

STaR-MT: Select, Translate, and Revise Pipeline for Terminology Aware Machine Translation

Guilherme Aren Marchetti¹, Gil Rocha¹, Henrique Lopes Cardoso¹, Rui Sousa-Silva²

¹LIACC, Faculdade de Engenharia, Universidade do Porto

²CLUP, Faculdade de Letras, Universidade do Porto

Correspondence: up202408600@up.pt

Abstract

To assess the progress in terminology-aware machine translation, the Terminology Shared Task is hosted alongside the Conference on Machine Translation (WMT). In this paper, we describe our submission to the shared task: STaR-MT. This is a three-step pipeline composed of (S)election, (T)ranslation, (a)nd (R)evision, focused on terminology-aware machine translation (MT). Our methodology adopts a lightweight agent-based design that selects relevant entries or examples and presents them to the translation agent, which uses a general-purpose LLM with a large context window to produce the initial output. Next, the revision agent uses language-specific models to improve translation quality for selected target languages. Experiments on the FLORES+ dataset with different model sizes in the pipeline suggest that machine translation pipelines can benefit from combining large general-purpose translation models with language-specific revision models, at least in some language pairs. While simple in architecture, this approach allows us to generate translations at the document level, with no need for pre-processing or paragraph splitting, and with minimal errors across most languages.

1 Introduction

The performance of Large Language Models (LLMs) in machine translation tasks has significantly improved in recent years (Lyu et al., 2024). Despite this progress, important challenges remain. One such challenge is terminology-aware machine translation, in which systems must not only produce fluent, accurate translations but also preserve domain-specific terminology consistently across sentences and documents. This requirement is especially important in professional translation settings, as economy, precision, and adequacy are three basic principles of Terminology (Cabr , 1999) even when the overall translation is understandable. Recent findings from the WMT Terminology Shared

Task show that terminology-aware translation still lags behind standard sentence-level translation, particularly in more complex document-level scenarios (Semenov et al., 2025). This is understandable, considering that terminology relies to a large extent on conceptualization systems, with which LLMs still struggle.

To assess progress in this area, the Terminology Shared Task is organized as part of the Conference on Machine Translation (WMT). In its fourth edition (Charkiewicz et al., 2026), the shared task expanded its language coverage to include two lower-resourced languages, Basque and Polish, in addition to English and Traditional Chinese. Participants could submit systems to two tracks. The first track focused on sentence-level translation with access to an explicit terminology dictionary, while the second track addressed document-level translation with only a small number of translation examples per language pair.

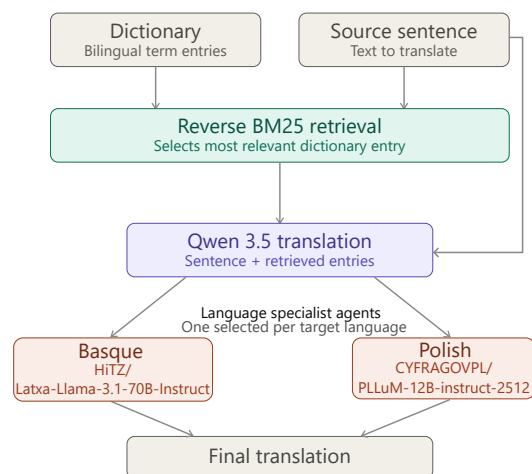


Figure 1: Diagram of the STaR-MT pipeline, comprised of three steps: Select, Translate, and Revise.

In the first track, the goal was to perform sentence-level terminology-aware translation us-

ing a provided dictionary. The track included two translation directions: Spanish→Basque and English→Polish. For each language pair, the input sentences were divided into two subject domains. The Spanish→Basque pair covered Automation and Energy, while the English→Polish pair covered Mechanical Engineering and Medicine. Each language-domain combination had its own terminology dictionary, containing approximately 1,000 to 3,000 entries. To evaluate whether systems could correctly use the intended terminology, the dictionaries also included a “random” translation alongside the correct one. Participants were required to generate three translations for each sentence: one using the correct terminology entry, one using the random terminology entry, and one without any terminology information.

