# Expert Margin Optimization: Enhancing Multi-Domain Translation Capabilities of LLM with MoE-LoRA

Anonymous ACL submission

## Abstract

In the realm of machine translation utilizing Large Language Models (LLMs), the standard workflow involves Cross-Lingual Alignment learning followed by Instruction-tuning. Low-Rank Adaptation (LoRA) has been a widely-used and effective method for fine-tuning LLMs. However, LoRA alone exhibits limited benefits when confronted with multi-task or multi-domain scenarios. Given the prevalent existence of multi-domain challenges in machine translation, this paper focuses on enhancing the multi-domain translation capabilities of LLMs. We extend LoRA to Mixture of Experts (MoE) architecture, defined as MoE-LoRA, to address domain conflicts in multi-domain settings. Our approach involves introducing MoE-LoRA solely at higher layers to target specific domain-related knowledge acquisition, preceded by General Cross-Lingual Alignment during the training process. Particularly, we propose a methodology called Expert Margin Optimization (EMO) to facilitate the transfer of additional knowledge from other domains to enhance the inputs specific to a domain. Experimental validations conducted on the English-to-German and English-to-Chinese translation directions using the Llama2-7B and Llama3-8B models demonstrate consistent improvements in BLEU and COMET scores, highlighting the efficacy of our proposed approach.

## 1 Introduction

Generative (decoder-only) large language models (LLMs), such as GPT models (Brown et al., 2020), PaLM (Chowdhery et al., 2023), LLaMA (Touvron et al., 2023a,b), and others, have garnered significant attention and shown substantial progress in the field, capturing the fascination of researchers. Leveraging fine-tuning techniques, LLMs have exhibited remarkable performance across a spectrum of NLP tasks. Among these techniques, LoRA (Hu et al., 2022) (Low-Rank Adaptation) has emerged as an efficient and widely adopted method for fine-tuning. LoRA involves freezing the pretrained model weights and incorporating trainable rank decomposition matrices within each layer of the Transformer architecture, leading to a significant reduction in the number of trainable parameters for downstream tasks. However, as underscored by recent studies (Biderman et al., 2024), "LoRA Learns Less and Forgets Less." While LoRA excels in preserving the inherent capabilities of the base model and mitigating forgetfulness, it tends to acquire comparatively less performance improvement than full fine-tuning.

Machine translation (MT) predominantly relies on encoder-decoder architectures, as evidenced by prominent models like M2M (Fan et al., 2021), MT5 (Xue et al., 2021), and NLLB (Costa-jussà et al., 2022). Although these models have reached a high level of maturity, avoiding overfitting and achieving robust out-of-domain performance remain significant challenges in machine translation. Notably, the state-of-the-art (SOTA) model NLLB covers only four major domains such as news, formal speech, and health. Recently, the field has shifted towards utilizing LLMs for MT. While employing small yet high-quality instruction data for supervised fine-tuning (SFT) of LLMs has shown effectiveness in various NLP tasks (Taori et al., 2023; Touvron et al., 2023b), it poses challenges in the context of translation. Fine-tuning LLMs with a small amount of high-quality translation instruction data still falls short when compared to state-of-the-art encoder-decoder MT models (Yang et al., 2023; Zeng et al., 2023; Zhang et al., 2023).

Guo et al. (2024) introduced a Three-Stages Translation Pipeline (TP3) to boost translation capabilities of LLMs, emphasizing the significance of its second stage, Continual Pre-training with Interlinear Text Format Documents, achieving comparable quality to the leading model NLLB. TP3 also integrates LoRA to enhance its efficiency, with this stage being considered as Cross-Lingual Alignment learning. Building upon this work, we delve into the challenges posed by multi-domain scenarios.

In this study, we extend the concept of LoRA to the Mixture of Experts (MoE) architecture. A typical MoE setup includes multiple expert networks and a router. Here, we view LoRA as an expert, training multiple LoRAs to learn Cross-Lingual Alignment from data across various domains. We hypothesize the existence of commonalities in alignment among different domains, emphasizing the need to learn this shared knowledge first. To address this, we configure the lower layers of the model to utilize a unified LoRA, referred to as General LoRA, while reserving MoE LoRA for the higher layers. Experimental outcomes validate the efficacy of our approach in effectively managing domain conflicts. Our training process encompasses General Cross-Lingual Alignment Pre-training, Domain-Motivated Experts Pre-training and Instruction-Tuning, and Expert Margin Optimization (EMO). For detailed insights into our training strategies, please refer to Section 3.3. For the inference phase, we introduce two modes: one leveraging all trained MoE LoRAs and the other utilizing only the top-$k$ MoE LoRAs to reduce computational overhead.

