pingan-team at SemEval-2025 Task 2: LoRA-Augmented Qwen2.5 with Wikidata-Driven Entity Translation

DiYang Chen
Ping An Insurance Company of China, Ltd
Shenzhen, China
ytimespace@gmail.com

Abstract

This paper presents our solution for SemEval-2025 Task 2 on entity-aware machine translation. We propose a parameter-efficient adaptation framework using Low-Rank Adaptation (LoRA) to fine-tune the Qwen2.5-72B model, enabling effective knowledge transfer while preserving generalization capabilities. To address data scarcity and entity ambiguity, we design a Wiki-driven augmentation pipeline that leverages Wikidata's multilingual entity mappings to generate synthetic training pairs. Our system achieves state-of-the-art performance across 10 languages, securing first place in the competition. Experimental results demonstrate significant improvements in both translation quality (COMET) and entity accuracy (M-ETA).

1 Introduction

The accurate translation of named entities remains a critical challenge in modern machine translation systems, particularly when processing rare, ambiguous, or culture-specific references. This paper presents our approach to Task 2 of SemEval-2025 (Conia et al., 2025), a shared task aimed at developing robust machine translation systems for complex entity translation between English and 10 target languages: Arabic (Ar), French (Fr), German (De), Italian (It), Japanese (Ja), Korean (Ko), Spanish (Es), Thai (Th), Turkish (Tr), and Traditional Chinese (Zh-TW).

Named entities - including personal names, organizations, geographical locations, and culture-specific items (CSIs) such as movie titles, literary works, and commercial products - present unique translation difficulties. While neural machine translation (NMT) systems have achieved remarkable progress in general domain translation, their performance significantly degrades when encountering low-frequency or domain-specific entities. This limitation stems from multiple factors: the inherent ambiguity of proper nouns across linguistic contexts, the lack of transliteration conventions for emerging entities, and the cultural specificity embedded in certain references.

The SemEval-2025 Task 2 challenge specifically addresses these limitations by constructing a testbed containing carefully curated named entities across three complexity categories: 1) Rare entities with limited parallel corpus occurrences 2) Cross-lingual ambiguous terms 3) Novel entities absent from training data.

Our investigation makes two primary contributions: First, we implement a parameter-efficient adaptation framework for the Qwen2.5-72B (Qwen et al., 2025) model through Low-Rank Adaptation (LoRA) tuning, enabling effective knowledge transfer while preserving the base model's generalization capabilities. This approach addresses the computational challenges of fine-tuning ultra-large language models (LLMs) for domain-specific named entity translation tasks. Second, we design a Wikidata Data-driven augmentation pipeline that systematically injects cross-lingual entity knowledge into the training process (see figure 1). By leveraging Wikidata's multilingual entity mappings and property graphs, our method automatically generates synthetic training pairs.

2 Related Work

Prior work on entity-aware translation has focused on retrieval-augmented methods (Conia et al., 2024) and cross-lingual alignment techniques (Wang et al., 2023). Our approach is based on the task framework proposed by (Conia et al., 2025), which formalizes the challenges of translating rare, ambiguous and novel entities. Recent advances in parameter-efficient adaptation (Hu et al., 2021) and synthetic data generation (Li et al., 2022) inspired our LoRA-based fine-tuning strat-

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 2065–2070

egy and Wikidata augmentation pipeline. While traditional NMT systems struggle with low-frequency entities (Vaswani et al., 2017), our work extends the capabilities of large language models through targeted adaptation, addressing limitations in cross-cultural transliteration and contextual disambiguation.

3 System Description

3.1 System Components

Pre-processing: Add information on Wiki data based on Wiki id to the prompt (see Appendix 1). To address the inconsistent language labeling of Traditional Chinese translations in Wikidata (e.g., *zh*, *zh-tw*, *zh-hant*, *zh-yue*), we implemented two preprocessing methods: 1) Prioritizing *zh-tw* translations extraction from Wikidata, with fallback to *zh* labels when unavailable; 2) Utilizing the *zhconv* toolkit to convert Simplified Chinese entities to their Traditional Chinese equivalents.

Model: The Qwen2.5-72B-DA model (see 3.3) utilized by our system builds upon the Qwen2.5-72B pre-trained language model, a 72 billion-parameter decoder-only transformer optimized for instruction-following tasks and released by Alibaba Cloud as an open-source large language model (LLM), enhanced through the synergistic integration of Low-Rank Adaptation (LoRA) and data augmentation techniques, enabling efficient domain-specific fine-tuning.

