IRB-MT at WMT25 Terminology Translation Task: Metric-guided Multi-agent Approach

Ivan Grubišić*
Division of Electronics
Ruđer Bošković Institute
Zagreb, Croatia
ivan.grubisic@irb.hr

Damir Korenčić*
Division of Electronics
Ruđer Bošković Institute
Zagreb, Croatia
damir.korencic@irb.hr

Abstract

Terminology-aware machine translation (MT) is needed in case of specialized domains such as science and law. Large Language Models (LLMs) have raised the level of state-of-the-art performance on the task of MT, but the problem is not completely solved, especially for use-cases requiring precise terminology translations. We participate in the WMT25 Terminology Translation Task with an LLM-based multi-agent system coupled with a custom terminology-aware translation quality metric for the selection of the final translation. We use a number of smaller open-weights LLMs embedded in an agentic "translation revision" workflow, and we do not rely on data- and compute-intensive fine-tuning of models. Our evaluations show that the system achieves very good results in terms of both `MetricX`-24 and a custom `TSR` metric designed to measure the adherence to predefined term mappings.

1 Introduction

When translating texts from specialized technical domains such as medicine, finance, or law, it is important to translate technical terms accurately and consistently (Castilho and Knowles, 2025; Oncevay et al., 2025). To this end, the translation systems can be provided with an existing list of terms and their translations. While the LLMs have emerged as state-of-the-art models for MT (Kocmi et al., 2024), they are rarely evaluated on specialized domains that require strict adherence to terminology. The goal of the WMT25 Terminology Translation Task (Semenov et al., 2025) is to determine how well do the modern MT systems tackle this challenge.

Recently, a number of capable multilingual LLMs that support instruction following were made available to the community (Team et al., 2025; Yang et al., 2025; Martins et al., 2025). In parallel, research in multi-agent workflows showed that embedding individual LLMs in multi-step workflows leads to performance gains. These workflows can be generic, such as self-refine (Madaan et al., 2023), or task-oriented workflows in which agents are assigned natural task-specific roles (Wu et al., 2024; Briakou et al., 2024).

Our goal was to propose a resource-efficient solution based on smaller instruction- and reasoning-capable multilingual LLMs with open weights, embedded in an agentic workflow for performance improvement. We hypothesize that such a workflow could lead to solid performance for a number of language pairs, as the multilinguality of modern LLMs facilitates translation, and their instruction-following capabilities enable the implementation of complex terminology- and revision-related instructions. Such a system does not require datasets and compute for model adaptation and fine-tuning – only relatively modest inference-time compute to run the agents is needed. Participation in the WMT25 Terminology Task enables us to compare such a system with various other approaches.

WMT25 Terminology Task datasets are divided into `Track1` and `Track2`, consisting of texts from the information technology and financial domains, respectively (Semenov et al., 2025). `Track1` datasets contain paragraph-level texts and cover en-de, en-es, and en-ru language pairs. `Track2` datasets contain long document with document-level terminology mappings and cover en-zh and zh-en pairs. Predefined source→translation term mappings are included in the datasets and they come in two flavors: "proper" terminologies covering technical terms, and "random" terminologies with random words. The idea is to measure the influence of the predefined terminology on the performance. For the same reason, an additional "no terminology" setup is included in the task.

Our system, named `MeGuMa`, is an agentic translation system that operates in three phases: 1) translation with individual LLMs; 2) translation revision

*Equal contribution.

Proceedings of the Tenth Conference on Machine Translation, pages 1302–1334
November 8-9, 2025

based on reasoning LLMs; 3) the selection phase using a custom terminology-aware translation quality metric. Revision agents act as senior translators who revise the work of junior translators (individual LLMs). In the revision phase, a reviser LLM examines a number of phase one translations and composes the final translation. Variants of Gemma3, Qwen3, and EuroLLM models are employed as base translators, while Gemma3 and Qwen3 models set up for "thinking" mode are used for revision. The final translation is selected from all the generated translations using a combination of two translation measures: MetricX (Juraska et al., 2024), and a custom TSR metric that measures the adherence to predefined source→translation term mappings.

The final system achieves good results. For Track1 all MetricX scores (range between 0 and 25, lower is better) are below 2.00, while the TSR scores are above 0.85 (indicating that 85% of source terms are correctly translated). For Track2 all MetricX scores are below 2.5 (below 2.0 for the "proper terms" setup), and the TSR scores range between 0.69 and 0.80 except for the random terms setup with scores between 0.40 and 0.50.

