Enhancing Translation Ability of Large Language Models by Leveraging Task-Related Layers

Pei Cheng[1], **Xiayang Shi**[1*], **Yinlin Li**[2]
[1]Zhengzhou University of Light Industry, Zhengzhou, Henan, China
[2]Key Laboratory of Multimodal Artificial Intelligence Systems,
Institute of Automation, Chinese Academy of Sciences, Beijing, China
cp2199405327@gmail.com, aryang123@163.com, yinlin.li@ia.ac.cn

Abstract

Fine-tuning Large Language Models (LLMs) for machine translation is effective but costly. It also increases the risk of overfitting and catastrophic forgetting, especially when training data is limited. To tackle these challenges, we propose a novel method that involves adjusting task-related layers in large models to better harness their machine translation capabilities. This method aims to retain the model's knowledge on other tasks while optimizing performance on translation tasks. By revealing the structure and characteristics of attention weights through singular value decomposition (SVD), we can make fine adjustments to specific layers, leveraging the model's potential for more accurate and efficient translations. Our method not only addresses computational resource consumption and catastrophic forgetting but also offers a new perspective on utilizing the capabilities of large models effectively. Experimental validation shows that adjusting task-related layers significantly improves performance on translation tasks while maintaining stability and accuracy on other tasks. This finding provides valuable insights for fine-tuning and applying large models, advancing the field of machine translation.

Keywords: large language models, machine translation, fine-tuning

1. Introduction

In recent years, LLMs have demonstrated substantial potential and value across various domains of Natural Language Processing (NLP), including text generation, text summarization, sentiment analysis, and question-answering systems, among others. LLMs such as GPT-3 (Brown et al., 2020), BLOOM (Scao et al., 2022) and LLaMA (Touvron et al., 2023), by learning from extensive text data, have achieved the capability to understand and generate complex human languages, thereby attaining remarkable results in a myriad of NLP tasks. Although the exemplary performance of large language models in most NLP applications, how to implement machine translation with LLMs still encompasses numerous unresolved challenges and unexplored issues.

Although adapting machine translation tasks through fine-tuning pre-trained large models (PLMs) is a very effective approach, it also increases the risk of model overfitting and leads to catastrophic forgetting problems, especially when the available training data is relatively scarce. To mitigate this issue, some research works have adopted methods like adapter (Alam and Anastasopoulos, 2022) and prompt (Zhang et al., 2023a) to enhance the capabilities of the models. These methods have made certain optimizations regarding the issue of resource consumption and have significantly improved the capabilities of the models. However, they mainly focus on adjustments at the data level and do not adjust the corresponding parameters based on the machine translation task itself, leaving potential room for improvement in model performance on specific tasks. This may lead the model to generate translations that seem reasonable but are actually inaccurate (Zhang et al., 2023b).

Additionally, apart from neural machine translation based on encoder-decoder, LLMs perform machine translation usually understand the mapping relationships between different languages by learning a large amount of multilingual text data and generate corresponding translations. However, due to the specific definitions and contexts of cross-language interactions in the model's training data often being vague and variable, the model cannot clearly define specific cross-language interaction rules.

To address above issues, we propose a novel method, which involves adjusting the task-related layers in large models to harness the model's machine translation capabilities better and improve the performance on specific tasks. This method aims to retain the model's knowledge on other tasks while optimizing the performance on a specific translation task. The core idea is to reveal the inherent structure and characteristics of attention weights through SVD (Garneau et al., 2020), understand the behavior of the model at each layer, and make fine adjustments to specific layers, thereby better leveraging the model's potential to achieve more accurate and efficient translations. This method not

LREC-COLING 2024, pages 6110–6121
20-25 May, 2024.

only helps address the issues of computational resource consumption and catastrophic forgetting but also offers a new research direction and perspective on how to utilize the capabilities of LLMs more effectively. Through experimental validation, we found that this strategy of adjusting task-related layers indeed significantly improves the performance of large models on specific translation tasks while maintaining stability and accuracy on other tasks. This finding provides valuable insights for the fine-tuning and application of large models and is expected to further propel advancements in the field of machine translation.

In summary, the main contributions of this paper are as follows:

- We propose a novel adjustment strategy focusing on task-related layers, allowing for more precise extraction of the model's machine translation capabilities.

- By adjusting task-related layers, we effectively address the issue of catastrophic forgetting, especially its impact on the model's in-context ability, while preserving the model's knowledge on other tasks.

- While enhancing model capabilities, our method places a strong emphasis on the efficient utilization of computational resources, providing a pathway for efficient model training and application under resource-constrained scenarios.

2. Related Work

2.1. Large Language Models

At present, the foundational structure of LLMs is Transformer (Vaswani et al., 2017). LLMs have shown great potential in many applications of NLP. Many relatedworks that using a decoder-only language model can perform multi-task learning on unsupervised monolingual corpora (Radford, 2019). (Ren et al., 2023) discovered the scaling law of LLMs, indicating that as the neural network parameters increase, the capabilities of LLMs also enhance. When parameters reach a certain extent, the model will bring astonishing emergent abilities (Zhang et al., 2022), which is only present in large models. An increasing amount of work has started to focus on the scaling of large language models such as GPT-3 (Brown et al., 2020), BLOOM (Scao et al., 2022) and LLaMA (Touvron et al., 2023). They have shown promising results across a variety of different NLP tasks.