Post-processing: To ensure output consistency, we implement a regex-based filtering mechanism that removes non-final translation segments. When performing translation tasks, the Qwen2.5-72B model occasionally exhibits a tendency to generate hybrid-language preliminary explanations before producing the target translation. This phenomenon occurs in all languages. As illustrated in Appendix 3, the model initially outputs a Traditional Chinese sentence containing an English entity, followed by an explanation for retaining the English entity, and finally provides the correct Traditional Chinese translation. To address this issue, we employ a regular expression-based approach that segments paragraphs using newline characters and selects sentences containing the target entity at the paragraph end as the final output. Notably, while the final sentence typically contains the correct translation in most cases, we observe that in rare instances the valid translation appears in the penultimate sentence.

3.2 Hyperparameters

In our fine-tuning setup for the Qwen2.5-72B-Instruct model, we employ parameter-efficient LoRA (Low-Rank Adaptation) with rank 8 applied to all trainable layers, optimized through DeepSpeed ZeRO-3 for memory-efficient distributed training. The training configuration adopts a global batch size of 64 (8 per-device batch size with 8 gradient accumulation steps), cosine learning rate scheduling with an initial rate of 1e-4 and 0.1 warmup ratio, running for 3 epochs to balance convergence and computational cost. We set the sequence length cutoff at 2048 tokens to match the model's context window. This configuration leverages adaptive mixed-precision training (bfloat16) and LORA's low-rank reparameterization to maintain the base model's linguistic capabilities while efficiently adapting to downstream tasks.

3.3 Data Augmentation

Data augmented (DA) is a widely used strategy to mitigate data scarcity in low-resource environments. The insufficient training data issue for certain languages is addressed through the application of this technique.

We trained the initial model, Qwen2.5-72B-LoRA, using the training set provided in the task along with the Qwen2.5-72B model. We compared the translation capabilities of Qwen2.5-Max, Deepseek-R1 (DeepSeek-AI, 2025), qwen-mt, and qwen-mt-turbo in Chinese and Korean. After comparative evaluation revealed the superior cost-effectiveness of the baseline model over alternative approaches, we strategically repurposed the trained Qwen2.5-72B-LoRA model for synthetic data generation.

Then, for each target language, we first randomly sampled data from Wikidata and filtered out entities lacking translations in the specific target language, resulting in entity pairs (original entity and its translated counterpart). These entities were then formatted into prompts for generation using our Qwen2.5-72B-LoRA model(see Appendix 2). To ensure the richness of the generated data, the sampling parameters with a temperature value of 0.7 and a top-p value of 0.9 were used. We subsequently filtered out samples where 1) the English sentence didn't contain the original English entity, or 2) the target language sentence lacked the corresponding translated entity. The filtered data is incorporated into the training set through reference

