Enhancing Named Entity Translation from Classical Chinese to Vietnamese in Traditional Vietnamese Medicine Domain: A Hybrid Masking and Dictionary-Augmented Approach

Nhu Pham[1,2*] **and Uyen Nguyen**[1,2*] **and Long Nguyen**[1,2†] **and Dien Dinh**[1,2]
[1]Faculty of Information Technology, University of Science, Ho Chi Minh City, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam
{pvqnhu21, npbuyen21}@apcs.fitus.edu.vn
{nhblong, ddien}@fit.hcmus.edu.vn

Abstract

Vietnam's traditional medical texts were historically written in Classical Chinese using Sino-Vietnamese pronunciations. As the Vietnamese language transitioned to a Latin-based national script and interest in integrating traditional medicine with modern healthcare grows, accurate translation of these texts has become increasingly important. However, the diversity of terminology and the complexity of translating medical entities into modern contexts pose significant challenges. To address this, we propose a method that fine-tunes large language models (LLMs) using augmented data and a Hybrid Entity Masking and Replacement (HEMR) strategy to improve NE translation. We also introduce a parallel NE translation dataset specifically curated for traditional Vietnamese medicine. Our evaluation across multiple LLMs shows that the proposed approach achieves a translation accuracy of 71.91%, demonstrating its effectiveness. These results underscore the importance of incorporating NE awareness into translation systems, particularly in low-resource and domain-specific settings like traditional Vietnamese medicine.

1 Introduction

Classical Chinese script, pronounced through Sino-Vietnamese pronunciations, was the primary writing system for Vietnamese texts until the 20th century. It played a central role in documenting traditional Vietnamese medicine. With the shift to the Latin-based Vietnamese national script, the ability to access and interpret these historical texts has steadily declined. As the Vietnamese government increasingly looks to incorporate traditional medicine into modern healthcare systems, the preservation and digitization of these historical documents has become more urgent than ever.

A critical aspect of translating traditional Vietnamese medicine is the accurate translation of named entities (NEs), particularly medicinal materials. These entities often possess culturally and linguistically specific names in Vietnamese that do not directly correspond to their Chinese equivalents. Mistranslations can occur due to variations in naming conventions. For example,"黄斤" (Kudzu powder) can be translated as "Cát căn", "Hoàng cân", or "Sắn dây". Among these, "Sắn dây" is the most widely understood in contemporary usage, but many traditional medical texts still prefer the earlier terms. The challenge stems not only from the diversity of names but also from the difficulty of accurately translating these terms in modern contexts.

Given this challenge, it is essential to develop machine translation systems for Classical Chinese–Vietnamese texts with a strong focus on NE accuracy. In recent years, the emergence of large language models (LLMs) has significantly improved multilingual translation capabilities, including for Chinese and Vietnamese. LLMs have shown great potential in the field of Traditional Chinese Medicine by analyzing classical texts, supporting clinical decision-making, and bridging traditional and modern medical paradigms (Zhang et al., 2025). Rikters and Miwa (2024) fine-tuned T5 models using SpaCy for entity recognition combined with XML-based tagging, while Liang et al. (2024) proposed a data augmentation framework based on translation difficulty and context diversity scores. These advancements highlight the potential of LLMs as a promising approach to address the complexities of NE translation in traditional Vietnamese medicine.

In this paper, we aim to tackle the challenge of NE translation from Classical Chinese to Vietnamese within the domain of traditional Vietnamese medicine. Our approach uses fine-tuned LLMs, enhanced through data augmentation tech-

[*]Equal contribution.
[†]Corresponding author.

Proceedings of the 18th International Natural Language Generation Conference, pages 408–418
October 29–November 2, 2025.

niques and an entity masking and replacement strategy based on a specialized dictionary. To achieve these goals, we present the following contributions:

- We explore the use of data augmentation techniques to improve the translation accuracy of named entities in LLMs.
- We propose a systematic method called Hybrid Entity Masking and Replacement (HEMR), which combines entity masking and replacement using a Vietnamese medicinal material dictionary and LLMs.
- We develop a dictionary of medicinal materials and traditional medicine terms with an evaluation dataset for NE translation between Classical Chinese and Vietnamese.

2 Background

Chinese script, Sino-Vietnamese, and Modern Vietnamese Translation: Literary Chinese, written in Chinese script and read with Sino-Vietnamese pronunciations, was once a major writing system in Vietnam (Nguyễn, 1979). Most texts, including those on traditional medicine, were composed this way. With the adoption of the Latin-based Vietnamese National Script, Vietnam shifted to a phonographic system that preserves about 60% of Sino-Vietnamese pronunciations (Lê, 2002). However, this transition reduced the population's ability to read Chinese script, limiting access to historical texts. Although we can transliterate Classical Chinese into the National Script, the text often stays largely incomprehensible because much of the vocabulary is obsolete or has been replaced.

Traditional Vietnamese medicine: Traditional Vietnamese Medicine (TVM) is a medical system that incorporates practices from indigenous Vietnamese knowledge and external influences, including Traditional Chinese Medicine (TCM). Practices like Dật Gió ("wind snatching") and Cạo Gió ("wind scraping," similar in function to TCM's Gua Sha) appear in both systems, yet Vietnamese practitioners typically regard them as indigenous traditions (Ahn et al., 2006).