However, the massive scale of LLMs makes fine-tuning very difficult. To overcome this challenge, some research has proposed adapter-based fine-tuning methods. The adapters of LLMs refer to neural modules integrated into LLMs, which contain a small number of additional trainable parameters, allowing for effective fine-tuning on specific tasks without affecting the pre-trained parameters of LLMs. For example, the introduction of parameter-efficient fine-tuning (PEFT) (Houlsby et al., 2019; Lester et al., 2021; Mangrulkar et al., 2022; Fu et al., 2023) allows the model to achieve performance fully equivalent to full-parameter fine-tuning by adjusting only a small number of parameters. In the PEFT, Series Adapter adds the bottleneck feed-forward layer in series to each multi-head and feed forward layer of a Transformer block (Houlsby et al., 2019). And the Parallel Adapter integrates bottleneck feed-forward layers in Parallel with the multi-head and feed-forward layers of a Transformer block in LLMs (Pfeiffer et al., 2020). LoRA introduces trainable low-rank decomposition matrices within the existing layers of LLMs, enabling the model to adapt to new data while keeping the original LLMs fixed to retain previous knowledge.

2.2. LLMs for machine translation

Using LLMs for machine translation with prompt-tuning is attracting increasing attention. (Lin et al., 2022) evaluate GPT-3 and XGLM-7.5B on 182 directions. (Bang et al., 2023) evaluate ChatGPT on 12 directions. The capability of the prompt-tuning LLMs largely depends on its surface representation, small modifications to the prompt can stimulate the abilities inherent in the model, leading to highly variable performance. Prompting LLMs to implement MT is far from the encoder-decoder NMT. In NMT, the target language label is usually appended to the source input to indicate the translation direction (Johnson et al., 2017; Zhang et al., 2020). Additionally, incorporating retrieved phrases and sentences into the input can enhance the translation quality (Li et al., 2022; Garcia and Firat, 2022). (Agrawal et al., 2023) explored strategies for selecting specific input examples and observed that input-relevant examples based on n-gram overlap significantly enhanced the capability of the prompt.

Still, these methods still rely on fine-tuning the model with conventional methods or prompting frozen LLMs, rather than fine-tuning for a specific task. Our research aims to identify the key model parameters for machine translation tasks by observing changes in model parameters, and to fine-tune the model accurately while reducing computational overhead.

