Proceedings of the 5th International Conference on Computing Innovation and Applied Physics
DOI: 10.54254/2753-8818/2026.31854

A Two-Stage Fine-Tuning Method for Large Language Models Towards Interpretable Medical Translation

Jingjing Li[1*], Yuan Hong[1]

[1]*School of Artificial Intelligence and Computer, North China University of Technology, Beijing, China*

**Corresponding Author. Email: 1115095230@qq.com*

Abstract. Translation in the medical field is an important research direction for large language models. Medical translation texts are usually very professional and have different meanings for the same word. We have very high requirements for the trustworthiness of translation models. But the existing methods are not good at making professional term accuracy and semantic consistency unified. Medical experts can hardly trust the decisions made by the models. For this reason, this study puts forward an interpretable framework with two-stage LoRA fine-tuning based on LLaMA-3-Chinese-8B-Instruct-v3—TSX-MedTrans. The first stage is term-level fine-tuning to make the representation of professional terms better. The second stage is sentence-level fine-tuning to make the context semantic coherence better. This framework also adds an interpretability module. It is used to show the attention distribution visually, so as to check the model's decision-making mechanism. The experimental results show that TSX-MedTrans improves the translation quality a lot: after term-level fine-tuning, BLEU rises from 29.624 to 34.728, and COMET rises from 0.836 to 0.891; after sentence-level fine-tuning, BLEU rises from 34.583 to 38.232, and COMET rises from 0.846 to 0.876.

Keywords: Medical Translation, Large Language Model, TSX-MedTrans Framework, LoRA, Fine-Tuning

1. Introduction

AI technology is developing very fast. It is making big changes in many fields. Large Language Models (LLMs) have become a very important foundation in the Natural Language Processing (NLP) field. Machine translation is one of the typical uses of LLMs. It is very important in medical scenes, such as international medical communication, scientific research cooperation and clinical decision-making. The accuracy and reliability of medical text translation are closely related to medical quality. So, to check the model's output and do trustworthy reviews, LLMs for medical translation need to be highly interpretable and transparent.

Researchers have studied the use of LLMs in professional fields like law and medicine. In the law field, Briva-Iglesias [1] pointed out that consistent terms and understanding the context are important. Zhang [2] found that fine-tuning a small number of parameters of medium-sized models (such as LLaMA-2-13B) can greatly improve translation performance. But existing models still have

DOI: 10.54254/2753-8818/2026.31854

obvious shortcomings in legal term explanation, generalization across legal systems and deep reasoning. Their reliability needs to be improved [3-5]. In the medical field, Galiero [6] stressed that high-quality domain corpus is still the key to improving translation performance. Chen [7] proved that domain-specific training is effective. Zheng [8] combined methods such as medical dictionary enhancement, RAG retrieval and LoRA fine-tuning. This made the success rate of medical term translation rise a lot to 78%. Besides, Xu [9] and Mujadia [10] put forward two methods respectively. Xu [9] proposed a two-stage method that uses monolingual corpus to enhance cross-language ability and aligns with parallel corpus. Mujadia [10] proposed a two-stage framework combining full-parameter fine-tuning and LoRA. Both methods effectively improved the translation quality of low-resource language pairs. In general, although LLMs show great potential in translation tasks of professional fields such as law and medicine, they still have obvious limitations in term accuracy and cross-language robustness. There is still a gap from the expected goals.

For the high interpretability demand of translation results in vertical fields, many researchers have put forward various strategies to improve model interpretability. Leiter [11] proposed a two-layer framework of "decision explanation - model explanation". He systematically classified interpretability methods into five categories. And he pointed out that existing evaluation indicators still have obvious shortcomings in fidelity and credibility. Besides, Futeral [12] realized the dynamic attention intensity measurement and key head recognition in multi-modal alignment through a three-level attention view system. Palikhe [13] systematically summarized the interpretability research of LLMs based on Decoder-only architecture. He used the BertViz tool to directly analyze and visualize the Self-Attention Mechanism inside the model, and revealed the focus of the model when generating text. These studies show that attention visualization is widely used in model explanation. It also has generality and flexibility in cross-modal and cross-task scenes.

However, the progress of existing research is still mainly focused on the overall translation quality and general interpretability. There is a lack of fine-grained traceability and mechanistic verification for medical terms. First, existing research is still insufficient in term-level accuracy. Medical terms are highly professional and polysemous. Models trained with general fine-tuning or small-scale domain corpus are still easy to have ambiguity or mistranslation when dealing with key terms. Especially in complex texts such as long sentences, compound sentences or multi-term combinations, such errors will damage the overall semantic coherence and reduce the usability of translations in clinical or scientific research scenes. Second, although there are many fine-tuning strategies, including one-stage or two-stage fine-tuning, most methods are still insufficient in the collaborative optimization of terms and contextual semantics. For this reason, this study focuses on the medical Chinese-English translation task, and proposes an interpretable framework integrating parameter-efficient optimization and translation decision analysis — TSX-MedTrans.