The second track introduced two changes that substantially increased the task’s difficulty. First, the input texts were much longer. While the first track focused on sentence-level translation, the second track required translating at the document level, with several texts exceeding 20,000 words. Second, the explicit terminology dictionaries were removed. Instead, participants were given only a few translation examples from the relevant domain. This made it necessary for systems to identify and reuse useful translation patterns without direct access to a complete term base. The second track also introduced a third translation direction, Traditional Chinese→English. The subject domains for this track were Mechanical Tools and Railways for Spanish→Basque, Mechanical Tools and Medicine for English→Polish, and Finance for Traditional Chinese→English.

In this paper, we describe our submission to the shared task: STaR-MT, a terminology-aware machine translation pipeline composed of three stages (i.e., Selection, Translation, and Revision). Motivated by findings from the previous edition of the shared task, where LLM-based approaches were used in the majority of submissions (Semenov et al., 2025), we focus exclusively on open-weight models. The pipeline first selects the most relevant terminology entries or translation examples using a BM25 index. It then prompts a multilingual LLM to generate an initial translation conditioned on the selected information. Finally, when beneficial for the target language, a language-specific model revises the initial translation to improve its adequacy in the target language. The same general architecture is used for both tracks, with the selec-

tion component adapted to the available resources in each setting (i.e., the translation dictionary for track 1 and translation examples for track 2).

Our internal evaluation on the FLORES+ dataset (NLLB Team et al., 2024; Burchell et al., 2024; Dale et al., 2025) shows that this is a promising architecture. By comparing different model sizes and revision configurations across the language directions included in the shared task, we find that language-specific models can improve translations produced by more general-purpose models in some settings. It is interesting to note that although the larger language-specific models produce the best revisions, the general-purpose model can still benefit from revisions made by smaller specialist models, albeit to a reduced degree. This is especially relevant for lower-resourced or morphologically rich target languages, where dedicated models may provide linguistic knowledge that complements the broader multilingual capabilities of large LLMs.

2 Related Work

Several approaches have been proposed to improve the performance of LLMs in machine translation. Although these approaches are often discussed separately, they are not mutually exclusive and can be combined within a single translation pipeline.

Prompt engineering methods rely on LLMs’ ability to follow natural-language instructions. In MT, such instructions can specify the desired style, register, domain, target audience, or degree of structural preservation. These methods are attractive because they do not require updating the model parameters and can therefore be applied directly to general-purpose LLMs. However, prompt engineering also requires substantial experimentation, since small changes in prompt wording may lead to unpredictable changes in the final output. Despite this limitation, previous work has shown that explicitly providing information about the translation purpose and target audience can improve translation quality (Yamada, 2023).

A closely related strategy is in-context learning, also referred to as few-shot prompting (Brown et al., 2020). Instead of modifying the model parameters, this approach includes translation examples directly in the prompt, allowing the model to infer the expected translation pattern from the provided context. Prior work has shown that translation examples can steer LLM outputs and improve translation behavior (Alves et al., 2023). Be-

yond sentence-level examples, prompts can also include terminology constraints, style guidelines, or domain-specific instructions. This is particularly relevant for specialized translation tasks, where preserving terminology is often as important as producing fluent target-language text. In-context information has also been shown to help lower-resource languages, for example, through the use of loanword or terminology constraints (Ali et al., 2025).

More recently, LLMs have also been used as components of agent-based systems. This line of work has gained popularity as LLMs have become increasingly capable of instruction following, reasoning, planning, and tool use (Gao et al., 2024; Arora et al., 2023). In translation, agent-based methods enable decomposition of the process into multiple stages, with different agents responsible for distinct aspects of the workflow (Wu et al., 2024). For example, one agent may generate translation guidelines, another may enforce terminology and localization constraints, and a third may revise or proofread the output. This structure resembles professional translation workflows, where translation and revision are typically performed as separate steps.

Our methodology follows this general direction, but adopts a lightweight agent-based design. Rather than relying on a large number of interacting agents, we use a three-step pipeline composed of selection, translation, and revision. The selection step retrieves relevant terminology entries or translation examples, connecting retrieval-augmented generation with in-context learning. The translation step uses a general-purpose LLM with a large context window to produce the initial output from a terminology-enriched prompt. Finally, the revision step uses language-specific models to improve translation quality for selected target languages. In this way, the proposed STaR-MT pipeline combines concepts from prompt engineering, retrieval-augmented generation, and agent-based translation, while keeping the architecture simple.