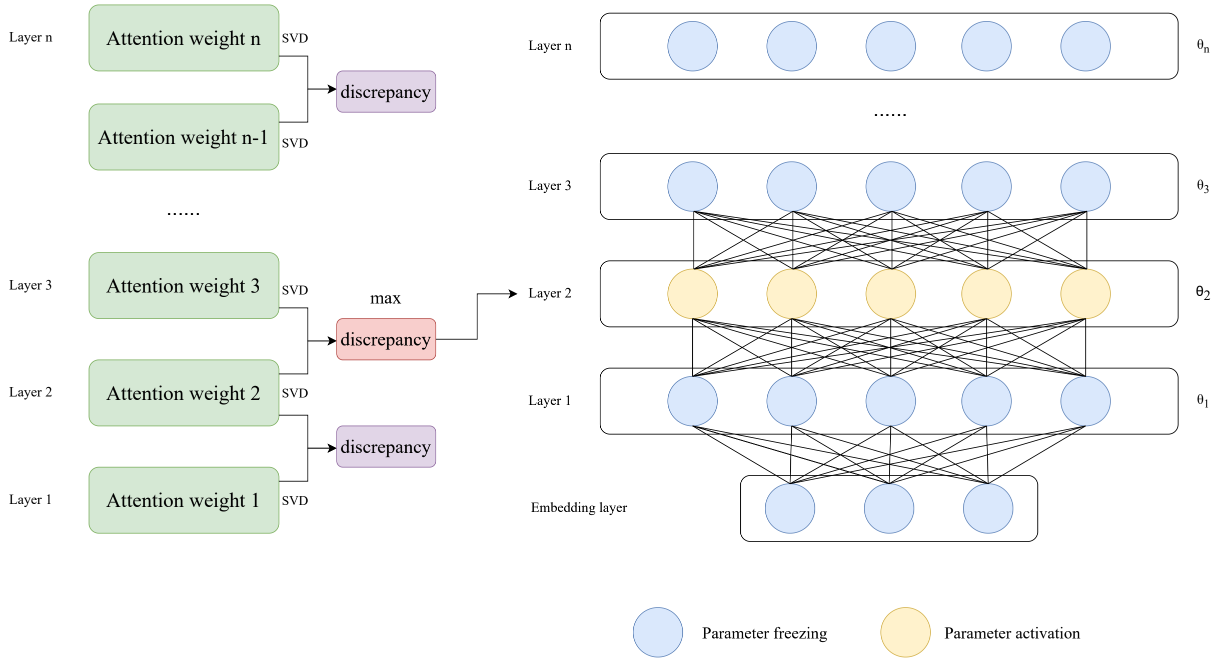


Figure 1: The consists of two stages

3. Proposed Method

Although the LLMs can implement translation, performing translation under a decoder-only architecture, the definition and context of cross-lingual interactions often become ambiguous and variable, making it challenging for the model to define specific cross-lingual interaction rules. In this paper, we propose an adaptive parameter unfreezing strategy that pivots on the quantified discrepancy between adjacent layers within the decoder. The core around identifying and unfreezing the parameters of the layer that manifests the maximal discrepancy in its attention weights, thereby directing the training to focus on these potentially critical parameters. The method consists of two steps, which we show in Figure 1. The first step involves determining the specific layers to unfreeze through SVD analysis. The second step is to unfreeze the parameters for fine-tuning.

3.1. Identification of Special Layers

Let $\mathsf{A}^{(l)}$ denote the attention weights of the l^{th} layer in the decoder, the Singular Value Decomposition (SVD) of which can be articulated as:

$$\mathsf{A}^{(l)} = \mathsf{U}^{(l)}\Sigma^{(l)}(\mathsf{V}^{(l)})^{\mathsf{T}} \tag{1}$$

where $\mathsf{U}^{(l)}$, $\Sigma^{(l)}$ and $\mathsf{V}^{(l)}$ represent the left singular vectors, singular values, and right singular vectors, respectively, of the attention weights $\mathsf{A}^{(l)}$ at layer l. Given that the singular values $\Sigma^{(l)}$ capture the energy or importance of different components in the attention mechanism, the discrepancy $D(l, l+1)$ between two adjacent layers l and $l+1$ can be quantified by comparing their respective singular values. One possible metric for such quantification can be given as:

$$D(l, l+1) = \|\Sigma^{(l)} - \Sigma^{(l+1)}\|_{\mathsf{F}} \tag{2}$$

where $\|\cdot\|_{\mathsf{F}}$ denotes the Frobenius norm.

To assess the significance of the discrepancy, we evaluate $D(l, l+1)$ in valid dataset. Formally, we define $\bar{D}(l, l+1)$, the average discrepancy over n examples between layers l and $l+1$ as:

$$\bar{D}(l, l+1) = \frac{1}{n}\sum_{i=1}^{n} D_i(l, l+1) \tag{3}$$

where $D_i(l, l+1)$ is the discrepancy between layers l and $l+1$ for the i^{th} example. The layer to unfreeze, l^* is determined by:

$$l^* = \arg\max_{l} \bar{D}(l, l+1) \tag{4}$$

Within the context of multilingual machine translation, the universality and flexibility of this adaptive unfreezing strategy become especially pertinent, as different target languages may exhibit varying patterns of inter-layer disparities. Consequently, we extend this layer unfreezing approach to multilingual scenarios, conducting a specific analysis and adjustment of the inter-layer discrepancies for each language pair, to facilitate individualized optimization for each target language. For a given source language translating into multiple target languages, we individually evaluate the inter-layer discrepancies for each target language and decide the layer to unfreeze for that specific language pair. Formally, let L denote the set of target languages, for each target language $lang \in L$, we compute its inter-layer discrepancies $\bar{D}_l(l, l+1)$ and determine the layer l^*_{lang} to unfreeze:

$$l^*_{lang} = \arg\max_{l} \bar{D}_l(l, l+1) \tag{5}$$

In this manner, the model can dynamically adjust its parameter optimization process according to the unique characteristics and specific inter-layer disparity patterns of each target language. This multilingual adaptive layer unfreezing strategy is anticipated to further enhance the performance and generalization capability of the translation model across various target languages.

prompt

Translate the following sentence from [src] to [tgt]:

[src] : [source sentence]

[tgt] : [target sentence]

Figure 2: The template for prompting

3.2. Fine-tuning Specific Parameters

Assuming the model parameters are denoted as Θ = $\{\theta_1, \theta_2 \cdots \theta_n\}$. n represents the total number of layers in the model. Based on the specific layers provided in Section 3.1, we can unfreeze particular parameters represented by θ_i for training. During training, the data needs to be formatted in a specific manner. Embed the source language X and the target language Y into the template T. we adopt the following template in Figure 2. where [src] and [tgt] denote source and target language name of the language pair, [source sentence] represents X, [target sentence] represents Y. Perform Causal Language Modeling tasks during training, Y is used as the supervisory signal. There's a slight difference during the inference phase, there is no Y. The prompt first converts each source input X into a prompt according to template T, then generates a translation Y by providing the prompt to the LLM. In this study, we consider zero-shot prompting for translation.

4. Experimental Setup

Settings We experiment with bloomz-7b1-mt, a LLM with 7B parameters pretrained on 46 languages monolingual corpora. For the maximum token length, we uniformly set it to 2048. Our implementation is based on the pre-trained models in transformers (Wolf et al., 2020), and optimized with ZeRO2 (Rajbhandari et al., 2020) and accelerate (Gugger et al., 2022) during training.