However, many materials used in Traditional Vietnamese Medicine (TVM) are not included in classical Chinese pharmacopeia. There are 70 officially classified medicinal plants that are native to Vietnam. In addition, Vietnamese often use distinct names or naming conventions for medicinal materials, which differ from their Chinese counterparts. These names are often specific to Vietnamese practice and require accurate handling in translation. To address this, datasets must include domain-specific terms and reflect the linguistic and botanical differences present in TVM. Accurate translation of medicinal material names and TVM texts is essential not only for preserving the original meaning but also for ensuring correct interpretation and application of medical knowledge.

Entity-Aware Machine Translation Entity-Aware Machine Translation (EAMT) is a specialized branch of Natural Language Processing (NLP) and Machine Translation (MT) that focuses on the accurate and contextually appropriate translation of named entities (NEs), such as person names, organizations, locations, and dates. They are essential to preserve contextual meaning and factual precision during translation.

In the domain of Classical Chinese–Vietnamese traditional medicine texts, accurate translation of named entities—such as medicinal herbs, disease names, anatomical terms, and measurement units—is critical for preserving semantic fidelity.

3 Related work

Traditional approaches to NE handling in neural machine translation (NMT) typically involve explicitly introducing NE information into the translation pipeline to improve translation accuracy. Early research focused on integrated neural models and attention mechanisms. Sennrich and Haddow (2016), and Niehues and Cho (2017), enhanced NMT using linguistic annotations and NE embeddings. Ugawa et al. (2018) extended token embeddings with entity-specific vectors, while Modrzejewski et al. (2020) incorporated source-side linguistic factors into Transformer networks. These methods improved contextual representation but relied heavily on large amounts of labeled data, which poses scalability challenges.

To address these limitations, Xie et al. (2022) proposed an end-to-end NMT architecture with entity classifiers integrated into both encoder and decoder. An adaptive loss function emphasized NE translation accuracy. This built-in NE awareness lessens dependence on external systems and reduces error propagation, leading to better performance on six benchmark tasks.

More recently, the field has shifted towards data-centric and LLM-based approaches. Rikters and Miwa (2024) fine-tuned T5 models for NE-aware

translation using SpaCy for entity recognition and XML-based entity tagging. This method increased NE preservation and improved BLEU scores, especially for rare entities. Liang et al. (2024) introduced a data augmentation framework based on translation difficulty and context diversity scores. Their strategy improved overall translation and NE accuracy across WMT news and terminology datasets. Rikters and Miwa (2024) further explored instruction tuning for NE-aware tasks using customized prompts on pre-annotated data. This instruction-based approach offers flexibility for domain-specific scenarios and aligns with recent trends in large language model fine-tuning.

4 Methodology

To address the challenge of NE-aware translation from Classical Chinese script to Vietnamese in the domain of traditional Vietnamese medicine, we conducted a series of experiments guided by the following research questions.

RQ$_1$: *What is the current accuracy of standard fine-tuning approaches—without the aid of external dictionaries—in translating named entities from Classical Chinese texts into Vietnamese within the domain of traditional Vietnamese medicine?*

RQ$_2$: *To what extent does fine-tuning with augmented data with curated dictionary improve the translation accuracy of named entities in the domain of traditional Vietnamese medicine?*

RQ$_3$: *To what extent does applying fine-tuning with NE masking and replacement using a hybrid approach—combining translation with augmentation data—improve the translation accuracy of named entities in the domain of traditional Vietnamese medicine?*

4.1 Dataset

There are currently no existing parallel Chinese-script–Vietnamese datasets in the domain of Traditional Vietnamese Medicine, nor any datasets that include NE annotations in this domain. The creation of our dataset serves as a foundational step toward advancing research in Chinese-script to Vietnamese translation for this field. Based on the characteristics of classical medicine texts, we propose the following five named entity labels, as detailed in Table 1.

4.1.1 Dataset Construction Process

Our dataset consists of two main components:

Label	Full Name
PLT_MM	Plant Medicinal Material
ANI_MM	Animal Medicinal Material
MIN_MM	Mineral Medicinal Material
PER	Person
LOC	Location
TRE	Prescription Name
CMB	Titles of Classical Books

Table 1: Named Entity Taxonomy

Medicinal Material and Terminologies Dictionary: Entries were compiled from two authoritative sources. These will serve as the reference dictionary for our NE tagging process.

- WHO International Standard Terminologies on Traditional Medicine in the Western Pacific Region (Thuật Ngữ Y Học Cổ Truyền của Tổ chức Y tế Thế Giới Khu vực Tây Thái Bình Dương) (World Health Organization, 2011).
- Guide to Medicinal Properties (Dược Tính Chỉ Nam), a comprehensive reference on herbal substances and their uses (Minh, 1967).

