# MoE-CT: A Novel Approach For Large Language Models Training With Resistance To Catastrophic Forgetting

**Anonymous ACL submission**

## Abstract

The advent of large language models (LLMs) has predominantly catered to high-resource languages, leaving a disparity in performance for low-resource languages. Conventional Continual Training (CT) approaches to bridge this gap often undermine a model's original linguistic proficiency when expanding to multilingual contexts. Addressing this issue, we introduce a novel MoE-CT architecture, a paradigm that innovatively separates the base model's learning from the multilingual expansion process. Our design freezes the original LLM parameters, thus safeguarding its performance in high-resource languages, while an appended MoE module, trained on diverse language datasets, augments low-resource language proficiency. Our approach significantly outperforms conventional CT methods, as evidenced by our experiments, which show marked improvements in multilingual benchmarks without sacrificing the model's original language performance. Moreover, our MoE-CT framework demonstrates enhanced resistance to forgetting and superior transfer learning capabilities. By preserving the base model's integrity and focusing on strategic parameter expansion, our methodology advances multilingual language modeling and represents a significant step forward for low-resource language inclusion in LLMs, indicating a fruitful direction for future research in language technologies.

## 1 Introduction

Large language models (LLMs) (Ouyang et al., 2022; Brown et al., 2020b) have achieved remarkable progress in recent years, particularly in areas such as language generation (Radford et al., 2019; Brown et al., 2020a), machine translation (Vaswani et al., 2017; Wu et al., 2019), text summarization (See et al., 2017; Rush et al., 2015), and language understanding (Devlin et al., 2019a; Peters et al., 2018). However, the majority of these models have

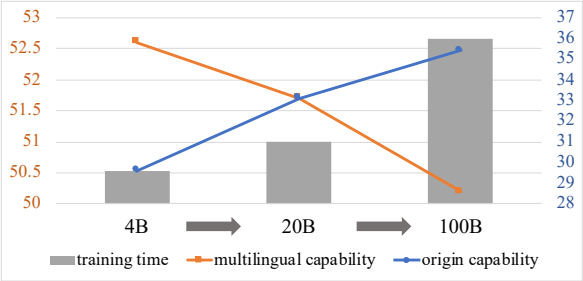


Figure 1: The abscissa represents the number of tokens from the original data incorporated during the Continual Training (CT) process. Although an increased volume of original data may decelerate the model's forgetting, it can significantly impede the enhancement of multilingual capabilities.

focused on resource-rich languages such as English, leaving substantial potential for performance improvements in low-resource languages. To alleviate above mentioned issues, researchers have aimed to Parameter-Efficient Fine-Tuning (PEFT) methods. Specifically, continual training (CT) has been proposed, which have proven to be effective in enhancing the performance of low-resource languages by training on specific language data. Besides, other researchers have resorted to method of Low-Rank Adaptation (LoRA) (Hu et al., 2021). By introducing a low-rank structure to reduce the model parameters for efficient updating. Thereby, it achieves a successful success between maintaining model performance and saving computation abd storage. Despite these developments, research on how to extend the multilingual capabilities of large models are still limited. We conducted an in-depth analysis of existing capability extension technologies.

1. Firstly, **it is difficult to obtain original training data.** In conventional training methods, the original training data accounts for a significant proportion, so that the lack of data poses a significant challenge for the training

in use. In addition, due to the different distributions between the training data of different stages, the absence of original training data will significantly exacerbate the catastrophic forgetting problems.

2. Secondly, **adding large amounts of raw data will limit the improvement of multilingual capabilities.** As illustrated in Figure 1, with the volume of original language data greatly alleviated the issue of catastrophic forgetting, significantly improving the generation capability in original data distribution. When the volume of original language data is five times greater than oth new one, the catastrophic forgetting issue is almost disappeared. However, it limits the model's final performance on multilingual tasks. Furthermore, a large amount of original data significantly increases the training costs of the model.

In response to the aforementioned issues, we propose a method that utilizes a Mixture of Experts (MoE) (Fedus et al., 2021; Lepikhin et al., 2020)approach. The Mixture of Experts model integrates multiple experts into the model architecture, where each expert is tasked with learning a specific task or feature subspace, thereby enhancing the model's learning capacity and generalization performance. Our specific approach involves extending additional expert networks on top of the pretrained model to more efficiently learn new multilingual capabilities. To effectively prevent catastrophic forgetting, we froze most of the parameters of the original LLM. In addition, we employed a frozen shared feed-forward network (shared-ffn) to preserve the original knowledge, and implemented a gating mechanism to dynamically merge the original knowledge with the newly acquired knowledge from the expert networks. The crux of this method is that it enables us to effectively retain the capabilities of the base model, even with only a limited amount of original data. Our approach loosens the restrictions on the proportion of multilingual data, allowing it to play a more significant role during training. This not only raises the ceiling for the model's performance on multilingual tasks but also reduces training costs due to a substantial reduction in the overall volume of data.

In our proposed method, we selected Qwen as our foundational model and, based on this, expanded the MoE network with 2-8 experts per layer. We extracted Chinese and English data, as well as a substantial quantity of multilingual data, from Wikipedia and mC4 to serve as the continued training data. Our multilingual capability enhancements were validated using datasets from the xwinograd (Tikhonov and Ryabinin, 2021), XCOPA (Ponti et al., 2020), and pasw-X (Yang et al., 2019) tasks designed for multilingual understanding, while the MMLU (Hendrycks et al., 2020)and C-Eval (Huang et al., 2023)datasets were employed to verify the model's retention of original capabilities. Empirical evidence demonstrates that our MoE expansion strategy outperforms conventional continual training (CT) and LoRA in terms of multilingual improvement. Moreover, in tests of original capabilities, our approach exhibits a notable increase in resistance to forgetting when compared to conventional CT. Our scheme has achieved commendable results across various sizes of the Qwen model series.

