E^3: Optimizing Language Model Training for Translation via Enhancing Efficiency and Effectiveness

Linqing Chen[*1] and **Weilei Wang**[1] and **Dongyang Hu**[2]
[1]PatSnap Co., LTD., Suzhou, China
[2]Soochow University, Suzhou, China
{chenlinqing, wangweilei}@patsnap.com

Abstract

In the field of Natural Language Processing (NLP), Large-scale Language Models (LLMs) have demonstrated exceptional capabilities across a variety of tasks, including question answering, classification, and particularly, natural language understanding. The integration of neural machine translation with LLMs presents significant potential, transforming the paradigms of cross-lingual communication and information exchange. This study investigates the foundational aspects of LLMs' translation abilities and identifies effective training methodologies to equip them with multilingual capacities. We specifically explore the optimal timing for introducing translation capabilities to LLMs via supervised tasks, considering the inherent bilingual nature of machine translation. Key questions explored include whether it is more beneficial to integrate multiple languages during the pre-training or supervised fine-tuning (SFT) stages, how variations in language ratios influence LLMs' translation abilities, and whether longer or shorter texts are more effective for training these models. This research conducts a thorough investigation by training multiple LLMs from scratch with parameter scales in the billions and enhances the robustness of our findings by upgrading the language capabilities of pre-trained open-source models with parameter scales reaching tens of billions. The aim is to provide a detailed analysis that elucidates the complexities of augmenting machine translation capabilities within LLMs.

1 Introduction

The Large Language Model (LLM) has proven itself capable across a broad spectrum of tasks, including named entity extraction, text classification, and document understanding. Moreover, the LLM has demonstrated the ability to address highly complex tasks using the innovative Chain of Thoughts (CoT) methodology, as pioneered by Zhang et al. (Zhang et al., 2022). Among the various capabilities of the LLM, this paper focuses on one particular area that has significantly intrigued the authors: its exceptional proficiency in machine translation. This proficiency promises transformative potential for the field, with extensive implications for cross-lingual communication and the broader dissemination of information.

The evolutionary trajectory of Large Language Models (LLMs) has undergone significant cyclical transformations, each contributing substantially to their growth and capabilities. The journey began with the introduction of the Transformer model for neural machine translation, a foundational breakthrough as described by (Vaswani et al., 2017). Building on this, BERT (Devlin et al., 2019) leveraged the Transformer's encoder to achieve exceptional performance in text classification and cloze tests.

Following the advent of BERT, a series of influential models emerged, notably GPT (Radford et al., 2019), which is characterized by its unique decoder-only architecture. Furthermore, models like T5 (Raffel et al., 2019) and mT5 (Xue et al., 2020), with their comprehensive encoder-decoder frameworks, have continuously expanded the boundaries of what is possible with LLMs. Most of these models are either directly based on the Transformer architecture or represent substantial enhancements to it, reaffirming the critical role of the Transformer in the advancement of the field. This progression of models, each aug-

*Corresponding author.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1023-1034, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers

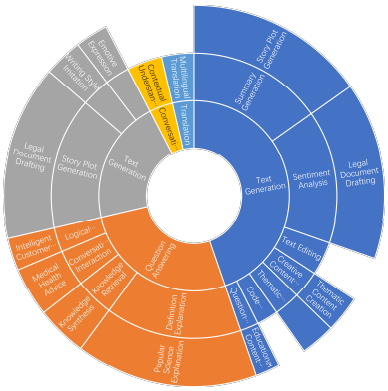

Figure 1: Capabilities of large scale language model.

menting the work of its predecessors, has driven LLMs to their current state of remarkable performance and vast potential in natural language processing (NLP).

This paper aims to delve deeper into these evolving aspects of natural language processing, particularly focusing on the translation proficiency of Pre-trained Language Models (PLMs). Our objectives are threefold: to explore the foundational aspects of translation efficiency in PLMs, to investigate the impact of diverse training methodologies and data ratios on PLMs' translation capabilities, and to propose effective strategies for training PLMs to enhance their translation performance. We explore several pertinent questions in the context of neural machine translation using LLMs:

- **Multiple Languages Alignment**: Should the alignment of language semantic spaces be performed during the pre-training stage or the supervised fine-tuning (SFT) stage?