Our main contributions are summarized as follows:

- To the best of our knowledge, we are the first to introduce MoE-LoRA to Translation upon LLM to alleviate domain conflicts.
- We propose a training strategy that involves initial general pre-training followed by domain-specific training, further enhancing domain performance. By utilizing MoE LoRAs only in the higher layers, we reduce computational complexity while preserving a General LoRA within these LoRAs to retain the acquired common knowledge.
- We introduce a novel Expert Margin Optimization strategy aimed at training a more effective Router policy, ensuring that the combined output of multiple experts consistently outperforms that of a single specific expert.
- Experimental validation conducted on the Llama2-7B and Llama3-8B confirms the effectiveness of our approach, underpinning the robustness of our methodology.

## 2 Background

### 2.1 TP3

Guo et al. (2024) propose a novel training paradigm, consisting of Three-Stages Translation Pipeline (TP3), to boost the translation capabilities of LLMs. The training paradigm includes:

Stage 1: Continual Pre-training using Extensive Monolingual Data. This stage aims to expand the multilingual generation capabilities of LLMs. While it is inherently related to machine translation tasks, it is not essential.

Stage 2: Continual Pre-training with Interlinear Text Format Documents. They construct interlinear text format from sentence-aligned bilingual parallel data and utilize them for continual pre-training of LLMs. Experimental results demonstrate the critical importance of this stage, resulting in a significant improvement in translation quality, particularly for English-Other translations.

Stage 3: Leveraging Source-Language Consistent Instruction for Supervised Fine-Tuning. In this stage, they discover that setting instructions consistent with the source language benefits the supervised fine-tuning process.

As Stage 1 is high compute and not essential, we only use there Stage 2 and Stage 3 in our paper. We consider there Stage 2 as Cross-Lingual Alignment learning.

## 3 Method

In this section, we will begin by outlining the model architecture. Following that, we will delve into the challenges of MoE. Finally, we will introduce the training and inference procedures in detail.

### 3.1 Model Architecture

Our model is illustrated in Figure 1. We employ LoRA for acquiring knowledge of Cross-Lingual Alignment and the instruction for the translation task. In each of the lower $m$ layers, we designate a single LoRA, known as the General LoRA, while in the higher $n$ layers, we expand LoRA into MoE architecture, referred to as MoE-LoRA.

For MoE-LoRA, we define a total of $d + 1$ experts, where one aligns with the lower layers to capture General knowledge, and the other $d$ experts are dedicated to learning the knowledge of d different domains. Additionally, for MoE-LoRA, we train a gate function as the Router. We employ a linear transformation to learn a weight vector, de-

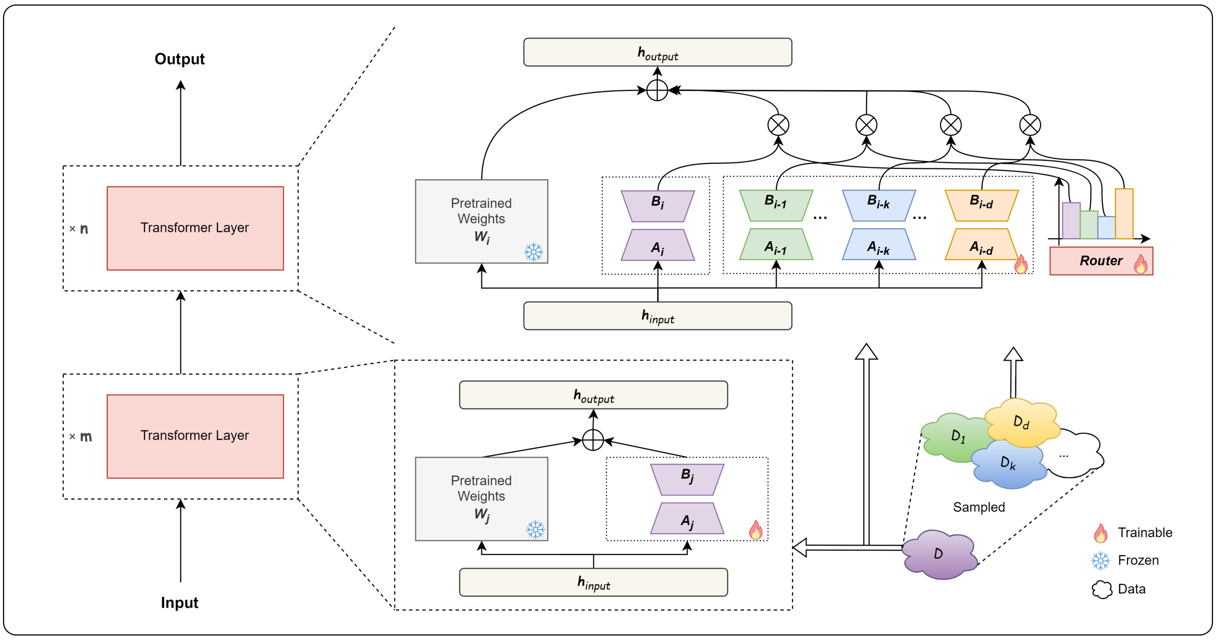


Figure 1: The overall of our model architecture.

noted as $w \in \mathcal{R}^{d+1}$, representing the contribution weights for these experts.

