# FII the Best at SemEval 2025 Task 2: Steering State-of-the-art Machine Translation Models with Strategically Engineered Pipelines for Enhanced Entity Translation

**Delia-Iustina Grigoriță** [1] **Tudor-Constantin Pricop** [1] **Sergio-Alessandro Șuteu** [1]
**Daniela Gîfu**[1,2] **Diana Trandabăț**[1]
[1]Faculty of Computer Science,
“Alexandru Ioan Cuza” University of Iasi
[2]Institute of Computer Science,
Romanian Academy - Iasi Branch
{deliaiustinagrigorita, pricoptudor2001, suteu.sergio}@gmail.com
daniela.gifu@iit.academiaromana-is.ro
diana.trandabat@info.uaic.ro

## Abstract

Entity-Aware Machine Translation (EAMT) aims to enhance the accuracy of machine translation (MT) systems in handling named entities, including proper names, domain-specific terms, and structured references. Conventional MT models often struggle to accurately translate these entities, leading to errors that affect comprehension and reliability. In this paper, we present a promising approach for SemEval 2025 Task 2, focusing on improving EAMT in ten target languages. The methodology is based on two complementary strategies: (1) multilingual Named Entity Recognition (NER) and structured knowledge bases for preprocessing and integrating entity translations, and (2) large language models (LLMs) enhanced with optimized prompts and validation mechanisms to improve entity preservation. By combining structured knowledge with neural approaches, this system aims to mitigate entity-related translation errors and enhance the overall performance of MT models. Among the systems that do not use gold information, retrieval-augmented generation (RAG), or fine-tuning, our approach ranked $1^{st}$ with the second strategy and $3^{rd}$ with the first strategy.

## 1 Introduction

Entity-aware machine translation (EA-MT) aims to improve MT accuracy for named entities, including proper names, dates, and domain-specific terms (Gifu & Vasilache 2014). These are crucial in fields like technical documentation, legal texts, and medical literature (Gîfu & Cioca 2013), yet translating them remains challenging despite modern advancements.

Early rule-based MT struggled with named entities due to rigid linguistic rules (Slocum 1985). Statistical Machine Translation (SMT) in the 1990s improved overall quality but still faced issues with proper nouns and domain-specific terms (Wang et al. 2022). Phrase-Based SMT (PB-SMT) in the 2000s enhanced phrase-level translations but remained inconsistent with named entities and long-distance dependencies (Koehn et al. 2003, Lopez 2008).

Neural Machine Translation (NMT) and Transformer-based models like BERT and GPT (Vaswani 2017) have enhanced fluency and contextual awareness. Yet, challenges remain in entity preservation, cultural adaptation, and low-resource language support (Zaki 2024, Gifu & Covaci 2025, Lupancu et al. 2023).

For SemEval 2025 Task 2 (Conia et al. 2025) on EA-MT, we developed two systems to improve entity-centric translation across ten languages. Our approach combines:

1. Multilingual Named Entity Recognition (NER) and structured knowledge bases– We preprocess source text by identifying named entities, aligning them with external structured resources (e.g., Wikidata), and reintegrating their transla tions while preserving contextual accuracy

2. Large Language Models (LLMs) with optimized prompt engineering and validation mechanisms – We leverage LLMs to refine translations, ensuring that named entities are preserved, properly adapted, and fluently integrated into the

*Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 558–565
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

target language. Beyond technical implementation, we systematically evaluate our models using both standard MT metrics (e.g.,BLEU, METEOR) and specialized entity-aware evaluation techniques that assess entity preservation and translation accuracy. Given the linguistic diversity of the task, our system is designed to handle complex challenges such as morphological variations, transliteration issues, and script-based differences in languages like Japanese, Chinese, Arabic, and Thai.

The remainder of this paper is structured as follows. Section 2 provides a background on the evolution of MT, from early rule-based systems to state-of-the-art transformer models. Section 3 details the system architecture design for EA-MT, outlining the experimental setup and datasets used. Section 4 presents the results and the comparative evaluation. Finally, Section 5 discusses conclusions and future directions for entity-aware MT research.

The complete implementation of our system is available on GitHub[1]

## 2 Background

From the 1960s to the 1980s, early machine translation (MT) systems were rule-based, offering structured translations but struggling with named entities, proper nouns, and idiomatic expressions (Hutchins 1986, Song & Xu 2024).