To sum up, our contributions are as follows:

- Our method introduces a novel MoE training paradigm for multilingual large model training, eliminating the need for extensive pretraining data while achieving compatibility between original and multilingual capabilities.

- We employed additional expert networks to learn multilingual competencies, and by increasing the proportion of multilingual data, we have surpassed the upper limits of multilingual capabilities inherent in conventional CT methods.

- Our approach leads the way in enhancing multilingual abilities and resistance to forgetting when compared to standard CT, demonstrating generalizability and transferability across models of varying scales.

## 2 Related Work

**Multilingual Pretraining and Fine-tuning in Language Models.**Developments in deep learning have significantly improved performance across a broad spectrum of natural language processing (NLP) tasks. Early efforts utilizing neural architectures such as recurrent neural networks (RNNs) (Zaremba et al., 2014)and long short-term memory (LSTM) (Sutskever et al., 2011) networks have laid the foundation for understanding and generating

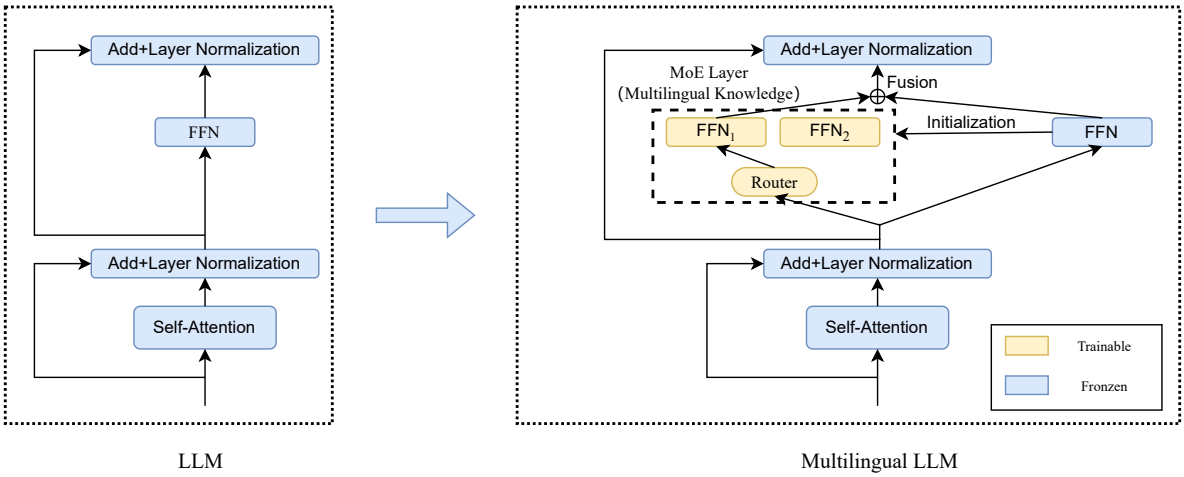


Figure 2: The diagram on the right represents the training structure of MoE-CT, where the blue area indicates that parameters are frozen, and the yellow area indicates that parameters are trainable. The parameters for all experts and the shared feed-forward network (shared-ffn) are initialized from the feed-forward network (ffn) of the original model.

language representations (Cho et al., 2014). The advent of self-attention mechanisms and transformer architectures (Vaswani et al., 2017) has shifted the focus towards models that can be pretrained on large unlabeled corpora for better generalization across NLP tasks.Recent research has expanded the scope of these models to accommodate multiple languages, yielding multilingual models such as mBERT (Devlin et al., 2019b) and XLM (Lample and Conneau, 2019), which are capable of learning cross-lingual representations. By pretraining on text from various languages, these models leverage shared linguistic properties and can be fine-tuned for downstream tasks in different languages.

**Sparsely Gated Networks** Efficiency in neural network scaling has gained paramount importance as the size of state-of-the-art language models has ballooned, resulting in steep computational costs. Sparsely gated networks have emerged as a promising direction for realizing model scaling without a linear increase in resource demands. At the core of this approach lies the concept of conditional computation, where different parts of the network are activated based on the input. The pioneering work of (Shazeer et al., 2017) demonstrated the potential of sparsely gated mixture-of-experts (MoE) layers in language models, where only a subset of experts is chosen for each input, drastically reducing computation during inference. This is achieved through gating mechanisms that learn to distribute the computation across a diverse set of experts, each specializing in different aspects of the data. Concurrently, a growing body of work is focusing on extending the Mixture of Experts (MoE) architecture with sparse activations (Hestness et al., 2017; Shazeer et al., 2018; Kudugunta et al., 2021).Switch Transformer (Fedus et al., 2021) and GLaM (Du et al., 2021)showcased a model with orders of magnitude more parameters than traditional models, but with fewer activated parameters per example, leading to improved performance and training efficiency. Similarly, the GShard framework (Lepikhin et al., 2020) demonstrated the scalability of MoE for large-scale multilingual machine translation tasks, reinforcing the viability of sparse gating for extensive language coverage.