- **Efficient Training**: Which type of contextual input—long or short—enhances the training efficiency and effectiveness of Large-scale Language Models for translation tasks?

- **Language Proportion Impact**: How does the composition of different languages in the corpus during LLMs' pre-training process affect their performance in translating various languages?

2 Machine Translation in the Large Language Model (LLM) Era

The exploration of machine translation driven by Language Models (LMs) has been revitalized by the advent of Large Language Models (LLMs), building on a rich history of significant research. Pioneering studies such as Brants et al. (2007) (Brants et al., 2007) integrated large-scale statistical language modeling into machine translation, laying the groundwork for subsequent advancements. This approach was refined by Mohit et al. (2009) (Mohit et al., 2009), who tailored LMs to improve the translation of complex phrases. The shift to neural language models, notably researched by Vaswani et al. (2013) (Vaswani et al., 2013), emphasized their potential to enhance translation accuracy and fluency. Lee (2020) (Lee, 2020) effectively bridged theoretical research with practical applications, particularly in quality estimation tasks, demonstrating the real-world impact of these developments.

In more recent studies, Han et al. (2022) (Han et al., 2022) examined the superiority of extra-large pre-trained language models (xLPLMs) over smaller counterparts in domain-specific translation tasks, providing vital insights into model selection for specific contexts. Concurrently, Tan et al. (2022) (Tan et al., 2022) introduced Multi-Stage Prompting, an innovative method that utilizes pre-trained LMs for translation, illustrating the continued evolution in LM applications in translation tasks. Zhu (2023) (Zhu et al., 2023) extended this investigation, exploring the capability of LLMs in translating a diverse array of languages and thus contributing to our understanding of the scalability and adaptability of LLMs in multilingual translation scenarios.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1023-1034, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers

3 Experimentation

We explore and analyze the performance of translation of multiple Large Language Models (LLMs) from scratch, with parameter scales reaching the order of several billion. Simultaneously, given that training models with hundreds of billions of parameters incurs costs in the order of millions of dollars, when investigating questions on an even larger scale of Large Language Models (LLMs), we opt to leverage open-source LLMs.

3.1 Training LLM from Scratch

The development of expansive Large Language Models (LLMs) is a resource-intensive endeavor, requiring substantial investments in both time and financial resources. Previous research (Henighan et al., 2020) has systematically demonstrated that insights garnered from experiments on smaller-scale LLMs can reliably inform the development of their larger-scale versions. This is substantiated by the rigorous empirical investigations conducted by leading organizations such as OpenAI (OpenAI, 2023) and Google (Longpre et al., 2023). These studies primarily focus on Pretrained Language Models (PLMs) with parameters numbering in the billions, which, although substantial, are comparatively modest against the backdrop of the most extensive LLMs. Such research underscores the practical relevance and extrapolative potential of smaller LLM findings to larger model architectures, thereby justifying continued investment in their development.

Opting for models of a more manageable scale confers distinct advantages, notably in accelerating the pace of model iteration. This acceleration facilitates a more efficient exploration of the research landscape in natural language processing. Furthermore, this strategy provides us with a valuable opportunity to engage in the crucial task of contrasting and evaluating diverse phenomena within the field. Such activities are essential for broadening the collective understanding and advancing the state of knowledge in this rapidly evolving domain.

3.1.1 Training Data

Detailed information regarding the datasets used is thoroughly documented in Table 1.

Stage	Data	Size(GB)	Line(M)
Pre-Training	Wikipedia EN	20.0	8.5
	Wikipedia ZH	2.4	2.0
	news2016zh	7.5	1.3
Translation	translation2019	1.6	5.1
	LDC zh-en news	0.3	0.8

Table 1: Training Data.

Pre-Training Data For the initial pre-training of our Pre-trained Language Models (PLMs), we utilize both the Chinese Wikipedia[0] and the English Wikipedia[1] as the foundational corpora. Due to the relatively limited size of the Chinese Wikipedia dataset, we enhance our Chinese pre-training corpus with additional datasets, specifically the news2016zh (brightmart, 2019) and the translation2019zh[2], to enrich the training data significantly. This supplementation of the Chinese corpus is crucial for maintaining a proportional balance in the volume of language data between Chinese and English, particularly in ablation studies designed to evaluate the impact of dataset size on model performance.