The 1990s introduced Statistical Machine Translation (SMT), leveraging probabilistic models to improve flexibility, yet still facing challenges with rare terms and domain-specific terminology. The 2000s saw Phrase-Based SMT (PB-SMT), enhancing contextual coherence but retaining difficulties with named entities (Zens & Ney 2004, Pal et al. 2004).

Neural Machine Translation (NMT) emerged in the 2010s, using deep learning to improve fluency and entity handling, though challenges persisted in low-resource languages and domain adaptation (Vaswani et al. 2017, Koehn & Knowles 2017).

Today, Transformer-based models like GPT and BERT push translation accuracy forward, excelling in contextual understanding but still struggling with cultural adaptation and low-resource languages (Devlin et al. 2019, Wang et al. 2022). Large Language Models (LLMs), such as GPT-4, now rival leading NMT systems, though performance varies across language pairs (Manakhimova et al. 2023).

Despite advances, translating low-resource languages remains a challenge, necessitating refined techniques like back-translation and transfer learning (Zeng 2023, Her & Kruschwitz 2024). Hybrid methodologies integrating rule-based, statistical, and neural approaches continue to be explored for further improvements (Wang et al. 2022).

## 3 Dataset and Methods

### 3.1 Dataset

The dataset contains sentence pairs aligned between English and 10 target languages, with named entities linked to Wikidata IDs for multilingual NER tasks. However, entity tagging is incomplete, often marking only some entities in a sentence while leaving others untagged, impacting annotation reliability and depth for tasks like translation.

For example, consider the following sentence pair:

**Source (English):** *"Which actor was Stephenie Meyer's first choice to play Edward Cullen in the movie Twilight?"*

**Target (Example Language):** *"Quale attore era stata la prima scelta di Stephanie Meyer per interpretare Edward Cullen nel film Twilight?"*

This sentence contains three distinct entities:

- **Stephenie Meyer** (author, `Q160219`)
- **Edward Cullen** (fictional character)
- **Twilight** (movie)

However, in the dataset, only **Stephenie Meyer** is tagged with the corresponding Wikidata ID `Q160219`, while**Edward Cullen** and **Twilight** are not tagged.

This inconsistency in entity recognition results in incomplete annotations, which directly impacts the utility of the dataset. This limitation is particularly critical when it comes to translation tasks, as missing entities such as **Edward Cullen** and **Twilight** could significantly alter the understanding of the original sentence in the target language.

### 3.2 Methods

#### 3.2.1 First approach

In our initial approach, we used the mBERT model, trained on the `WikiNEuRal:`

[1] https://github.com/deliagrigorita/FII-the-best-SemEval2025

Multilingual NER (Tedeschi et al. 2021) dataset, to extract named entities from the source text. This dataset is considered state-of-the-art for Multilingual Named Entity Recognition (NER) and is automatically derived from Wikipedia.

We generated two vectors: one containing the extracted named entities and another with corresponding translations retrieved via the Wikidata API, which offers accurate, human-curated translations for many entities.

To preserve the positions of the entities within the text, we replaced each named entity in the original text with a placeholder ([TAG-HOLDER]). The modified text was then used for subsequent processing: translation using the deep translator, deep_translator - GoogleTranslator, after which the placeholders were replaced with the Wikidata translations.

For entities not found in Wikidata, we kept them in the original language as a fallback. While this method delivered reasonable results, we identified a potential issue: translating a sentence with placeholders rather than the full context might disrupt grammatical conventions in the target language (e.g., misgendering articles in languages like Italian for person names).

To address this, we refined the process by replacing the translator with the Gemini API, utilizing the free Gemini 1.0 Pro version. This allowed us to leverage prompt engineering, providing the original entities, their Wikidata translations, and a request for grammatically accurate translation. This approach yielded superior results that aligned with the grammar of the target language. It also opens the possibility of experimenting with various LLMs to determine which delivers the best outcomes.

In Figure 1, we present the architecture of the first approach.

During the extraction, we observed that the model occasionally permuted certain special characters in the extracted named entities (NEs). For instance, in the extracted named entity *St Anne’s Cathedral*, the corresponding Wikidata translation would appear as ”St Anne’s Cathedral,” with spaces added around the apostrophe. We identified this as a consistent issue where punctuation marks, apostrophes, and other special characters were misrepresented. To address this, we implemented a normalization step to remove these inconsistencies, ensuring that such cases, along with others involving special characters, were corrected. After performing this normalization, we observed a marked improvement in the evaluation results, as the translation output became more accurate and consistent.