Our contributions are: a new agentic system for terminology-aware machine translation, a new TSR metric for approximating the adherence to predefined terminology, and a TTQ metric that aggregates MetricX and TSR in order to combine translation quality and the adherence to predefined terminology. We make the code and the models' output freely available.[1]

2 Datasets

Three different collections of datasets were released for the Terminology MT task: 1) DEV, a set of three datasets for sentence- and paragraph-level terminology translation task; 2) Track1, a set of three datasets for sentence- and paragraph-level terminology translation task; and 3) Track2, a set of ten test datasets for document-level terminology translation task. Each of these two tracks focus on a different domain, with Track1 dealing with the *information technology* domain, while Track2 consists of texts and terminology from the *finance* domain (Semenov et al., 2025).

The datasets in DEV and Track1 are quite similar, with the same language pairs and one unique dataset per pair: English to German, English to Spanish, and English to Russian. Track2 contains ten unique datasets – five datasets for English to Traditional Chinese translation and five datasets for Traditional Chinese to English.

Each text from DEV, Track1, and Track2 datasets is associated with a set of terms and their translations. These term mappings define the terminology-aware translation tasks. For DEV and Track1 a term dictionary is provided at the sentence/paragraph level, and it contains source_term : target_term entries. For Track2 each document is provided with a dictionary of source terms mapped to a list of viable target terms. Each Track1 and Track2 has three variants of term mappings: 1) no terms; 2) proper terms; and 3) random terms. These three variants differ in their inputs. The *no terms* variant uses only input texts, the *proper* terminology variant adds specialized dictionaries for domain-specific terms, and the *random* terminology variant uses dictionaries with randomly selected words from the texts to compare the impact of accurate versus arbitrary terminology on system performance (Semenov et al., 2025).

While the DEV and Track1 datasets contain 500 smaller texts with an average of approximately 10 whitespace-separated tokens (see the Appendix, Table 9, 10, 11 and 12), each Track2 dataset contains between 9 and 13 long texts with approximately 50 paragraphs per text, where each paragraph has approximately 50 tokens on average (see the Appendix, Table 16, 14 and 15).

To reduce the required computational resources and increase the quality of translations and evaluations, we convert Track2 translation datasets from the document-level to the paragraph-level format of Track1. This process requires two steps: text splitting and term splitting. Following a data analysis showing that Track2 texts contain paragraphs separated by two or more newline characters, we first use regular expressions to split texts into paragraphs. However, since the term mappings are provided on the document-level, another step is needed to create paragraph-level term dictionaries containing only the source terms that occur in a paragraph.

To achieve this, for each paragraph we examine all source terms in the document-level term dictionary and check if a source term is present in the paragraph text. However, we need to prevent subterms (substrings of longer terms) to be detected in places where their super-term exists. To this end

[1] https://github.com/igrubi/irb-mt-wmt2025

we employ these steps: 1) lowercase the text and all terms; 2) sort the list of all document terms by decreasing length (longest terms at the start of the list); 3) iterate through the term list and check if a term is present in the paragraph text. If a term is detected, we save it to the paragraph-level dictionary and remove all occurrences of the term from the paragraph text. This way a sub-term will not be included after a super-term. With this procedure we create a paragraph-level term dictionary that contains only a relevant subset of document-level terms.

With this procedure we only reduce the set of dictionary keys to the paragraph-level. The dictionary values remain the same as in the original `Track2` datasets, i.e., each source term is still mapped to a list of viable target terms. This is the key difference between the generated `Track2` paragraph-level term dictionaries and the `Track1` dictionaries, and we tackle these two cases differently with translation and revision agents.

Finally, in order to recreate the whole translated document later, we give each paragraph a unique identifier that contains information on the exact location of the paragraph within the document.

After the final submission we discovered that there was a bug in the function that we used for splitting Chinese terms. Namely, when string-matching a term in a text, we required the term to be surrounded by word-separating whitespace. While this is a proper approach for European languages, it is not for Traditional Chinese because often there is no such separation between words. As a result, we lost a good number of the paragraph terms for the datasets where the source language is Chinese. For details, see the Appendix, Subsection A.3, difference between Table 16 and 17.

3 The System

Our translation system produces an output in three phases: 1) individual translator LLMs generate initial translations, 2) reviser LLMs generate improved translations from the initial ones, and 3) all of the candidate translations are pooled, and the best one is selected based on a custom quality metric.

In the development phase we benchmarked a number of translator-LLM candidates on a development set consisting of a subset of pairs from both the `DEV`, `Track1` and `Track2` datasets. Most of the evaluated LLMs demonstrated good results and were included into the final system.

We start the detailed description of the complete system with metrics, used both for the benchmarking of LLM and for the selection of final system outputs.