Fine-Tuning Data To achieve Chinese-to-English (ZH-EN) translation capabilities in our models, we have selected the Linguistic Data Consortium (LDC) corpora as our training data for translation tasks. We chose this corpus due to its unique properties; it not only serves as a resource for sentence alignment but also for document alignment. Inspired by Zhang et al. (Zhang et al., 2018), we utilize different forms

[0] https://dumps.wikimedia.org/zhwiki/
[1] https://dumps.wikimedia.org/enwiki/
[2] https://github.com/brightmart/nlp_chinese_corpus

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1023-1034, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers

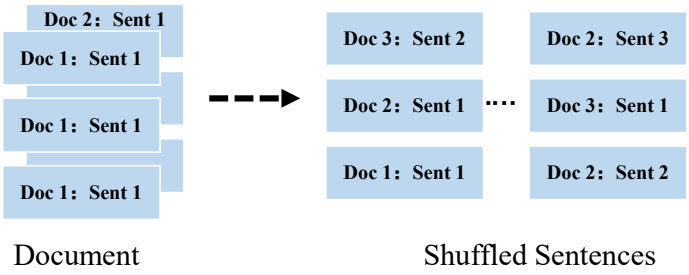

Figure 2: In the context of translation training data, retaining the inherent document structure is termed as document-level translation data. Conversely, when parallel sentence pairs within such data are isolated from their document context, this process results in what is known as sentence-level parallel data.

of this parallel corpus under various experimental setups. When evaluating the effectiveness of short-text or context-free translation data, we employ the sentence-aligned version of the corpus. Conversely, for assessing the performance with long-text or rich-context translation data, we use the document-aligned version of the corpus. Both forms are consistent in content and volume, but differ in their alignment, which is illustrated in Figure 2. When using the sentence-aligned version, we segment the originally document-aligned corpus into sentences.

3.1.2 Pre-Training

Hyper Parameters	Pre-training	Finetuning
Model	1.3B	1.3B
mini batch size	16	8
global batch size	128	64
Min_LR	1e-5	1e-6
Max_LR	3e-5	1e-5
Max_len	2048	4096
TP	8	8
PP	8	8

Table 2: Details of training settings.

We have adopted an architectural framework akin to that of LLaMA (Touvron et al., 2023a) and have undertaken the training of our model from scratch. The AdamW optimizer (Loshchilov and Hutter, 2018) is employed for training, with hyperparameters β_1 and β_2 set to 0.9 and 0.95, respectively. We apply a weight decay of 0.1 and enforce a gradient norm clipping of 0.5. Unless otherwise stated, the base model in this subsection is configured with 1.3 billion parameters. The maximum sequence lengths are set to 2048 tokens for pre-training and 4096 tokens for fine-tuning. In the pre-training phase, when supervised tasks like machine translation are incorporated, the computation of the input's loss is excluded to prevent the model from predicting sentences in the source language during training. Machine translation evaluations in this paper are reported using the BLEU score (Papineni et al., 2002), unless specified otherwise. On average, training each model requires one week on 8 A800 GPUs. Further details regarding training settings are presented in Table 2.

3.1.3 Fine-Tuning

In the experiments conducted in this subsection, we employ Supervised Fine-Tuning (SFT) to endow the model with translation capabilities. The SFT used here differs slightly from that in other research works. On one hand, we require SFT to enable the model to understand the task of mapping the source language to the target language. On the other hand, we do not need a diverse set of SFT tasks since capabilities beyond translation are not required. For the same reasons, we also forgo the Reinforcement Learning from Human Feedback (RLHF) process.

Method	No.	CTX	MT02	MT03	MT04	MT05	MT08	Average
NTP+SUP	#1	Non-Concat	42.39	42.30	41.99	42.35	31.88	40.07
	#2	Concat	43.10	43.27	42.33	42.59	32.20	40.98
NTP+MIX	#3	Non-Concat	43.84	44.10	41.93	42.89	32.90	40.81
	#4	Concat	44.95	44.80	**44.63**	**44.71**	**36.50**	**43.01**
Holistic MIX	#5	Non-Concat	43.55	44.27	41.85	42.90	32.64	41.04
	#6	Concat	**44.01**	**44.89**	44.57	44.61	36.03	42.82

Table 3: NTP is an abbreviation for 'next token prediction,' and SUP stands for 'supervised task training'. 'Concat' and 'Non-concat' represent whether to concatenate supervised training data.