## 3.2 Challenge of MoE

One challenge of the Mixture of Experts (MoE) model is the phenomenon where the gating network displays no preference for any particular expert, resulting in a seemingly random routing process. To address this issue, a straightforward approach is to individually train each expert for specific tasks or domains, and then train a task- or domain-motivated gate as the router. However, this approach essentially reduces the model to a single expert, negating the contributions of different experts. This contradicts the original intent of employing MoE, which aims to decompose large problems into smaller sub-problems, effectively solve these sub-problems with different experts, and then combine their outputs. Therefore, learning an effective routing strategy poses a significant challenge within the MoE architecture.

## 3.3 Training

In the field of translation, it is common practice to conduct pre-training in a general domain before training in a specific domain to prevent overfitting of the model to the particular domain. Therefore, our training process begins with General Cross-Lingual Alignment Pre-training, followed by Domain-Motivated Experts Pre-training and Instruction-Tuning. Finally, we propose a strategy called Expert Margin Optimization to train the Router.

### 3.3.1 General Cross-Lingual Alignment Pre-training

In the General Cross-Lingual Alignment Pre-training phase, we utilize multiple LoRAs modules at each layer of the LLM, collectively referred to as the General LoRA. The weights assigned to the LoRA modules at the $i$-th layer are denoted as $A_i$ and $B_i$ respectively. These modules undergo training on a mixed dataset sampled from various domains. Data processing involves the use of an Interlinear Text Format similar to Stage 2 of TP3. For detailed formatting guidelines, please consult Appendix A.

### 3.3.2 Domain-Motivated Experts Pre-training and Instruction-Tuning

During the Domain-Motivated Experts Pre-training and Instruction-Tuning process, we commence with establishing LoRA experts in the higher $n$ layers, each corresponding to a specific domain. Concurrently, we preserve the General LoRA in these layers as a General Expert, yielding a total of $d+1$ experts per layer, where $d$ represents the total number of domains. The weights of these experts are initialized based on those of the General LoRA. Each Domain LoRA expert is denoted by $A_{i-k}$ and $B_{i-k}$, where $i$ denotes the layer number and $k$ signifies the domain index.

Subsequently, we proceed with Continued Pre-training for each expert using domain-specific data. The training is conducted sequentially with the same data format as in the prior phase. In this stage, the lower $m$ layers of the pre-trained General LoRA are frozen from parameter updates, and a slightly reduced learning rate is employed for training.

In the Instruction-Tuning phase, we focus on training with a compact, high-quality dataset comprising translation instructions spanning various domains. It is important to note that instruction data from different domains is utilized to train the respective domain experts, while a combination of data from all domains is employed to train the General Expert. Throughout this phase, both the weights of the General LoRA and the MoE LoRA are updated simultaneously.

#### 3.3.3 Expert Margin Optimization

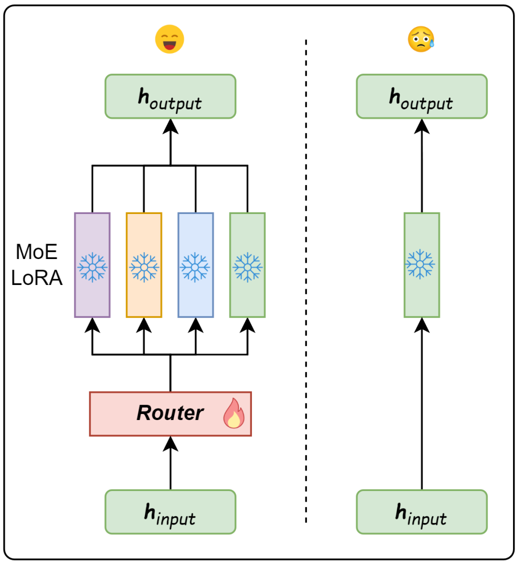


Figure 2: EMO.