Datasets We work on six languages: English(en), Arabic(ar), Spanish(es), Portuguese(pt), Chinese(zh), French(fr) on opus-100 [1]. We have constructed a training set centered around English, comprising a total of ten language directions, each including both forward and reverse, with 100000 entries per direction, amounting to a total of 1000000 entries. The validation set is used to evaluate attention weights during parameter tuning. Finally, we perform major analysis on opus-100 test set.

We evaluate translation performance using BLEU from SacreBLEU (Post, 2018).

5. Experiments

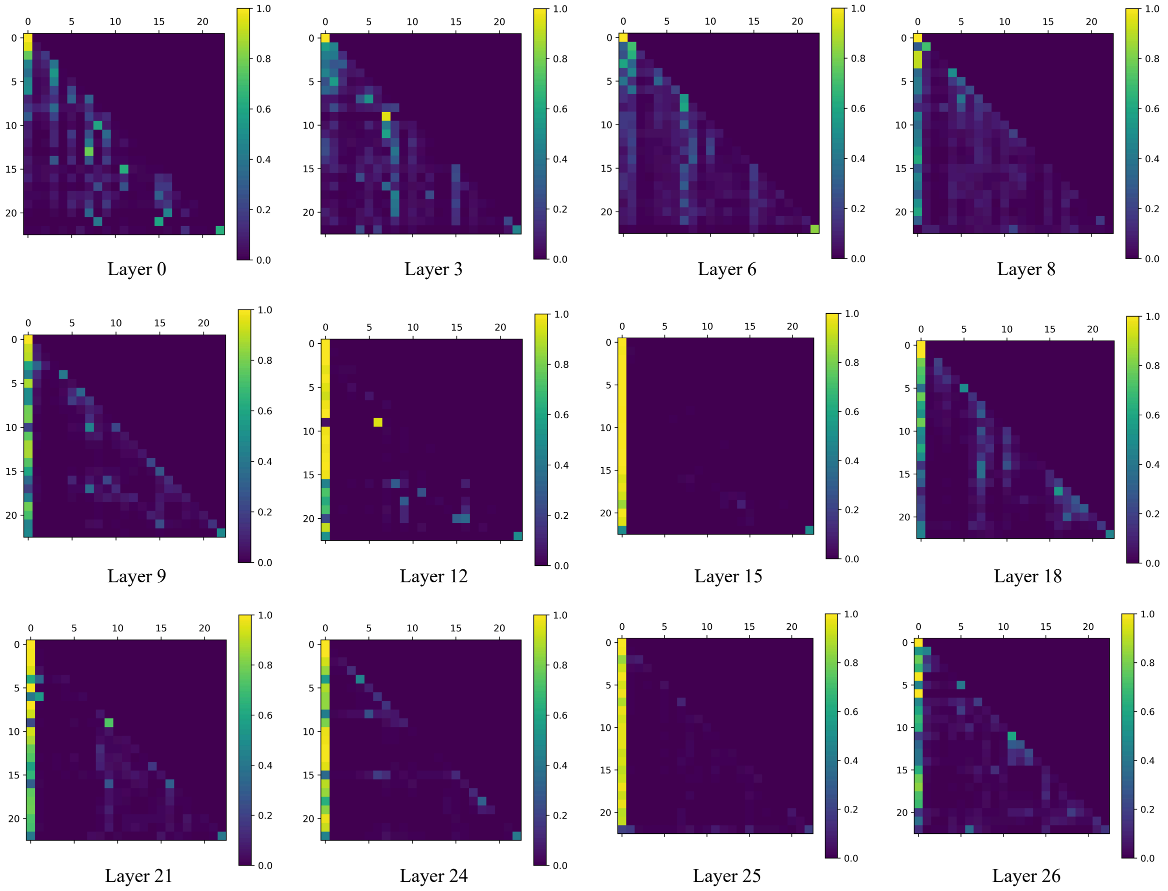
In this section, we provide a detailed account of a series of experiments conducted for our proposed model and method, including Main Result, ablation studies, visual analysis, low-resource language, and domain evaluation, aimed at comprehensively evaluating the effectiveness of our proposed training approach.

5.1. Main Result

As shown in Figure 3, we visualize the attention weights during the translation process from English to French. This attention matrix displays significant variations across multiple layers in images, especially between two particular layers. We observed that the attention distribution of the model undergoes drastic changes between these two layers, which might imply that the model has learned crucial mapping relationships or representations between different language pairs in these two layers. We also find that the model's focus points (parts with higher attention weights) swiftly shift from one structure to another distinct structure, which might represent a pivotal turning point in the model learning syntactic structures or semantic mappings. This finding is particularly noteworthy because models typically comprehend input information at a deep level through appropriate hierarchical decomposition.

