training for 1 epoch, which was sufficient to observe significant convergence. This training was conducted on 8 Nvidia A800 GPUs, utilizing DeepSpeed ZeRO stage 2 for model parallel training.

4.4 Evaluation

During the inference phase, to assess the ICL capabilities of the LTM, we engaged in both zero-shot and few-shot translation. Utilizing the methodology delineated in §3.3, we crafted demonstrations for reasoning. For all few-shot reasoning activities, we standardized the number of demonstrations K to 5. Within the realm of domain translation, due to the test data's absence of logical contextual linkage, we selected the top five demonstrations with the highest R-BM25 scores as our few-shot examples. For the document-level Translation, we adjusted U to 2 and V to 3, thereby opting for 2 retrieval-based demonstrations alongside 3 document-level demonstrations. To generate the most optimal outcomes, we employed beam search, setting the beam size at five. In evaluating the inference results, we follow the common practice of adopting COMET (Rei et al., 2020) for appraising the quality of domain translation and utilized document-level sacreBLEU (d-BLEU, Liu et al., 2020) for measuring the outcomes of our document-level translation efforts.

4.5 Results

Table 2 displays the main results of demonstrations-Aware LTM and Vanilla LTM. According to the main results, we have several findings:

Performance of tuning-based LTM may be hurt by demonstrations. Comparing the performance of zero-shot LTM and LTM with demonstrations, we observe that prompting-based LTM can effectively utilize demonstrations to improve translation performance. However, demonstrations significantly affect the performance of tuning-based LTM on almost all datasets, except for the IT dataset. This suggests that although tuning-based LTM improves translation through parameter adjustments during training, the inclusion of demonstrations during inference can introduce biases, leading to

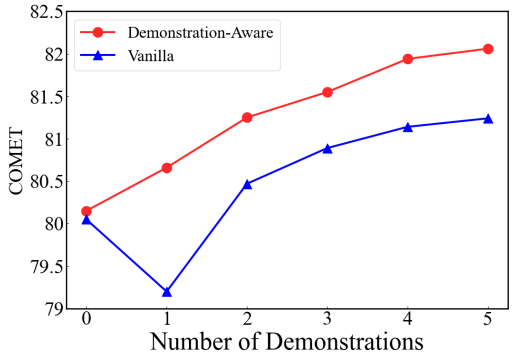

Figure 1: COMET scores across different quantities of demonstrations in the domain translation for Vanilla LTM-7B and Demonstration-Aware LTM.

performance degradation. These results indicate a critical limitation of tuning-based LTMs in their ability to effectively leverage demonstrations.

Demonstrations-aware LTM can learn from demonstrations significantly and stably. The demonstration-aware LTM consistently outperforms baseline models on all datasets, demonstrating significant and stable improvements in ICL ability brought by demonstrations. In domain and document-level translation, the demonstration-aware LTM effectively integrates high-quality retrieved examples with self-generated contextual translations, resulting in more accurate and contextually appropriate outputs. This demonstration-aware model training strategy enriches the LTM's ICL ability, enabling it to better utilize demonstrations during inference.

Demonstrations-aware LTM boost in-context ability across different model sizes. We observe that demonstrations-aware LTM beat vanilla LTM on both the 7B variant and the 13B variant on all the datasets. These results demonstrate the stable superiority of demonstrations-aware LTM across different model sizes.

5 Analysis

5.1 Performance across X-Shot Demonstrations

Since the number of demonstrations may vary according to different conditions, here arises a question: *Can demonstration-aware LTM achieve consistent performance gains on different numbers of demonstrations?* To answer this question, we conducted tests with 1 to 5 demonstrations in the domain translation. Figure 1 displays the results of the comparison between Vanilla LTM-7B and Demonstration-Aware LTM at the 7B scale across 1 to 5-shot demonstrations. We can observe that Vanilla LTM-7B equipped with the 1-shot demonstration would even drop the performance in contrast to the zero-shot Vanilla LTM-7B, indicating the failures of Vanilla LTM-7B when exploiting ICL. Demonstration-aware LTM outperforms the Vanilla LTM-7B on all settings, indicating that demonstration-aware LTM can boost translation quality consistently on different numbers of demonstrations. To provide a more comprehensive analysis, we further evaluated the domain translation scores across all possible combinations of Vanilla LTM-7B with 1 to 5-shot demonstrations. The averaged COMET scores across all five domains are reported in Appendix A.2, offering detailed insights into the performance variations.