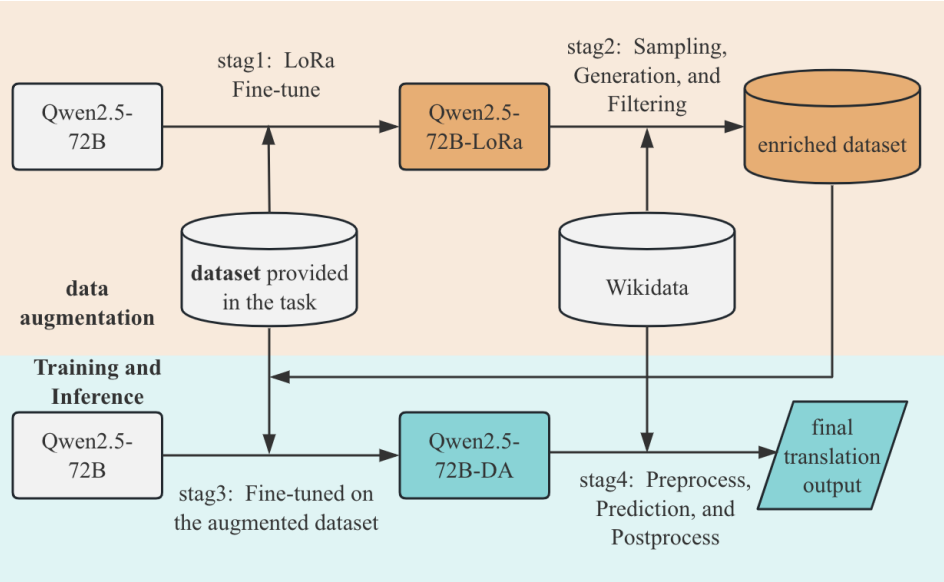

Figure 1: Data Augmentation and Training Pipeline

inference prompts(see Appendix 1). Finally, we combined this augmented dataset with the original training data to fine-tune Qwen2.5-72B, yielding our final translation model Qwen2.5-72B-DA.

4 Experimental Setup

The experiments were conducted on 8×NVIDIA A100 GPUs, which each contain 80GB HBM2 memory. The experiments is evaluated using the official metric of the competition: the harmonic mean of translation quality (COMET) and the precision of the translation of the entity (M-ETA). COMET is a metric for assessing machine translation quality by comparing system outputs to human reference translations. It utilizes a pre-trained model to generate translation quality scores, quantifying the semantic and linguistic fidelity of translations. M-ETA evaluates the accuracy of entity translations. Given a gold-standard set of entity translations and a system's predicted entities, M-ETA calculates the proportion of correctly translated entities. The final overall score will be the harmonic mean of the two scores:

$$\text{FinalScore} = \frac{2 \cdot \text{COMET} \cdot \text{M-ETA}}{\text{COMET} + \text{M-ETA}} \quad (1)$$

5 Results and Analysis

5.1 Determining Base Model

To identify an optimal open-source foundation model for fine-tuning and establish performance baselines for subsequent enhancement comparisons, we conducted a systematic comparative analysis of inference capabilities across multiple candidate models. All models were evaluated under identical experimental conditions, using a standardized prompt template. As shown in Table 1, the Qwen2.5-72B model demonstrated superior performance across benchmark evaluations, leading to its selection as our base architecture. Interestingly, our comparative analysis revealed that the DeepSeek-R1 distilled model underperformed the original model in translation tasks, a phenomenon potentially attributable to knowledge distillation effects that may enhance model capabilities for complex reasoning tasks at the expense of linguistic transfer proficiency.

	Average across all languages		
Model	**M-ETA**	**Comet**	**Overall**
DeepSeek-R1-Distill-Qwen-32B	87.51	91.56	89.44
Qwen2.5-32B	87.72	91.87	89.71
DeepSeek-R1-Distill-Llama-70B	87.31	93.10	90.06
Phi-4	87.81	93.33	90.45
Llama-3.3-70B	88.40	93.83	90.98
Qwen2.5-72B	**88.71**	**94.01**	**91.24**

Table 1: The scores of several current top native open source models on the task. These models use the 'instruct' version.

System	ZH		KO	
	M	C	M	C
Deepseek-R1	81.00	93.75	-	-
Qwen2.5-Max	80.00	94.19	90.00	95.31
qwen-mt-turbo	-	-	**91.00**	94.92
qwen-mt-plus	-	-	**91.00**	95.10
Qwen2.5-72B-LoRA	**81.26**	**94.44**	90.24	**95.44**

Table 2: Comparative evaluation of Qwen2.5-72B-LoRA against commercial LLMs on traditional Chinese (ZH) and Korean (KO) machine translation, measured through M-ETA and COMET metrics.

5.2 Determining Data Augmentation Model

To address data imbalance issues, we adopted a closed source LLM-based data enhancement strategy to improve system capabilities. Our selection encompassed four state-of-the-art commercial models: Qwen2.5-Max (the most powerful LLM in Qwen series), Qwen-MT-Plus (Qwen's premium machine translation model), Qwen-MT-Turbo (Qwen's efficiency-optimized MT model), and DeepSeek-R1 (a high-performance reasoning LLM comparable to OpenAI's o1 model, accessed via API due to computational constraints precluding local deployment of its 671B variant).

Limited to two linguistically distinct languages, Traditional Chinese and Korean, our evaluation revealed critical performance boundaries. The Qwen-MT series exhibited subpar performance in Traditional Chinese translation tasks, whereas DeepSeek-R1 showed constrained multilingual proficiency beyond Chinese-English language pairs, as documented in their respective technical reports. For controlled comparison, we included Qwen2.5-72B-LoRA(our LoRA fine-tuned Qwen2.5-72B), the top-performing model on task leaderboards without data augmentation.