Traditional Medicine Remedies: This part compile from classical Vietnamese medical documents:

- Treatise of Medical Knowledge of Hải Thượng Lãn Ông (Hải Thượng Lãn Ông Y Tông Tâm Lĩnh) (Lê, 1998).
- Dialogue of the Fisherman and the Woodcutter on the Art of Medicine (Ngư Tiều Vấn Đáp Y Thuật) (Nguyễn Đình Chiểu, 2003).

For the annotation process, we adopt a hybrid human-AI approach, applied independently to the Classical Chinese and Vietnamese texts to ensure consistent and accurate entity labeling:

- **Stage 1: Automated Pre-annotation**
  - **Automatic Entity Recognition**: Performed using the Gemini LLMs model to identify potential named entities.
  - **Dictionary Cross-Referencing**: Identified entities are automatically matched against the dictionary for validation.
  - **Confidence Scoring**: Each entity is given a confidence score that indicates the model's level of certainty.

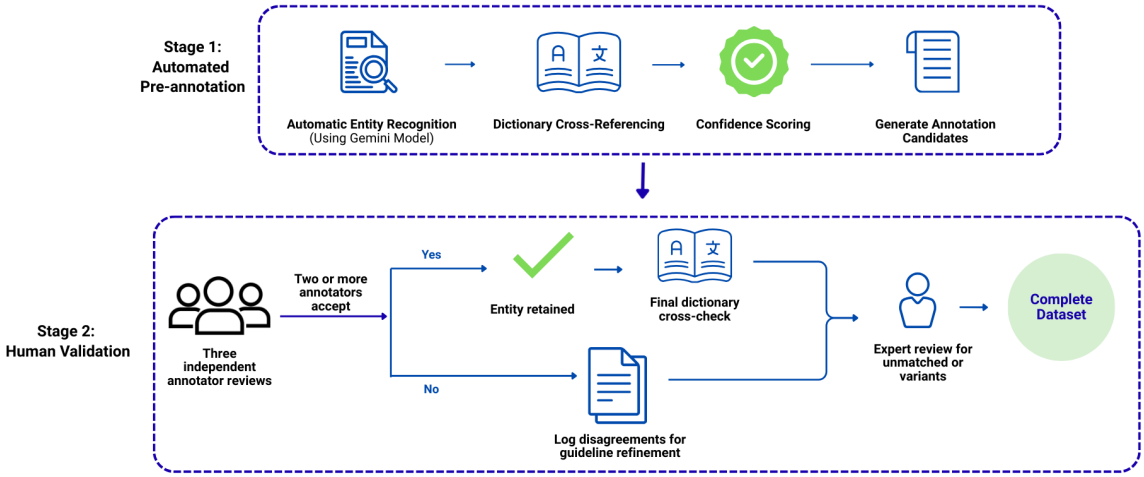

Figure 1: Two-stage Annotation Workflow

  - **Annotation Candidate Generation**: Preliminary entity annotations are created for human evaluation.

- **Stage 2: Human Validation**
  - **Human Review**: A team of three annotators independently reviews all annotations to ensure accuracy and consistency.
  - **Annotation Decisions**: Annotators accept or reject each candidate entity.
  - **Majority Voting**: We retain an entity if at least two out of the three annotators agree on its correctness.
  - **Disagreement Logging**: Cases of disagreement are documented to inform future guideline and expert reviews.

- **Dictionary Validation**: All accepted entities were cross-validated against the *Dược Tính Chỉ Nam* dictionary (Minh, 1967) to ensure terminological consistency and medical accuracy. Entities not found in the dictionary were subjected to additional experts review to distinguish between legitimate historical variants and annotation errors.

4.1.2 Dataset Statistics

The final dataset is divided into two components: a Medicinal Material and Terminologies Dictionary containing **8,228 entries**, and a Traditional Medicine Remedies corpus consisting of **17,960 aligned pairs**. The overall distribution of NE labels is summarized in Table 2. The LOC and TRE values are currently 0 due to the absence of credible sources for reliable dictionary extraction. Further expansion in this area is planned for future work.

Label	(1)	(2)	Total
PLT_MM	11,252	18303	26,951
ANI_MM	2,652	2344	4,882
MIN_MM	1,850	3643	5,371
PER	138	662	804
LOC	**0**	63	63
TRE	**0**	2265	2,131
CMB	324	909	1,233
Total	**16,216**	**25,219**	**41,435**

Table 2: Named Entity Counts.
(1): Medicinal Material and Terminologies Dictionary,
(2): Traditional Medicine Remedies.

4.2 Proposed Approach

In this section, we present our proposed approaches to improving NE translation in the domain of traditional Vietnamese medicine. We describe two key components: a data augmentation strategy and a hybrid entity masking and replacement method to preserve the accuracy of specialized terms during translation.

4.2.1 Data Augmentation

LLMs can effectively learn from limited domain-specific data through fine-tuning. Building on this, we propose a data augmentation method that enhances NE translation by systematically replacing annotated entities with alternatives from curated dictionaries. The augmentation process is:

- **Identification of Sentences**: Select sentences

with annotated entities, keeping only those confirmed by at least two annotators.

- **Entity Substitution**: For each identified entity, randomly select an alternative term from the corresponding dictionary. Replace the original mention in both the Classical Chinese sentence and the Vietnamese translation while preserving the entity label.