**Continual Learning for Large Language models** Advancements in Natural Language Processing (NLP) have led to the exploration of various strategies to support continual learning capabilities. These strategies encompass: i) techniques that leverage the concept of replay to retain knowledge (ROBINS, 1995; Rebuffi et al., 2016; Shin et al., 2017; Lopez-Paz and Ranzato, 2017; Chaudhry et al., 2018); ii) strategies grounded in regularization to prevent overwriting of existing information (Kirkpatrick et al., 2016; Li and Hoiem, 2016); and iii) designs that restructure the neural architecture itself to accommodate new learning(Rusu et al., 2016; Lee et al., 2017; Mallya and Lazebnik, 2017; Wen et al., 2020) . The domain of NLP has

witnessed a burgeoning interest in these continual learning paradigms, as evidenced by a series of recent contributions(Wang et al., 2019a; Biesialska et al., 2020; Sun et al., 2019a; Huang et al., 2021; Hussain et al., 2021; Ahrens et al., 2021; Jin et al., 2021; Lin et al., 2022). These include embedding alignment with episodic memory (Wang et al., 2019b); improvements to memory-based parameter adaptation via sparse experience replay (de Masson d'Autume et al., 2019); approaches to lifelong language modeling (Sun et al., 2019b); and the adoption of meta-learning frameworks that integrate sparse replay (Holla et al., 2020). (Chen et al., 2023) marks the first application of the Mixture of Experts (MoE) structure to expand NLP tasks, mitigating the issue of catastrophic forgetting encountered in cross-task learning. Traditional approaches have predominantly focused on models of small to medium scale, where recent research has made strides in mitigating catastrophic forgetting to a certain degree. However, these strategies are often not applicable to large models with substantial parameter counts. In the continued training phase, such large models typically confront a more pronounced issue of catastrophic forgetting. Fine-tuning these models to adapt to specific tasks can lead to rapid loss of previously acquired general knowledge, even with minor parameter updates.In light of this, our work is dedicated to exploring how to balance the trade-off between stability (retaining old knowledge) and plasticity (acquiring new knowledge). By investigating and developing strategies that enable large models to maintain previously learned information while also learning new tasks effectively, we aim to contribute to the field's understanding of how to construct artificial intelligence systems capable of continual learning without sacrificing their pre-trained knowledge.

# 3 Method

As discussed above, it is evident that the multilingual capabilities of large-scale models remain constrained.Our goal is to maintain a robust performance in the generation of high-resource languages, effectively mitigating the risk of catastrophic forgetting, while simultaneously enhancing the multilingual capabilities of the model.

To achieve this, we introduce a novel architectural paradigm named MoE-CT, which substantially integrates diverse knowledge domains into our model, by employing an MoE (Mixture of Experts) approach tailored specifically for the multilingual context.

## 3.1 Mixture of Experts

In the described model, we have a collection of N feed-forward neural networks, which are all structurally equivalent and operate independently, denoted as a set of experts $E_i i = 1^N$. These experts are integrated with a dedicated gating mechanism, denoted as $G(\cdot)$, which functions as a decision-maker. This gating mechanism is responsible for determining the contribution of each expert's output in the final response of the system. Specifically, if we consider $h$ to be the resultant vector from an attention mechanism in any block of the model, then the final output $y$ from the mixture of experts (MoE) layer is derived by taking a weighted sum of each expert's output. Mathematically, this can be described by the equation:

$$y = \sum i = 1^N G(h)_i \cdot E_i(h)$$

In this expression, $E_i(h)$ represents the output of the $i$-th expert network, while $G(h)_i$ signifies the weight allocated to that output by the gating function. The gating function itself is defined using a softmax operation applied to the dot product of the input $h$ and a set of learned parameters encapsulated in the matrix $W_g$, which is expressed as:

$$G(\cdot) = Softmax(h \cdot W_g)$$

In essence, the matrix $W_g$ holds the parameters that the gating function adapts during training to optimize the routing of information through the various experts in the MoE architecture.

## 3.2 Routing mechanism design

Our goal is to store multilingual knowledge in the MoE layer and store English/Chinese knowledge in the old FFN layer of pretrained LLM. We hope that the sparse extended multilingual LLM can combine the outputs of the two layers to achieve the best world of multilingual and English/Chinese abilities. Therefore, we have added a fusion module with different structures to explore the best combination of MoE and FFN layers. The output of the fusion module can be described as follows:

$$Fusion(x, y) = w \cdot x + (1 - w) \cdot y \qquad (1)$$

In the equation, $w$ represents the weight of the MoE layer, which, when combined with the weight of the shared-ffn, sums to 1.

| | en | zh | ar | id | vi | de | fr | ja | th | pt | es | it |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Tokens | 2B | 2B | 2B | 2B | 2B | 2B | 2B | 2B | 2B | 2B | 2B | 2B |

Table 1: The composition of the continual training dataset, there are 24B tokens in total, and each language is allocated an equal proportion of the data.

| *Type* | *Expert* | *Params* | *Act-params* | *Layers* | *Hidden size* | *Heads* |
|---|---|---|---|---|---|---|
| Dense | - | 1.8B | 1.8B | 24 | 2048 | 16 |
| MoE | 2 | 3.4B | 2.6B | 24 | 2048 | 16 |
| Dense | - | 7B | 7B | 32 | 4096 | 32 |
| MoE | $2\sim 8$ | 16B$\sim 40B$ | 11B | 32 | 4096 | 32 |

Table 2: The size and architecture of our MoE model and dense model

$$output = Fusion\{\sum_{i=1}^{K} w_i FFN_i(h_{in}), FFN(h_{in})\} \tag{2}$$

where $w_i$ is the weight of $i$-th expert determined by router, and $K$ is the Top-$K$ experts selected by the router.