The framework uses a two-stage fine-tuning method and interpretability analysis to systematically improve the translation quality and credibility of the model. The first stage uses term datasets for LoRA fine-tuning, making the model learn the feature distribution of medical terms. The second stage of fine-tuning uses sentence-level datasets to improve the consistency of the model's semantics and context. This strategy greatly improves the accuracy and consistency of medical translation.

An interpretability module is introduced into the framework. Through attention weight visualization and interpretability alignment analysis, it can directly show the attention distribution of the model in cross-language term alignment, and provide traceable evidence for the verification of generation results.

DOI: 10.54254/2753-8818/2026.31854

After training with the TSX-MedTrans framework, the model's translation generation results are traceable and verifiable, which strongly supports medical experts to carry out translation review and reliability verification. This interpretability not only strengthens the trust foundation of LLMs to assist manual translation in the medical field, but also provides a feasibility study for the application of trustworthy AI in key medical scenes such as clinical decision-making.

2. The TSX-MedTrans framework

With limited computing resources, to make the model light and optimize its training performance, this study chooses LLaMA-3-Chinese-8B-Instruct-v3 as the base model. We use the efficient LoRA fine-tuning method to train it. LLaMA-3-Chinese-8B-Instruct-v3 is based on LLaMA-3-8B-Instruct. It has an expanded Chinese vocabulary and is retrained with a large number of Chinese corpus. So its ability to understand and generate Chinese is much better. LLaMA-3-Chinese-8B-Instruct-v3 is better than LLaMA-3-8B-Instruct in multilingual processing and resource use efficiency.