3.1 Metrics

General translation quality. To evaluate the general (not terminology-aware) translation quality, we use `MetricX` metric (Juraska et al., 2024) (the "metricx-24-hybrid-xl-v2p6" variant). `MetricX` was chosen since its scores of translation quality are highly correlated with human judgments (Juraska et al., 2024), and it does not require a reference translation in order to assess translation quality. This makes `MetricX` applicable for assessing translation performance on `Track1` and `Track2` test datasets.

Terminology translation success rate. When performing terminology-aware translation, it is important that the terms from the source language are translated to the target language according to the provided dictionary with correct term mappings. To measure this we designed a custom metric that does not require a reference translation and is therefore usable both for benchmarking on the test set and for filtering multiple generated translations.

The metric, dubbed the Terminology Success Rate (`TSR`) metric, directly compares source-language terms present in the source text with the target-language terms present in the translation. The input to the metric consists of the source and target texts, and of the dictionary of term translations. The metric relies on lemmatizers and sentence segmenters for both source and target languages, and a sentence-level alignment module.

In the first step, source and target texts are segmented into sentences. In the second step two sets of sentences are aligned using the SentAlign method that relies on a multilingual sentence embedding module and an alignment optimization algorithm (Steingrimsson et al., 2023). The output of SentAlign is a list of pairs of matching sentence blocks – one or more source sentences can be aligned to one or more target sentences, and a sentence can be without a match. Given a pair of aligned (sub)texts, and a (source_term, target_term) pair from the dictionary such that the source term appears in the source text, a pair-level score is calculated as the percentage of matched target terms in relation to the matched source terms. This score,

capped to 100%, approximates the coverage of the source terms occurring in source text by the target terms occurring in the translation. The final score is calculated by averaging all per-term scores for each pair of aligned texts, and then averaging over all pairs in the alignment.

Matching of terms to their occurrences in texts is preceded by lowercasing and lemmatization in order to account for different surface forms that a term can assume. To avoid over-counting of terms the matching process takes into account the term overlap – longer terms are matched first and each subsequent term is matched only to a positions where it does not overlap with any previously matched term.

Terminology translation quality. Both previously described translation metrics are equally important to assess the final terminology-aware translation quality. For this reason we combine these metric into a new metric that we name Terminology Translation Quality (TTQ).

To achieve this, we convert MetricX from the 25-0 scale (where a lower score means better translation) to the 0-1 scale (where a higher score means better translation), using the following equation:

$$\tilde{\text{MetricX}} = (25 - \text{MetricX})/25$$

Then the final TTQ metric is calculated by averaging the arithmetic, geometric, and harmonic means of converted MetricX and TSR.

While the arithmetic mean is influenced by both scores equally ignoring the relation between their values, the harmonic and geometric mean progressively penalize cases with larger differences giving more weight to smaller values. Since we want the TTQ metric to capture both of these characteristics, we define it as the average of all three means. In this way, we enforce the discriminatory strength of the arithmetic mean in those cases where one score is extremely small. The TTQ metric produces a stable and fair score capable of discriminating between similar solutions.

3.2 LLM Benchmarking and Selection

Translation is performed by LLM-based agents in two phases - *translation* and *revision*. We use pretrained, relatively small (8-27 billion parameters) and open-weights LLMs. No further adaptation or fine-tuning is performed, and no additional data is used.

As candidates for the translation we considered the following LLMs: Gemma3 (12B and 27B) (Team et al., 2025), Qwen3 (8B and 14B – both thinking and non-thinking variants) (Yang et al., 2025), and EuroLLM (9B) (Martins et al., 2025). These are recent LLMs, built using state-of-the-art approaches, that have multilingual and instruction-following capabilities.

As a first step, we performed a test of basic translation capabilities on a subsample of DEV and Track2 datasets. We took 10 longest texts from each DEV dataset, and 10 longest paragraphs for each Track2 dataset, and evaluated the LLMs with both MetricX and TSR. The models and their variants tested were: gemma3_27b, gemma3_12b, qwen3_14b-think, qwen3_8b-think, qwen3_14b, qwen3_8b, and eurollm_9b.

The results, displayed in Tables 1 and 2, show that all models have competitive performance for almost all language pairs. Some of the models appear to be biased towards MetricX, like eurollm_9b, some others are more inclined to TSR, like qwen3_8b-think, while some models are balanced between these two metrics, like gemma3_27b. Only eurollm_9b shows slightly poorer translation performance from English to Chinese and vice versa in both metrics (Table 2). For more details, see the Appendix, Section D.

Based on these results we decided to use all the model variants except eurollm_9b as basic translators for both Track1 and Track2. We only use eurollm_9b as one of the translators for the European languages of Track1.