3 Methodology

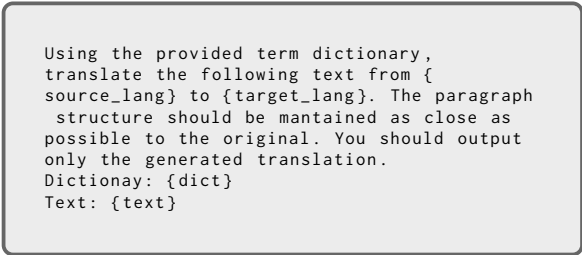
Our STaR-MT methodology is composed of three different steps (Figure 1):

Selection. The pipeline first selects the terminology entries or translation examples that are most relevant to each source sentence. We implement this selection using a reverse BM25 retrieval strat-

egy. Rather than building the BM25 index over the translation dictionary, we index the source-language sentences that need to be translated. Each dictionary entry is then treated as a query and searched against this sentence index. If a dictionary entry ranks a given source sentence highly, the entry is considered relevant to that sentence. For each sentence, we keep the top-ranked dictionary entries according to their BM25 scores and include them in the prompt provided to the translation agent.

Translation. The initial translation is produced by a single LLM-based agent. This agent receives a code-switched prompt: the main instructions are written in English, while the selected terminology entries and translation examples are preserved in their original languages, as shown in Figure 2. The model is instructed to translate the input while preserving, as much as possible, the original sentence and paragraph structure.

We use a model from the Qwen 3.5 family (Qwen Team, 2026). This choice is motivated primarily by its large context window of approximately 262k tokens, which allows the model to process the long documents from Track 2 alongside retrieved terminology and translation examples without splitting the input into multiple smaller prompts. In addition, models from this family support all target languages considered in this work, including Basque, although performance for Basque is comparatively lower.



```
Using the provided term dictionary,
translate the following text from {
source_lang} to {target_lang}. The paragraph
structure should be maintained as close as
possible to the original. You should output
only the generated translation.
Dictionary: {dict}
Text: {text}
```

Figure 2: Prompt used for the initial translation using few-shot examples and code-switching. The {dict} parameter will be replaced by either pair of words from track 1, or the translation samples from track 2. The {text} parameter contains the text to be translated from the source language.

Revision. The final step revises the initial translation generated by the Qwen model. For this stage, we employ language-specific models for each target language. These models were further trained by their respective teams on curated datasets. For

Basque, we use a model from the Latxa family (Sainz et al., 2025); for Polish, we use a model from the PLLuM family (Kocoń et al., 2025). English translations were revised using the Gemma 4 family only during development, as our evaluation showed that this step decreased translation quality in English, as discussed in Section 3.1.

Each revision agent uses a prompt written in the corresponding target language, ensuring that all text presented to the model is in a single language. The revision prompt was originally written in English and translated using the DeepL platform¹. Figure 3 contains the original revision prompt in English, while Figures 4 and 5 depict the Basque and Polish versions, respectively.

This text was originally translated from Chinese. Its structure and wording should be checked for compliance with standard {subject} sector terminology. The structure of the original paragraphs should be preserved. Output only the revised version. Text to be checked: {translation}.

Figure 3: Prompt for the English revision.

Testu hau jatorriz gazteleraz itzuli zen. Egitura eta hitzak berrikusi, {subject} sektorean maiz erabiltzen den terminologiaren arabera egon dadin. Jatorrizko paragrafoen egitura mantendu behar da. Berrikusi beharreko testua: {translation}

Figure 4: Prompt for the Basque revision. Note that the {subject} and the {translation} placeholders will be replaced with their respective content in Basque.

Tekst ten został pierwotnie przetłumaczony z języka angielskiego. Należy sprawdzić jego strukturę i sformułowania pod kątem zgodności ze standardową terminologią sektora {subject}. Należy zachować strukturę oryginalnych akapitów. Tekst do sprawdzenia: {translation}

Figure 5: Prompt for the Polish revision. Note that the {subject} and the {translation} placeholders will be replaced with their respective content in Polish.

Based on the evaluation described in Section 3.1, we selected the following model configuration for

¹<https://www.deepl.com/>

the final task submissions:

- Initial translation: Qwen 3.5 27B²
- Basque revision: Latxa 70B³
- Polish revision: PLLuM 12B⁴
- English revision: none (Gemma 4 family used during evaluation only)

Even though the PLLuM 70B performed better than the 12B variant, given the time-sensitive nature of producing the final translations, we have selected the 12B variant, as it showed very similar performance with a shorter generation time. Both revision models use beam-search decoding with three beams and no sampling.