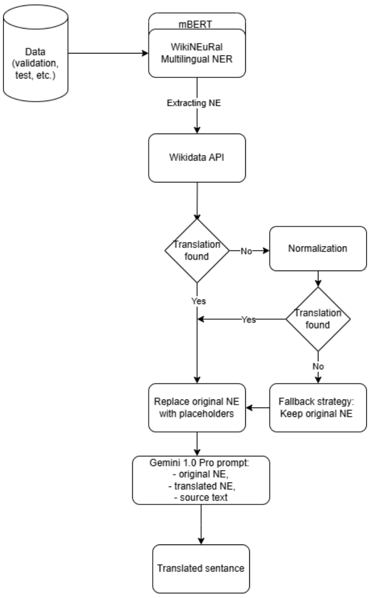

Figure 1: The architecture of the first approach

### 3.2.2 Second approach

The second implementation presents another approach to Entity-Aware Machine Translation (EAMT), leveraging large language models (LLMs) for high-fidelity text translation while ensuring the preservation and correct translation of named entities. The system follows a structured pipeline that isolates named entities, processes their translations separately, and reintegrates them into the translated text.

The translation of common words within a sentence is performed directly inside the LLM, as it is powerful enough to handle basic translations accurately. However, the translation of named entities utilizes external resources to ensure higher precision, as named entities require a human touch to maintain accuracy. Additionally, named entities evolve more frequently than common words, making it necessary to rely on up-to-date external resources such as structured knowledge bases (Conia et al. 2024).

Named Entity Recognition (NER) is the first step, where named entities are identified and extracted from the source text using a combination of entity recognition models and regex-based pattern matching. Once the named entities are detected, they are separated by specific tags in the source text to maintain the underlying linguistic structure during translation. The masked text is then processed independently using a large-scale LLM, such as Qwen 2.5 Instruct (Team 2024), which has demonstrated powerful translation skills in the required languages. Extracted named entities are handled separately, with their translations obtained from structured knowledge bases such as Wikidata. Once translated, the named entities are reintegrated into the translated text at their corresponding positions, ensuring fluency and semantic coherence.

The named entity translation module follows multiple strategies, including knowledge-based lookup by querying structured data sources like Wikidata and cross-lingual LLM-based heuristics, where the original entity may be retained or transliterated if no reliable translation is available. To enhance efficiency, the implementation integrates optimization techniques such as parallel processing to handle multiple sentences concurrently, using vLLM's fast inference framework.

In Figure 2, we present the architecture of the second approach.

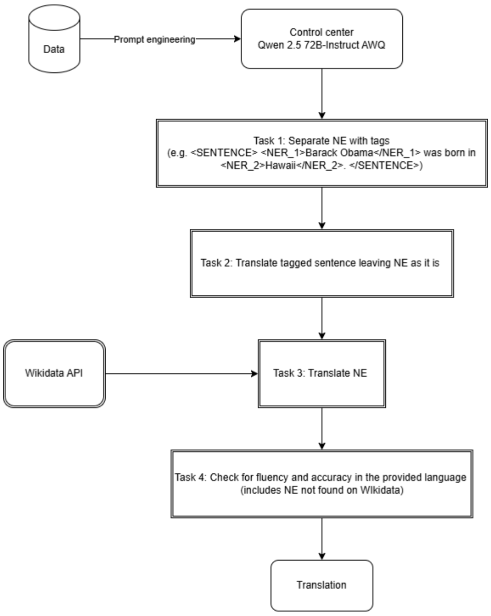
Data
Prompt engineering
Control center
Qwen 2.5 72B-Instruct AWQ
Task 1: Separate NE with tags
(e.g. <SENTENCE> <NER_1>Barack Obama</NER_1> was born in <NER_2>Hawaii</NER_2>. </SENTENCE>)
Task 2: Translate tagged sentence leaving NE as it is
Wikidata API
Task 3: Translate NE
Task 4: Check for fluency and accuracy in the provided language
(includes NE not found on WIkidata)
Translation

Figure 2: The architecture of the second approach

The effectiveness of the translation pipeline relies on well-structured prompts designed to guide the LLM in performing translations with high fidelity. Initially, prompts are crafted to explicitly instruct the LLM to focus on translating only the non-entity words while preserving placeholders for named entities. These prompts are refined iteratively to optimize clarity and accuracy, ensuring that the model correctly understands the distinction between common words and named entities. Additional prompt tuning techniques are employed, such as providing context-specific examples to enhance translation performance and prevent ambiguity. The prompt design also incorporates validation mechanisms, where the model's responses are analyzed, and adjustments are made dynamically to improve consistency in entity-aware translations.