As detailed in Table 2, Qwen2.5-72B-LoRA achieved superior performance in traditional Chinese while maintaining competitive results in Korean. This performance-cost equilibrium led to the selection of Qwen2.5-72B-LoRA for the final implementation. Our analysis suggests that domain-adapted fine-tuning effectively compensates for data sparsity without requiring extensive augmentation pipelines.

5.3 Main Results

During the validation phase, the Qwen2.5-72B-LoRA (without data augmentation) achieved state-of-the-art performance with a score of 91.79, securing the top position on the task leaderboard (see Table 3). In the subsequent post-validation phase, we implemented a data augmentation strategy by automatically generating sentences and their corresponding translations based on the source and target entities of Wikidata using Qwen2.5-72B-LoRA. The augmented dataset, formed by merging these synthetic instances with the original training data, was utilized to re-fine-tune the Qwen2.5-72B model via LoRA, yielding the enhanced Qwen2.5-72B-DA variant. This optimized model demonstrated superior efficacy, attaining an improved score of 91.93.

5.4 Analysis

As shown in Table 3, languages including Arabic, German, Italian, Japanese, Korean, and Thai achieve the most significant performance gains in our system. As revealed in Table 4, these improvements stem primarily from the enhancements in COMET (C) scores. The M-ETA (M) metric is primarily influenced by the presence of entities in translated sentences. However, since we directly retrieve corresponding target-language entities via Wiki IDs, and given that many entities in the test set lack target-language entries in Wikidata, it becomes challenging to improve M-ETA scores substantially. In contrast, the COMET (C) scores are model-generated evaluations that can better capture grammatical and semantic improvements through training. This explains why experiments demonstrates consistent improvements in COMET (C) scores across nearly all languages, which consequently drives the overall performance gains. The distinct evaluation mechanisms of these two metrics account for the observed differential improvement patterns.

6 Conclusion

Our work demonstrates that combining parameter-efficient adaptation with structured knowledge injection significantly improves entity translation accuracy. The LoRA-tuned Qwen2.5-72B model outperforms both commercial MT systems and open-source LLMs. The Wikidata augmentation strategy proves particularly effective for handling culture-specific items and rare entities. Future work

rank	System	AR	DE	ES	FR	IT	JA	KO	TH	TR	ZH	Avg
5	Phi4-FullFT	92.09	89.61	92.3	92.72	94.31	93.53	92.98	91.27	89.35	87.02	91.54
4	GPT-4o-WikiData-RAG	**93.24**	89.46	92.42	92.5	94.33	92.55	92.92	92.46	88.82	**87.51**	91.62
3	Qwen2.5-Max-Wiki	92.89	89.92	**92.63**	92.43	94.3	93.55	92.88	92.28	89.2	87.11	91.72
1	**Qwen2.5-72B-LoRA**	92.68	90.03	92.54	**92.92**	94.39	93.34	92.77	92.35	**89.54**	87.36	91.79
-	**Qwen2.5-72B-DA**	93.22	**90.35**	92.52	92.53	**94.54**	**93.86**	**93.15**	**92.61**	89.31	87.28	**91.93**

Table 3: Comparison of our system with top-performing systems in the leaderboard of SemEval-2025 Task 2 across different languages. Language codes: Arabic (AR), German (DE), Spanish (ES), French (FR), Italian (IT), Japanese (JA), Korean (KO), Thai (TH), Turkish (TR), and Chinese (ZH).

	AR		DE		ES		FR		IT		JA		KO		TH		TR		ZH		Avg	
System	M	C	M	C	M	C	M	C	M	C	M	C	M	C	M	C	M	C	M	C	M	C
Qwen2.5-72B	**91.7**	93.6	84.9	93.4	90.0	94.9	91.3	93.3	92.9	95.2	91.4	95.4	90.0	94.1	**91.3**	93.0	82.7	93.4	80.9	93.9	88.7	94.0
Qwen2.5-72B-LoRa	**91.7**	93.6	86.4	94.0	**90.1**	95.1	**91.6**	94.3	93.0	95.8	91.4	95.4	90.2	95.4	91.2	93.5	**84.1**	**95.7**	**81.3**	94.4	**89.1**	94.7
Qwen2.5-72B-DA	91.6	**94.5**	**86.5**	**94.6**	90.0	**95.2**	90.6	**94.5**	**93.1**	**96.1**	**91.7**	**96.1**	**90.7**	**95.8**	91.1	**94.1**	83.9	95.5	81.1	**94.5**	89.0	**95.1**