3.1.4 Experimental Results

Languages Alignment Stage At what stage does aligning different languages enhance the multilingual capabilities of models? Contemporary state-of-the-art Large Language Models (LLMs) such as GPT (Radford et al., 2019) and LLaMA (Touvron et al., 2023a) typically do not utilize parallel corpora during the pre-training phase. To explore the effect of language alignment on the translation proficiency of LLMs, we initiated an investigation involving the training of two distinct types of base models. Table 3 lists all experimental results about language alignment in different stages. NTP is an abbreviation for 'next token prediction,' and SUP stands for 'supervised task training.'

- **NTP+SUP**, represents a two-step training process where pure token prediction pretraining is performed first, followed by supervised translation task training, with both training phases kept entirely separate.
- **NTP+MIX**, denotes a two-step training approach where pure token prediction pretraining is conducted initially, followed by mixed supervised and unsupervised task training.
- **Holistic MIX**, refers to a unified training approach that combines pure token prediction pre-training with translation task data in a completely integrated manner.
- **Concat**, denotes the method of concatenating supervised training data. In the 'Concat' configuration, supervised translation data is concatenated to fully utilize tokens within sequences of up to 2048 in length during the pre-training phase.
- **Non-Concat**, setting does not concatenate the data, with each sequence containing only one source-target sentence pair.

The experimental results presented in Table 3 clearly demonstrate that Model #3 achieves superior translation performance compared to Model #1 and Model #5, while Model #4 outperforms Model #2 and Model #6 in terms of translation quality. Keeping the volume of data constant, it is apparent that segregating supervised and self-supervised training phases does not contribute positively to the translation efficacy of Large Language Models (LLMs). In contrast, integrating self-supervised and supervised training approaches—particularly after completing the self-supervised next token prediction task—significantly improves translation performance. This improvement suggests that mixed training methodologies may enable models to concurrently develop foundational background knowledge and enhance their capabilities in aligning semantic spaces of two different languages during translation task training.

Long or Short? Further analysis of the experimental results in Table 3 indicates that Model #2 exhibits superior translation performance compared to Model #1, while Model #4 outshines Model #3 in terms of translation quality. With the data volume held constant, it is found that concatenating supervised translation data significantly improves the translation capabilities of Large Language Models (LLMs). These findings suggest that concatenated supervised translation training data not only optimizes the use of sequence length, thereby reducing the number of data samples needed, but also allows for a higher

number of sentence pairs to be processed simultaneously. This efficient approach to data utilization evidently enhances the translation performance of LLMs.

A closer examination of the experimental results unveils an intriguing phenomenon. Despite being based on differing training methodologies, Model #3 exhibits only marginal improvements in translation performance over Model #2 and, in some cases within individual sub-test sets, performs worse. This observation underscores the importance of fully utilizing sequence lengths in training data and ensuring data richness within these sequences to effectively boost the translation performance of Large Language Models (LLMs). Consequently, for the training and development of LLMs, the employment of apt data utilization strategies and alignment methods is critical for enhancing translation capabilities.

Translation Performance It is worth noting that the authors of this paper have observed that the performance of simply trained PLMs on translation tasks does not surpass that of dedicated translation models. This finding exceeds initial expectations and aligns with the research conducted by (Zhu et al., 2023). In both concatenation and non-concatenation translation tasks, PLMs, functioning solely as decoders, have not achieved superior performance compared to specialized translation models. We speculate that this limitation could be attributed to the decoder-only architecture of PLMs. However, it is conceivable that with further increases in model parameters and training data or the adoption of models that incorporate both encoder and decoder components, such as mT5, PLMs may eventually surpass dedicated translation models.

Model	MT02	MT03	MT04	MT05	MT08	BLEU
#1 ZH-EN 2:1	37.09	36.92	36.54	34.68	27.74	34.89
#2 ZH-EN 1:1	36.58	36.40	35.87	36.23	30.08	34.05
#3 ZH-EN 1:2	37.17	36.85	36.31	34.61	27.72	34.50

Table 4: The impact of baseline models trained with different language ratios on downstream translation tasks is investigated. ZH-EN x:y denotes baseline models with a Chinese-English language ratio of x:y.