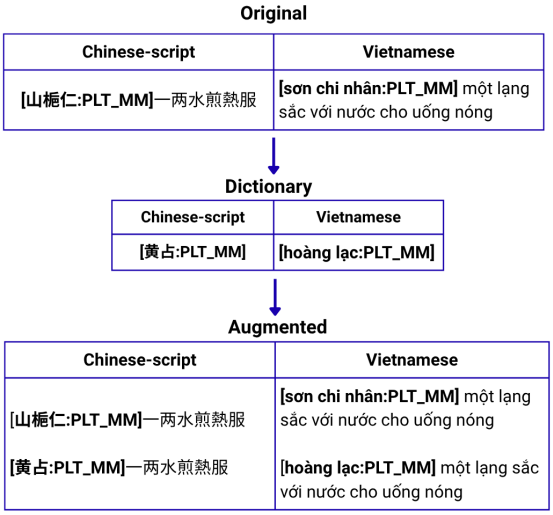

Figure 2: Augmented Process

Original	Augmented	Total
14,519	7,314	21,833

Table 3: Statistics of the training dataset before and after augmentation

4.2.2 Hybrid Entity Masking and Replacement

To preserve the semantic integrity of specialized terms, we introduce a structured masking and demasking approach applied to our previously augmented dataset. The process starts by identifying named entities in the source text and replacing them with placeholder tokens, preventing distortion by general-purpose translation models. We refer to this method as **Hybrid Entity Masking and Replacement (HEMR)** (Figure 3). HEMR operates through the following resolution process:

- **Dictionary Lookup** – Each entity is first checked against the curated dictionary containing medicinal plant names, mineral, and animal-based medicinal materials.
- **Automate Translation** – If no match is found, we fallback to the Gemini translation API to produce a candidate translation for the entity. The prompt is detailed in Appendix D.
- **Sino-Vietnamese Conversion** – Where applicable, entities are rendered in their Sino-Vietnamese forms to maintain their historical and cultural authenticity.

Once translated, the placeholders are replaced with their resolved entity translations, restoring a complete and coherent target-language sentence. This hybrid pipeline ensures fidelity to domain-specific concepts while leveraging the flexibility and fluency of modern language models.

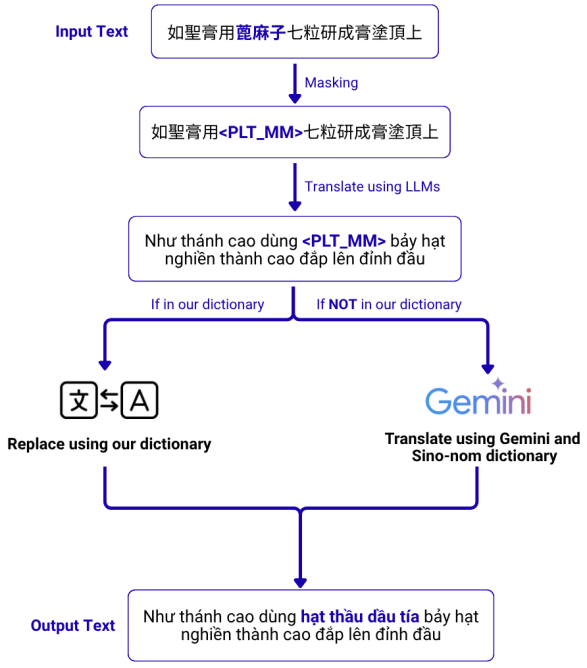

Figure 3: Hybrid Entity Masking and Replacement

We explore several language models to assess the current state of Classical Chinese to Vietnamese translation. We prioritize models that demonstrate strong multilingual capabilities, particularly with support for Vietnamese and Traditional Chinese. The models we selected—Phi-4 (Abdin et al., 2024), Qwen-3 (Yang et al., 2025), LLaMA-3.1[1], and NLLB-200[2]—were chosen based on their performance in multilingual and low-resource settings, availability of open-source checkpoints, and compatibility with fine-tuning pipelines. Notably, models such as Qwen-3 and NLLB-200 have been

[1]https://huggingface. co/meta-llama/Llama-3.1-8B
[2]https: //huggingface.co/facebook/nllb-200-3.3B

specifically trained with an emphasis on Asian languages, rendering them particularly well-suited for this task. Detailed information on the configuration and model resources can be found in Appendix C.

5 Results and Analysis

In this section, we present the experimental results and analysis of three approaches: standard fine-tuning, fine-tuning with augmented data, and the hybrid entity masking and replacement method.

5.1 Standard Fine-tuning

Case	Model	BLEU	BERTScore	METEOR
(1)	Qwen-3-14B	**5.41**	**48.56**	**31.92**
	Phi-4-14B	0.87	34.66	12.73
	LLaMA-3.1-8B	2.75	40.81	14.94
	NLLB-200-3.3B	0.74	25.84	9.82
(2)	Qwen-3-14B	12.73	60.68	36.46
	Phi-4-14B	**23.94**	**68.38**	**51.30**
	LLaMA-3.1-8B	13.26	60.87	36.53
	NLLB-200-3.3B	14.73	61.96	42.72

Table 4: Comparison of model performance in (1) before fine-tuning and (2) after fine-tuning without dictionary (scaled to 1–100).