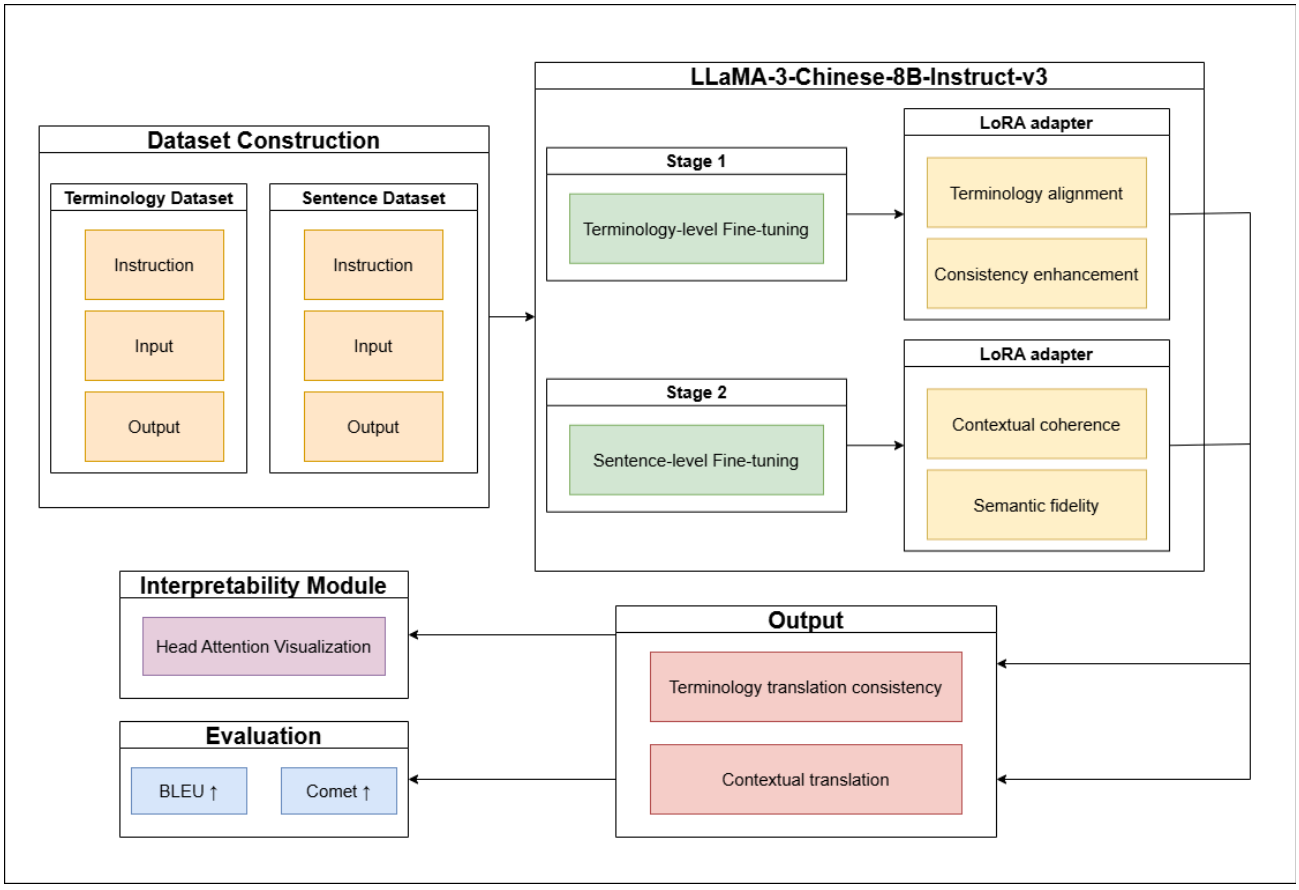

Figure 1. TSX-MedTrans flow chart: an interpretable two-stage lora fine-tuning framework

This study puts forward an interpretable framework TSX-MedTrans based on two-stage LoRA fine-tuning. It is used to improve the model's adaptability and interpretability in the medical field. Figure 1 is the flow chart of the framework. The framework uses a layered fine-tuning strategy. It makes the model connect the term-level knowledge learning and sentence-level semantic understanding well. It ensures both the accuracy of medical translation and the consistency of context.

DOI: 10.54254/2753-8818/2026.31854

The first stage is fine-tuning for term-level data. The model is trained on high-quality Chinese-English medical term pairs. This can improve its ability to recognize and align proper nouns in the medical field. The goal of this stage is to make the model build more stable term representations in the encoding and decoding process. In this way, the model can keep the consistency of key concepts when doing cross-language mapping. Then, the second stage of fine-tuning is for the sentence-level dimension. On the basis of the first stage fine-tuning, the model is trained on sentence-level corpus with complex semantic relations and context dependencies. This can improve its language generation ability in medical narration, diagnosis description and term co-occurrence scenes. This stage focuses on optimizing the model's context modeling and semantic fidelity. When the model translates long sentences and multi-term texts, it can keep the overall semantic fluency and also maintain the term accuracy obtained in the first stage.

To verify the actual contribution of two-stage fine-tuning in term translation, a term-level interpretability module is added to TSX-MedTrans. This module is based on the BertViz visualization tool. It visualizes and analyzes the model's attention mechanism to medical terms during translation. By inputting medical field texts, we can observe the attention distribution of the model when generating professional terms in the target language. We can also analyze its stability and transferability in complex contexts. This module is not only used to verify the effectiveness of the fine-tuning strategy, but also provides empirical support for the model's transparency and interpretability.

To fully evaluate the performance of TSX-MedTrans in medical translation tasks, this study uses two indicators, BLEU [14] and COMET [15], for quantitative evaluation. BLEU evaluates the generated results based on n-gram matching. It mainly measures the local accuracy and fluency of the translation. It can reflect the model's precision in term translation and phrase structure. COMET evaluates the overall semantic consistency between the translation and the reference text through a pre-trained language model. Its results are highly consistent with human evaluation. It can capture the semantic fidelity in long sentences and complex contexts. The combined evaluation results of the two indicators can not only clarify the translation accuracy of the model in generating professional term translations, but also reflect the semantic consistency and context coherence of the model's generated results in the overall context. In this way, we can more comprehensively and reliably quantify the improvement effect of the TSX-MedTrans framework on the medical text translation ability of the base model.

In the first stage, this study builds a term-level Chinese-English aligned dataset. The data comes from Common Clinical Medical Nouns (2023 Edition) released by the National Health Commission of the People's Republic of China. This dataset contains a total of 52,778 Chinese-English medical term pairs. It covers 32 clinical departments such as ophthalmology, otolaryngology, stomatology, emergency department and cardiology department. We divide this dataset into training set, validation set and test set at a ratio of 90:5:5. It is used for term-level fine-tuning and evaluation.