3.1 Evaluation

Since the shared task did not include a held-out test set, we evaluated our pipeline using the FLORES+ dataset (NLLB Team et al., 2024; Burchell et al., 2024; Dale et al., 2025). FLORES+ provides parallel evaluation data for all language pairs considered in the task, making it suitable for measuring general translation quality across the required directions. However, the dataset does not include task-specific terminology annotations or terminology constraints. Therefore, this evaluation does not measure the effectiveness of the terminology selection component described in Section 3. Instead, it focuses on the two generative stages of the pipeline: the initial translation step and the subsequent revision step.

For this evaluation, we employed the development split of FLORES+, with 997 source sentences per language direction. We evaluated the same translation directions required by the shared task: Spanish→Basque, English→Polish, and Traditional Chinese→English. For each language pair, the source sentence was translated by the initial translation model and passed to a second model for revision. The revised output was then compared against the FLORES+ reference translation.

Translation quality was measured using sentence-level chrF++ (Popović, 2017) and SacreBLEU (Post, 2018) averaged over sentences. We evaluated three variants from the Qwen 3.5 family for the initial translation: a small 9B variant, a 27B variant, and a 35B Mixture-of-Experts

²Qwen/Qwen3.5-27B

³HiTZ/Latxa-Llama-3.1-70B-Instruct-v2

⁴CYFRAGOVPL/PLLuM-12B-instruct-2512

		Qwen3.5-9B		Qwen3.5-27B		Qwen3.5-35B-A3B	
		chrF	SacreBLEU	chrF	SacreBLEU	chrF	SacreBLEU
		t-value (p-value)	t-value (p-value)	t-value (p-value)	t-value (p-value)	t-value (p-value)	t-value (p-value)
PL	First Translation	47.97	20.20	49.13	21.31	47.72	19.97
	Revision (Plumm 12)	48.40 t= 3.77 (1.72E-04)	20.81 t= 3.51 (4.73E-04)	49.25 t= 1.64 (1.01E-01)	21.58 t= 1.95 (5.11E-02)	48.06 3.17 (1.55E-03)	20.30 t= 2.00 (4.61E-02)
	Revision (Plumm 70)	48.82 t= 6.41 (2.24E-10)	21.30 t= 5.16 (2.98E-07)	49.66 t= 4.27 (2.17E-05)	21.97 t= 3.62 (3.04E-04)	48.71 6.55 (9.16E-11)	21.05 t= 5.24 (1.95E-07)
	Self-Revision	47.23 t= -2.31 (2.13E-02)	19.18 t= -1.74 (8.24E-02)	47.89 t= -5.08 (4.49E-07)	19.59 t= -4.04 (5.78E-05)	47.35 -0.67 (5.06E-01)	18.97 t= -1.53 (1.26E-01)
	First Translation	40.45	7.39	44.22	9.61	42.43	8.29
EU	Revision (Latxa 8B)	40.62 t= 6.93 (7.39E-12)	7.44 t= 2.00 (4.61E-02)	44.55 t= 2.93 (3.44E-03)	10.19 t= 2.00 (4.61E-02)	42.97 t= 3.69 (2.39E-04)	8.97 t= 2.00 (4.61E-02)
	Revision (Latxa 70B)	43.36 t= 15.23 (3.20E-47)	9.85 t= 10.03 (1.33E-22)	45.64 t= 10.96 (1.77E-26)	11.38 t= 9.86 (5.91E-22)	44.87 t= 12.60 (7.12E-34)	10.68 t= 8.90 (2.57E-18)
	Self-Revision	41.91 t= 8.92 (2.27E-18)	8.67 t= 5.57 (3.26E-08)	43.79 t= 1.98 (4.75E-02)	9.27 t= 2.97 (3.05E-03)	43.19 t= 5.76 (1.12E-08)	9.45 t= 5.53 (4.00E-08)
	First Translation	54.80	27.03	56.18	28.17	55.22	27.23
EN	Revision (Gemma 4 12B)	52.15 t= -13.40 (9.17E-38)	24.32 t= -9.81 (9.74E-22)	53.27 t= -15.32 (9.41E-48)	25.35 t= -11.31 (5.39E-28)	52.52 t= -12.70 (2.34E-34)	24.52 t= -9.09 (5.17E-19)
	Revision (Gemma 4 31B)	52.47 t= -11.37 (3.06E-28)	24.65 t= -8.79 (6.57E-18)	53.17 t= -14.08 (3.27E-41)	25.32 t= -10.32 (8.29E-24)	52.72 t= -11.11 (4.02E-27)	25.03 t= -7.74 (2.41E-14)
	Self-Revision	51.08 t= -16.79 (6.93E-56)	22.78 t= -13.03 (6.00E-36)	52.58 t= -17.38 (2.67E-59)	24.19 t= -13.47 (4.22E-38)	51.85 t= -14.51 (2.08E-43)	23.33 t= -11.66 (1.58E-29)
	First Translation	54.80	27.03	56.18	28.17	55.22	27.23