We hold all trained LoRAs and pre-trained model parameters constant, focusing exclusively on optimizing the gating function's parameters. In the previous stage, each expert was independently engaged in translating task within a specific domain, having been thoroughly trained. In this phase, we aim to maximize learning from other experts while being cautious of potential quality degradation in the current domain due to inputs from experts in different domains, see Figure 2. As a result, we impose a constraint to ensure that the benefits derived from multiple experts are at least as good as those from a single expert. Thus, given a set of source sentences $x$, targets $y$ and the model $\mathcal{F}$ with a learnable parameters $\theta$, we introduce Expert Margin Optimization, with the following loss function:

$$\begin{aligned} &\min_{\theta} \mathcal{L}(\mathcal{F}_\theta) \\ &s.t.\ \mathbb{E}_{(x,y)\sim\mathcal{D}}[\log \mathcal{F}_\theta(y|x) - \mathcal{G}(y|x)] < \epsilon \end{aligned} \tag{1}$$

where $\epsilon$ is a small positive constant and $\mathcal{G}$ represents $\mathcal{F}$ with the single domain expert.

Then the final EMO loss is as following:

$$\min_{\theta} \underbrace{\mathcal{L}(\mathcal{F}_\theta)}_{\mathcal{L}_{prefer}} \underbrace{-\mathbb{E}_{(x,y)\sim\mathcal{D}}[\log \mathcal{F}_\theta(y|x)]}_{\mathcal{L}_{NLL}} \tag{2}$$

which includes one preference learning term $\mathcal{L}_{prefer}$ and one negative log likelihood term $\mathcal{L}_{NLL}$.

### 3.4 Inference

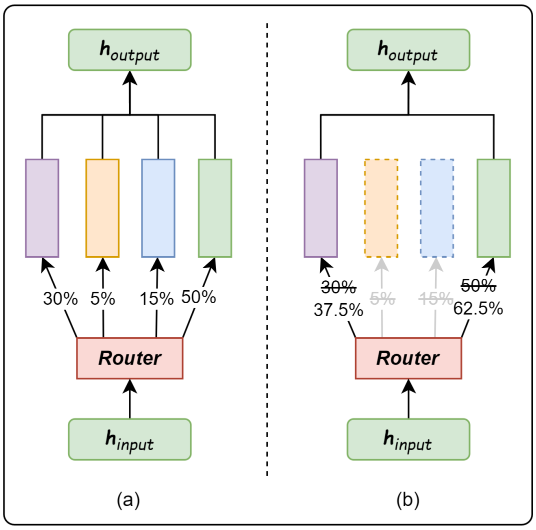


Figure 3: Infer mode.

Inspired by Wu et al. (2024a), we introduce two inference modes in our study. In the first mode, we utilize all trained LoRAs with a learned gating function, retaining their unique features with assigned weights as illustrated in Figure 3(a). In the second mode, we only retain the top-$k$ LoRAs and readjust the distributed weights. These dual modes enable our model to adapt to various situations, providing a versatile and adaptable strategy for composing effective LoRAs as shown in Figure 3(b).

## 4 Experiments

### 4.1 Datasets

When it comes to the datasets, we have utilized the parallel data for English-German and English-Chinese provided by WMT22 [1]. Following the approach of Aharoni and Goldberg (2020), we employed Bert for unsupervised domain clustering, resulting in the division of the data into four domains: Subtitle, IT, Medical, and Law.

To mitigate discrepancies arising from varying data volumes, we conducted domain-experts pre-training on 500w paired sentences for each domain. Using COMET(Rei et al., 2020) from Unbabel/wmt22-cometkiwi-da, we scored and ranked the data for each domain, selecting the top 1w sentences with the highest scores as high-quality instruction data.

Our test set comprises 5k sentences per domain. Similar to the methodology of Aharoni and Goldberg (2020), none of the sentences from the test set, pre-training data, and instruction data overlap. This precaution is crucial due to the susceptibility of neural models to memorization and hallucination, as observed by Müller et al. (2020).

### 4.2 Evaluation Metrics

For automatic evaluation, we utilize SacreBLEU, which implements BLEU(Papineni et al., 2002), and COMET(Rei et al., 2020) from Unbabel/wmt22-comet-da. SacreBLEU calculates similarity based on n-gram matching, while COMET leverages cross-lingual pretrained models for evaluation.

### 4.3 Compared Baselines

- **Base Model** ($\mathcal{B}$): A model trained directly using multi-domain instruction data.

- **Single Domain Models** ($\mathcal{S}_{Sub}$, $\mathcal{S}_{IT}$, $\mathcal{S}_{Med}$, $\mathcal{S}_{Law}$): Individual domain-specific translation models pre-trained and fine-tuned on data from the Subtitle, IT, Medical, and Law domains. Not utilizing the MoE architecture, each layer consists of a single LoRA.

- **Combined Domain Model** ($\mathcal{C}$): A unified translation model pre-trained and fine-tuned on a blend of data from the Subtitle, IT, Medical, and Law domains. Similar to the Single Domain Models, this model does not employ the MoE architecture and features a single LoRA per layer.