To prevent the problem of *catastrophic forgetting* during multilingual training, most of the parameters in our sparse-extended multilingual LLM will be fronzen, and only the MoE layer, embedding layer and fusion module are trainable, as shown in Figure 2, the blue modules are frozen, and the yellow modules are trainable.

### 3.3 An innovative model training methodology

As shown in Figure 2, our method extends large language models from dense structures to sparse structures by adding MoE layer parallel to the Feed Forward Layer. The MoE layer includes multiple FFN layers, but only a small portion will be activated, and the activation strategy is determined by the router module. We use the parameters of FFN layer from pretrained LLM to initialize multiple FFN layers in the MoE layer, which can significantly accelerate the training process and provide valuable knowledge transfer from old one to new one..

Throughout the continuation training process, we exclusively train the expert networks and the embedding layer within the expanded MoE model architecture, while all other structural parameters and shared-FFN parameters are kept fixed. Through experimental validation, we have ascertained that such a training strategy optimally retains the model's original capabilities while effectively expanding its multilingual capacity.

## 4 Experiment

### 4.1 Training Datasets

The composition of our data and the proportion of each language variety are reflected in Table 1. Our continued training data set comprises a total of 24 billion tokens, of which 20 billion are multilingual data encompassing ten languages (Arabic, Indonesian, Vietnamese, German, French, Japanese, Thai, Portuguese, Spanish, and Italian), with each language accounting for 2 billion tokens. Additionally, there are 2 billion tokens of Chinese data and 2 billion tokens of English data. All of the data were extracted from mC4 (Xue et al., 2020) and Wikipedia.

In addition, we prepared an extra 50 billion tokens for Chinese and 50 billion tokens for English to ensure that the proficiency in these languages does not diminish during the standard continual training (CT) process. Empirical evidence suggests that for the expanded multilingual dataset of 20 billion tokens, we need to incorporate at least five times more Chinese and English data to maintain their resistance to forgetting. However, within our Mixture of Experts (MoE) expanded architecture, we found that only 4 billion tokens of Chinese and English data are required to achieve resistance to forgetting in these languages, which substantially reduces the cost of training data.

### 4.2 Architecture Setting

We leverage Qwen (Bai et al., 2023) , a large language model pre-trained using a large amount of Chinese and English data as our base model, exhibits industry-leading performance in both Chinese and English. Due to limitations in training resources, we employ the 1b8 and 7b versions of Qwen's models as our foundational architectures.

| Task category | Task | Test | Lang. | Metric | Prompt |
|---|---|---|---|---|---|
| NLU | XNLI | 5,010 | 4 | Acc. | [Premise], right? {Yes/Also/No}, [Hypothesis] |
| | XCOPA | 500 | 7 | Acc. | [Prefix] {because/therefore} {choice1/choice2} [Suffix] |
| | PAWS-X | 2,000 | 7 | Acc. | [Sentence1], right? {Yes/No}, [Sentence2] |
| Knowledge | C-Eval | 1590 | 1 | Acc. | [Question]{Choices}[Answer] |
| | MMLU | 14213 | 1 | Acc. | [Question]{Choices}[Answer] |
| MT | WMT/IWSLT | 991-3,002 | 9 | BLEU | [INPUT] Translate this sentence from [SRC] to [TGT]. |

Table 3: Multilingual benchmark

| Model | Multilingual Ability | | | | | | | Chinese Ability | English Ability |
|---|---|---|---|---|---|---|---|---|---|
| | XCOPA | PAWS-X | XNLI | MT (en2xx) | MT (xx2en) | MT (en2zh) | MT (zh2en) | C-eval | MMLU |
| Qwen-1b8 | 54.7 | 53.0 | 38 | 15.1 | 21.0 | 35.6 | 22.1 | 48.5 | 37.0 |
| Qwen-1b8-CT | 62.3 | 51.0 | 44.7 | 19.1 | 20.9 | 27.2 | 15.2 | 29.7 | 29.5 |
| Qwen-1b8-MoE-CT (Experts=1) | 62.4 | 51.7 | 44.2 | 19.0 | 21.1 | 29.7 | 17.8 | 40.5 | 32.2 |
| Qwen-1b8-MoE-CT (Experts=2) | 62.5 | 52.9 | 44.4 | 19.5 | 21.9 | 30.8 | 18.5 | 41.5 | 33.4 |

Table 4: Main results on Qwen-1b8.When compared to Qwen-CT, the Qwen-MoE-CT exhibits a more robust resistance to forgetting and also demonstrates a greater enhancement in multilingual capabilities.

Table 2 shows the parameter structure of Qwen as a dense model, as well as the parameter structure after extending Qwen based on MoE. "Expert" represents the number of experts utilized in the MoE architecture. "Params" refers to the total number of parameters in the model. "Act-params" denotes the quantity of parameters activated during model inference. "Layers" indicates the number of layers in the model. "Hidden size" is the hidden dimension of the feed-forward layers, and "Heads" is the number of self-attention heads.

### 4.3 Multilingual Benchmark

We used six test datasets to evaluate the multilingual and Chinese/English capabilities of our models, and we divided them into three task categories: **NLU**, **Knowledge task** and **Machine Translation**. Please refer to Table 3 for detailed information.

**NLU** For Natural Language Understanding (NLU) task category, we choose three multilingual test task: XNLI, XCOPA and PAWS-X.

**Knowledge** The Knowledge task category includes two test dataset, C-Eval and MMLU, which are used to evaluate the Chinese and English capabilities of our models respectively.