5.2 Effect of Mixing Two Types of Demonstrations during Training

Models	Domain	Document
Sentence	82.00	20.62
Document	81.16	20.21
Mixture	**82.06**	**20.94**

Table 3: COMET and d-BLEU scores on models using different training data with diverse types of demonstrations. "Mixture" represents our proposed method that mixing both the sentence and document-level demonstrations during training.

To validate the effect of mixing sentence-level and document-level demonstrations, we trained demonstration-aware LTMs using each type of demonstration independently and assessed the differences in model performance. Table 3 displays the results. We can observe that the model trained with only sentence-level demonstrations outperforms the model trained only with document-level demonstrations. However, the model trained with the mixture of two types of demonstrations performs better than the model trained with a single type of demonstration, indicating that mixing two types of demonstrations during training could help the training process. This could be explained as that the mixture of two types of demonstrations could prevent the model from over-optimizing to only one type of demonstration.

Models	Domain		Document	
	5-shot	0-shot	5-shot	0-shot
Full	82.08	80.16	21.04	19.64
LoRA	82.06	80.05	20.94	19.80

Table 4: COMET and d-BLEU scores for full training and LoRA training.

5.3 Full Fine-tuning vs. LoRA

To confirm the efficiency brought by lightweight modular LoRA, we also implement the variant of demonstration-aware LTM trained with full-parameter fine-tuning. Table 4 shows that the LoRA variant achieves competitive performance compared to the full-tuned variant with slight score differences, indicating that the LoRA variant is efficient enough. Therefore, we recommend the LoRA variant as our default method.

5.4 Necessity of Using Hybrid Demonstrations on Document-Level Translation

Models	Fiction	GuoFeng
Retrieval-based	18.14	22.93
Document-level	17.73	21.47
Hybrid	**18.86**	**23.02**

Table 5: d-BLEU scores for document-level translation using various types of demonstrations. "Hybrid" represents our proposed method that concatenate the retrieval-based demonstrations and document-level demonstrations during inference.

To ensure the necessity of introducing retrieval-based demonstrations for document-level translation, we conducted tests on document-level translation using retrieval-based and document-level demonstrations of the same length separately. The results in table 5 display that our hybrid demonstrations are the optimal choice. Using only the document-level demonstrations falls behind the retrieval-based demonstrations, which could be attributed to the document-level demonstration's lack of specific information to guarantee the translation quality. We conclude that introducing hybrid demonstrations is important for document-level translation.

5.5 Robustness of Tolerating Noisy Demonstrations

The robustness of models should be paid attention to since noisy demonstrations may damage the performance of demonstration-aware model inference. We test models in the medical domain and simulate the noisy demonstrations with random word substitution. Accordingly, the clean demonstrations are the original demonstrations. Besides, we also implement the model variants trained with consistency loss with data augmentation since consistency loss is a widely used method to enhance the robustness of prior works. Specifically, the consistency loss with data augmentation on demonstrations could be formulated as:

$$L_{con} = \frac{1}{2} KL(p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{C}^K) \| p(\boldsymbol{y}|\boldsymbol{x}, \widetilde{\boldsymbol{C}}^K)) + \frac{1}{2} KL(p(\boldsymbol{y}|\boldsymbol{x}, \widetilde{\boldsymbol{C}}^K) \| p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{C}^K)) \tag{10}$$

where $\widetilde{\boldsymbol{C}}^K$ represents the demonstrations processed by data augmentation. With weight α for consistency loss, the final training criterion with consistency loss is:

$$L_{final} = L_{ce} + \alpha L_{con} \tag{11}$$

Considering the possible data augmentations, we implement three types of models trained with consistency loss and present specific examples for each type of data augmentation in Appendix A.3:

- **Token Augmentation** is the model variant augmented by token-level substitution on demonstrations. After tokenizing the data, we create a copy for noise addition purposes. For the word vectors corresponding to the positions of demonstrations, we randomly replace word vectors based on a Bernoulli distribution with a 1% probability, resulting in approximately 10% of the data being altered. Considering the magnitudes of our cross-Entropy loss and consistency loss, we set α to 0.02 to maintain a reasonable loss ratio.