- **MoE Domain Model** ($\mathcal{M}$): This model is constructed by combining the LoRAs from $\mathcal{S}_{Sub}$, $\mathcal{S}_{IT}$, $\mathcal{S}_{Med}$ and $\mathcal{S}_{Law}$) in an MoE architecture, followed by Expert Margin Optimization training. Two key distinctions from our approach: firstly, this model combines independently trained domain models, whereas our method involves first training a general model and then domain-specific training on top of the general model, with a retained general expert; secondly, while every layer in this model is a MoE-LoRA, our approach utilizes MoE-LoRA only in the higher layers, resulting in fewer parameters and reduced computational overhead.

- **Ours**: As defined above.

### 4.4 Setup

We conducted experiments using Hugging Face Transformers with open-source LLMs from the Llama family (Touvron et al., 2023a,b), specifically leveraging Llama2-7b and Llama3-8b with matched parameters as our base models.

Our experiments were based on the llama-recipes project code. The original code supported only the StepLR learning rate update strategy, where the learning rate was updated after each epoch, suitable for the Instruction-tuning phase but too slow for extensive pre-training on large datasets. To address this, we expanded the code to flexibly accommodate strategies like WarmupLR, ConstantLR, and step-level updates.

During the general pre-training stage, we employed the WarmupLR strategy with 2000 warmup steps, a learning rate of 1e-4, using a batch strategy with packing, batch size of 8, a context length of 4096, and trained for 2 epochs. For the domain pre-training stage, the learning strategy was switched to ConstantLR with a learning rate of 1e-6, while other settings remained consistent with the previous stage.

In the Instruction-Tuning and Expert Margin Optimization phases, the batch strategy shifted to padding, allowing for a reduced context length of 512, enabling a larger batch size of 64. We used the StepLR strategy with a learning rate of 1e-4. For comparative experiments, we adjusted the number

[1] https://www.statmt.org/wmt22/translation-task.html#download

| Models | MoE | Subtitle | | IT | | Medical | | Law | |
|---|---|---|---|---|---|---|---|---|---|
| | | BLEU | COMET | BLEU | COMET | BLEU | COMET | BLEU | COMET |
| | | Our Recipe with Backbone Model: Llama2-7B | | | | | | | |
| B | ✗ | 20.99 | 79.75 | 18.23 | 75.16 | 14.22 | 69.88 | 16.23 | 70.3 |
| $S_{Sub}$ | ✗ | **24.80** | **81.97** | 20.8 | 78.94 | 15.56 | 72.37 | 17.44 | 73.35 |
| $S_{IT}$ | ✗ | 23.02 | 80.25 | **22.91** | **81.53** | 15.05 | 73.09 | 16.82 | 72.75 |
| $S_{Med}$ | ✗ | 20.08 | 74.96 | 17.20 | 73.99 | **18.71** | **75.93** | 15.43 | 70.18 |
| $S_{Law}$ | ✗ | 19.92 | 76.78 | 17.04 | 74.86 | 12.75 | 69.06 | **20.77** | **79.06** |
| C | ✗ | 23.75 | 80.70 | 21.55 | 81.13 | 17.38 | 75.64 | 18.76 | 77.41 |
| M | ✓ | **25.01** | **82.13** | **23.3** | **81.88** | **19.24** | **76.31** | **21.19** | **79.19** |
| Ours | ✓ | **26.41** | **83.46** | **24.21** | **82.41** | **19.78** | **78.20** | **21.55** | **80.61** |
| - EMO | ✗ | **25.59** | **82.36** | **23.15** | **82.68** | **19.00** | **77.89** | **21.36** | **78.89** |

Table 1: Overall results for the English-German translation direction.