The idea of a revision phase comes from the translation-revision system, which is a successful practice performed by professional translators for years (Arthern, 1978). Although basic translation can be carried out by less experienced personnel, the revision requires an experienced translator to produce final high-quality translations of the texts. This is the reason why we speculated that the revision agent should be a larger and more capable LLM. This intuition was confirmed by the pilot experiment that we conducted on the subsample of the DEV and Track2 datasets, which showed that reviser agents based on smaller models, such as revis-qwen3_8b-think and revis-eurollm-think, produce a large number of errors.

Therefore, we decided to use as revis-

agents	ende			enes			enru		
	MetricX	TSR	TTQ	MetricX	TSR	TTQ	MetricX	TSR	TTQ
trans-eurollm	**1.17**	0.44	0.65	**3.39**	0.78	0.82	3.10	0.44	0.62
trans-qwen3_8b	1.36	0.56	0.73	3.67	0.88	0.86	4.77	0.68	0.74
trans-qwen3_8b-think	1.80	0.62	0.76	3.53	0.93	0.89	4.50	0.77	0.80
trans-qwen3_14b	1.37	0.56	0.73	3.53	0.90	0.88	3.03	0.69	0.78
trans-qwen3_14b-think	1.23	0.55	0.72	3.86	**0.95**	**0.90**	3.82	**0.85**	**0.85**
trans-gemma3_12b	1.61	0.64	0.77	3.52	**0.95**	**0.90**	3.00	0.74	0.81
trans-gemma3_27b	1.36	**0.75**	**0.84**	3.49	0.93	**0.90**	**2.66**	0.80	**0.85**

Table 1: Combined mean scores from English to German (ende), Spanish (enes), and Russian (enru), showing MetricX, TSR and TTQ scores evaluated on subset of DEV datasets.

agents	enzh			zhen		
	MetricX	TSR	TTQ	MetricX	TSR	TTQ
trans-eurollm	4.06	0.30	0.50	3.50	0.50	0.65
trans-qwen3_8b	3.82	0.40	0.58	3.22	0.51	0.67
trans-qwen3_8b-think	3.90	0.41	0.59	3.26	**0.53**	**0.68**
trans-qwen3_14b	**3.51**	0.41	0.59	3.34	0.52	0.67
trans-qwen3_14b-think	**3.51**	0.41	0.59	**3.20**	0.52	0.67
trans-gemma3_12b	4.40	**0.45**	**0.61**	3.31	0.51	0.66
trans-gemma3_27b	3.93	**0.45**	**0.61**	3.43	0.52	0.67

Table 2: Combined mean scores for Track2 subset translating English to Traditional Chinese (enzh) and Traditional Chinese to English (zhen), showing MetricX, TSR and TTQ scores evaluated on subset of Track2 datasets.

ers only agents based on the largest variants of Gemma3 and Qwen3 models that showed solid performance in all languages: revis-gemma3_27b-think, revis-gemma3_12b-think, and revis-qwen3_14b-think. Unlike the standard translation-revision system where the reviser receives only one translation, our revision agents receive as input translations from all of the basic translation agents, and produce their final translations from the entire input. To increase the reasoning power, and thus the quality of revised texts, all of the revision models were prompt-induced to first think before producing the final solution. Similar to *chain-of-thought* prompting (Wei et al., 2022).

3.3 Context Engineering

Our system is based on prompt-guided LLM agents. Therefore, the output of these agents is highly dependent on the context of the task, i.e. the way the context is compiled and formatted for the LLM.

As there are three different cases of term data in the test datasets, we created an appropriate prompt for each case: 1) Case 1, when no term data is provided; 2) Case 2, when a single translated term is given for each source term (Track1); and 3) Case 3, when multiple viable translated terms are provided for each source term (Track2). The final prompt versions are listed in the Appendices: for *translation prompts* look at B and for *revision prompts* at C.

Each prompt consists of two parts: the system prompt and the user prompt. For translation-only agents, the system prompt defines a role of the agent, explains the task, describes the key requirements of the task, and provides the term dictionary if the term data is available. The user prompt provides the text that is to be translated.

For the reviser agents the system prompt is quite similar to the translation-only one. However, it is expanded with the original source text and with all translations produced by the translation agents. The user prompt only gives a short summary of the task. An additional difference from the translation-only prompt is the "thinking command". This command is used to induce the agent to think about the given translations and about the potential enhancements prior to giving the final revised translation.