Table 1: Translation quality achieved by different pipeline configurations, measured by sentence-level chrF++ and SacreBLEU averaged over sentences on the FLORES+ dataset. The t-value columns contain the results from the paired-samples t-test (997 samples), along with their respective p-values. A positive t-value represents an improvement over the unrevised translation. The highest score for each language is noted in bold.

variant with 3B active parameters. For the revision step, we experimented with two model variants for each target language, yielding 6 pipeline configurations evaluated per language. In addition to the aggregated average value for each metric, we also performed a paired-samples t-test, using the unrevised translation as the baseline, to determine whether the revisions offered any benefit to the

translation. The experiments’ results are shown in Table 1.

To further investigate the contribution of specialized revision models, we also include a Self-Revision (Madaan et al., 2023) configuration in Table 1. In this setting, the revision step is performed by a Qwen 3.5 27B model rather than a language-specific model, since it consistently generated the

best initial translation. This comparison allows us to distinguish between two possible sources of improvement. The first is the revision process itself, which may improve quality simply by giving the model a second chance to revise the translation with a monolingual target-language prompt. The second is the use of specialized models, which may provide additional benefits because they were trained on or adapted to the target language. If Self-Revision improves over the initial translation, this suggests that the revision setup alone is beneficial. Otherwise, this indicates that language-specific models offer additional gains.

The results reveal several relevant patterns. First, the performance of the initial translation models suggests that the number of active parameters is more important than the total parameter count alone. This can be observed by comparing the Qwen 3.5 variants used for the initial translation step. Although the 35B A3B variant has a larger total parameter count than the other evaluated models, it performs among the lowest across the tested language pairs. In contrast, the Qwen 3.5 27B model consistently achieves stronger results across the initial translation configurations. This suggests that, for this task, a dense model with more active parameters may be preferable to a larger mixture-of-experts model with fewer active parameters during inference.

Second, the revision step improves translation quality for the two lower-resource target languages when performed by specialized language models. For Spanish→Basque, the Latxa-based revision model improves the output generated by the initial Qwen translation model. Similarly, for English→Polish, the PLLuM-based revision model also increases translation quality, but to a smaller degree. It is interesting to note that for these two language pairs, both model sizes positively affected the initial translation, although the larger models showed greater improvement. This suggests that even smaller, target-language-specialized models can complement larger general-purpose models to some extent, particularly when the target language benefits from dedicated linguistic knowledge or language-specific adaptation.

The opposite trend is observed in Traditional Chinese→English translation. In this direction, revision with smaller models consistently reduces translation quality. This is likely because English is already well represented in the pretraining and instruction-tuning data for most general-purpose

LLMs. As a result, the smaller revision models do not provide additional target-language expertise, introducing unnecessary edits that degrade the translation. Based on this observation, we removed the revision step for the Chinese→English direction in the final task submissions.

3.2 Manual evaluation on the Shared Task dataset

Since there is no evaluation split in the shared task dataset, we evaluated quality through manual inspection of samples from the generated translations (2 from the Chinese→English, 10 from Spanish→Basque, and 6 from the English→Polish) and the validation script provided by the organizers, which only validates paragraph structure. Even through this simple evaluation method, we gathered some interesting insights.

Starting with the English translations, we observed that these had the fewest issues. The only issue detected was paragraph structure when compared with the reference. As the original Chinese texts employed no line breaks and the translated text included some paragraphs to follow more standard English newspaper article conventions, the validation identified a misalignment of paragraphs, which were easily corrected by manually excluding line breaks. Content-wise, while we lack the required expertise to evaluate financial terminology, all sampled translations seemed coherent and on topic.