The sentence-level Chinese-English parallel dataset used in the second stage is built from two authoritative medical corpora. It contains a total of 62,390 parallel corpus in the medical field. The first corpus is the WMT19 Biomedical Translation Task Dataset (hereinafter referred to as WMT19 BioMT Dataset). This dataset is released by the WMT 2019 International Machine Translation Competition and supported by institutions such as ACL. WMT19 BioMT is built based on the titles and abstracts of scientific papers in the Medline literature database. It provides high-quality English-Chinese parallel corpus. The second corpus is the ParaMed Dataset. It is a parallel corpus for Chinese-English translation in the medical field. This corpus is crawled from the website of The New England Journal of Medicine and has been included in the WMT22 medical field dataset. Same

DOI: 10.54254/2753-8818/2026.31854

as the first stage, we divide this sentence-level dataset into training set, validation set and test set at a ratio of 90:5:5. It is used for the second stage of fine-tuning and evaluation.

To fully evaluate the improvement of the first and second stage fine-tuning on the base model's translation ability in the medical field, we combine the test sets divided from the two-stage datasets to build the final evaluation test set. This evaluation test set contains a total of 5,757 Chinese-English parallel corpus. It is used for the objective quantitative evaluation of model performance.

3. Experimental results and discussion

3.1. Training settings

In the first stage of fine-tuning, we set the LoRA rank to 16. We do low-rank adaptation on the model weights, and only need to update about 0.1% of the parameters. The MLP layer plays a core role in language modeling and knowledge storage [16]. Fine-tuning this layer helps the model learn and integrate domain-specific knowledge [17,18]. So we insert LoRA adapters at three positions of the MLP module: gate_proj, up_proj and down_proj. This lets the model capture and solidify static knowledge such as medical terms better.

In the second stage, we focus on the model's translation adaptation ability at the sentence level. The self-attention layer is mainly responsible for modeling context dependencies and cross-language alignment. That is, it confirms the correspondence between source language words and target language words [19,20]. So in this stage, we insert LoRA adapters at two positions of the attention layer: q_proj and v_proj. This design can strengthen the model's alignment and fluency modeling. It can also keep good training stability while ensuring low video memory usage and few parameters [21]. Considering the larger semantic span of the data in this stage, we adjust the LoRA rank to 8, and still only update about 0.1% of the parameters.

The training hyperparameter settings are basically the same for the first stage (term-level fine-tuning) and the second stage (sentence-level fine-tuning). Specifically, both stages use the same hyperparameters: batch size is 16, warmup ratio is 0.03, maximum sequence length is 1024. We train the model for 5 epochs in each stage, and select the model with the lowest validation set loss as the final model.

3.2. Experimental results

3.2.1. Stage 1

After the first stage of fine-tuning with term-level data, LLaMA-3-Chinese-8B-Instruct-v3 shows a significant improvement in both BLEU and COMET indicators. Specifically, the BLEU value rises from 29.624 to 34.728. This means the model's translation accuracy of professional terms and local structures is significantly enhanced through term-level fine-tuning. The COMET score rises from 0.836 to 0.891. This shows term-level fine-tuning effectively improves the overall semantic consistency between the model's translations and the reference translations. The improvement of BLEU and COMET fully proves that term-level fine-tuning is effective in enhancing the model's domain term knowledge and context alignment ability.

3.2.2. Stage 2

After the second stage of fine-tuning with sentence-level data, LLaMA-3-Chinese-8B-Instruct-v3's translation performance is further improved. Specifically, the BLEU value rises from 34.583 to

DOI: 10.54254/2753-8818/2026.31854

38.232. This means the base model's semantic alignment at the sentence level and translation quality are further enhanced through sentence-level fine-tuning. At the same time, the COMET score rises from 0.846 to 0.876. This shows sentence-level fine-tuning makes the model further improve the overall semantic consistency and context coherence.

To further verify the comprehensive advantages of the TSX-MedTrans framework in improving model translation performance, we conduct a systematic comparison of current representative translation LLMs on a unified test set covering term-level and sentence-level data (see Table 1). Among them, the multilingual translation model BigTranslate-13B uses general language understanding and generation abilities to achieve high-quality cross-language translation. m2m100-1.2B is a typical baseline model in the research of general-domain multilingual translation tasks, and is widely used in cross-language transfer and low-resource translation research. The base model LLaMA-3-Chinese-8B-Instruct-v3 is trained with the Teaching with Comparison (TIM) framework to get the model TIM-LLaMA3-Chinese-8B-Instruct-v3, which has better zero-shot translation ability. In addition, we also select the current state-of-the-art general large language model Qwen2-7B-Instruct as the SoTA reference.