Language Ratio We employed different ratios of Chinese and English Wikipedia data to investigate the impact of non-translation task data proportion during the pretraining stage on the performance of downstream translation tasks. The model named "EN-ZH 1:2" represents a base model where English Wikipedia data accounts for 1/3, and Chinese data accounts for 2/3 of the total. The proportions of the other two models can be inferred from their names accordingly. During the supervised fine-tuning stage, we introduced LCD zh-en news translation tasks to assess the influence of different language proportions during the pretraining stage on the performance of downstream translation tasks.

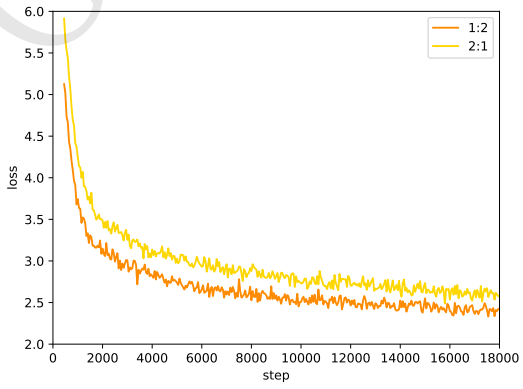

Figure 3: Illustration of two different models('ZH-EN 2:1 and 'ZH-EN 1:2') loss.

Figure 3 illustrates the loss curves of two base models with Chinese-English data ratios of 1:2 and 2:1 during the training process. It can be observed that the base model with a higher proportion of Chinese

data converges more slowly. Initially, the authors of this paper believed that the slower loss reduction observed in the base model with a Chinese-to-English ratio of 2:1 might result in relatively poorer translation performance, because the Chinese language is harder to learning for LLM. The experimental data in Table 4 indicates that, in terms of zh-en translation tasks, while different language ratios may affect the convergence speed or rate of loss reduction of the base models, **there is no significant difference observed in downstream translation tasks**.

3.2 Post-Training on Open-Source LLM

In the preceding sections, we engaged with a range of intriguing questions and observed phenomena based on several models, each possessing 1.3 billion parameters, trained from scratch. This naturally leads us to a critical inquiry: Are these phenomena universal, and do they persist across larger-scale LLMs? To address these questions within reasonable time and budget constraints, we extended the Chinese language capabilities of the open-source 13 billion parameter model and reaffirmed the phenomena previously observed in the 1.3 billion parameter model.

3.2.1 Training Data

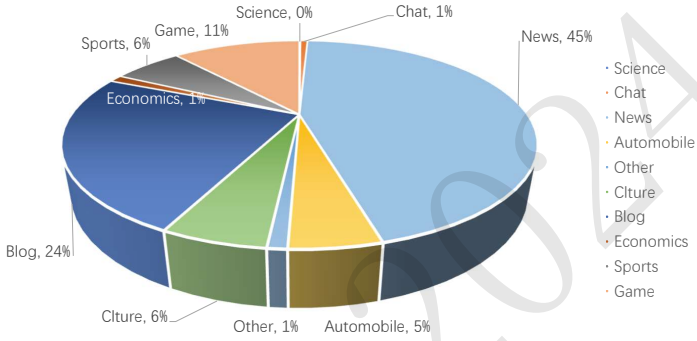

Figure 4: The composition of training data types employed to enhance the Chinese language capabilities of the model.

We examine the impact of various training methods on the translation performance by focusing on enhancing the English-to-Chinese translation capabilities of our base models. To this end, we have randomly sampled 40 billion tokens from the Wudao corpus (Yuan et al., 2021), aiming to improve the English-to-Chinese translation proficiency. The distribution of data types from this corpus is depicted in Figure 4.

3.2.2 Post-Training

We have trained the model using the 13 billion parameter version of LLaMA2 (Touvron et al., 2023b). Unless otherwise noted, the maximum sequence length for data is set at 4096 tokens. During the pre-training phase, when supervised tasks such as machine translation are integrated, the computation of the input's loss is omitted to prevent the model from merely reproducing sentences from the source language. Unless specified differently, the evaluation of machine translation performance in this study is based on the BLEU score. For the training of these models, an average duration of 10 days on 64 A800 GPUs per model is required.