3.4 The Final Translation System

Our system is based on multi-agent approach with three steps: 1) translation; 2) revision; 3) selection. The translation step is performed by a number of agents based on open-weight LLMs that each translate an input text from a source to a target language. The revision step is performed by three LLM agents. Each reviser agent receives all the translations from the translation step and produces a revised translation. In the selection step we use the TTQ metric to evaluate all of the produced translations (from both the translation and the revision step), and select the best one as the final system output. Additionally, to enable a unified approach for both tracks, we perform a data preparation step in which we break Track2 document-level texts into paragraph-level texts corresponding to the texts of Track1. During this process document-level terminology dictionary is projected to each of the paragraphs. The details of the process are explained in Section 2.

4 Results

The final system produces high-quality translations with low MetricX scores (Tables 3 and 5) and high TSR scores (Tables 4 and 6). TSR scores are significantly higher for Track1 than for Track2– this could be caused by both the language differences (European languages vs. Chinese) and by differences in the complexity of the terminology dictionaries (single-choice vs. multi-choice terms).

pair	noterm	proper	random	mean
ende	0.36	0.89	0.76	0.67
enes	1.24	1.73	1.63	1.53
enru	0.77	1.38	1.23	1.13

Table 3: MetricX scores of final solution for Track1.

pair	proper	random	mean
ende	0.86	0.87	0.87
enes	0.88	0.88	0.88
enru	0.94	0.93	0.94

Table 4: TSR scores of final solution for Track1.

When compared to the translation produced by all of the LLM agents used (both translators and revisers) the final system's translations have the best MetricX and TSR scores for almost all Track1 and Track2 datasets. The only exceptions are: 1) translations from English to German for the "proper term" case where a standalone trans-eurollm has the best MetricX score (Track1, see the Appendix, Table 23); and 2) translations from English to Traditional Chinese for the "random term" case where trans-qwen3_14b has the best MetricX scores (Track2, see the Appendix, Table 30).

The final system uses a metric-guided approach based on the custom TTQ metric used to select the best output from a pool of translations. The selection step is an important part of the proposed translation pipeline since it significantly increases translation quality scores, raising them above the scores of the best individual translators and the scores of the best reviser agents (see the Appendix, Section E). The reason why the selector chooses the best translation from the pool of translations of all agents and not only the reviser agents is shown in Tables 7 and 8. These tables show the frequency with which an agent's output was chosen for the final solution. Although the reviser agents have a higher chance of being chosen by the selector (17.13% vs. 6.89% in Track1, and 14.3% vs. 9.42% in Track2), in total about half of all samples are selected from individual translator agents (48.2% in Track1 and 56.5% in Track2).

In addition, there is an interesting trend in the results. MetricX scores are increasing (i.e. translation quality drops) for random and proper term cases compared to the no-term case (Tables 3 and 5; for more information, see the Appendix, Subsections E.1.1, E.2.1, E.2.2 and E.2.3). We hypothesize that forcing a predefined translation of a set of terms reduces the general translation quality because translators need to balance between two different objectives: 1) general translation and 2) terminology constraint. However, this hypothesis requires more comprehensive research to be proven.

Upon submission, our system was evaluated[2] and compared to other participating systems (20 systems in Track1 and 4 systems in Track2) (Semenov et al., 2025). ChrF2++ was used to measure translation quality and a custom terminology success rate metric (which we label Term-Acc) was used to measure adherence to the predefined terminology (Semenov et al., 2025).

[2] https://github.com/wmt-conference/wmt25-terminology/

agents	2015	2017	2019	2021	2023	mean
enzh.noterm	1.49	1.40	1.37	1.41	1.46	1.43
enzh.random	2.39	2.28	2.26	2.23	2.22	2.28
enzh.proper	1.88	1.70	1.73	1.77	1.84	1.78
zhen.noterm	1.22	1.17	1.12	1.14	1.14	1.16
zhen.random	1.54	1.58	1.53	1.46	1.52	1.53
zhen.proper	1.59	1.57	1.47	1.58	1.53	1.55

Table 5: `MetricX` scores of final solution for `Track2`.

agents	2015	2017	2019	2021	2023	mean
enzh.random	0.75	0.70	0.72	0.73	0.75	0.73
enzh.proper	0.78	0.77	0.78	0.81	0.79	0.79
zhen.random	0.39	0.48	0.48	0.48	0.46	0.46
zhen.proper	0.69	0.74	0.75	0.80	0.75	0.75

Table 6: `TSR` scores of final solution for `Track2`.

agents	ende	enes	enru	mean
`trans-eurollm`	9.7%	12.4%	11.0%	11.0%
`trans-qwen3_8b`	5.4%	5.1%	4.9%	5.1%
`trans-qwen3_8b-think`	5.2%	6.3%	6.2%	5.9%
`trans-qwen3_14b`	6.2%	7.0%	6.0%	6.4%
`trans-qwen3_14b-think`	4.8%	4.0%	4.0%	4.3%
`trans-gemma3_12b`	8.6%	6.5%	9.8%	8.3%
`trans-gemma3_27b`	6.8%	7.3%	7.4%	7.2%
`revis-qwen3_14b-think`	11.5%	10.0%	10.6%	10.7%
`revis-gemma3_12b-think`	**24.8%**	**25.3%**	**24.8%**	**24.9%**
`revis-gemma3_27b-think`	16.6%	15.8%	15.1%	15.8%