**MT** For Machine Translation tasks, we selected 9 language test datasets from WMT and IWSLT to evaluate the translation ability from these languages to English and English to these languages.

### 4.4 Main Results

The main results are shown in Table 4 and Table 5, we conducted experiments on Qwen-1b8 and Qwen-7b models to verify the general effectiveness of our method. As shown in Table 4, Qwen-1b8-CT is a multilingual continue-training version of Qwen-1b8, which has a significant improvements of multilingual abilities over Qwen-1b8 model, improves the accuracy of XCOPA from 54.7% to 62.3%, XNLI from 38.0% to 44.7%. For machine translation task, multilingual continue-training brings in an average of 4.0 BLEU on En-XX directions, but no improvement is found on XX-EN directions.

Although continue-training strategy can improve multilingual ability of LLM, it always causes problem of *catastrophic forgetting*, which damages the original abilities of the model. We can find that Qwen-1b8-CT model has a significant drop on English-to-Chinese and Chinese-to-English translation performance, from 35.6 to 27.2 and 22.1 to 15.2, respectively. For Chinese and English ability evaluation, Qwen-1b8-CT achieves a performance decline from 48.5 to 29.7 and 37.0 to 29.5 on C-Eval and MMLU test set, respectively. What's more, simply continue train the LLM on multilingual corpus may not improve multilingual abilities, as we find a performance decline on PAWS-X test dataset (53.0 vs 51.0). *Catastrophic forgetting* problem can also be found on Qwen-7b model in Table 5. We have also experimented with the use of LoRA-CT, setting the LoRA dimension to 8, as shown in Table 5, although LoRA-CT can effectively alleviate the issue of catastrophic forgetting, its limited number of changed parameters results in a significant performance gap in extended multilingual capabilities when compared to conventional CT.

By using the sparse MoE architecture for continual training of LLM, the best results can be achieved in both multilingual and Chinese English ability. In Table 5, we can find that Qwen-7b-

| Model | Multilingual Ability | | | | | | | Chinese Ability | English Ability |
|---|---|---|---|---|---|---|---|---|---|
| | XCOPA | PAWS-X | XNLI | MT (en2xx) | MT (xx2en) | MT (en2zh) | MT (zh2en) | C-eval | MMLU |
| Qwen-7b | 61.6 | 60.1 | 41.6 | 24.0 | 30.2 | 40.1 | 29.0 | 57.4 | 42.7 |
| Qwen-7b-CT | 69.8 | 60.1 | 46.8 | 27.5 | 30.3 | 37.7 | 26.8 | 50.1 | 39.2 |
| Qwen-7b-LoRA-CT | 62.3 | 59.6 | 44.0 | 24.4 | 29.1 | 40.0 | 27.8 | 55.4 | 43.3 |
| Qwen-7b-MoE-CT (Experts=2) | 69.5 | 60.6 | 46.0 | 27.8 | 30.3 | 39.3 | 27.5 | 53.6 | 41.7 |
| Qwen-7b-MoE-CT (Experts=4) | 69.3 | 60.5 | 46.0 | 28.3 | 30.5 | 39.8 | 27.7 | 55.9 | 42.1 |
| Qwen-7b-MoE-CT (Experts=8) | 69.6 | 60.6 | 46.5 | 28.4 | 30.7 | 39.8 | 28.1 | 55.5 | 42.3 |

Table 5: Main results on Qwen-7b.Our MoE architecture is equally applicable to various model sizes, and we have achieved substantial resistance to forgetting on Qwen-7b, alongside a comprehensive enhancement of multilingual capabilities that surpasses those of conventional CT and LoRA-CT.

MoE-CT performs equivalently to Qwen-7b-CT in multilingual tasks, and significantly prevents the performance degradation of Chinese and English capabilities, where performance on C-eval only decreases from 57.4 to 55.5 and MMLU decreases from 42.7 to 42.3. Our method also alleviates the decline in the performance of English-to-Chinese and Chinese-to-English translation, and has a significant improvement in the translation ability on English-XX directions from 24.0 to 28.4. The details of translation results on Qwen-1b8 and Qwen-7b can be found in Table 9 and Table 10. In our experiments, we have identified a significant data ratio problem when employing conventional continual training (CT) methods to combat catastrophic forgetting. Specifically, to preserve the proficiency in Chinese and English, the volume of data for these languages must be at least five times greater than that of the multilingual dataset. However, as indicated in Table 6, such an excessive reliance on Chinese and English data significantly hampers the enhancement of multilingual capabilities. Moreover, it substantially increases the training cost, posing a major obstacle to the expansion of large language models.

In contrast to the aforesaid conventional CT, our proposed MoE expansion technique significantly reduces the dependence on Chinese and English data, with these languages' data requiring only a one-fifth proportion of the multilingual dataset. As demonstrated in Table 6, due to the reduced incorporation of Chinese and English data in our MoE framework, the model can more effectively assimilate multilingual knowledge. Consequently, the amplification of multilingual abilities is more pronounced compared to the conventional CT methods. This approach highlights the efficiency of our MoE expansion method in achieving a better and more efficient balance between preserving original language capabilities and enhancing expanded multilingual proficiencies.

### 4.5 Ablation Study

In this study, we conducted ablation experiments on the freezing strategy, routing mechanism, and the number of experts used in the MoE expanded architecture to verify their effects on multilingual improvement and resistance to forgetting of Chinese-English bilingual capabilities.