Table 1. Performance comparison of four representative large language models and the LLaMA-3-Chinese-8B-Instruct-v3 model (before and after fine-tuning with the TSX-MedTrans framework) on the same evaluation dataset. The bold data represents the performance of the model fine-tuned with the TSX-MedTrans framework in this study

Model Name	BLEU ↑	COMET ↑
BigTranslate-13B	33.151	0.733
m2m100-1.2B	34.257	0.732
TIM-LLaMA3-Chinese-8B-Instruct-v3	30.329	0.737
LLaMA3-Chinese-8B-Instruct-v3	34.583	0.846
Qwen2-7B-Instruct	31.393	0.794
TSX-MedTrans-8B-LoRA(OURS)	38.232	0.876

From the evaluation results in Table 1, we can see that TSX-MedTrans-8B-LoRA achieves a leading performance in both BLEU and COMET indicators. Specifically, its BLEU value reaches 38.232, which is about 5.081 and 3.975 higher than that of the multilingual translation models BigTranslate-13B (33.151) and m2m100-1.2B (34.257) respectively. Compared with TIM-LLaMA3-Chinese-8B-Instruct-v3 (30.329) fine-tuned with the TIM framework and LLaMA3-Chinese-8B-Instruct-v3 (34.583) without two-stage fine-tuning, the BLEU improvement is more significant. This fully shows that the proposed two-stage fine-tuning strategy can effectively enhance the model's translation ability.

In terms of semantic consistency, the COMET score of TSX-MedTrans reaches 0.876, which is about 0.14 higher than that of BigTranslate-13B (0.733) and m2m100-1.2B (0.732), and is significantly better than traditional neural machine translation models. At the same time, its score is also higher than that of the SoTA model Qwen2-7B-Instruct (0.794), showing the comprehensive advantages of TSX-MedTrans in overall semantic understanding and translation quality.

DOI: 10.54254/2753-8818/2026.31854

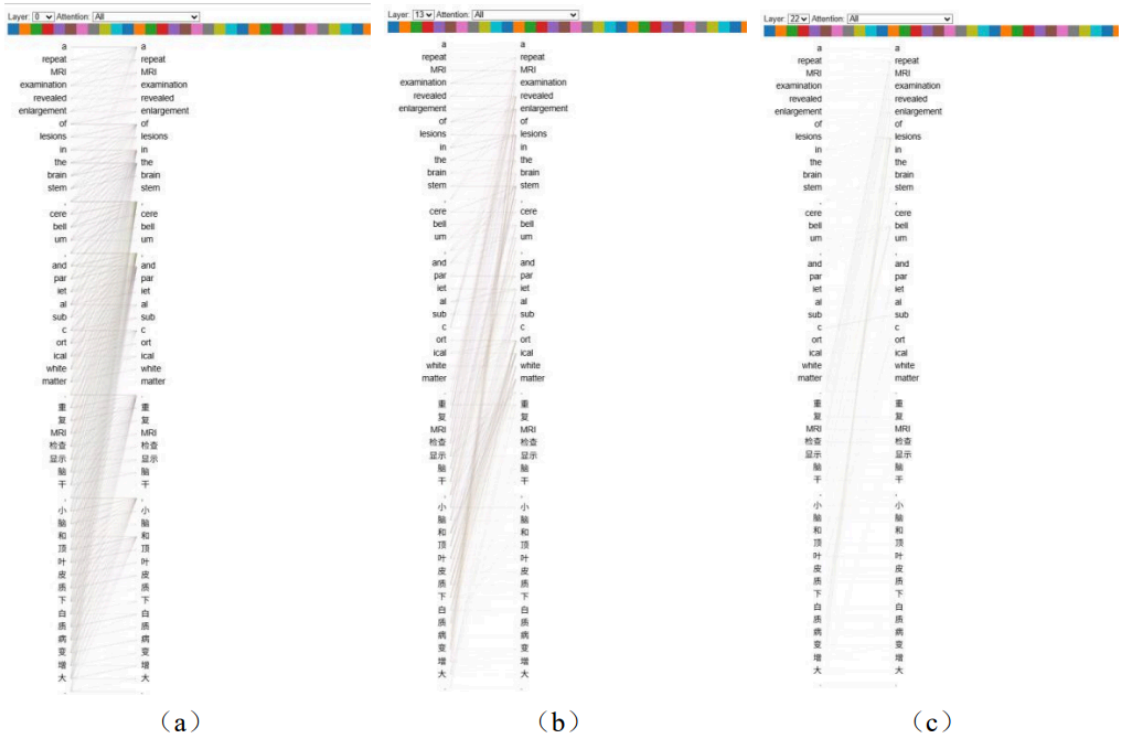

Figure 2. This figure shows the attention weight distribution of Layer 1, Layer 14 and Layer 23. The English text is the source language sentence, and the Chinese text is its target language translation. (a) is the attention weight distribution of Layer 1, (b) is the attention weight distribution of Layer 14, (c) is the attention weight distribution of Layer 23

In general, the experimental results verify the performance improvement advantages of the TSX-MedTrans framework for the base model in medical translation tasks. Through phased fine-tuning at the term and sentence levels, the framework significantly improves the coherence and semantic consistency of translations while maintaining training efficiency. It provides a scalable and interpretable solution for the adaptation of large language models in professional domain translation tasks.