Table 7: Aggregated total frequencies of agents selected for the final solution across English to German, Spanish, and Russian translations (`Track1`).

agents	enzh-n	enzh-p	enzh-r	zhen-n	zhen-p	zhen-r	mean
`trans-qwen3_8b`	7.4%	6.0%	8.6%	8.6%	9.7%	8.0%	8.1%
`trans-qwen3_8b-think`	6.1%	6.4%	7.3%	7.1%	6.0%	7.3%	6.7%
`trans-qwen3_14b`	6.8%	6.8%	7.0%	9.4%	11.3%	10.7%	8.7%
`trans-qwen3_14b-think`	8.7%	8.2%	7.1%	9.5%	8.5%	8.0%	8.3%
`trans-gemma3_12b`	11.6%	14.4%	15.8%	15.2%	9.8%	12.6%	13.2%
`trans-gemma3_27b`	10.6%	11.7%	15.4%	9.7%	10.5%	11.4%	11.5%
`revis-qwen3_14b-think`	12.6%	11.7%	8.2%	**15.8%**	**17.4%**	**16.0%**	13.6%
`revis-gemma3_12b-think`	**17.3%**	**20.0%**	**16.1%**	13.5%	14.8%	12.6%	**15.7%**
`revis-gemma3_27b-think`	18.4%	14.3%	14.1%	10.7%	11.3%	13.0%	13.6%

Table 8: Aggregated total frequencies of agents selected for final solutions across different translation directions and term conditions (`Track2`).

In Track1, our system has an average ChrF2++ of 67.2 (6th place) and a high average Term-Acc of 97.4 (4th place). In terms of Pareto optimality between ChrF2++ and Term-Acc, our system is near-optimal, with only two systems having Pareto dominance over it: o3-term-guide and duterm.

In Track2, our system has an average ChrF2++ of 54.3 (3rd place) and a competitive average Term-Acc of 79.5 (2nd place), with only one system, CommandA_WMT, having Pareto dominance over it.

In our final submission there is an error for the proper and random cases of Track2 translations from Traditional Chinese to English. The error originates in the data preparation process (more information can be found at the end of Subsection 2). We estimated the drop in scores as the result of this error (see the Appendix, Subsections E.2.4 and E.2.8). The expected drop is around 0.30 MetricX scores and around 0.03 and 0.11 TSR scores for random and proper term cases.

Expectedly, this error effects the official results of our system: there is a large gap between the Term-Acc score for en-zh (96.6) and the Term-Acc score for zh-en (62.4) (Semenov et al., 2025). To examine the impact of the error, we evaluated[3] the debugged version of the system and obtained the zh-en Term-Acc of 96.6 (see E.4 for more details). Our repository[4] contains details of the error, reproducibility instructions, and the outputs of both error-containing and error-free versions of the system.

Conclusion and Future Work

Our system uses a multi-agent approach with three steps: 1) translation; 2) revision; 3) selection. The idea of a revision comes from the translation-revision system, which is a successful practice carried out by professional translators for years (Arthern, 1978). Here, we expand this practice that uses one human translator and one human reviser to multiple LLM translators and multiple LLM revisers. In addition, we use metric-guided selection as the final step of our system workflow. Here, we evaluate each translation with the custom TTQ metric and select the best one as the final output. Finally, the final solution yields the best performance compared to all translations and revisions for all language pairs and terminology constraint cases.

Internal evaluations show that the proposed system has very good translation scores in terms of both metrics: MetricX (which measures the general translation quality) and TSR (which measures the terminology success rate).

The contributions of this work are a novel agentic approach for terminology-aware machine translation, a novel TSR metric for measuring the adherence to the predefined term translations, and a novel TTQ metric that aggregates MetricX and TSR in a score that combines translation quality and correctness of terminology translation.

Future work will focus on expanding the evaluation to more language pairs in order to test the robustness of the system. Additionally, evaluation on translation datasets that contain reference translations would expectedly provide a more precise assessment of the system's performance.

Qualitative analysis of system outputs, based on evaluations by humans or by the top LLM systems, could lead to valuable insights and improvements. Measuring how much the system improves the productivity of translators in a real-world setting would also be valuable. To enable a more fine-grained analysis of the we make available the outputs of all the constituent LLMs.[4]

The TSR metric should be evaluated against human quality scores on a diverse set of languages pairs in order to verify its quality and robustness. The assessment of the quality and reliability of the TTQ metric also requires validation against human annotations.