- **Sentence Augmentation** is the model variant trained with sentence replacement on demonstrations. To achieve sentence-level robustness, we create a copy of the original demonstration data, in which 10% of the target sentences in the demonstrations are replaced with either their source sentences or randomly with other target sentences. We then feed both sets of data into the model for training. Considering the significant difference in data distribution caused by sentence-level noise, we set the parameter α to 0.001 for consistency learning.

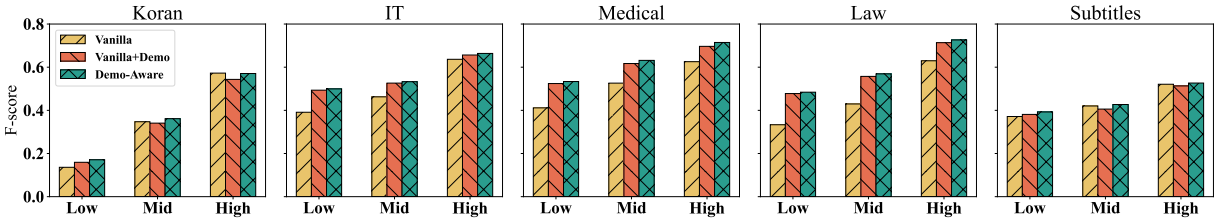

Figure 2: Word translation accuracy versus frequency on domain translation.

Models	Clean	Noisy
Vanilla LTM-7B	83.30	-
+ Demonstrations	84.49	75.37
Token Augmentation	84.90	75.72
Sentence Augmentation	84.72	75.16
Shot Augmentation	84.97	75.28
Demonstration-Aware LTM	**85.10**	**75.73**

Table 6: Evaluating the performance and robustness of models using COMET scores across various approach variants on the medical domain translation task.

- **Shot Augmentation** is the model trained to be consistent among samples equipped with different numbers of demonstrations. We set the parameter α to 0.0001 for consistency learning.

According to the results in Table 6, by comparing the decrease in COMET scores when using noisy demonstrations, we find that only the token augmentation LTM, which shows a decrease of 9.18 points, exhibits a slight improvement in robustness compared to our demonstration-aware LTM, which shows a decrease of 9.38 points. However, this improvement comes at the cost of overall translation performance. Additionally, training with these augmentation methods proves to be more complex and time-consuming. Overall, our proposed demonstration-aware LTM strikes the right balance between performance and practicality.

5.6 Performance Stability across Word Frequency

To understand the improvement brought by our method, we divided the word frequency into three intervals: Low ($[0, 50)$), Middle ($[50, 1000)$), and High ($[1000, +\inf)$). The word frequency is calculated from the training set of the WMT19 De-En translation benchmark, which is treated as the general domain data. We display the statistics of word translation accuracy versus word frequency of the domain translation in Figure 2. We find that incorporating demonstrations helps translate rare words according to the behaviors on the Low bucket. This phenomenon may be due to the necessary specific information provided by demonstrations when translating rare words. We can also observe that Vanilla LTM-7B's performance issues due to demonstrations mostly occur with middle and high-frequency word translations, especially in the Koran and Subtitles datasets. In contrast, the demonstration-aware LTM avoids these performance issues and shows consistent improvements across all word frequencies and datasets.

6 Conclusion

To enhance the demonstration learning capability of tuning-based LTM, this paper introduces the demonstration-aware LTM. We crafted sentence-level and document-level demonstrations to better enable the LTM to perceive and utilize demonstrations. Following these developments, we also employed retrieval and self-generation techniques to create high-quality demonstrations that fully leverage the LTM's capacity for model inference. Our experimental results in domain translation and document-level translation demonstrate that our proposed method enables tuning-based LTM to learn from demonstrations significantly and stably.