3.2.3 Fine-Tuning

Similar to Section 3.1, the models in this subsection, which continue training on LLaMA2, utilize Supervised Fine-Tuning (SFT) to enable the model to comprehend the machine translation task. This approach specifically suppresses other capabilities of the model to prevent interference with the automated evaluation of its translation performance.

Model	Concat	MT02	MT03	MT04	MT05	MT08	Average
#1	Holistic Mix	48.88	48.20	44.58	48.01	40.97	46.33
#2	NTP + Sup	47.90	47.30	43.88	47.20	40.08	45.07
#3	NTP + Mix	**49.59**	**48.91**	**45.22**	**48.69**	**41.52**	**47.00**
Model	Non-Concat	MT02	MT03	MT04	MT05	MT08	Average
#4	Holistic Mix	46.41	46.24	44.53	46.44	40.50	45.37
#5	NTP + Sup	45.59	46.21	45.47	45.43	35.00	44.01
#6	NTP + Mix	46.89	46.64	44.90	46.89	40.98	45.80

Table 5: NTP is an abbreviation for 'next token prediction,' and SUP stands for 'supervised task training.'. 'Concat' and 'Non-concat' represent whether to concatenate supervised training data.

3.2.4 Experimental Results

In our research, we aimed to enhance the Chinese language capabilities of open-source Large Language Models (LLMs) by adhering to established experimental protocols similar to those used for training models from scratch. The training of translation tasks was segmented into distinct phases, and a comparative analysis of the outcomes was conducted.

Languages Alignment Stage We investigated the optimal timing for imparting translation capabilities to LLMs through supervised tasks. According to the results presented in Table 5, Model #1 showed superior translation performance compared to Model #2, and similarly, Model #4 outperformed Model #5. These training experiments, designed to enhance translation capabilities in open-source LLMs, are consistent with our previous findings. Notably, maintaining a constant volume of training data and integrating the next token prediction (NTP) task with supervised translation training has proven effective in enhancing the translation performance of LLMs. Additionally, our experiments revealed that LLMs, particularly those with billions of parameters, can sometimes misinterpret translation tasks when trained on large and diverse datasets, leading to undesired outputs. To address this, we maintained a constant volume of training data but extracted a small subset for translation training. Following the MIX training phase, we conducted targeted machine translation training to refine the model's translation capabilities. This focused approach resulted in improved translation outcomes.

Long or short? How to train the translation capability of Large-scale Language Models more efficiently. The experiments in the Table 5 are broadly categorized into two main groups: concatenating or not concatenating the translation data. It is evident that, on the whole, the translation approach using concatenated training data outperforms the non-concatenated approach. Furthermore, based on our calculations, the time consumed for training with concatenated data is slightly lower, further substantiating its efficiency.

3.3 Ablation Experiments

In this section, we explore additional critical factors influencing the training of Large Language Models (LLMs) with translation capabilities. Specifically, we examine the effects of the length of translation pairs during the training stage and the model's performance in domains requiring specialized knowledge.

3.3.1 Length of Training Language Paris

We utilize context-aware BLEU (d-BLEU) to assess the impact of text length on the translation performance of Large Language Models (LLMs). The ablation experiments in this section build upon the models from the language ratio impact experiments detailed in Section 3.1.4. These models were trained using varying language ratios in sentence pairs to examine their translation capabilities. Although the maximum length of training sequences is set to 4096, many tokens are often filled with the special token [pad], denoting **None**, due to the varying lengths of sentences.

Context Aware Template In the ctx-awar experiments, each sentence is precisely designated as the target for translation. Concurrently, the sentences preceding and following the target sentence are seam-

lessly concatenated to provide the necessary contextual backdrop. To construct the input, we adeptly utilize a structured template: “Translate [sentence] based on the contextual foundation: [context].”