**Training strategy.** In our experiments on the Qwen-1.8B scale model, we attempted to freeze different parts of the MoE model to determine which components could retain the model's original capabilities. As indicated in Table 7, continual training all parameters or attention layers will result in a significant decrease in the model's original ability, while training expert layers and embedding layers can balance the original ability and expanded ability. Therefore, in the final model, we chose to train the parameters of the experts and the embedding layer.

**Routing mechanism.** In our experiments, we attempted various combination methods for the outputs of the shared feed-forward network and the MoE layer. Initially, we utilized a weighted sum approach for integrating the two, assigning MoE output weights from 0.1 to 0.9. Our experiments revealed that a lower MoE weight corresponded to a more pronounced resistance to catastrophic forgetting, yet resulted in a smaller improvement in multilingual capabilities. Conversely, a higher MoE weight weakened the resistance to catastrophic forgetting while yielding a greater enhancement in multilingual capabilities. As shown in table 8, when the weights for both the MoE and the shared-ffn are set to 0.5, an optimal balance is achieved between the enhancement of multilingual capabilities and the resistance to forgetting.

**The number of experts.** As shown in Table 4 and Table 5, two experts already exhibits considerable resistance to forgetting of Chinese-English bilingual capabilities. To verify whether a greater

| Model | Continue training Tokens | | | Chinese Ability | English Ability | Multilingual Ability |
|---|---|---|---|---|---|---|
| | EN | ZH | Multilingual | C-eval | MMLU | Avg |
| Qwen-1b8 | - | - | - | 48.5 | 37 | 48.6 |
| Qwen-1b8-CT | 50B | 50B | 20B | 39.5 | 31.3 | 50.2 |
| Qwen-1b8-CT | 2B | 2B | 20B | 29.7 | 29.5 | 52.6 |
| Qwen-1b8-MoE-CT | 2B | 2B | 20B | **41.5** | **33.4** | **53.3** |
| Qwen-7b | - | - | - | 57.4 | 42.7 | 54.4 |
| Qwen-7b-CT | 50B | 50B | 20B | 53.3 | 41.1 | 57.6 |
| Qwen-7b-CT | 2B | 2B | 20B | 50.1 | 39.2 | 58.9 |
| Qwen-7b-MoE-CT | 2B | 2B | 20B | **55.5** | **42.3** | **58.9** |

Table 6: The data ratios required for conventional CT and those necessitated by the MoE expansion indicate that the MoE architecture can achieve better resistance to forgetting in Chinese and English, as well as enhanced multilingual capabilities, with only a minimal amount of Chinese and English data.

| Trainable parameters | original | expanded |
|---|---|---|
| Qwen-1b8 | 42.7 | 48.6 |
| all | 30.8 | 52.8 |
| attention | 31.3 | 51.5 |
| embedding | 35.3 | 52.4 |
| experts | 35.9 | 52.7 |
| embedding&experts | 37.5 | 53.3 |

Table 7: The Impact of Training Strategies on Original and Expanded Capabilities in the MoE architecture, all means that all parameters participate in the training

| | MoE weights | original | expanded |
|---|---|---|---|
| Weighted | 0.1 | 37.8 | 52.1 |
| | 0.9 | 34.4 | 53.3 |
| | 0.5 | 37.5 | 53.3 |

Table 8: The impact of different combinations of MoE layers and shared FFN on original and expanded capabilities in the Qwen-1B8 model, where MoE weight represents the proportion of the MoE layer's output in the total output.

number of experts would yield further enhancements in multilingual capabilities, we expanded the model to include 4 and 8 experts at the Qwen-7B scale. However, according to the experimental results, increasing the number of experts does not significantly improve multilingual abilities. In light of this, we hypothesize that the quantity of continued training data may be insufficient, preventing the multiple experts from converging to their optimal performance. Consequently, we plan to utilize a larger corpus of continued training data, provided sufficient training resources, to support an increased number of experts in future experiments.

## 5 Conclusion

In conclusion, our research presents a significant advancement in the field of multilingual language modeling, particularly addressing the challenges posed by catastrophic forgetting in large language models (LLMs). Through the introduction of the MoE-CT structure, we have demonstrated a novel approach that not only enhances the extension of LLMs to low-resource languages but also preserves the original linguistic competencies in high-resource languages. Our experiments on models such as Qwen-1b8 and Qwen-7b have validated the effectiveness of MoE-CT, marking clear improvements in multilingual benchmarks while maintaining or even improving performance in the original languages.

The MoE-CT structure showcases a delicate balance between stability and plasticity, ensuring that the base model's parameters remain undisturbed and that the newly introduced MoE layers absorb the additional linguistic knowledge. This balance is crucial for achieving a harmonious integration of multilingual capabilities without the detriment of pre-existing language proficiencies. Our findings indicate that MoE-CT can achieve substantial resistance to forgetting with a minimal amount of pre-training data, which is a considerable stride towards reducing training costs and resources.

The implications of our work are manifold. Primarily, it facilitates the creation of more inclusive language technologies that do not favor solely high-resource languages. Furthermore, it paves the way for future research into continual learning for LLMs, emphasizing the importance of models that can continually evolve and adapt to new language without losing previously established knowledge.

## Limitations

In summary, we have proposed the MoE-CT architecture to address the issue of catastrophic forgetting encountered by LLMs during the expansion of multilingual capabilities. Due to the limitations of computational resources, we have not attempted to extend the MoE-CT architecture to other open-source models beyond Qwen, which may not fully demonstrate the catastrophic forgetting challenges faced by all LLMs. Therefore, our future work will explore whether this structure can be adapted to a wider range of open-source models.