| Models | MoE | Subtitle | | IT | | Medical | | Law | |
|---|---|---|---|---|---|---|---|---|---|
| | | BLEU | COMET | BLEU | COMET | BLEU | COMET | BLEU | COMET |
| | | Our Recipe with Backbone Model: Llama2-7B | | | | | | | |
| B | ✗ | 22.48 | 78.93 | 20.71 | 75.66 | 16.82 | 71.32 | 18.89 | 70.84 |
| $S_{Sub}$ | ✗ | **26.56** | **82.09** | 23.67 | 79.07 | 18.30 | 74.14 | 20.17 | 73.88 |
| $S_{IT}$ | ✗ | 24.42 | 79.30 | **25.45** | **81.74** | 17.88 | 74.47 | 19.61 | 73.16 |
| $S_{Med}$ | ✗ | 21.56 | 74.32 | 19.84 | 74.17 | **21.28** | **77.87** | 18.21 | 70.74 |
| $S_{Law}$ | ✗ | 21.71 | 75.68 | 19.81 | 74.89 | 15.34 | 70.98 | **23.47** | **79.84** |
| C | ✗ | 24.33 | 80.00 | 24.06 | 81.51 | 19.6 | 76.98 | 21.28 | 77.92 |
| M | ✓ | **26.78** | **82.19** | **25.64** | **81.82** | **21.83** | **78.16** | **23.95** | **79.93** |
| Ours | ✓ | **27.26** | **82.80** | **26.10** | **82.35** | **22.28** | **79.05** | **24.15** | **80.97** |
| - EMO | ✗ | **26.56** | **82.46** | **25.88** | **81.65** | **21.38** | **78.54** | **24.09** | **80.48** |
| | | Our Recipe with Backbone Model: Llama3-8B | | | | | | | |
| B | ✗ | 26.52 | 81.06 | 24.98 | 80.93 | 20.65 | 77.25 | 22.49 | 79.01 |
| $S_{Sub}$ | ✗ | **26.77** | **82.18** | 25.27 | 81.29 | 19.48 | 76.32 | 21.19 | 77.02 |
| $S_{IT}$ | ✗ | 26.35 | 81.36 | **25.83** | **81.6** | 18.91 | 75.52 | 20.71 | 75.17 |
| $S_{Med}$ | ✗ | 25.43 | 80.53 | 24.57 | 80.42 | **21.36** | **77.91** | 20.42 | 74.74 |
| $S_{Law}$ | ✗ | 25.86 | 80.69 | 24.54 | 80.86 | 18.33 | 75.87 | **23.66** | **80.01** |
| C | ✗ | 26.61 | 81.36 | 25.20 | 81.40 | 20.90 | 77.63 | 22.68 | 79.03 |
| M | ✓ | **26.87** | **82.20** | **26.23** | **82.00** | **22.04** | **78.39** | **24.06** | **80.12** |
| Ours | ✓ | **27.53** | **82.81** | **26.64** | **82.11** | **22.28** | **79.07** | **24.3** | **80.21** |
| - EMO | ✗ | **26.70** | **82.50** | **26.11** | 81.54 | **21.89** | **78.91** | **24.21** | 79.94 |

Table 2: Overall results for the English-Chinese translation direction.

of epochs or steps based on the dataset's size to ensure a consistent total number of trained tokens.

### 4.5 Results and Analysis

As shown in Table 1 and Table 2, our training strategy ultimately achieved the best results compared to other methods, demonstrating the effectiveness of our approach.

The Single Domain Models $\mathcal{S}_{Sub}$, $\mathcal{S}_{IT}$, $\mathcal{S}_{Med}$, $\mathcal{S}_{Law}$ performed well on their respective domain test sets but showed relatively weaker performance in other domains, with some results even inferior to the Base Model $\mathcal{B}$. While the Combined Domain Model ($\mathcal{C}$) demonstrated more balanced results overall, it consistently underperformed compared to the specific Single Domain Models in their respective domains. These observations underscore the existence of conflicts between domains. The

MoE Domain Model ($\mathcal{M}$) outperformed the Single Domain Models, highlighting the effectiveness of MoE in mitigating domain conflicts.

We conducted experiments using Llama2-7B and Llama3-8B as Backbone Models in English-to-Chinese translation. Our training strategies yielded the best results, showcasing the versatility of our approach. It is worth noting that Llama3-8B offers significant enhancements in multilingual capabilities compared to Llama2-7B, resulting in our method showing relatively lower gains when evaluated on Llama3-8B.

In conclusion, these experiments collectively establish our approach as an effective method for enhancing LLM multi-domain translation capabilities.

#### 4.5.1 Measuring the Effectiveness of General Pre-training

As shown in Table 1 and Table 2, - *EMO* represents our approach of utilizing domain experts to generate outputs for each domain's test set. As mentioned earlier, domain experts undergo general pre-training before domain-specific training. Remarkably, our findings reveal that these results outperform the performance of the Single Domain Models in each domain, thus validating the effectiveness of General Pre-training.

#### 4.5.2 Measuring the Effectiveness of EMO

As demonstrated in Table 1 and Table 2, following the application of EMO, the BLEU scores decrease by 0.2 to 1.2 on test sets with varying language directions, while the COMET scores decrease by 0.2 to 0.6. These results serve as evidence of the effectiveness of the EMO strategy. Furthermore, The MoE Domain Model ($\mathcal{M}$), which combines LoRA parameters from Single Domain Models to form MoE LoRA before undergoing EMO training, yields superior results across various domains compared to the Single Domain Model. This further validates the effectiveness of the EMO strategy.

#### 4.5.3 Understanding the Top-$k$ Inference

As depicted in Figure 4 and Figure 5, we have computed the variations in BLEU and COMET scores when employing different k values in the English-German direction. It is evident that the results are noticeably lower when $k$=1. However, when $k$=2, the results are already quite satisfactory. Further increasing k leads to fluctuations in results, with a marginal overall improvement observed.