Table 4: Results across languages with M-ETA (M) and Comet (C) scores. Language codes: Arabic (AR), German (DE), Spanish (ES), French (FR), Italian (IT), Japanese (JA), Korean (KO), Thai (TH), Turkish (TR), and Chinese (ZH).

should investigate neural entity linking frameworks to supplant structured knowledge base alignment, enabling dynamic adaptation to emerging entities beyond predefined Wikidata schemas.

Limitations

(1) **Cross-lingual entity alignment challenges**: Systematic misalignment between Wikidata reference translations and human-annotated entities in linguistically distant languages (e.g., Traditional Chinese) degrades M-ETA performance. Our mitigation strategy—optimizing for COMET score improvements—partially compensates but fails to resolve the underlying knowledge base inconsistencies.
(2) **Structured knowledge dependency**: Reliance on Wikidata ID matching creates deployment bottlenecks, as real-world applications rarely provide structured knowledge base identifiers.

Acknowledgments

This research was enabled by computational infrastructure support from Ping An Life Insurance (Group) Co., Ltd. of China. We additionally acknowledge the SemEval-2025 Task 2 organizers for establishing the multilingual evaluation framework that catalyzed this investigation.

References

Simone Conia, Daniel Lee, Min Li, Umar Farooq Minhas, Saloni Potdar, and Yunyao Li. 2024. Towards cross-cultural machine translation with retrieval-augmented generation from multilingual knowledge graphs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16343–16360, Miami, Florida, USA. Association for Computational Linguistics.

Simone Conia, Min Li, Roberto Navigli, and Saloni Potdar. 2025. SemEval-2025 task 2: Entity-aware machine translation. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*. Association for Computational Linguistics.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

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

Min Li, Liangyou Li, and Philipp Koehn. 2022. Low-frequency entity translation in neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3021–3033.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz

Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Yuxuan Wang, Ning Ding, Xin Zhou, Hai-Tao Zheng, and Zhiyuan Liu. 2023. Cross-lingual entity linking via adversarial domain adaptation. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence (AAAI)*, pages 13289–13297.

A Appendix

Listing 1: Entity-aware Translation Prompt

```
Please translate the following English sentence
    into {target_lang}, requiring an accurate
    translation of the entity.
The sentence to be translated is:
'{sentence}'
You must translate sentences using the entities
    mentioned below:
'{source_entity}': '{target_entity}'
Please output the translated sentence directly.
```

Listing 2: Data Augmentation Prompt

```
### Role and Task Specification
Generate 3 natural English sentences containing
    '{source_entity}', then translate them into
    {target_language} using the exact
    translation '{target_entity}'.

### Requirements
1. Apply different tenses (past/present/future)
    and voices (active/passive)
2. Include these styles:
   - 1 conversational sentence (daily)
   - 1 formal sentence (news/tech)
   - 1 literary expression
3. Return JSON array format with fields:
   - "source": Original English sentence
   - "target": {target_language} translation
   - "style_tag": (daily/news/tech/literary)

### Example Input:
{{
  "source_entity": "Quantum Computing",
  "target_entity": "量子計算",
  "target_language": "Chinese"
}}

### Example Output:
[
    {{
        "source": "The breakthrough in quantum
            computing could revolutionize data
            encryption.",
        "target": "量子計算領域的突破可能徹底改變
            數據加密技術。",
        "style_tag": "tech"
    }},
    # ...other examples...
]
```

Listing 3: Examples with explanations of the process and mixed language output.

```
{
    "id": "0404e153f1dbd9d8",
    "source_language": "English",
    "target_language": "zh",
    "text": "What are some notable traits or
        characteristics of You Sanjie?",
    "prediction": "尤三姐有哪些 notable traits 或
        特性？\n\n(Note: \"notable traits 或特
        性\" can be more naturally phrased as \"
         notable 的特質或特
        點\" in Traditional Chinese, but I kept
        it closer to the original English
        structure for clarity.) \n\nFor a more
        natural translation:\n尤三姐有哪些顯著的
        特質或特點？"
}
```