Appendix A contains the prompts that were used to guide Qwen for translation.

## 4 Results

In this section, we present the evaluation results of our two proposed strategies. The evaluation was conducted using two main metrics: COMET and M-ETA across ten target languages: Arabic (AE), German (DE), Spanish (ES), French (FR), Italian (IT), Japanese (JP), Korean (KR), Thai (TH), Turkish (TR), and Traditional Chinese (TW).

- COMET (Cross-lingual Optimized Metric for Evaluation of Translation) is a neural-based metric that assesses machine translation quality using contextual embeddings to compare source, translation, and reference sentences.
- M-ETA (Manual Entity Translation Accuracy) measures entity translation accuracy by computing the proportion of correctly translated entities against a gold standard.

The final evaluation score is calculated as the harmonic mean of the COMET and M-ETA scores, ensuring a balanced assessment that accounts for both overall translation quality and entity preservation.

The first strategy demonstrated varying levels of performance across languages, as shown in Table 1.

Spanish (es_ES) achieved the highest final score of 79.1, followed closely by Arabic (ar_AE) and French (fr_FR) with final scores of 77.54 and 77.5, respectively. The lowest performance was observed for Chinese (zh_TW), which obtained a final score of 40.71 due to a significantly lower M-ETA score (26.46). Other languages, such as Turk-

| Languages | M-ETA Score | COMET Score | Final Score |
|---|---|---|---|
| Arabic (ar_AE) | 68.11 | 90.01 | 77.54 |
| German (de_DE) | 62.63 | 89.13 | 73.56 |
| Spanish (es_ES) | 69.91 | 91.06 | 79.1 |
| French (fr_FR) | 68.11 | 89.89 | 77.5 |
| Italian (it_IT) | 67.67 | 88.5 | 76.7 |
| Japanese (ja_JP) | 66.68 | 91.82 | 77.26 |
| Korean (ko_KR) | 64.11 | 90.72 | 75.13 |
| Thai (th_TH) | 55.41 | 85.2 | 67.15 |
| Turkish (tr_TR) | 56.9 | 90.19 | 69.77 |
| Chinese (zh_TW) | 26.46 | 88.29 | 40.71 |

Table 1: First Strategy Results

ish (tr_TR) and Thai (th_TH), also showed moderate performance, with final scores of 69.77 and 67.15, respectively.

The second strategy, shown in Table 2, yielded higher final scores across most languages compared to the first strategy. Italian (it_IT) achieved the highest final score of 83.4, followed by Spanish (es_ES) with 81.22 and French (fr_FR) with 80.52. The lowest performance was again observed for Chinese (zh_TW); however, the final score (74.19) showed a significant improvement over the first strategy. Additionally, languages such as Turkish (tr_TR) and Thai (th_TH) exhibited better scores than in the first strategy, with final scores of 77.77 and 75.16, respectively.

| Languages | M-ETA Score | COMET Score | Final Score |
|---|---|---|---|
| Arabic (ar_AE) | 66.42 | 91.35 | 76.91 |
| German (de_DE) | 66.98 | 91.3 | 77.27 |
| Spanish (es_ES) | 72.35 | 92.58 | 81.22 |
| French (fr_FR) | 72.46 | 90.59 | 80.52 |
| Italian (it_IT) | 75.79 | 92.71 | 83.4 |
| Japanese (ja_JP) | 67.03 | 93.56 | 78.11 |
| Korean (ko_KR) | 66.02 | 92.78 | 77.14 |
| Thai (th_TH) | 65.25 | 88.62 | 75.16 |
| Turkish (tr_TR) | 67.56 | 91.63 | 77.77 |
| Chinese (zh_TW) | 62.5 | 91.25 | 74.19 |