Then, we utilize SVD to determine which layer's parameters to unlock. This algorithm is based on a theory that layers where significant changes occur in the attention matrix might be crucial for the model to learn the mapping between the source and target languages. Our strategy is to update and optimize the weights at this layer, allowing the model to gain more flexibility in learning the mappings between different language pairs at this stage. By unfreezing these layers, the model can further enhance

[1] https://opus.nlpl.eu/opus-100.php



Layer 0 Layer 3 Layer 6 Layer 8

Layer 9 Layer 12 Layer 15 Layer 18

Layer 21 Layer 24 Layer 25 Layer 26

Figure 3: Attention weights in English to French machine translation.

its understanding of the more complex structural mappings between source and target languages, thereby improving overall translation performance. As shown in Table 1, compared to three models on multilingual translation tasks: pre-trained models, full-parameter fine-tuning models and random parameter unfreezing models, the specific parameter unfreezing model showed a significant improvement in performance in most language pairs. Compared to full-parameter fine-tuning models that require fine-tuning 7B parameters, specific parameter unfreezing only needs to fine-tune a small number of parameters to achieve similar effects.

However, in the translation task for Arabic, we observed an phenomenon: although the translation performance for most languages improved, the specific parameter unfreezing model did not show a significant enhancement in performance for Arabic compared to the random parameter unfreezing model. Here, we attempt to explore possible reasons. First, the linguistic characteristics of Arabic itself might be a significant factor. The differences in word order between Arabic and many other target languages (especially those using the Latin alphabet) may require the model to do more "reorganization" work during the transformation process. This might become a challenge in certain situations, particularly under our strategy of fine-tuning only some layers of the model by unfreezing them. Secondly, the quality of the pre-trained model and the compatibility with the data might also be a crucial factor. If the pre-trained model performs weakly in Arabic compared to other languages, then achieving desired results might be challenging even with unfreezing and fine-tuning. In this case, we might need more professional and precise pre-training and fine-tuning strategies, such as utilizing richer and more diverse Arabic pre-training data. Even though our layer-unfreezing strategy demonstrates advantages on most language pairs, the case with Arabic reminds us: in practical applications, we need to fully consider the characteristics of the target language, the quality of the pre-trained model, and the selection of source-target language pairs.

5.2. Ablation study

In this Section, we conducted another set of ablation experiments: random layer unfreezing. Compared to the previous methods, we did not choose

Direction	Pre-trained	Fine-tuning	random	specific activation
en-fr	19.4	31.6	25.5	31.3
fr-en	23.7	31.7	27.3	32.6
en-ar	10.9	18.0	16.3	17.9
ar-en	27.9	32.9	34.5	32.7
en-es	15.8	35.2	28.8	35.3
es-en	22.7	28.9	26.8	29.2
en-pt	12.4	30.5	24.8	29.1
pt-en	23.5	31.5	29.5	30.7
en-zh	3.8	23.7	18.6	22.2
zh-en	15.8	27.8	25.7	26.8

Table 1: BLEU scores for different language pairs, with bold text indicating the best translation performance.

layers to unfreeze based on explicit criteria or attention matrix analysis, but randomly selected one or several layers to unfreeze and fine-tune at each training step. Judging from the experimental performance in Table 1, the results of this method are slightly inferior to full unfreezing.

In machine learning, goal-directed optimization is often more effective than random or directionless optimization. Randomly unfreezing layers might mean that the model, while being fine-tuned, lacks a clear and targeted direction for improvement. It may involve parts of the model that are less needed, or overlook areas that require urgent optimization. Since the layers to be unfrozen are randomly selected, it might affect the stability of model training to some extent. Certain layers may involve core parameters of the model, and changes to them during the training process might render the model's learning process unstable. Layers of the model are usually not entirely independent; the update of one layer might depend on the parameter state of other layers. Randomly unfreezing some layers might break this possibility of coordinated updates. The partially unfrozen layers might not be the focal point of the experiment, meaning that even if some layers are optimized, it might not be well transmitted throughout the entire network.

In summary, random layer unfreezing might bring about performance improvement under certain circumstances because some optimization can still be beneficial. However, due to its lack of directionality and possible stability issues, it typically struggles to outperform methods that involve selectively unfreezing layers or completely unfreezing them with a targeted approach.