4. Interpretability

To meet the demand for high reliability and interpretability of model translation results in the medical field, this study further uses the BertViz tool to visually analyze the attention weights during the model generation process. This can directly show the model's attention distribution in intra-sentence semantic modeling and term translation. Different colors stand for different attention heads, and the thickness of the lines shows the strength of the attention scores.

Layer 1 is a shallow structure. It mainly focuses on function words like "a" and "of", and the local connections between adjacent tokens. It also builds basic word mapping relationships (as shown in Figure 2(a)). Layer 14 is a middle layer. It focuses on the overall semantic correspondence of term phrases such as "cerebellum" and "parietal subcortical white matter" (as shown in Figure 2(b)). Layer 23 is a deep structure. It is responsible for controlling the core semantics and the overall expression of the translation, to make sure medical information is accurate and fluent (as shown in Figure 2(c)).

DOI: 10.54254/2753-8818/2026.31854

In addition, Figure 3(a) and Figure 3(b) show the weight distribution of different attention heads in Layer 14. We can see that each attention head pays high attention when identifying medical terms such as "cerebellum" and "parietal subcortical white matter". This shows the model has formed a relatively stable pattern in term recognition and context alignment. By comparing the attention weights of LLaMA-3-Chinese-8B-Instruct-v3 before and after fine-tuning under the TSX-MedTrans framework (see Figure 3(b) and Figure 3(c)), we find that the attention distribution is more concentrated after two-stage fine-tuning, and the term correspondence is clearer.

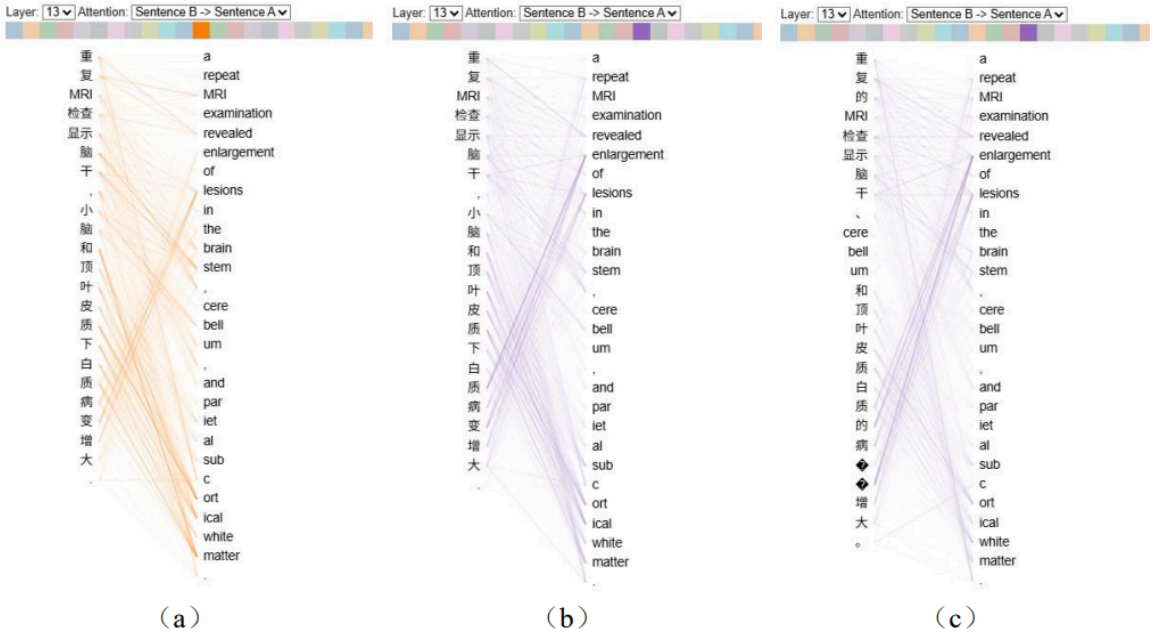

Figure 3. This figure shows the weight distribution of Attention Head 12 and Attention Head 15 in Layer 14. In the three subgraphs, the English in the right column is the tokens of the source language sentence, and the Chinese in the left column is the tokens of the target language translation. (a) Weight distribution of Attention Head 12; (b) Weight distribution of Attention Head 15; (c) Weight distribution of Attention Head 15 before two-stage fine-tuning. In (c), some Chinese characters are split into multiple tokens and cannot be displayed as recognizable characters, resulting in garbled symbols

4.1. Ablation experiment

To further analyze the effect of two-stage fine-tuning, we compare the performance of the base model in three stages on the same evaluation dataset: the unfinetuned base model, the base model after the first stage of term-level fine-tuning, and the base model after the first stage of term-level fine-tuning plus the second stage of sentence-level fine-tuning. The comparison results are shown in Figure 4. The results show that the model's translation ability in the medical field improves step by step with the progress of phased fine-tuning.