Limitation

While the proposed demonstration-aware LTM can effectively enhance the model's ability to perceive demonstrations, leading to improved translation performance, it still has some limitations: (1) For the selection of the two types of demonstrations in the training data, we merely mixed them based on a certain probability without investigating the impact of different data ratios; (2) We trained our model using multiple languages, but did not fully evaluate our model across a diverse set of languages; (3) We selected only one type of tuning-based LTM as our Vanilla LTM, which is not sufficient to prove the widespread applicability of our method.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China (Grant Nos. 62206076, 62336008, 62261160648), Guangdong Basic and Applied Basic Research Foundation (Grant No. 2024A1515011491), Shenzhen Science and Technology Program (Grant Nos. ZDSYS20230626091203008, KJZD20231023094700001, RCBS20221008093121053), Shenzhen College Stability Support Plan (Grant Nos. GXWD20220811173340003, GXWD20220817123150002), the Science and Technology Development Fund, Macau SAR (Grant Nos. FDCT/060/2022/AFJ, FDCT/0070/2022/AMJ), the Multi-year Research Grant from the University of Macau (Grant No. MYRG-GRG2023-00006-FST-UMDF), and Tencent AI Lab Rhino-Bird Gift Fund (Grant No. EF2023-00151-FST). Xuebo Liu was sponsored by CCF-Tencent Rhino-Bird Open Research Fund. We would like to thank the anonymous reviewers and meta-reviewers for their insightful suggestions.

References

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

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *ArXiv preprint*, abs/2302.04023.

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 *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2024. Exploring human-like translation strategy with large language models. *Transactions of the Association for Computational Linguistics*, 12:229–246.

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. *ArXiv preprint*, abs/2302.09210.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

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

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Sai Koneru, Miriam Exel, Matthias Huck, and Jan Niehues. 2023. Contextual refinement of translations: Large language models for sentence and document-level post-editing. *ArXiv preprint*, abs/2310.14855.

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.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

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*, abs/2305.06575.

Zhuoyuan Mao and Yen Yu. 2024. Tuning llms with contrastive alignment instructions for machine translation in unseen, low-resource languages. *ArXiv preprint*, abs/2401.05811.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-task generalization via natural language crowdsourcing instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3470–3487, Dublin, Ireland. Association for Computational Linguistics.

Keqin Peng, Liang Ding, Yancheng Yuan, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2024. Revisiting demonstration selection strategies in in-context learning. *arXiv preprint arXiv:2401.12087*.

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. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5622–5633, Singapore. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Nathaniel R Robinson, Perez Ogayo, David R Mortensen, and Graham Neubig. 2023. Chatgpt mt: Competitive for high-(but not low-) resource languages. *ArXiv preprint*, abs/2309.07423.

Weiting Tan, Haoran Xu, Lingfeng Shen, Shuyue Stella Li, Kenton Murray, Philipp Koehn, Benjamin Van Durme, and Yunmo Chen. 2023. Narrowing the gap between zero-and few-shot machine translation by matching styles. *ArXiv preprint*, abs/2311.02310.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

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*, abs/2307.09288.

Longyue Wang, Zhaopeng Tu, Yan Gu, Siyou Liu, Dian Yu, Qingsong Ma, Chenyang Lyu, Liting Zhou, Chao-Hong Liu, Yufeng Ma, et al. 2023. Findings of the wmt 2023 shared task on discourse-level literary translation: A fresh orb in the cosmos of llms. *ArXiv preprint*, abs/2311.03127.

Yutong Wang, Jiali Zeng, Xuebo Liu, Fandong Meng, Jie Zhou, and Min Zhang. 2024. Taste: Teaching large language models to translate through self-reflection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Jason Wei, Maarten Bosma, Vincent Y. 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 *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. *ArXiv preprint*, abs/2304.12244.

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*.

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

Jiali Zeng, Fandong Meng, Yongjing Yin, and Jie Zhou. 2023. Tim: Teaching large language models to translate with comparison. *ArXiv preprint*, abs/2307.04408.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023a. 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. 2023b. Bayling: Bridging cross-lingual alignment and instruction following through interactive translation for large language models. *ArXiv preprint*, abs/2306.10968.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.