5.3. Visual Analysis

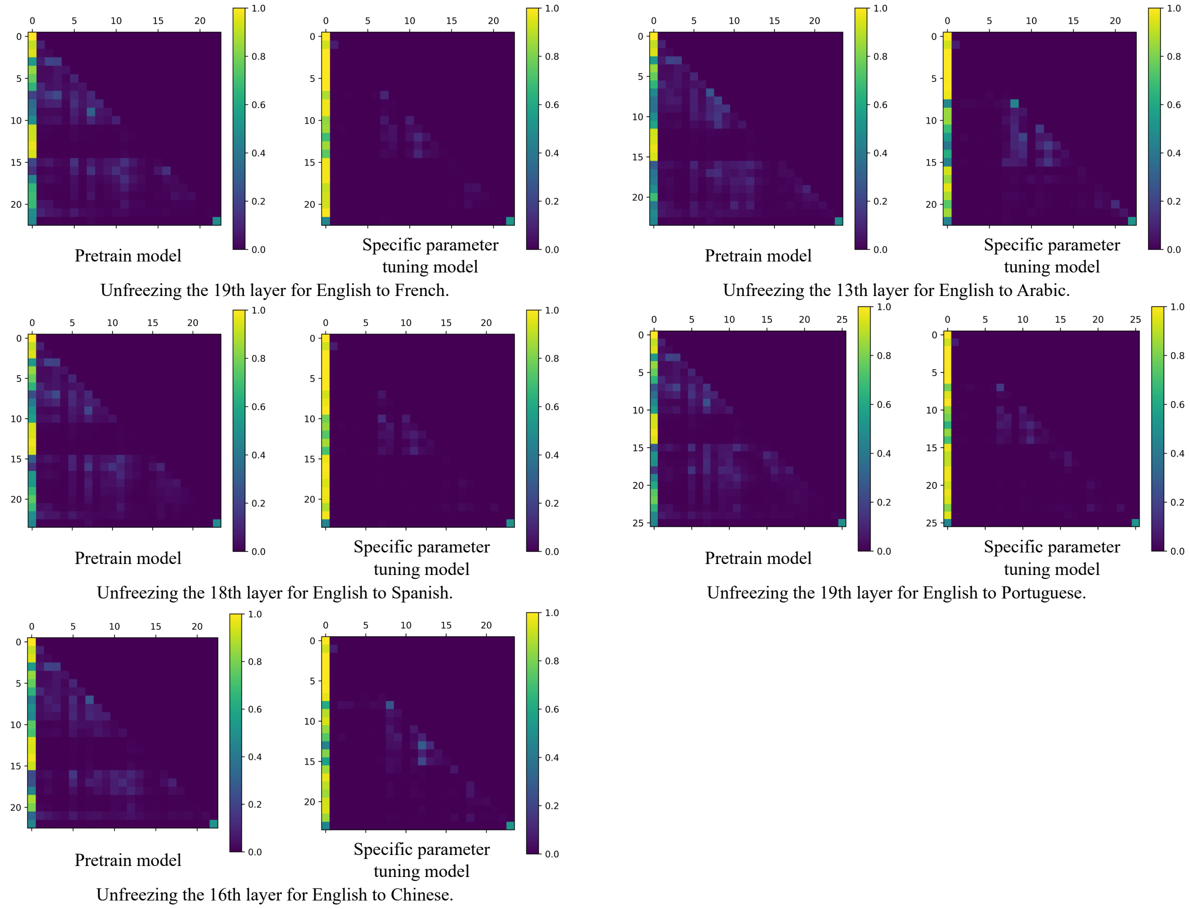
In this subsection, we will further analyze the changes brought about by training after unfreezing specific layers through visualization. We will use five translations from English to other languages as samples, to compare the differences in weights for each language direction between the pre-trained model and the model after training with specific unfrozen layers.

Direction	Pretrain model	Fine-tuning	specific activation
en-id	13.7	24.4	24.6
en-ca	15.8	17.8	17.6
en-hi	6.2	14.9	15.6
en-he	3.9	19.7	19.5

Table 2: BLEU scores for low-resources language pairs.

In the Figure 4, we observe a notable phenomenon, that is, after unfreezing specific parameters, the model's attention matrix undergoes a significant change when handling machine translation tasks. The first word of each sentence is **Translate**.Through visualization, we can see that in the heatmap, the word **Translate** attracts the majority of the model's attention weights, forming a very distinct bright spot. Compared to the model's distributed attention under general circumstances, the focused attention of this particular layer demonstrates how the model tightly locks its focus on this directive word while completing the translation task.

This phenomenon may be a strategy gradually learned by the model during the training process, through optimizing the objective function, that is, to focus attention as much as possible on the keywords describing the nature of the task when dealing with translation tasks, in order to execute the task more precisely and efficiently. The model focuses the majority of its attention on the word **Translate**. This helps the model maintain a very clear and stable objective when performing tasks, ensuring the model understands that its primary role is machine translation, and minimizing the interference of other non-essential information. Such a strategy may also enhance the robustness of the model to a certain extent. When faced with input texts of various types and styles, it can eliminate potential disturbances in the language-to-be-translated related to machine translation tasks.



Unfreezing the 19th layer for English to French.

Unfreezing the 13th layer for English to Arabic.

Unfreezing the 18th layer for English to Spanish.

Unfreezing the 19th layer for English to Portuguese.

Unfreezing the 16th layer for English to Chinese.

Figure 4: Comparison before and after training.

5.4. Low-resource language

In machine translation tasks, many language pairs lack bilingual parallel sentences (Zhu et al., 2023a), resulting in unequal resources for different languages within a model, which is also the case in LLMs. In this experiment, we focused on four language pairs: English to Indonesian (en-id), English to Catalan (en-ca), English to Hindi (en-hi), and English to Hebrew (en-he), each of which represents less than 2% of the training set in the original BLOOM task. Based on the performance exhibited by the model through the previous specific parameter unfreezing strategy, we explored whether similar trends could be observed or new patterns could be found in these low-resource language pairs. The results are shown in Table 2.