The BLEU score of the base model without TSX-MedTrans fine-tuning is 34.983, and the COMET score is 0.846. After the first stage of term-level fine-tuning (+ Stage 1 Fine-tuning), the performance of the base model is significantly improved compared with before fine-tuning: the BLEU score rises to 36.986, and the COMET score rises to 0.862. This shows that term-level fine-tuning can effectively enhance the model's learning and mastery of medical domain knowledge and improve its term translation accuracy. On this basis, after the second stage of sentence-level fine-

DOI: 10.54254/2753-8818/2026.31854

tuning (+ Stage 1 + Stage 2 Fine-tuning), the model performance is further improved: the BLEU score rises to 38.232, and the COMET score reaches 0.876. Both indicators are significantly higher than those in the previous two stages. This result fully shows that the phased fine-tuning strategy based on term level and sentence level can improve the base model's ability to model medical terms and syntactic logic, thus improving medical translation performance.

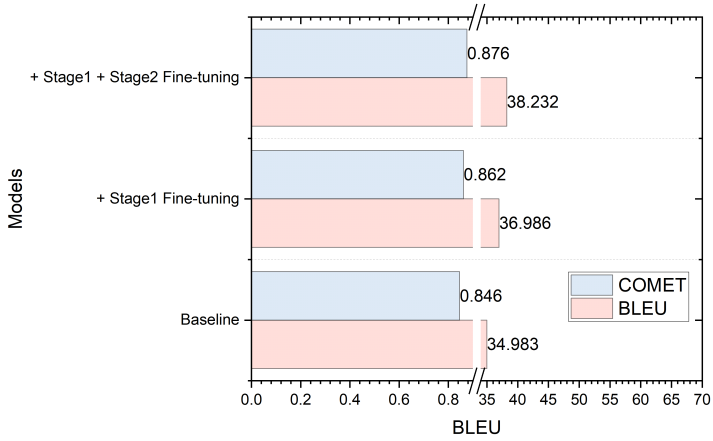

Figure 4. Performance evaluation results of three scenarios (before two-stage fine-tuning, only after the first stage of fine-tuning, after the complete two-stage fine-tuning) on the same evaluation dataset, based on the base model LLaMA-3-Chinese-8B-Instruct-v3

5. Conclusion

Aiming at the problems of insufficient term accuracy and weak semantic consistency in medical field translation tasks, this study proposes an interpretable framework TSX-MedTrans based on two-stage LoRA fine-tuning. This framework takes LLaMA-3-Chinese-8B-Instruct-v3 as the base model. Through the phased fine-tuning strategy at term level and sentence level, it improves the model's adaptation to professional knowledge and context semantic modeling.

The experimental results show that TSX-MedTrans significantly improves the translation quality in both stages. After the first stage of term-level fine-tuning, the BLEU value rises from 29.624 to 34.728, and the COMET score rises from 0.836 to 0.891. This proves that the term-level fine-tuning method is effective in enhancing the model's knowledge of professional terms in the field and reducing lexical ambiguity. After the second stage of sentence-level fine-tuning, the BLEU value rises from 34.583 to 38.232, and the COMET score rises from 0.846 to 0.876. This further shows the role of the sentence-level fine-tuning method in improving the overall semantic consistency and translation fluency.

To further explore the model's translation generation path, an interpretability module is added to TSX-MedTrans. Through this module, we visually analyze the attention weights of the model in the process of term translation and semantic modeling. The results show that after the first stage of term-level fine-tuning, the model's focus on key medical terms in the middle and high-level attention heads is significantly enhanced. And the second stage of sentence-level fine-tuning makes the model form a more concentrated attention distribution in the whole sentence, thus improving context consistency and semantic coherence. This result reveals the complementary mechanism of two-stage fine-tuning at different levels, and also verifies the effectiveness of TSX-MedTrans in model interpretability. Ablation experiments further confirm that missing either the term-level adaptation or

DOI: 10.54254/2753-8818/2026.31854

the sentence-level fine-tuning stage will lead to insufficient performance improvement. It shows that the two-stage fine-tuning strategy is indispensable for the final performance improvement.