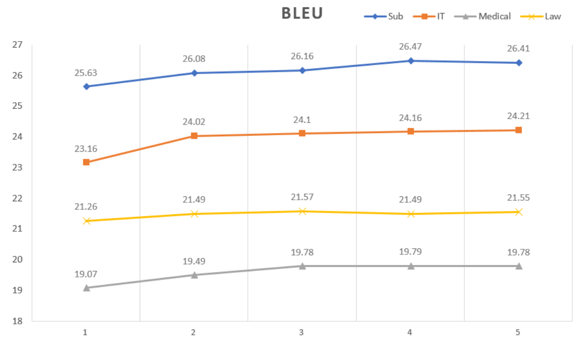


Figure 4: BLEU under different $k$.

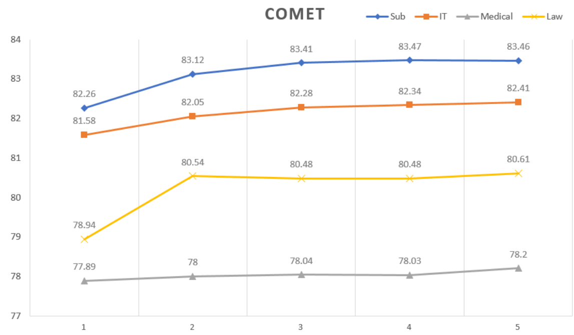


Figure 5: COMET under different $k$.

#### 4.5.4 Understanding the Router

| Expert | Subtitle | IT | Medical | Law |
|---|---|---|---|---|
| General | 27.03% | 26.32% | 19.74% | 21.21% |
| Subtitle | 45.05% | 21.93% | 5.92% | 7.88% |
| IT | 22.52% | 43.86% | 5.26% | 7.27% |
| Medical | 2.25% | 4.39% | 65.79% | 3.03% |
| Law | 3.15% | 3.51% | 3.29% | 60.61% |

Table 3: The overall results.

We were intrigued by how much knowledge each domain actually acquired from different experts, prompting an analysis of the results at the final layer of the Router. By examining token-level probability distributions across experts, the findings are summarized in Table 3.

It was observed that for each domain, the probability of selecting an expert from the same domain was the highest. Additionally, all domains were found to source knowledge from the General Expert. Specifically, the Medical and Law domains exhibited a stronger inclination towards choosing experts within their respective fields. On the other hand, the Subtitle and IT domains displayed a tendency to mutually select experts.

## 5 Conclusion

In conclusion, our study presented a approach by incorporating MoE-LoRA into Translation upon LLM, which effectively mitigates domain conflicts. By proposing a training methodology that combines general pre-training with domain-specific training and strategically placing MoE LoRAs in higher layers while maintaining a General LoRA for preserving common knowledge, we have achieved significant improvements in domain performance while reducing computational complexity. Additionally, the introduction of the Expert Margin Optimization strategy has led to the successful training of a robust Router policy that consistently outperforms individual experts in diverse scenarios. These findings highlight the efficacy of our methods in enhancing model performance and addressing domain-specific challenges in natural language processing tasks. In summary, our research offers valuable insights and techniques that may potentially contribute to the progress of machine learning and language processing in the future.

## 6 Related Work

There are some studies on the combination of MoE and LoRA.

**When MOE Meets LLMs**(Liu et al., 2024): This study focuses on the integration of MoE and LoRA, where a Router is utilized to learn a Task ID. During inference, only one expert is activated for final prediction.

**MoELoRA**(Luo et al., 2024): Another research effort incorporates Contrastive Learning into MoE training, aiming to encourage each expert to capture distinct knowledge representations.

**MoLE**(Wu et al., 2024b): A different study explores the use of a top-k strategy during prediction, as briefly mentioned in our preceding discussion.

**Mix-of-show**(Gu et al., 2023): This work delves more into the realm of image processing, employing LoRA as a feature extractor within MoE to facilitate the intricate fusion of multiple features.

## 7 Limitations

However, it is important to acknowledge the limitations of our research. One limitation is the scope of our dataset, which may not fully represent all possible scenarios. These limitations should be taken into consideration when interpreting the results and implications of our study.

## References


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

Dan Biderman, Jose Gonzalez Ortiz, Jacob Portes, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, Cody Blakeney, and John P. Cunningham. 2024. Lora learns less and forgets less. *CoRR*, abs/2405.09673.

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*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2023. Palm: Scaling language modeling with pathways. *J. Mach. Learn. Res.*, 24:240:1–240:113.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Y. Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau

Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *CoRR*, abs/2207.04672.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22:107:1–107:48.