After implementing the specific parameter unfreezing strategy on the four language pairs, we observed that the model did indeed achieve a certain degree of performance improvement on all language pairs. However, at the same time, we also found that this improvement is basically on par with or slightly lacking compared to the level achieved with full parameter fine-tuning. The model's capability on low resources is inherently inferior to high-resource languages(Li et al., 2023). Although specific parameter unfreezing can accelerate the model's training speed and enhance performance to a certain extent, it also means that the model can only learn and optimize in the parts of the parameters that are unfrozen, failing to effectively unlock the model's full capability. Consequently, this imposes certain limitations on the model's overall adaptability and learning ability.

5.5. Catastrophic Forgetting

In the deep learning, catastrophic forgetting refers to the phenomenon where a model forgets previously learned knowledge when learning new knowledge. In the scenario of multilingual machine translation, this forgetting is particularly evident when the model is fine-tuned on a specific language pair, which may lead to a significant decline in performance on other non-fine-tuned target languages. In this experiment, we focus on exploring and addressing this issue by implementing two strategies: full-parameter fine-tuning and specific layer unfreezing, and validating them across multiple lan-

Trained Language Pairs	Untrained Language Pairs	Fine-tuning	specific activation
en-fr	en-ar	7.4	9.5
	en-pt	9.2	11.5
	en-es	12.9	14.9
	en-zh	3.7	4.2
en-ar	en-fr	13.8	15.2
	en-pt	9.1	10.9
	en-es	12.4	14.5
	en-zh	2.1	2.2
en-pt	en-fr	17.5	18.2
	en-ar	6.87	9.7
	en-es	13.7	15.1
	en-zh	2.6	3.9
en-es	en-fr	15.8	17.9
	en-ar	8.6	8.9
	en-pt	10.1	11.7
	en-zh	2.8	3.1
en-zh	en-fr	15.6	17.9
	en-ar	7.4	8.6
	en-pt	10.9	11.1
	en-es	11.5	14.1

Table 3: Evaluation Results for Untrained Language Pairs Under Single-Language Training.

guage pairs.

The experiment was set up with five language pairs. Each of these five pairs was individually subjected to full-parameter fine-tuning and specific layer unfreezing fine-tuning, and subsequently tested on the remaining four languages. The experimental results are shown in Table 3. Through the experimental results, we can observe that the full-parameter fine-tuning performs worse in the cross-language generalization experiment compared to the specific layer unfreezing strategy. Full-parameter fine-tuning may lead the model to over-optimize for a specific language pair, sacrificing its generalization capability across other language pairs. Comprehensive parameter updates might reshape the model's language representation and generalization capabilities, and this specialized learning might not be beneficial for maintaining performance on other language pairs. Since all parameters are retuned, the model might lose the multi-language universal knowledge or specific language grammatical and semantic rules accumulated during the pre-training phase, while gradually adapting to the new language pair.

The strategy of specific layer unfreezing can often better preserve the language representation capabilities of a pre-trained model because it only updates parameters in certain layers, allowing the model to acquire new language feature learning while retaining original multilingual knowledge. This method can achieve more refined fine-tuning objectives by selectively unfreezing certain crucial layers, for example, those more closely related to specific tasks or language features, thereby improving performance on the target language pair while maintaining some level of generalization capability on other language pairs.

5.6. Multilingual Translation Hierarchical Analysis

In this section, we analyzed the factors that affect the changes of model weights. In general, models may tend to learn to capture universal language features at lower layers and capture more specific language features at higher layers (Collobert et al., 2011). The difficulty of translation may vary for different languages. Some language pairs may require more lower-level features to handle grammar structures and basic translation, while others may require more higher-level features to handle semantics and context. This could lead to variations in attention weights concentrated at different layers. On the other hand, the distribution of the data in the pre-training corpus can also affect the changes in attention weights. In language pairs with abundant resources, the model may learn higher-level features more quickly, leading to attention weight changes that are more concentrated at higher layers. Conversely, in language pairs with limited resources, the model may rely more on lower-level features, resulting in attention weight changes that are more concentrated at lower layers. In Table 4, we present the data distribution of BLOOM during model training, as well as the target layers unfrozen by our unfreezing strategy and the final results ob-

Direction	Distribution(%)	Layer(th)	BLEU
en-ar	4.6	13	31.3
en-fr	12.9	19	17.9
en-es	10.8	18	35.3
en-pt	4.9	19	29.1
en-zh	16.2	16	22.2

Table 4: The distribution of **en-xx** target language data in BLOOM, the unfreezing positions, and the BLEU scores obtained after training with special unfreezing.

tained. Arabic and Chinese exhibit significant gap compared to English. However, due to the larger volume of Chinese data, the layers where weight changes occur tend to be positioned higher than those in Arabic. The data volume for Portuguese is approximately similar to Arabic, but Portuguese is more similar to English. Therefore, the model tends to learn higher-level features, and the layers where weight changes occur are also positioned higher. In summary, the proximity of language pairs and the data distribution in pre-trained models are key factors that determine where attention weight changes are concentrated. This difference can help identify the appropriate unfreezing strategy, allowing for the selection of unfrozen layers based on the specific requirements of each language pair.