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

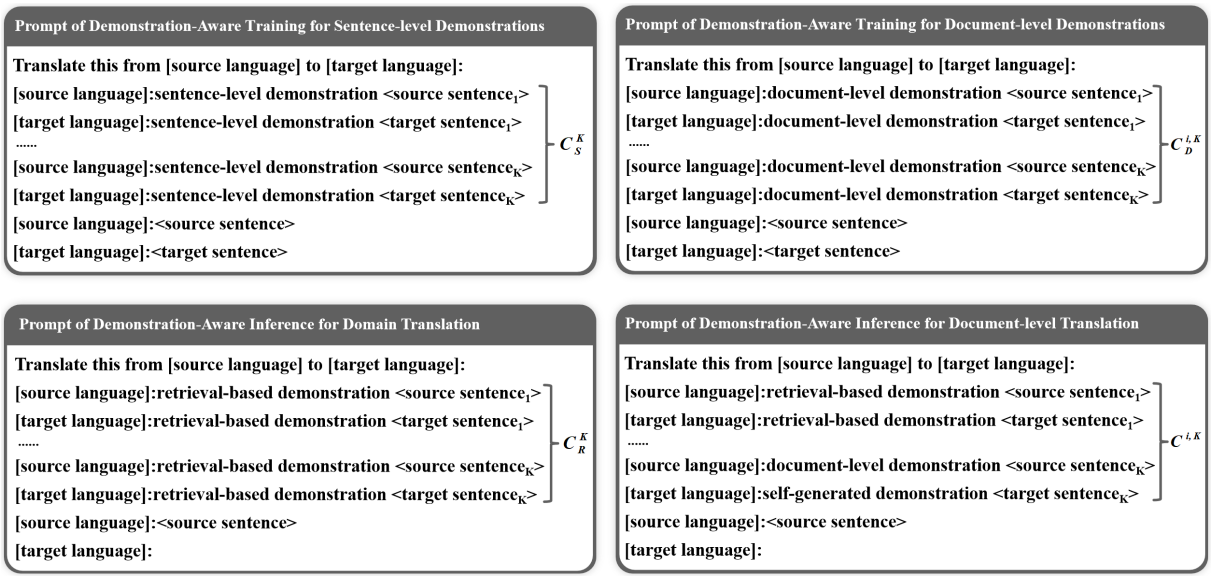

Figure 3: Prompts for four types of demonstrations. **Model Training**: C_S^K are demonstrations randomly selected from the training dataset, and $C_D^{i,K}$ are demonstrations that precede source sentence according to the sentence order. **Model Inference**: C_R^K refers to the demonstrations selected through retrieval methods for their high similarity to the source sentence. $C^{i,K}$ is a hybrid of demonstrations based on similarity retrieval and self-generated contextual demonstrations at the inference stage.

A Appendix

A.1 Prompts for Four Types of Demonstrations

Table 1 in Section 3 presents our proposed four types of demonstrations. The specific prompt forms for these four types of demonstrations during the training and inference stages are shown in Figure 3. For each type of demonstration, we use the same translation instructions to ensure that the LTM clearly understands the translation task and direction. These prompt forms guide the LTM to effectively utilize the demonstrations, thereby achieving better translation performance.

A.2 Performance across Different Combinations of X-Shot Demonstrations

Regarding the number of demonstrations, we analyze the impact of different quantities of demonstrations on LTM performance in Section 5.1, with the results presented in Figure 1. The findings indicate that both Vanilla LTM-7B and Demonstration-aware LTM perform better with multiple demonstrations than in a zero-shot scenario, achieving optimal results with five demonstrations. Building upon this, we further analyze the effect of the number of demonstrations on Vanilla LTM-7B. We evaluate the domain translation scores across all possible configurations of Vanilla LTM-7B with varying numbers and compositions of demonstrations and report the averaged COMET scores over all 5 domains. The results are presented in Table 7.