Yuchao Gu, Xintao Wang, Jay Zhangjie Wu, Yujun Shi, Yunpeng Chen, Zihan Fan, Wuyou Xiao, Rui Zhao, Shuning Chang, Weijia Wu, Yixiao Ge, Ying Shan, and Mike Zheng Shou. 2023. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. *Preprint*, arXiv:2305.18292.

Jiaxin Guo, Hao Yang, Zongyao Li, Daimeng Wei, Hengchao Shang, and Xiaoyu Chen. 2024. A novel paradigm boosting translation capabilities of large language models. *CoRR*, abs/2403.11430.

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.

Qidong Liu, Xian Wu, Xiangyu Zhao, Yuanshao Zhu, Derong Xu, Feng Tian, and Yefeng Zheng. 2024. When moe meets llms: Parameter efficient fine-tuning for multi-task medical applications. *Preprint*, arXiv:2310.18339.

Tongxu Luo, Jiahe Lei, Fangyu Lei, Weihao Liu, Shizhu He, Jun Zhao, and Kang Liu. 2024. Moelora: Contrastive learning guided mixture of experts on parameter-efficient fine-tuning for large language models. *Preprint*, arXiv:2402.12851.

Mathias Müller, Annette Rios, and Rico Sennrich. 2020. Domain robustness in neural machine translation. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas, AMTA 2020, Virtual, October 6-9, 2020*, pages 151–164. Association for Machine Translation in the Americas.

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, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

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 2020, Online, November 16-20, 2020*, pages 2685–2702. Association for Computational Linguistics.

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, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Xun Wu, Shaohan Huang, and Furu Wei. 2024a. Mixture of lora experts. *CoRR*, abs/2404.13628.

Xun Wu, Shaohan Huang, and Furu Wei. 2024b. Mixture of lora experts. *Preprint*, arXiv:2404.13628.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 483–498. Association for Computational Linguistics.

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

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

Shaolei Zhang, Qingkai Fang, Zhuocheng Zhang, Zhengrui Ma, Yan Zhou, Langlin Huang, Mengyu Bu, Shangtong Gui, Yunji Chen, Xilin Chen, and Yang Feng. 2023. Bayling: Bridging cross-lingual alignment and instruction following through interactive translation for large language models. *Preprint*, arXiv:2306.10968.

# A Appendix

**English Version:**

**Sharp concrete diamond, Dismantle of ferro-concrete designs, Installation of reinforced-plastic windows, Punching holes in the w / w ceilings, Diamond drilling of apertures, Installation of metaloplastic designs (windows, doors), Dismantle of metaloplastic designs (windows, doors).**

**Within apprupt's Performance Network your brand message will be carried to the most relevant and valuable audiences for your specific brand.**

**By introducing greater flexibility in how the Fund is used and by reducing the number of redundancies from 1 000 to 500, it will become an ever more effective instrument for helping to tackle the effects of the economic down-turn.**

**...**

**German Version:**

**Diamant-Betonschneiden, Demontage von Stahlbetonkonstruktionen, Montage von Metall-Kunststoff-Fenstern, Stahlbeton-Lochung, Diamantbohren, Montage von Kunststoff-Metall-Verbundkonstruktionen (Fenster, Türen), Demontage von Kunststoff-Metall-Verbundkonstruktionen (Fenster, Türen).**

**Brands können so gezielt innerhalb des Performance Networks positioniert und an die jeweils relevanten Nutzergruppen kommuniziert werden.**

**Durch die Einführung von mehr Flexibilität in der Nutzung der Fonds und durch die Reduzierung der Zahl der Entlassungen von 1 000 auf 500 wird er ein noch effizienteres Instrument zur Bekämpfung der Folgen des Wirtschaftsabschwungs werden.**

**...**

**English Version:**

**Results The key nursing methods were to promote nurses fundamental diathesis to carry out mental care for the patients and the nursing during acetazolamide stress brain perfusion SPECT.**

**Genome and transcriptome analyses will complement the proteome and genetic information available today.**

**In any kind of need, please do not hesitate to contact us as we are always at your complete disposal.**

**Petals spatulate-emarginate, ca. 1 mm.**

**It works in all conditions and can work in extreme hot or cold without any problem**

**...**

**Chinese Version:**

**结果提高护理人员的基本素质；做好患者的心理护理和乙酰唑胺负荷脑血流灌注显像中的护理是关键性的护理措施。**

**基因组和转录组分析无疑将补充现有的蛋白质组和遗传学知识。**

**如果有任何需要的那种、请不要犹豫与我们联系、我们一直在您完成处置。**

**花瓣匙形微缺，约1毫米。**

**它可以在任何情况下，可以工作在极端高温或低温没有任何问题**

**...**

Figure 6: Data format.