6. Conclusion

In this paper, we explore the challenges and unresolved issues in implementing machine translation with LLMs. We propose a novel method that involves adjusting task-related layers in large models to better harness their machine translation capabilities. Our approach aims to retain the model's knowledge on other tasks while optimizing performance on a specific translation task. We achieve this by revealing the structure and characteristics of attention weights through SVD, understanding the model's behavior at each layer, and making fine adjustments to specific layers. This method not only addresses computational resource consumption and catastrophic forgetting but also offers a new perspective on utilizing large models more effectively.

Acknowledgements

This work is partly supported by the Foundationand Cutting-Edge Technologies Research Program of Henan Province(Grant No.232102210107) and the National Natural Science Foundation of China Youth Foud (Grant No.11701525).

7. References

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2023. In-context examples selection for machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8857–8873, Toronto, Canada. Association for Computational Linguistics.

Md Mahfuz Ibn Alam and Antonios Anastasopoulos. 2022. Language adapters for large-scale MT: the GMU system for the WMT 2022 large-scale machine translation evaluation for african languages shared task. In *Proceedings of the Seventh Conference on Machine Translation, WMT 2022, Abu Dhabi, United Arab Emirates (Hybrid), December 7-8, 2022*, pages 1015–1033. Association for Computational Linguistics.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *CoRR*, abs/2302.04023.

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. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel P. Kuksa. 2011. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, 12:2493–2537.

Zihao Fu, Haoran Yang, Anthony Man-Cho So, Wai Lam, Lidong Bing, and Nigel Collier. 2023. On the effectiveness of parameter-efficient fine-tuning. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 12799–12807. AAAI Press.

Xavier Garcia and Orhan Firat. 2022. Using natural language prompts for machine translation. *CoRR*, abs/2202.11822.

Nicolas Garneau, Mathieu Godbout, David Beauchemin, Audrey Durand, and Luc Lamontagne. 2020. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings: Making the method robustly reproducible as well. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 5546–5554. European Language Resources Association.

Sylvain Gugger, Lysandre Debut, Thomas Wolf, Philipp Schmid, Zachary Mueller, Sourab Mangrulkar, Marc Sun, and Benjamin Bossan. 2022. Accelerate: Training and inference at scale made simple, efficient and adaptable. `https://github.com/huggingface/accelerate`.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

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.

Zhiqiang Hu, Yihuai Lan, Lei Wang, Wanyu Xu, Ee-Peng Lim, Roy Ka-Wei Lee, Lidong Bing, and Soujanya Poria. 2023. Llm-adapters: An adapter family for parameter-efficient fine-tuning of large language models. *CoRR*, abs/2304.01933.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda B. Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Trans. Assoc. Comput. Linguistics*, 5:339–351.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. Generalization through memorization: Nearest neighbor language models. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 3045–3059. Association for Computational Linguistics.

Shangjie Li, Xiangpeng Wei, Shaolin Zhu, Jun Xie, Baosong Yang, and Deyi Xiong. 2023. MMNMT: modularizing multilingual neural machine translation with flexibly assembled moe and dense blocks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 4978–4990. Association for Computational Linguistics.

Yafu Li, Yongjing Yin, Jing Li, and Yue Zhang. 2022. Prompt-driven neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2579–2590, Dublin, Ireland. Association for Computational Linguistics.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona T. Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 9019–9052. Association for Computational Linguistics.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, and Sayak Paul. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`.

Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023. Adaptive machine translation with large language models. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation, EAMT 2023, Tampere, Finland, 12-15 June 2023*, pages 227–237. European Association for Machine Translation.

Jonas Pfeiffer, Ivan Vulic, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: an adapter-based framework for multi-task cross-lingual transfer.

In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 7654–7673. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Alec Radford. 2019. Language models are unsupervised multitask learners.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: memory optimizations toward training trillion parameter models. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, SC 2020, Virtual Event / Atlanta, Georgia, USA, November 9-19, 2020*, page 20. IEEE/ACM.

Xiaozhe Ren, Pingyi Zhou, Xinfan Meng, Xinjing Huang, Yadao Wang, Weichao Wang, Pengfei Li, Xiaoda Zhang, Alexander Podolskiy, Grigory Arshinov, Andrey Bout, Irina Piontkovskaya, Jiansheng Wei, Xin Jiang, Teng Su, Qun Liu, and Jun Yao. 2023. Pangu-σ: Towards trillion parameter language model with sparse heterogeneous computing.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, and et al. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. *CoRR*, abs/1409.3215.

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. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Minghan Wang, Jinming Zhao, Thuy-Trang Vu, Fatemeh Shiri, Ehsan Shareghi, and Gholamreza Haffari. 2023. Simultaneous machine translation with large language models. *CoRR*, abs/2309.06706.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023a. Prompting large language model for machine translation: A case study. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 41092–41110. PMLR.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: open pre-trained transformer language models. *CoRR*, abs/2205.01068.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023b. Siren's song in the AI ocean: A survey on hallucination in large language models. *CoRR*, abs/2309.01219.

Shaolin Zhu, Chenggang Mi, Tianqi Li, Yong Yang, and Chun Xu. 2023a. Unsupervised parallel sentences of machine translation for asian language pairs. *ACM Trans. Asian Low Resour. Lang. Inf. Process.*, 22(3):64:1–64:14.

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