Table 2: Second Strategy Results

While both strategies performed well in handling named entities in translation, the second strategy generally produced higher final scores across most languages. Improvements in M-ETA and COMET scores were particularly noticeable for Italian (it_IT), French (fr_FR), and Chinese (zh_TW). However, variations still exist among different languages, indicating that certain language pairs may require further refinement. Future work will explore the potential benefits of merging these two strategies to leverage their strengths and further enhance translation performance. [2]

## 5 Conclusion

In this work, we explored entity-aware machine translation (EA-MT) by proposing two approaches aimed at improving the translation of named entities across multiple languages. Our first approach relied on the mBERT model for Named Entity Recognition (NER) combined with Wikidata-based entity translations, while our second approach leveraged large language models (LLMs) with structured prompt engineering to enhance translation accuracy.

Our experiments demonstrated that accurately recognizing and preserving named entities is crucial for high-quality translation. We identified several challenges, such as inconsistent entity annotations in the dataset and grammatical disruptions caused by placeholder-based translations. To mitigate these issues, we refined our methodology by incorporating normalization techniques and utilizing Wikidata as a reliable source for entity translations. The second approach, which integrated LLMs for translation while maintaining entity integrity, proved to be more effective in producing fluent and semantically accurate translations.

## 6 Future Work

While our proposed strategies have shown promising results, there is still room for improvement in enhancing the quality of the final translation. Firstly, let's consider the strategy that relied on Gemini 1.0. Although useful, this model occasionally struggled to fully adhere to prompt instructions, resulting in deviations from expected outputs. Additionally, as Gemini 1.0 is now being discontinued, transitioning to more advanced models has become a necessity.

To address these issues, future iterations of our first strategy will incorporate a more advanced large language model (LLM) with superior capabilities. By leveraging a model with improved contextual awareness and better alignment to user

[2] In the final leaderboard, the submissions can be found under the names *FII-UAIC-SAI* for the second strategy and *FII the Best* for the first strategy.

prompts, we expect a significant boost in translation accuracy across multiple languages.

Another area for future improvement is to integrate both strategies into a unified system, leveraging their strengths to enhance translation performance.

## References

Conia, S., Lee, D., Li, M., Minhas, U. F., Potdar, S. & Li, Y. (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', Association for Computational Linguistics, Miami, Florida, USA.

Conia, S., Li, M., Navigli, R. & Potdar, S. (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.

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2019), Bert: Pre-training of deep bidirectional transformers for language understanding, Association for Computational Linguistics, pp. 4171–4186.

Gîfu, D. & Cioca, M. (2013), 'Online civic identity. extraction of features', *Procedia-Social and Behavioral Sciences* **76**, 366–371.

Gifu, D. & Covaci, S.-V. (2025), 'Artificial intelligence vs. human: Decoding text authenticity with transformers', *Future Internet* .

Gifu, D. & Vasilache, G. (2014), 'A language independent named entity recognition system', *Alexandru Ioan Cuza" University Publishing House, Iaşi* pp. 181–188.

Her, W.-H. & Kruschwitz, U. (2024), 'Investigating neural machine translation for low-resource languages: Using bavarian as a case study', *arXiv preprint arXiv:2404.08259* .

Hutchins, W. J. (1986), *Machine Translation: Past, Present, Future*, Ellis Horwood.

Koehn, P. & Knowles, R. (2017), 'Six challenges for neural machine translation', *arXiv preprint arXiv:1706.03872* .

Koehn, P., Och, F. J. & Marcu, D. (2003), Statistical phrase-based translation, *in* '2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Langauge Technology (HLT-NAACL 2003)', Association for Computational Linguistics, pp. 48–54.

Lopez, A. (2008), 'Statistical machine translation', *ACM Computing Surveys (CSUR)* **40**(3), 1–49.

Lupancu, V.-C., Platica, A.-G., Rosu, C.-M., Gifu, D. & Trandabat, D. (2023), Fii_better at semeval-2023 task 2: Multiconer ii multilingual complex named entity recognition, *in* 'Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)', pp. 1107–1113.