The suffix numbers in the table represent the rank of the demonstrations selected based on their scores from the R-BM25 method, within the top 5 scoring demonstrations. Our results indicate that Vanilla LTM achieves better translation performance with an increased number of demonstrations. Moreover, selecting the top-ranked demonstrations, namely those with the highest R-BM25 scores, results in better translation. The performance of Vanilla LTM with multiple demonstrations surpasses that of the zero-shot scenario. This demonstrates that our method of selecting demonstrations and the numbers used is both reasonable and effective for Vanilla LTM, thus not leading to any bias. Furthermore, when using the same demonstrations, our demonstration-aware LTM shows consistently better performance across different numbers of demonstrations compared to Vanilla LTM. This further validates the effectiveness of our approach.

A.3 Examples of Noisy Demonstrations

In Section 5.5, to discuss the robustness of the method and the effect of using consistency learning, we adopted three forms of data augmentation. Table 8 presents examples of each augmentation. Token augmentation involves randomly replacing

Vanilla LTM				
4shot-1234	**4shot-1235**	**4shot-1245**	**4shot-1345**	**4shot-2345**
81.14	81.01	80.81	80.87	80.61
3shot-123	**3shot-124**	**3shot-125**	**3shot-134**	**3shot-135**
80.89	80.79	80.73	80.63	80.65
3shot-145	**3shot-234**	**3shot-235**	**3shot-245**	**3shot-345**
80.61	80.36	80.35	80.4	80.18
2shot-12	**2shot-13**	**2shot-14**	**2shot-15**	**2shot-23**
80.47	80.16	80.02	79.89	79.83
2shot-24	**2shot-25**	**2shot-34**	**2shot-35**	**2shot-45**
79.73	79.75	79.51	79.42	79.41
1shot-1	**1shot-2**	**1shot-3**	**1shot-4**	**1shot-5**
79.20	78.42	78.11	78.02	77.93
Demonstration-aware LTM				
1shot	**2shot**	**3shot**	**4shot**	**5shot**
80.66	81.25	81.55	81.94	82.06

Table 7: COMET scores across different quantities and combinations of demonstrations in the domain translation. The suffix numbers represent the rank of the demonstrations selected based on their scores from the R-BM25 method, within the top 5 scoring demonstrations.

certain tokens within demonstrations with a 1% probability using a Bernoulli distribution. In our examples, the token "的" was randomly replaced with "和", and the token "倾" was replaced with "可". Sentence augmentation replaces sentences within demonstrations, where 10% of the target sentences are replaced with their source sentences or randomly with other target sentences. For the examples we provided, the target sentence was replaced with the source sentence "It's been a long day without you, my friend" in the first demonstration, and the target sentence "与你重逢之时 我会敞开心扉倾诉所有" was randomly replaced with another target sentence "与回头凝望 我们携手走过漫长的旅程" in the second demonstration. Shot augmentation modifies the number of demonstrations by randomly increasing or decreasing the number of demonstrations. In our examples, we randomly deleted the first demonstration.

Role	Example
Demonstrations	EN: It's been a long day without you, my friend ZH: 没有老友你的陪伴 日子真是漫长 EN: And I'll tell you all about it when I see you again ZH: 与你重逢之时 我会敞开心扉倾诉所有

Augmentation Type	Example
Token augmentation	EN: It's been a long day without you, my friend ZH: 没有老友你和陪伴 日子真是漫长 EN: And I'll tell you all about it when I see you again ZH: 与你重逢之时 我会敞开心扉可诉所有
Sentence Augmentation	EN: It's been a long day without you, my friend ZH: It's been a long day without you, my friend EN: And I'll tell you all about it when I see you again ZH: 与回头凝望 我们携手走过漫长的旅程
Shot Augmentation	EN: And I'll tell you all about it when I see you again ZH: 与你重逢之时 我会敞开心扉倾诉所有

Table 8: Examples of the proposed three data augmentations. Token augmentation involves replacing tokens within demonstrations. Sentence augmentation replaces entire sentences within demonstrations, either substituting the target sentences with their source sentences or randomly with other target sentences. Shot augmentation modifies the number of demonstrations by randomly increasing or decreasing the number of demonstrations.