## Ethics Statement

Our work on the MoE-CT for LLMs considers several ethical concerns. Primarily, we aim to address linguistic biases by enhancing LLMs performance in low-resource languages, promoting inclusivity and cultural diversity. We recognize the risk of potential biases in model training and commit to their mitigation. We also acknowledge the responsibility to prevent the misuse of our model for deceptive purposes and advocate for its transparent and responsible use. Environmental impacts due to the high computational requirements of LLMs are also considered. Our model aims to reduce training resources, and we encourage sustainable practices in AI research.

## Acknowledgements

## References


Kyra Ahrens, Fares Abawi, and Stefan Wermter. 2021. DRILL: dynamic representations for imbalanced lifelong learning. *CoRR*, abs/2105.08445.

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

Magdalena Biesialska, Katarzyna Biesialska, and Marta R. Costa-jussà. 2020. Continual lifelong learning in natural language processing: A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6523–6541, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020a. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

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. 2020b. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. 2018. Efficient lifelong learning with A-GEM. *CoRR*, abs/1812.00420.

Wuyang Chen, Yanqi Zhou, Nan Du, Yanping Huang, James Laudon, Zhifeng Chen, and Claire Cui. 2023. Lifelong language pretraining with distribution-specialized experts. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 5383–5395. PMLR.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *CoRR*, abs/1406.1078.

Cyprien de Masson d'Autume, Sebastian Ruder, Lingpeng Kong, and Dani Yogatama. 2019. Episodic memory in lifelong language learning. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019a. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019b. BERT: Pre-training of

deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Nan Du, Yanping Huang, Andrew M. Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, Barret Zoph, Liam Fedus, Maarten Bosma, Zongwei Zhou, Tao Wang, Yu Emma Wang, Kellie Webster, Marie Pellat, Kevin Robinson, Kathy Meier-Hellstern, Toju Duke, Lucas Dixon, Kun Zhang, Quoc V. Le, Yonghui Wu, Zhifeng Chen, and Claire Cui. 2021. Glam: Efficient scaling of language models with mixture-of-experts. *CoRR*, abs/2112.06905.

William Fedus, Barret Zoph, and Noam Shazeer. 2021. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *CoRR*, abs/2101.03961.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *CoRR*, abs/2009.03300.

Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory F. Diamos, Heewoo Jun, Hassan Kianinejad, Md. Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. 2017. Deep learning scaling is predictable, empirically. *CoRR*, abs/1712.00409.

Nithin Holla, Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova. 2020. Meta-learning with sparse experience replay for lifelong language learning. *ArXiv*, abs/2009.04891.

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

Yufan Huang, Yanzhe Zhang, Jiaao Chen, Xuezhi Wang, and Diyi Yang. 2021. Continual learning for text classification with information disentanglement based regularization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2736–2746, Online. Association for Computational Linguistics.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models.

Aman Hussain, Nithin Holla, and Pushkar Mishra. 2021. Towards a robust experimental framework and benchmark for lifelong language learning. In *NeurIPS Datasets and Benchmarks*.

Xisen Jin, Mohammad Rostami, and Xiang Ren. 2021. Lifelong learning of few-shot learners across nlp tasks. *ArXiv*, abs/2104.08808.

James Kirkpatrick, Razvan Pascanu, Neil C. Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2016. Overcoming catastrophic forgetting in neural networks. *CoRR*, abs/1612.00796.

Sneha Kudugunta, Yanping Huang, Ankur Bapna, Maxim Krikun, Dmitry Lepikhin, Minh-Thang Luong, and Orhan Firat. 2021. Beyond distillation: Task-level mixture-of-experts for efficient inference. *arXiv preprint arXiv:2110.03742*.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *CoRR*, abs/1901.07291.

Jeongtae Lee, Jaehong Yoon, Eunho Yang, and Sung Ju Hwang. 2017. Lifelong learning with dynamically expandable networks. *CoRR*, abs/1708.01547.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. 2020. Gshard: Scaling giant models with conditional computation and automatic sharding. *CoRR*, abs/2006.16668.

Zhizhong Li and Derek Hoiem. 2016. Learning without forgetting. *CoRR*, abs/1606.09282.

Bill Yuchen Lin, Sida I. Wang, Xi Victoria Lin, Robin Jia, Lin Xiao, Xiang Ren, and Wen tau Yih. 2022. On continual model refinement in out-of-distribution data streams. In *Annual Meeting of the Association for Computational Linguistics*.

David Lopez-Paz and Marc'Aurelio Ranzato. 2017. Gradient episodic memory for continuum learning. *CoRR*, abs/1706.08840.

Arun Mallya and Svetlana Lazebnik. 2017. Packnet: Adding multiple tasks to a single network by iterative pruning. *CoRR*, abs/1711.05769.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237,

New Orleans, Louisiana. Association for Computational Linguistics.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, and Christoph H. Lampert. 2016. icarl: Incremental classifier and representation learning. *CoRR*, abs/1611.07725.

ANTHONY ROBINS. 1995. Catastrophic forgetting, rehearsal and pseudorehearsal. *Connection Science*, 7(2):123–146.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal. Association for Computational Linguistics.

Andrei A. Rusu, Neil C. Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. 2016. Progressive neural networks. *CoRR*, abs/1606.04671.

Abigail See, Peter Liu, and Christopher Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Association for Computational Linguistics*.