Ratio	No.	CTX	MT02	MT03	MT04	MT05	MT08	BLEU
ZH-EN 2:1	#1	+ sent-level	37.09	36.92	36.54	34.68	27.74	34.89
	#2	+ ctx-aware	36.52	35.85	35.02	35.79	27.10	34.44
ZH-EN 1:1	#3	+ sent-level	36.58	36.40	35.87	36.23	30.08	34.05
	#4	+ ctx-aware	36.80	36.55	36.28	36.57	29.65	34.33
ZH-EN 1:2	#5	+ sent-level	37.17	36.85	36.31	34.61	27.72	34.50
	#6	+ ctx-aware	37.01	36.84	36.52	36.77	27.15	34.86

Table 6: The impact of baseline models trained with different language ratios on downstream translation tasks is investigated. ZH-EN x:y denotes baseline models with a Chinese-English language ratio of x:y, ’+ sent-level’ refer to using sentence-level ZH-EN news data as the downstream translation task. ”ctx-aware” refers to downstream tasks using the same data with context-aware template, with details provided in the Section 3.3.1.

Results As indicated in Table 6, Models #2, #4, and #6 perform context-aware translation tasks on Models #1, #3, and #5, respectively, using context-aware templates. These templates are utilized to assess each model’s capability with long sequences. Our hypothesis is that if training with short pairs does not negatively impact the model’s translation performance, then the BLEU scores for context-aware translation tasks should be approximately 1 point higher than those for sentence-level tasks, due to the provision of contextual information. The experimental results, as shown in Table 3 Section 3.1, suggest a similar trend. The performance of Models #6, #4, and #2 is only marginally better than that of Models #5, #3, and #1. This indicates that limited exposure to shorter text sequences during the sentence-level translation training in Models #1, #3, and #5 has adversely impacted their ability to process longer texts, resulting in a decline in performance on context-aware tasks with lengthier sequences.

3.3.2 Translation in the Biomedical Domain

Model	Concat	ZH-EN	EN-ZH
#1	Holistic Mix	18.60	20.11
#2	NTP + Sup	18.19	19.12
#3	NTP + Mix	18.98	20.25
Model	Non-concat	MT02	MT03
#4	Holistic Mix	17.58	19.25
#5	NTP + Sup	17.33	19.10
#6	NTP + Mix	17.90	19.91

Table 7: NTP is an abbreviation for ’next token prediction,’ and SUP stands for ’supervised task training.’. ’Concat’ and ’Non-concat’ represent whether to concatenate supervised training data.

To validate the results obtained in earlier chapters from various perspectives, we conducted zero-shot translation using publicly available translation test sets that encompass extensive domain knowledge. Zero-shot machine translation in this context refers to an experiment where, unlike previous experiments, we did not first enhance the base model’s Chinese capabilities using next token prediction with Chinese data. Instead, we directly trained on very limited translation corpora using different training methodologies for domain-specific translation tasks. We performed bidirectional zero-shot experiments on the publicly available translation test set in the biomedical field (Bawden et al., 2020). The results, as shown in Table 7, are consistent with the early findings using models trained as described in Table 5.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1023-1034, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers

4 Related Work

In the study by (Brants et al., 2007), a comprehensive examination is provided regarding the advantages stemming from the integration of large-scale statistical language modeling within the domain of machine translation. Additionally, (Mohit et al., 2009) conduct a thorough investigation into the intriguing notion of adapting language models specifically tailored for phrases exhibiting suboptimal translation quality. The research conducted by (Vaswani et al., 2013) delves into the intricate realm of neural language models and their potential application within the domain of machine translation, shedding light on the innovative techniques that harness the power of these models. In a detailed analysis, (Lee, 2020) elucidate the contributions made by Bering Lab to the WMT 2020 Shared Task on Quality Estimation (QE), underscoring the significance of their submission in advancing the field.

In a quest to explore the nuances of performance differences, (Han et al., 2022) embark on a meticulous examination aimed at ascertaining whether extra-large pre-trained language models (xLPLMs) unequivocally surpass their smaller-sized counterparts for domain-specific machine translation tasks. Expanding the horizon of possibilities, (Tan et al., 2022) introduce the concept of Multi-Stage Prompting (MSP), a straightforward yet impactful approach designed to harness the capabilities of pre-trained language models for translation tasks. Delving into the intricacies of multilingual translation, (Zhu et al., 2023) undertake an in-depth exploration of the performance characteristics exhibited by large language models (LLMs), simultaneously delving into the various factors that exert influence on the translation prowess of these models.