In the Polish translations, we noticed greater divergences from the reference texts, particularly in paragraph structure and notation styles. The first major difference we observed in the translation process was a tendency to standardize the usage of line breaks ('`\n`'). While the reference texts employed a mixture of single and double '`\n`' to differentiate between new lines and paragraphs, the translated text used a single variant: some translations used only single '`\n`', while others used only double '`\n`'. Another noticeable change introduced by the translation process concerns paragraph headers. The reference texts appear to be book chapters or report sections, particularly those used for translations in the Medicine domain. The translation process tended to include '`***`' markers before the numbers denoting new chapters, report sections, or document titles, whereas the reference text used only numbers. While these divergences do not cause any semantic differences in the generated translation, they could negatively affect n-gram-based auto-

matic metrics. These alterations may result from language models attempting to conform the translated text to the standards used in certain markup languages. An example of these changes is illustrated in Figure 6.

CLINICAL GUIDELINES AND PROTOCOLS IN MULTIDISCIPLINARY DERMATOLOGY AND AESTHETIC SURGERY Document Type: Comprehensive Institutional Review and Clinical Standard Operating Procedure Department: Division of Dermatology [...] **KLINICZNE WYTYCZNE I PROTOKOŁY W WIELEKIERUNKOWEJ DERMATOLOGII I CHIRURGII ESTETYCZNEJ** **Rodzaj dokumentu:**Kompleksowa ocena instytucjonalna i kliniczna standardowa procedura operacyjna \n**Wydział:**Podział Dermatologii [...]
--

Figure 6: Example of markup notation included in the Polish translation for one of the entries from the track 1 dataset from the shared task. The top sentence fragment is from the English reference text, and the bottom is the generated Polish translation.

Finally, the Basque translation offered the biggest challenge for our pipeline. Unlike the previous languages, where line-break behavior differed from the reference text but was consistent, the behavior in the Basque translations was less predictable. We observed instances in which the translations replaced commas with line breaks, introduced semicolons to split long sentences where none existed in the reference, and made other structural changes. Besides these structural changes, we observed more concerning alterations, specifically with generation length. While we are not able to assess the fluency or grammatical quality of the Basque text, we identified several instances in which the translated texts were significantly longer than the original reference, more than could reasonably be attributed to grammatical differences between Spanish and Basque. This leads us to conclude that the models tended to expand sections of the text to include new information, which can be particularly problematic in translation. Interestingly, these anomalous behaviors were mostly observed in translations from track 1, with a single instance from track 2 (out of 5 inspected) showing a significant generation error: a word-repetition loop at the tail end of the text. These observations of error generation led us to exclude the Basque translations from our track 1 submission.

System	chrF++	Term success (lemma)
Agentic-OPUS	77.1	95.4
UniOR-Mixed	74.5	70.8
GRIAL-TA	71.0	57.3
STaR-MT (ours)	68.6	44.9

Table 2: Track 1 results for the English-Polish language pair of the STaR-MT system, compared with similarly ranked systems and the top-performing one.

3.3 Shared Task overall performance

After the Shared Task submission concluded, the organizer provided us with the official results for all submitted systems. As mentioned before, we have not submitted translations for the Spanish→Basque language pair for Track 1, so we report results only for the English→Polish pair, as shown in Table 2. The Track 2 results, which include evaluations for all three language pairs, are shown in Table 3. In both tables, we have also included the results for the track’s top-performing system for comparison.

When comparing only automatic quality metrics, our system performed in line with other submissions. While the final quality is lower than that of the top performer, it does not deviate from the average observed in the submitted systems, reported in Charkiewicz et al. (2026). On the other hand, our system performed poorly when evaluated using the term success metric.

Comparing the results from Tracks 1 and 2, it is possible to see that the term success rate for the English→Polish pair dropped. This drop in performance is expected, as it is easier for LLMs to leverage explicit dictionaries than for them to generalize from translation examples. Interestingly, performance across languages in Track 2 varies widely. While we have achieved less than 50% term success in English→Polish and Spanish→Basque pairs, term success surpassed 80% for the Chinese→English pair. While the root cause of this change in behavior is unclear, it may result from differences in instruction-following capabilities when targeting specific languages or from errors introduced during the revision process for lower-resource languages. In either case, this significant change in performance warrants further exploration.