Manakhimova, S., Avramidis, E., Macketanz, V., Lapshinova-Koltunski, E., Bagdasarov, S. & Möller, S. (2023), Linguistically motivated evaluation of the 2023 state-of-the-art machine translation: Can ChatGPT outperform NMT?, *in* P. Koehn, B. Haddow, T. Kocmi & C. Monz, eds, 'Proceedings of the Eighth Conference on Machine Translation', Association for Computational Linguistics, Singapore, pp. 224–245.
**URL:** *https://aclanthology.org/2023.wmt-1.23/*

Pal, S., Naskar, S. K., Pecina, P., Bandyopadhyay, S. & Way, A. (2004), Handling named entities and compound verbs in phrase-based statistical machine translation, Association for Computational Linguistics, pp. 46–54.

Slocum, J. (1985), 'A survey of machine translation: Its history, current status and future prospects', *Computational linguistics* **11**(1), 1–17.

Song, H. & Xu, H. (2024), A deep analysis of the impact of multiword expressions and named entities on chinese-english machine translations, Association for Computational Linguistics, pp. 6154–6165.

Team, Q. (2024), 'Qwen2.5: A party of foundation models'.
**URL:** *https://qwenlm.github.io/blog/qwen2.5/*

Tedeschi, S., Maiorca, V., Campolungo, N., Cecconi, F. & Navigli, R. (2021), Wikineural: Combined neural and knowledge-based silver data creation for multilingual ner, *in* 'Findings of the association for computational linguistics: EMNLP 2021', pp. 2521–2533.

Vaswani, A. (2017), 'Attention is all you need', *Advances in Neural Information Processing Systems* .

Vaswani, A., Shazeer, N. M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. & Polosukhin, I. (2017), 'Attention is all you need.', *Neural Information Processing Systems* .

Wang, H., Wu, H., He, Z., Huang, L. & Church, K. W. (2022), 'Progress in machine translation', *Engineering* **18**, 143–153.

Zaki, M. Z. (2024), 'Revolutionising translation technology: Acomparative study of variant transformer models-bert, gpt and t5', *Computer Science and Engineering–An International Journal* **14**(3), 15–27.

Zeng, H. (2023), Achieving state-of-the-art multilingual translation model with minimal data and parameters, *in* 'Proceedings of the Eighth Conference on Machine Translation', pp. 181–186.

Zens, R. & Ney, H. (2004), Improvements in phrase-based statistical machine translation, Association for Computational Linguistics, pp. 257–264.

## A Prompts Used to Guide Qwen for Translation

```
system_prompt = """You are an advanced
    language model skilled at
    identifying and isolating named
    entities in a sentence."""

user_prompt = """Given a sentence,
    perform the following tasks:
Identify the named entities in the
    sentence.
Encapsulate each named entity between <
    NER_{number}> and </NER_{number}>
    tags, where number indicates the
    order of the entity found.
Encapsulate the entire sentence, with
    the named entity tags included,
    between <SENTENCE> and </SENTENCE>
    tags.
Example:
Input: Barack Obama was born in Hawaii.
Output: <SENTENCE> <NER_1>Barack Obama</
    NER_1> was born in <NER_2>Hawaii</
    NER_2>. </SENTENCE>
Task:
Input: %(input_sentence)s
Output:"""
translate_system_prompt = """You are a
    highly skilled language model
    capable of translating text between
    languages with high accuracy.
Translate sentences into the specified
    target language while preserving
    their meaning and context.
Do not translate the parts of the
    sentence enclosed between <NER> and
    </NER> tags."""
translate_user_prompt = """Translate the
     following sentence into %(
    target_language)s:\n"
Sentence: %(sentence)s
Translation:"""

validate_system_prompt = """You are an
    expert in evaluating the fluency and
     naturalness of sentences in a
    specific language.
Your task is to determine whether a
    provided sentence sounds natural and
     fluent in the target language.
If the sentence is already fluent and
    natural, return it as is.
Do not provide explanations or reasoning
    .
If minor adjustments are needed for
    fluency, provide the refined
    sentence in the target language.
The target language is %(language)s.
"""
validate_user_prompt = """The following
    sentence is in %(language)s.
Please evaluate whether it sounds
    natural and fluent in the target
    language.
Translated Sentence: %(translated)s
    Final fluent sentence: """
```

Listing 1: Qwen prompts

Implementation details:

- Model: Qwen 2.5 Instruct - 72b - AWQ (Team 2024)
- Sampling parameters: temperature=0.3 (small value, follow instructions more closely), min_p=0.01 (filter unlikely tokens).
- Environment: GPU L4 x 4