Finally, we plan to improve our system with a more granular agentic workflow that incorporates additional specialized roles like pre-editor and post-editor. The key challenge of such improvements is boosting performance while reducing the execution time, influenced by both the number of workflow steps and by the size of the LLMs used.

Acknowledgements

This paper was supported by the European Union's NextGenerationEU program. We would like to thank Tomislav Šmuc, Ph.D., and Prof. Sonja Grgić, Ph.D., for support and valuable discussions. We acknowledge EuroHPC Joint Undertaking for awarding us access to MareNostrum5 hosted by BSC, Spain, under the project ID EHPC-DEV-2025D05-087.

[3] https://github.com/wmt-conference/wmt25-terminology/

[4] https://github.com/igrubi/irb-mt-wmt2025

References

Peter J. Arthern. 1978. Machine translation and computerised terminology systems - a translator's viewpoint. In *Translating and the Computer*, London, UK. Aslib Proceedings.

Eleftheria Briakou, Jiaming Luo, Colin Cherry, and Markus Freitag. 2024. Translating step-by-step: Decomposing the translation process for improved translation quality of long-form texts. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1301–1317, Miami, Florida, USA. Association for Computational Linguistics.

Sheila Castilho and Rebecca Knowles. 2025. A survey of context in neural machine translation and its evaluation. *Natural Language Processing*, 31(4):986–1016.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. Metricx-24: The google submission to the wmt 2024 metrics shared task.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinthór Steingrímsson, and Vilém Zouhar. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, 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 *Advances in Neural Information Processing Systems*, volume 36, pages 46534–46594. Curran Associates, Inc.

Pedro Henrique Martins, João Alves, Patrick Fernandes, Nuno M. Guerreiro, Ricardo Rei, Amin Farajian, Mateusz Klimaszewski, Duarte M. Alves, José Pombal, Nicolas Boizard, Manuel Faysse, Pierre Colombo, François Yvon, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2025. Eurollm-9b: Technical report.

Arturo Oncevay, Charese Smiley, and Xiaomo Liu. 2025. The impact of domain-specific terminology on machine translation for finance in European languages. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2758–2775, Albuquerque, New Mexico. 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*, Suzhou, China. Association for Computational Linguistics.

Steinthor Steingrimsson, Hrafn Loftsson, and Andy Way. 2023. SentAlign: Accurate and scalable sentence alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 256–263, Singapore. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy

Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025. Gemma 3 technical report.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

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.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report.

A Data analysis

A.1 DEV

DATASET	TEXTS
ende	500
enes	500
enru	500

Table 9: Number of texts in DEV. datasets

DATASET	MIN	MEAN	MAX	TOTAL
ende	2	9.98	49	4991
enes	3	10.79	46	5397
enru	2	9.43	50	4716
ALL	2	10.07	50	15104

Table 10: Number of tokens in texts in DEV datasets.

A.2 Track1

DATASET	TEXTS
ende	500
enes	500
enru	500

Table 11: Number of texts in Track1 datasets.

DATASET	MIN	MEAN	MAX	TOTAL
ende.noterm	2	10.40	50	5199
enes.noterm	3	10.60	41	5298
enru.noterm	3	9.02	44	4509
ALL	2	10.00	50	15006

Table 12: Number of tokens in texts in Track1 datasets.

A.3 Track2

DATASET	DOCUMENTS
2015.enzh	9
2016.zhen	10
2017.enzh	10
2018.zhen	10
2019.enzh	11
2020.zhen	11
2021.enzh	12
2022.zhen	12
2023.enzh	13
2024.zhen	13

Table 13: Number of documents in Track2 datasets.

DATASET	MIN	MEAN	MAX	TOTAL
2015.enzh	18	38.11	68	343
2016.zhen	17	37.00	63	370
2017.enzh	19	43.10	119	431
2018.zhen	19	41.60	91	416
2019.enzh	24	40.27	71	443
2020.zhen	21	46.64	121	513
2021.enzh	28	47.08	119	565
2022.zhen	23	43.25	111	519
2023.enzh	14	41.62	107	541
2024.zhen	15	44.62	91	580
ALL	14	42.53	121	4721

Table 14: Number of paragraphs per document in Track2 datasets.