Table 4 shows that fine-tuning markedly improves translation quality across all metrics. Phi-4-14B demonstrates the largest gains overall (BLEU rising from 0.87 to 23.94, BERTScore nearly doubling from 34.66 to 68.38, and METEOR increasing from 12.73 to 51.30). LLaMA-3.1-8B and Qwen-3-14B also improve substantially, with BLEU gains of about 10 points. Notably, NLLB-200-3.3B, despite its low baseline, reaches competitive performance after fine-tuning. Without

Label	Qwen-3-14B	Phi-4-14B	LLaMA-3.1-8B	NLLB-200-3.3B
PLT_MM	35.76	**70.73**	38.11	32.39
ANI_MM	27.18	**58.25**	18.45	15.52
MIN_MM	48.45	**71.43**	52.80	35.75
PER	14.81	**51.85**	25.93	3.57
LOC	0.00	**33.33**	**33.33**	0.00
TRE	15.58	**70.13**	37.66	22.68
CMB	14.29	**50.00**	11.90	0.00
Overall	33.62	**67.79**	36.88	29.72

Table 5: Named Entity Accuracy (%) after fine-tuning (percentages scaled 1–100).

the aid of a dictionary, Table 5 shows that Phi-4-14B consistently achieves the highest accuracy across all entity types, with an overall average of 67.79%. Qwen-3-14B, LLaMA-3.1-8B and NLLB-200-3.3B perform moderately, scoring 33.62%, 36.88% and 29.72%. Despite this, categories like LOC and CMB remain challenging for all models.

Based on the results presented in Table 4 and Table 5, we can answer $\mathbf{RQ_1}$ by observing that standard fine-tuning approaches - without relying on external dictionaries - already provide a boost in overall translation. However, accuracy across certain entity categories, such as LOC and CMB, remains consistently low for all models. This underscores the significant constraints inherent in existing fine-tuning approaches, which limit their capacity to fully capture domain-specific nuances and address the complex requirements of Classical Chinese–Vietnamese translation.

5.2 Fine-tuning with Augmented Data

Case	Model	BLEU	BERTScore	METEOR
(3)	Qwen-3-14B	**27.92**	**70.78**	**55.03**
	Phi-4-14B	27.02	69.96	53.29
	LLaMA-3.1-8B	25.35	68.94	51.49
	NLLB-200-3.3B	11.96	59.92	38.57
(4)	Qwen-3-14B	26.51	**70.17**	**53.73**
	Phi-4-14B	**26.98**	70.11	53.60
	LLaMA-3.1-8B	24.94	68.63	51.02
	NLLB-200-3.3B	15.36	62.59	43.48

Table 6: Comparison of model performance in (3) augmented data without tagging and (4) augmented data with tagging (scaled to 1–100).

Table 6 show that when finetune with tag, the performance slightly reduces overall, with BLEU dropping for most models. Phi-4-14B and Qwen-3-14B see small gains in BERTScore and METEOR, while LLaMA-3.1-8B shows minor declines across all metrics. NLLB-200-3.3B is the exception, with BLEU improving from 11.96 to 15.36.

Label	Qwen-3-14B	Phi-4-14B	LLaMA-3.1-8B	NLLB-200-3.3B
PLT_MM	**66.60**	66.40	44.20	35.6
ANI_MM	48.54	**49.51**	29.13	17.5
MIN_MM	**73.29**	72.04	59.01	47.8
PER	37.03	**40.74**	11.11	18.5
LOC	33.33	0.00	**66.67**	33.3
TRE	67.53	**68.83**	41.56	24.7
CMB	**57.14**	50.00	9.52	7.1
Overall	**64.43**	63.99	42.41	32.97

Table 7: Named Entity Accuracy (%) with Augmented Data (Without Tagging) (percentages scaled 1–100).

Label	Qwen-3-14B	Phi-4-14B	LLaMA-3.1-8B	NLLB-200-3.3B
PLT_MM	60.71	**70.53**	67.58	41.45
ANI_MM	**50.49**	48.54	46.60	14.56
MIN_MM	68.32	76.39	**72.67**	59.62
PER	**29.63**	18.51	44.44	29.62
LOC	33.33	33.3	0.00	33.33
TRE	51.95	61.04	**66.23**	42.85
CMB	35.71	**52.38**	42.86	23.80
Overall	58.03	**65.80**	64.00	40.60

Table 8: Named Entity Accuracy (%) with Augmented Data (With Tagging) (percentages scaled 1–100).

However, Tables 7 and 8 reveal that tagging improves NE translation accuracy, particularly for smaller or multilingual models. Phi-4-14B improves from 63.99% to 65.80%, and LLaMA-3.1-8B shows the largest gain, from 42.41% to 64.00%. NLLB-200-3.3B also benefits, especially on MIN_MM and TRE. Howerver for Qwen-3-14B there is a slight drop with tagging compared to when finetune with no tag. The accuracy for the LOC category remains inconsistent, likely due to the limited number of entities in the test set—only 8 in total. For example, Phi-4's performance fluctuates from 0% to 33.33% with tagging, while LLaMA's accuracy drops significantly from 66.67% to 0% when tagging is applied.