5 Conclusion

In this study, we conducted an extensive series of comparative experiments using five models trained from scratch, each with a parameter count of 1.3 billion, alongside several continued-training scenarios on open-source Large Language Models (LLMs) of 13 billion parameters. Our key findings reveal that integrating translation tasks during the pre-training phase of LLMs consistently enhances performance in handling longer text sequences, especially evident in context-aware translation tasks. Additionally, our results indicate that strategically segmenting translation training tasks from the base model training into distinct phases can significantly expedite the overall training process of the model. While variations in language ratios were observed to affect the convergence speed and training duration of the base model, they did not substantially impact the performance metrics of downstream translation tasks. This dynamic interaction between pre-training strategies, task integration, and language ratio effects underscores the potential for optimizing both the efficiency and effectiveness of neural machine translation systems within large-scale language models. Future efforts will focus on employing under-resourced languages to investigate methods for bridging 'Zero-Shot' language translation challenges.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1023-1034, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers

References

Rachel Bawden, Giorgio Maria Di Nunzio, Cristian Grozea, Inigo Jauregi Unanue, Antonio Jimeno Yepes, Nancy Mah, David Martinez, Aurélie Névéol, Mariana Neves, Maite Oronoz, Olatz Perez-de Viñaspre, Massimo Piccardi, Roland Roller, Amy Siu, Philippe Thomas, Federica Vezzani, Maika Vicente Navarro, Dina Wiemann, and Lana Yeganova. 2020. Findings of the WMT 2020 biomedical translation shared task: Basque, Italian and Russian as new additional languages. In *Proceedings of the Fifth Conference on Machine Translation*, pages 660–687, Online, November. Association for Computational Linguistics.

Thorsten Brants, Ashok C. Popat, Peng Xu, Franz J. Och, and Jeffrey Dean. 2007. Large language models in machine translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 858–867, Prague, Czech Republic, June. Association for Computational Linguistics.

brightmart. 2019. brightmart/nlp_chinese_corpus: release version 1.0, September.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805.

Lifeng Han, Gleb Erofeev, Irina Sorokina, Serge Gladkoff, and Goran Nenadic. 2022. Examining large pre-trained language models for machine translation: What you don't know about it. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 908–919, Abu Dhabi, United Arab Emirates (Hybrid), December. Association for Computational Linguistics.

T. J. Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B. Brown, Prafulla Dhariwal, Scott Gray, Chris Hallacy, Benjamin Mann, Alec Radford, Aditya Ramesh, Nick Ryder, Daniel M. Ziegler, John Schulman, Dario Amodei, and Sam McCandlish. 2020. Scaling laws for autoregressive generative modeling. *ArXiv*, abs/2010.14701.

Dongjun Lee. 2020. Two-phase cross-lingual language model fine-tuning for machine translation quality estimation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1024–1028, Online, November. Association for Computational Linguistics.

S. Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David M. Mimno, and Daphne Ippolito. 2023. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity. *ArXiv*, abs/2305.13169.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Behrang Mohit, Frank Liberato, and Rebecca Hwa. 2009. Language model adaptation for difficult to translate phrases. In Lluís Màrquez and Harold L. Somers, editors, *Proceedings of the 13th Annual conference of the European Association for Machine Translation, EAMT 2009, Barcelona, Spain, Map 14-15, 2009*. European Association for Machine Translation.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Kishore Papineni, Salim Roukos, Ward Todd, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL*, pages 311–318.

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

Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *ArXiv*, abs/1910.10683.

Zhixing Tan, Xiangwen Zhang, Shuo Wang, and Yang Liu. 2022. MSP: Multi-stage prompting for making pre-trained language models better translators. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6131–6142, Dublin, Ireland, May. Association for Computational Linguistics.

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

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

Ashish Vaswani, Yinggong Zhao, Victoria Fossum, and David Chiang. 2013. Decoding with large-scale neural language models improves translation. In *Conference on Empirical Methods in Natural Language Processing*.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*.

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*.

Sha Yuan, Hanyu Zhao, Zhengxiao Du, Ming Ding, and Jie Tang. 2021. Wudaocorpora: A super large-scale chinese corpora for pre-training language models. *AI Open*.

Jiacheng Zhang, Huanbo Luan, Maosong Sun, Feifei Zhai, Jingfang Xu, Min Zhang, and Yang Liu. 2018. Improving the transformer translation model with document-level context. In *Conference on Empirical Methods in Natural Language Processing*.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022. Automatic chain of thought prompting in large language models.

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*, abs/2304.04675.