System	chrF++				Term Success (lemma)			
	Avg	enpl	eseu	zhen	Avg	enpl	eseu	zhen
CUNI-UFAL	78.1	81.3	76.9	76.1	89.2	90.4	88.5	88.7
VNFusion	73.8	75.1	76.3	70.1	73.4	66.4	75.8	78.0
STaR-MT (ours)	67.3	68.1	63.6	70.1	53.1	42.8	35.9	80.6
STRA-MT	65.9	67.3	64.8	65.5	48.0	38.7	40.3	64.9

Table 3: Track 2 results for all language pairs of the STaR-MT system, compared with similarly ranked systems and the top-performing one.

4 Conclusion

The performance of LLMs on machine translation tasks has improved significantly in recent years (Lyu et al., 2024), but they still struggle in some settings. This is particularly the case for terminology-aware translation, where terminological consistency across long documents is critical in technical and scientific settings.

In this paper, we present our submission to the WMT 26 Terminology shared task. Our methodology adopts a lightweight agent-based design, using a three-step pipeline composed of selection, translation, and revision. The selection step retrieves relevant entries or examples. The translation step uses a general-purpose LLM with a large context window to produce the initial output from a terminology-enriched prompt. Finally, the revision step uses language-specific models to improve translation quality for selected target languages.

Overall, our results from testing the pipeline on the FLORES+ dataset suggest that machine translation pipelines can benefit from combining large, general-purpose translation models with language-specific revision models, at least for some language pairs. The large model provides strong multilingual translation capabilities and can handle long prompts, while the specialized model acts as a targeted post-editor for languages where dedicated adaptation improves fluency, adequacy, or morphological correctness. It is interesting to note that although employing larger language-specific models produces better revisions, the general-purpose model can still benefit from revisions made by smaller specialist models, albeit to a reduced degree. This effect is not universal, as shown by the Chinese→English results, but it appears to be beneficial for lower-resource or morphologically richer target languages, such as Basque and Polish.

Conversely, this pipeline style has some drawbacks. Since there is no specific step to validate the intermediate generation, errors can easily prop-

agate downstream. Some of these error propagations were observed during the generation of the Basque translations for track 1 of the shared task. Moreover, the term success rate was significantly lower than that achieved by other participating systems. While our system performed well for the Chinese→English pair, its performance dropped considerably for the other two. Future work will focus on exploring alternative pipeline architectures to mitigate the identified errors and investigating approaches to reduce the performance gap between language pairs.

Limitations

Due to hardware limitations, we did not evaluate the pipeline with models larger than 70B, as even with 4-bit quantization, they exceeded our memory budget. We expect that employing the larger Qwen3.5 122B A10B and 397B A17B in the initial translation could significantly improve the pipeline’s final translation quality, or alter the impact of employing specialized models.

Acknowledgments

This research was financially supported by the project NORTE IA+ | Transparent and Secure Artificial Intelligence for a More Human and Competitive Future (operation code NORTE2030-FEDER-02724200, <https://www.norteia.pt>), funded by NORTE 2030/FEDER. This research was also financially supported by UID/00027/2025 – Artificial Intelligence and Computer Science Laboratory (LIACC) with DOI <https://doi.org/10.54499/UID/00027/2025> and UID/00022/2025 – Center for Linguistics of the University of Porto (CLUP), DOI <https://doi.org/10.54499/UID/00022/2025>, funded by Fundação para a Ciência e a Tecnologia, I.P./ MECI through national funds.