Compared the NE accuracy result of standard fine-tuning (table 5) and augmented (table 8) reveals that Phi-4-14B experiences a slight decrease in accuracy when using data augmentation, achieving 65.80% compared to 67.79% with standard fine-tuning. In contrast, Qwen-3-14B benefits significantly from augmentation, improving from 33.62% to 58.03%. Similar trends are observed with LLaMA-3.1-8B and NLLB-200-3.3B, which improve from 36.88% to 64.00% and from 29.72% to 40.60%, respectively. But for a limited entity like LOC the accuracy still varies significantly. These results indicate that although augmentation offers limited gains for models with strong baseline performance, it can markedly improve the accuracy of models starting from a lower baseline.

Addressing **RQ$_2$**, we observe that fine-tuning with dictionary-augmented data improves the overall accuracy of NE translation. However, the BLEU scores still vary. For a strong baseline model like Phi-4-14B, the impact is negligible, while models such as Qwen-3-14B, LLaMA-3.1-8B, and NLLB-200-3.3B show improvements. Despite these gains, the accuracy for certain entity types—specifically PER, LOC, and CMB—remains below 40%, the observed gains support the promise of this approach as a foundation for further refinement.

5.3 Hybrid Entity Masking and Replacement

For the HEMR method, NE accuracy primarily depends on the hybrid replacement process. As a result, fine-tuning the model does not significantly impact NE accuracy, which remains consistent across different models. Variations in NE accuracy are mainly influenced by the translation quality of the Gemini model and mismatched tags when translating from models. However, the rate of mismatched tag generation (i.e., producing more or fewer tags than expected) is low, as shown in Table 9.

Model	Mismatched tags
Phi-4	6
LLaMA	8
Qwen-3	4
NLLB	4

Table 9: Number Generate mismatched tags

Model	BLEU	BERTScore	METEOR
Qwen-3-14B	28.17	71.27	55.86
Phi-4-14B	**29.16**	**71.45**	**56.23**
LLaMA-3.1-8B	25.79	69.08	52.01
NLLB-200-3.3B	18.06	64.03	42.98

Table 10: Model performance using Hybrid Entity Replacement (scaled to 1–100).

Table 10 presents the performance of models after applying HEMR, showing consistent improvements across all evaluation metrics. Phi-4-14B achieves the highest scores overall, with a BLEU of 29.16, BERTScore of 71.45, and METEOR of 56.23, reinforcing its strong baseline capabilities. Qwen-3-14B follows closely with comparable performance, especially in BERTScore (71.27) and METEOR (55.86), indicating semantic and syntactic preservation. LLaMA-3.1-8B demonstrates moderate gains across all evaluation metrics. In contrast, NLLB-200-3.3B, while remaining behind the other models overall, exhibits notable improvement, particularly in BLEU, which increases to 18.06.

Label	Qwen-3-14B	Phi-4-14B	LLaMA-3.1-8B	NLLB-200-3.3B
PLT_MM	74.66	74.27	74.66	74.85
ANI_MM	68.93	68.93	68.93	68.93
MIN_MM	82.61	82.61	80.75	82.61
PER	55.56	55.56	55.56	55.56
LOC	33.33	33.33	33.33	33.33
TRE	58.44	58.44	58.44	57.14
CMB	42.86	42.86	42.86	42.86
Overall	**71.91**	71.69	71.58	**71.91**

Table 11: Named Entity accuracy (%) after applying Hybrid Entity Masking and Replacement (percentages scaled 1–100).

We look at the overall accuracy of NE translation across all models.The accuracy exceeds 71% overall (Table 11). NLLB-200-3.3B and Qwen-3-14B achieve the highest performance at 71.91%, followed closely by Phi-4-14B at 71.69% and LLaMA-3.1-8B at 71.58%. These slight discrepan-

cies are primarily attributed to differences in mismatched tag generation. Among the entity types, MIN_MM consistently yields the highest accuracy across all models, exceeding 80%. In contrast, categories such as LOC and CMB, while showing improvement over prior approaches, remain challenging, with accuracies below 45%. The low performance in the LOC category is largely due to the absence of corresponding entries in the dictionary for replacement.

Evaluation of Each Stage in HEMR Process To further evaluate the effectiveness of the HEMR process, we systematically assessed the individual and combined contributions of Dictionary Lookup, Automated Translation, and Sino-Vietnamese Conversion to NE translation accuracy within the domain of traditional Vietnamese medicine.