Noam Shazeer, Youlong Cheng, Niki Parmar, Dustin Tran, Ashish Vaswani, Penporn Koanantakool, Peter Hawkins, HyoukJoong Lee, Mingsheng Hong, Cliff Young, Ryan Sepassi, and Blake A. Hechtman. 2018. Mesh-tensorflow: Deep learning for supercomputers. *CoRR*, abs/1811.02084.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *CoRR*, abs/1701.06538.

Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. 2017. Continual learning with deep generative replay. *CoRR*, abs/1705.08690.

Fan-Keng Sun, Cheng-Hao Ho, and Hung-Yi Lee. 2019a. Lamol: Language modeling for lifelong language learning. *arXiv preprint arXiv:1909.03329*.

Fan-Keng Sun, Cheng-Hao Ho, and Hung yi Lee. 2019b. Lamol: Language modeling for lifelong language learning. In *International Conference on Learning Representations*.

Ilya Sutskever, James Martens, and Geoffrey E Hinton. 2011. Generating text with recurrent neural networks. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 1017–1024.

Alexey Tikhonov and Max Ryabinin. 2021. It's All in the Heads: Using Attention Heads as a Baseline for Cross-Lingual Transfer in Commonsense Reasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3534–3546, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Hong Wang, Wenhan Xiong, Mo Yu, Xiaoxiao Guo, Shiyu Chang, and William Yang Wang. 2019a. Sentence embedding alignment for lifelong relation extraction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 796–806, Minneapolis, Minnesota. Association for Computational Linguistics.

Hong Wang, Wenhan Xiong, Mo Yu, Xiaoxiao Guo, Shiyu Chang, and William Yang Wang. 2019b. Sentence embedding alignment for lifelong relation extraction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 796–806, Minneapolis, Minnesota. Association for Computational Linguistics.

Yeming Wen, Dustin Tran, and Jimmy Ba. 2020. Batchensemble: An alternative approach to efficient ensemble and lifelong learning. *CoRR*, abs/2002.06715.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2019. Google's neural machine translation system: Bridging the gap between human and machine translation. arxiv 2016. *arXiv preprint arXiv:1609.08144*, 2.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. In *North American Chapter of the Association for Computational Linguistics*.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3687–3692, Hong Kong, China. Association for Computational Linguistics.

Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. 2014. Recurrent neural network regularization. *CoRR*, abs/1409.2329.

# A Appendix A

In this section, we provide the detailed information about translation results on Qwen-1b8 and Qwen-7b models. The results are shown in Table 9 and Table 10.

| Model | En-X | | | | | | | | X-EN | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Ar | De | Es | Fr | Ru | Ko | Zh | Ja | Ar | De | Es | Fr | Ru | Ko | Zh | Ja |
| Qwen-1b8 | 1.4 | 16.9 | 22.2 | 22.4 | 9.4 | 4.0 | 35.6 | 8.8 | 15.9 | 30.7 | 28.2 | 30.8 | 25.8 | 6.2 | 22.1 | 8.0 |
| Qwen-1b8-CT | 7.3 | 21.3 | 25.1 | 26.7 | 12.4 | 11.9 | 27.2 | 20.5 | 24.4 | 31.0 | 28.9 | 31.0 | 23.2 | 6.5 | 15.2 | 7.2 |
| Qwen-1b8-MoE-CT (Experts=1) | 7.1 | 20.7 | 24.9 | 26.0 | 11.8 | 12.6 | 29.7 | 18.8 | 22.8 | 31.3 | 28.5 | 30.8 | 24.0 | 6.0 | 17.8 | 7.9 |
| Qwen-1b8-MoE-CT (Experts=2) | 7.5 | 21.0 | 25.3 | 26.5 | 12.3 | 12.1 | 30.8 | 20.2 | 24.8 | 31.7 | 29.2 | 31.2 | 26.1 | 6.5 | 18.5 | 7.4 |

Table 9: Translation results on Qwen-1b8 in details.

| Model | En-X | | | | | | | | X-EN | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Ar | De | Es | Fr | Ru | Ko | Zh | Ja | Ar | De | Es | Fr | Ru | Ko | Zh | Ja |
| Qwen-7b | 7.1 | 27.7 | 30.6 | 35.9 | 16.9 | 13.4 | 40.1 | 20.0 | 31.5 | 41.3 | 34.7 | 39.6 | 34.3 | 14.0 | 29.0 | 17.2 |
| Qwen-7b-CT | 11.7 | 30.0 | 32.2 | 36.3 | 19.1 | 21.0 | 37.7 | 31.6 | 35.2 | 40.9 | 34.4 | 39.0 | 33.6 | 15.8 | 26.8 | 16.3 |
| Qwen-7b-MoE-CT (Experts=2) | 11.6 | 30.0 | 32.3 | 36.8 | 18.9 | 21.7 | 39.3 | 31.9 | 35.4 | 41.1 | 34.5 | 39.3 | 34.0 | 14.3 | 27.5 | 16.6 |
| Qwen-7b-MoE-CT (Experts=4) | 11.6 | 30.6 | 32.7 | 37 | 19.4 | 23.0 | 39.8 | 32.2 | 34.7 | 41.2 | 34.4 | 39.5 | 34.0 | 14.9 | 27.7 | 17.6 |
| Qwen-7b-MoE-CT (Experts=8) | 11.8 | 30.5 | 32.8 | 37.2 | 19.3 | 23.5 | 39.8 | 32.5 | 35.4 | 41.1 | 34.4 | 39.2 | 34.1 | 15.5 | 28.1 | 18.0 |

Table 10: Translation results on Qwen-7b in details.