In general, the TSX-MedTrans framework effectively improves the term accuracy and semantic consistency of the base model in medical translation tasks. It also reveals the attention distribution of large language models in professional domain knowledge modeling from the interpretive perspective. Future research will consider verifying the generalization ability of this framework on larger-scale and multi-domain medical corpus, and explore combining human preference learning mechanisms to further optimize the reliability and controllability of the model.

References

[1] Briva-Iglesias, Vicent, Camargo, Joao Lucas Cavalheiro, Dogru, Gokhan. Large language models" ad referendum": How good are they at machine translation in the legal domain?. arXiv preprint arXiv: 2402.07681, 2024.
[2] Zhang, Xuan, Rajabi, Navid, Duh, Kevin, et al. Machine translation with large language models: Prompting, few-shot learning, and fine-tuning with QLoRA. //Proceedings of the Eighth Conference on Machine Translation, 2023.
[3] Wang, Jiaqi, Zhao, Huan, Yang, Zhenyuan, et al. Legal evalutions and challenges of large language models. arXiv preprint arXiv: 2411.10137, 2024.
[4] Li, Haitao, Chen, Junjie, Yang, Jingli, et al. Legalagentbench: Evaluating llm agents in legal domain. arXiv preprint arXiv: 2412.17259, 2024.
[5] Gre czuk, Andrzej, Chomiak-Orsa, Iwona, Tryczy ska, Katarzyna. AI-Supported Translation Tools for Legal Texts: A Comparative Analysis. Procedia Computer Science, 246: 5545--5554, 2024.
[6] Galiero L. Evaluating Domain Adaptation in Neural Machine Translation and Large Language Models: Insights from the TICO-19 Benchmark [J].
[7] Chen, Linqing, Wang, Weilei, Bai, Zilong, et al. PharmaGPT: Domain-specific large language models for bio-pharmaceutical and chemistry. arXiv preprint arXiv: 2406.18045, 2024.
[8] Zheng, Jiawei, Hong, Hanghai, Liu, Feiyan, et al. Fine-tuning large language models for domain-specific machine translation. arXiv preprint arXiv: 2402.15061, 2024.
[9] Xu, Haoran, Kim, Young Jin, Sharaf, Amr, et al. A paradigm shift in machine translation: Boosting translation performance of large language models. arXiv preprint arXiv: 2309.11674, 2023.
[10] Mujadia, Vandan, Urlana, Ashok, Bhaskar, Yash, et al. Assessing translation capabilities of large language models involving english and indian languages. arXiv preprint arXiv: 2311.09216, 2023.
[11] Leiter, Christoph, Lertvittayakumjorn, Piyawat, Fomicheva, Marina, et al. Towards explainable evaluation metrics for machine translation. Journal of Machine Learning Research, 25(75): 1--49, 2024.
[12] Futeral, Matthieu, Schmid, Cordelia, Laptev, Ivan, et al. Tackling ambiguity with images: Improved multimodal machine translation and contrastive evaluation. arXiv preprint arXiv: 2212.10140, 2022.
[13] Palikhe, Avash, Yu, Zhenyu, Wang, Zichong, et al. Towards Transparent AI: A Survey on Explainable Large Language Models. arXiv preprint arXiv: 2506.21812, 2025.
[14] Papineni, Kishore, Roukos, Salim, Ward, Todd, et al. Bleu: a method for automatic evaluation of machine translation. //Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002.
[15] Rei, Ricardo, Stewart, Craig, Farinha, Ana C, et al. COMET: A neural framework for MT evaluation. arXiv preprint arXiv: 2009.09025, 2020.
[16] Geva, Mor, Schuster, Roei, Berant, Jonathan, et al. Transformer feed-forward layers are key-value memories. arXiv preprint arXiv: 2012.14913, 2020.
[17] Fomenko, Vlad, Yu, Han, Lee, Jongho, et al. A note on lora. arXiv preprint arXiv: 2404.05086, 2024.
[18] Chekalina, Viktoriia, Rudenko, Anna, Mezentsev, Gleb, et al. SparseGrad: A Selective Method for Efficient Fine-tuning of MLP Layers. arXiv preprint arXiv: 2410.07383, 2024.
[19] Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, et al. Attention is all you need. Advances in neural information processing systems, 30, 2017.
[20] Clark, Kevin, Khandelwal, Urvashi, Levy, Omer, et al. What does bert look at? an analysis of bert's attention. arXiv preprint arXiv: 1906.04341, 2019.
[21] Hu, Edward J, Shen, Yelong, Wallis, Phillip, et al. Lora: Low-rank adaptation of large language models. ICLR, 1(2): 3, 2022.