Config	Total Accuracy
(1)	65.62
(2)	61.43
(3)	7.81
(1) + (2)	73.64
(1) + (3)	66.38
(2) + (3)	61.30
(1) + (2) + (3)	**73.86**

Table 12: Overall accuracy across different processes. (1) Dictionary Lookup, (2) Automated Translation, (3) Sino-Vietnamese Conversion

In Table 12, Dictionary Lookup (1) performs strongly for categories with good lexical coverage, such as PLT_MM (76.03%) and MIN_MM (82.61%), but is ineffective for LOC (0%) and TRE (1.30%), where many names are not in dictionaries. Automated Translation (2) yields more balanced improvements, notably enhancing PER, LOC, and TRE accuracy, though it slightly lowers PLT_MM performance compared to Dictionary Lookup (1). Furthermore, about 0.3 to 0.5% of the names could not be translated with this method. Sino-Vietnamese Conversion (3) alone performs poorly overall, with all accuracies under 11%. The detailed accuracy scores for each entity type are provided in the Appendix.

However, combining methods consistently boosts performance. The (1) + (2) configuration already outperforms any single-stage baseline, demonstrating clear complementarity between dictionary lookup and translation. Adding Sino-Vietnamese Conversion yields further gains: the full process (1) + (2) + (3) achieves the highest overall accuracy (73.86%), indicating that even limited Sino-Vietnamese matches can provide valuable additional cues when integrated with other strategies. Combining Dictionary Lookup, Automated Translation, and Sino-Vietnamese Conversion achieves the best performance in both NE accuracy and overall translation quality. While individual components offer partial improvements, their integration yields substantial gains.

From table 8 and 11, HEMR consistently outperforms the tag-aware approach in terms of overall accuracy. Qwen-3-14B and NLLB-200-3.3B achieve the highest accuracy with HEMR at 71.91%, compared to 58.03% and 40.60% of augmented with tag training. While the tag-aware augmentation helps capture specific entity types, the application of HEMR leading to consistent gains in overall accuracy. These results demonstrate the complementary nature of the two methods and highlight the effectiveness of HEMR as a post-processing step for improving entity translation.

To answer **RQ$_3$**, where our proposed HEMR process achieves a notable accuracy of 71.91% with the Qwen-3-14B model. This demonstrates the effectiveness of our HEMR approach in improving NE translation within the domain of traditional Vietnamese medicine. While components such as Automated Translation and Sino-Vietnamese Conversion play a role in handling rare or previously unseen entities—especially those not present in training data—the core contribution still lies in dictionary-based lookup. These additional steps act as important fallbacks, but their success is inherently tied to the quality and completeness of the underlying dictionaries. Therefore, to further enhance translation accuracy, especially for low-resource or domain-specific terms, it is vital to continue expanding and refining the dictionary resources used in the HEMR.

6 Conclusion

This work explores the challenge of NE translation from Classical Chinese script to Vietnamese within the domain of traditional medicine. We investigate a dictionary-based data augmentation method and propose a Hybrid Entity Masking and Replacement (HEMR) approach. To support our evaluation, we introduce a parallel NE translation dataset and a curated dictionary of medicinal material terminologies. Experimental results show that our HEMR process leads to a significant im-

provement in translation accuracy. These findings highlight the potential of combining LLMs with external dictionaries to enhance translation quality in domain-specific settings such as traditional Vietnamese medicine. For future work, we intend to enhance overall translation performance and further expand the medicinal material dictionary to improve coverage and accuracy in domain-specific entity translation.

Limitations

While our proposed approach shows improved translation of named entities, several limitations remain. First, in the HEMR approach, the limited coverage of LOC and TRE entities requires further investigation and expansion. Second, our study does not yet address named entity recognition; the ability to identify named entities significantly impacts the overall effectiveness of this method.

Acknowledgments

This research is supported by research funding from Faculty of Information Technology, University of Science, Vietnam National University - Ho Chi Minh City.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, and 8 others. 2024. Phi-4 technical report. *Preprint*, arXiv:2412.08905.

Andrew C Ahn, Quyen Ngo-Metzger, Anna TR Legedza, Michael P Massagli, Brian R Clarridge, and Russell S Phillips. 2006. Complementary and alternative medical therapy use among chinese and vietnamese americans: prevalence, associated factors, and effects of patient–clinician communication. *American Journal of Public Health*, 96(4):647–653.

H.T. Lê. 1998. *Hải Thượng Y - Tôn tâm lĩnh :.: từ tập 5 đến tập 9. Tập thủ - Nội kinh- Mạch lạc (quan miện)-Tính dược. Quyển II (trọn bộ 5 quyển)*. Sách thuốc Việt Nam. Nxb. Đồng Tháp.

Tian Liang, Xing Wang, Mingming Yang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Addressing entity translation problem via translation difficulty and context diversity. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 11628–11638.

Đình Khẩn Lê. 2002. *Từ vựng gốc Hán trong tiếng Việt*. Đại học quốc gia Hồ Chí Minh, TP. Hồ Chí Minh.

Nguyễn Văn Minh. 1967. *Dược Tính Chỉ Nam*. Việt Nam Kỳ Lão Ái Hữu.

Maciej Modrzejewski, Miriam Exel, Bianka Buschbeck, Thanh-Le Ha, and Alex Waibel. 2020. Incorporating external annotation to improve named entity translation in nmt. In *Proceedings of the 22nd annual conference of the european association for machine translation*, pages 45–51.