References

- Felermينو D. M. A. Ali, Henrique Lopes Cardoso, and Rui Sousa-Silva. 2025. [Leveraging Loanword Constraints for Improving Machine Translation in a Low-Resource Multilingual Context](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27643–27657, Suzhou, China. Association for Computational Linguistics.
- Duarte Alves, Nuno Guerreiro, João Alves, José Pomal, Ricardo Rei, José de Souza, Pierre Colombo, and Andre Martins. 2023. [Steering Large Language Models for Machine Translation with Finetuning and In-Context Learning](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11127–11148, Singapore. Association for Computational Linguistics.
- Daman Arora, Himanshu Singh, and Mausam. 2023. [Have LLMs Advanced Enough? A Challenging Problem Solving Benchmark For Large Language Models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7527–7543, Singapore. Association for Computational Linguistics.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. [Language models are few-shot learners](#). In *Advances in neural information processing systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Laurie Burchell, Jean Maillard, Antonios Anastasopoulos, Christian Federmann, Philipp Koehn, and Skyler Wang. 2024. [Findings of the WMT 2024 shared task of the open language data initiative](#). In *Proceedings of the ninth conference on machine translation*, pages 110–117, Miami, Florida, USA. Association for Computational Linguistics.
- Teresa Cabré. 1999. *Terminology*. John Benjamins Publishing Company.
- Adrian Charkiewicz, Pinzhen Chen, Thierry Etchegoyhen, Harritxu Gete Ugarte, Kamil Guttman, Xu Huang, David Ponce, Artur Nowakowski, Frédéric Odeh, Arturo Oncevay, Kirill Semenov, Dawei Zhu, and Vilém Zouhar. 2026. Findings of the WMT26 terminology translation task. In *To appear in Proceedings of the Eleventh Conference on Machine Translation*. Association for Computational Linguistics.
- David Dale, Laurie Burchell, Jean Maillard, Idris Abdumumin, Antonios Anastasopoulos, Isaac Caswell, and Philipp Koehn. 2025. [Findings of the WMT 2025 shared task of the open language data initiative](#). In *Proceedings of the tenth conference on machine translation*, pages 495–502, Suzhou, China. Association for Computational Linguistics.
- Chen Gao, Xiaochong Lan, Nian Li, Yuan Yuan, Jingtao Ding, Zhilun Zhou, Fengli Xu, and Yong Li. 2024. [Large language models empowered agent-based modeling and simulation: a survey and perspectives](#). *Humanities and Social Sciences Communications*, 11(1):1259.
- Jan Kocoń, Maciej Piasecki, Arkadiusz Janz, Teddy Ferdinan, Łukasz Radliński, Bartłomiej Koptyra, Marcin Oleksy, Stanisław Woźniak, Paweł Walkowiak, Konrad Wojtasik, Julia Moska, Tomasz Naskręt, Bartosz Walkowiak, Mateusz Gniewkowski, Kamil Szyk, Dawid Motyka, Dawid Banach, Jonatan Dalasiński, Ewa Rudnicka, and 80 others. 2025. [PLLuM: a family of polish large language models](#). *arXiv preprint arXiv:2511.03823*. ArXiv: 2511.03823 [cs.CL].
- Chenyang Lyu, Zefeng Du, Jitao Xu, Yitao Duan, Minghao Wu, Teresa Lynn, Alham Fikri Aji, Derek F. Wong, and Longyue Wang. 2024. [A Paradigm Shift: The Future of Machine Translation Lies with Large Language Models](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1339–1352, Torino, Italia. ELRA and ICCL.
- Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. SELF-REFINE: iterative refinement with self-feedback. In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS ’23*, pages 46534–46594, Red Hook, NY, USA. Curran Associates Inc.
- NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Hefernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2024. [Scaling neural machine translation to 200 languages](#). *Nature*, 630(8018):841–846.
- Maja Popović. 2017. [chrF++: words helping character n-grams](#). In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.
- Matt Post. 2018. [A Call for Clarity in Reporting BLEU Scores](#). In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.
- Qwen Team. 2026. [Qwen3.5: Towards native multi-modal agents](#).

- Oscar Sainz, Naiara Perez, Julen Etxaniz, Joseba Fernandez de Landa, Itziar Aldabe, Iker García-Ferrero, Aimar Zabala, Ekhi Azurmendi, German Rigau, Eneko Agirre, Mikel Artetxe, and Aitor Soroa. 2025. [Instructing large language models for low-resource languages: a systematic study for Basque](#). In *Proceedings of the 2025 conference on empirical methods in natural language processing*, pages 29136–29160, Suzhou, China. Association for Computational Linguistics.
- Kirill Semenov, Xu Huang, Vilém Zouhar, Nathaniel Berger, Dawei Zhu, Arturo Oncevay, and Pinzhen Chen. 2025. [Findings of the WMT25 Terminology Translation Task: Terminology is Useful Especially for Good MTs](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 554–576, Suzhou, China. Association for Computational Linguistics.
- Minghao Wu, Jiahao Xu, and Longyue Wang. 2024. [TransAgents: Build Your Translation Company with Language Agents](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 131–141, Miami, Florida, USA. Association for Computational Linguistics.
- Masaru Yamada. 2023. [Optimizing Machine Translation through Prompt Engineering: An Investigation into ChatGPT’s Customizability](#). In *Proceedings of Machine Translation Summit XIX, Vol. 2: Users Track*, pages 195–204, Macau SAR, China. Asia-Pacific Association for Machine Translation.