T.C. Nguyễn. 1979. *Nguồn gốc và quá trình hình thành cách đọc Hán Việt*. Nhà xuất bản Khoa Học Xã Hội.

Nguyễn Đình Chiểu. 2003. *Ngư Tiều y thuật vấn đáp*. NXB Thuận Hóa.

Jan Niehues and Eunah Cho. 2017. Exploiting linguistic resources for neural machine translation using multi-task learning. In *Proceedings of the Second Conference on Machine Translation*, pages 80–89, Copenhagen, Denmark. Association for Computational Linguistics.

Matīss Rikters and Makoto Miwa. 2024. Entity-aware multi-task training helps rare word machine translation. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 47–54.

Rico Sennrich and Barry Haddow. 2016. Linguistic input features improve neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers*, pages 83–91, Berlin, Germany. Association for Computational Linguistics.

Arata Ugawa, Akihiro Tamura, Takashi Ninomiya, Hiroya Takamura, and Manabu Okumura. 2018. Neural machine translation incorporating named entity. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3240–3250.

Regional Office for the Western Pacific World Health Organization. 2011. *Thuật ngữ Y học cổ truyền của tổ chức Y tế thế giới khu vực Tây Thái Bình Dương*.

Shufang Xie, Yingce Xia, Lijun Wu, Yiqing Huang, Yang Fan, and Tao Qin. 2022. End-to-end entity-aware neural machine translation. *Mach. Learn.*, 111(3):1181–1203.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

R. Zhang, D. Tian, and Y. Wang. 2025. Large language models in traditional chinese medicine: A short survey and outlook. *AI Medicine*, 2(1):3.

A Dataset Contributors

- Contributor 1 - Associate Professor in Computer Science and Comparative Linguistics.
- Contributor 2 - PhD in Computer Science and Natural Language Processing (NLP).
- Contributor 3 - PhD student in Traditional Vietnamese Medicine
- Contributor 4 - Undergraduate Student majoring in Computer Science and NLP
- Contributor 5 - Undergraduate Student majoring in Computer Science and NLP

B Dataset Statistics

Table 1 shows the distribution of entity labels produced by the augmentation process.

Label	Number of Instances
PLT_MM	11070
ANI_MM	888
MIN_MM	2369
PER	2572
LOC	587
TRE	74
CMB	2826
Total	**20386**

Table 1: Distribution of entity labels generated by the augmentation process.

Table 2 shows the distribution of entity labels across the training, validation, and test sets after applying augmentation.

Label	Train	Validation	Test
PLT_MM	24,715	1,037	1,017
ANI_MM	2,740	172	206
MIN_MM	5,272	296	322
PER	3,316	48	54
LOC	636	8	6
TRE	1,846	204	155
CMB	3,583	68	84

Table 2: Total counts of each entity label in the train, validation, and test sets after augmentation.

C Experiment Setup

This section outlines the experimental setup for evaluating our translation approach.

C.1 Computational Resources

All experiments were performed on Google Colab Pro with an NVIDIA A100 GPU (40GB VRAM). Mixed-precision training was utilized to improve memory efficiency and speed up convergence.

C.2 Model Configurations

We fine-tuned the following models using parameter-efficient LoRA techniques to optimize training speed and reduce computational resource requirements without compromising performance:

Model	Parameters
Qwen-3-14B	14.0B
Phi-4-14B	14.0B
LLaMA-3.1-8B	8.0B
NLLB-200-3.3B	3.3B

Table 3: Model configurations

D Gemini Prompt for Entity Translation

To translate named entities into Vietnamese, we used the following prompt with the gemini-2.0-flash model.

Vietnamese Prompt

> Dịch tên thực thể sau sang tiếng Việt, chỉ trả về bản dịch, không giải thích. Nếu không dịch được, hãy trả về nguyên văn: {text}

English Translation

> Translate the following named entity into Vietnamese. Return only the translation. If not possible, return the original: {text}

E Named Entity Accuracy Across HEMR Configurations

This section summarizes how each HEMR configuration performs in translating named entities across multiple categories. Table 4 shows the per-entity accuracy and overall accuracy for each combination of Dictionary Lookup, Automated Translation, and Sino-Vietnamese Conversion.

Cfg	PLT	ANI	MIN	PER	LOC	TRE	CMB
(1)	76.03	60.19	82.61	29.63	1.30	33.33	65.62
(2)	67.82	39.81	64.47	**60.00**	**33.33**	**58.70**	33.33
(3)	10.02	6.80	6.21	3.70	0.00	2.60	2.38
(1)+(2)	**78.98**	**66.99**	82.61	51.85	54.55	**42.86**	73.64
(1)+(3)	76.62	61.17	82.61	29.63	0.00	3.90	35.71
(2)+(3)	68.09	40.00	64.36	56.30	**33.33**	56.36	33.81
(1)+(2)+(3)	**78.98**	66.02	**83.23**	55.56	**33.33**	55.84	**42.86**

Table 4: Named Entity Accuracy (%) across configurations for different entity types. Cfg: (1) Dictionary Lookup, (2) Automated Translation, (3) Sino-Vietnamese Conversion.