DATASET	MIN	MEAN	MAX	TOTAL
2015.enzh	1	43.00	232	14748
2016.zhen	2	76.21	559	28199
2017.enzh	1	43.26	366	18644
2018.zhen	2	72.33	521	30089
2019.enzh	1	44.37	301	19654
2020.zhen	2	67.22	559	34485
2021.enzh	1	39.51	340	22323
2022.zhen	2	69.03	559	35827
2023.enzh	1	42.42	340	22948
2024.zhen	2	64.67	374	37510
ALL	1	56.01	559	264427

Table 15: Number of tokens per paragraph in Track2 datasets.

DATASET	MIN	MEAN	MAX	TOTAL
2015.enzh.proper	0	3.86	33	1324
2015.enzh.random	0	8.56	52	2936
2016.zhen.proper	0	3.20	22	1182
2016.zhen.random	1	21.94	114	8118
2017.enzh.proper	0	3.27	18	1409
2017.enzh.random	0	8.65	59	3727
2018.zhen.proper	0	3.77	21	1569
2018.zhen.random	1	22.6	122	9411
2019.enzh.proper	0	3.01	19	1333
2019.enzh.random	0	9.71	62	4302
2020.zhen.proper	0	3.22	20	1650
2020.zhen.random	0	20.02	92	10269
2021.enzh.proper	0	3.63	16	2050
2021.enzh.random	0	8.14	52	4601
2022.zhen.proper	0	3.18	17	1648
2022.zhen.random	0	20.15	83	10458
2023.enzh.proper	0	3.64	18	1970
2023.enzh.random	0	8.02	49	4340
2024.zhen.proper	0	3.23	16	1871
2024.zhen.random	1	19.91	99	11552
TOTAL	0	9.09	122	85720
ALL	0	3.26	62	30787

Table 16: Number of terms per paragraph in `Track2` datasets.

DATASET	MIN	MEAN	MAX	TOTAL
2016.zhen.proper	0	0.36	6	134
2016.zhen.random	0	1.01	10	372
2018.zhen.proper	0	0.51	7	214
2018.zhen.random	0	0.68	11	283
2020.zhen.proper	0	0.43	6	220
2020.zhen.random	0	0.64	8	330
2022.zhen.proper	0	0.58	7	300
2022.zhen.random	0	0.54	6	281
2024.zhen.proper	0	0.52	11	301
2024.zhen.random	0	0.62	8	360
ALL	0	3.26	62	30787

Table 17: Number of terms per paragraph in `Track2` datasets affected by data preparation bug.

B Translation prompts

B.1 Case 1 - no terms prompt

B.1.1 System prompt:

You are a professional translator specializing in **{source_language}** to **{target_language}**.

Your task is to translate the provided **{source_language}** text into fluent and natural **{target_language}**.

Key requirements:

- Accurately convey the meaning and nuances of the original text, respecting **{target_language}** grammar, vocabulary, and cultural norms.
- Provide only the full **{target_language}** translation as output. Do not include any explanations, comments, or additional text.

B.1.2 User prompt:

Translate the following **{source_language}** text into **{target_language}**:
{text}

B.2 Case 2 - single-choice terms prompt (Track1)

B.2.1 System prompt:

You are a professional translator specializing in **{source_language}** to **{target_language}**.

Your task is to translate the provided **{source_language}** text into fluent and natural **{target_language}**.

Key requirements:

- Accurately convey the meaning and nuances of the original text, respecting the grammar, vocabulary, and cultural norms of **{target_language}**.
- Whenever a **{source_language}** term matches an entry in the dictionary below, replace it with the exact **{target_language}** translation from the dictionary.
- Translate all other text normally, without altering any words not found in the dictionary.
- Provide only the full translation in **{target_language}** as output. Do not include any explanations, comments, or additional text.

Dictionary:
{terms}

B.2.2 User prompt:

Translate the following **{source_language}** text into **{target_language}**:
{text}

B.3 Case 3 - multi-choice terms prompt (Track2)

B.3.1 System prompt:

You are a professional translator specializing in **{source_language}** to **{target_language}**.

Your task is to translate the provided **{source_language}** text into fluent and natural **{target_language}**.

Key requirements:

- Accurately convey the meaning and nuances of the original text, respecting the grammar, vocabulary, and cultural norms of **{target_language}**.
- For any term in the **{source_language}** text that matches a key in the provided dictionary, use exactly one translation from that term's list (choose the best fitting translation in context).
- Translate all other text normally, without altering any words not found in the dictionary.
- Provide only the full translation in **{target_language}** as output. Do not include any explanations, comments, or additional text.

Dictionary:
{terms}

B.3.2 User prompt:

Translate the following **{source_language}** text into **{target_language}**:
{text}

C Revision prompts

C.1 Case 1 - no terms prompt

C.1.1 System prompt:

You are a professional senior translator specializing in **{source_language}** to **{target_language}**.

You will be given an original text in **{source_language}** followed by several translations into **{target_language}** produced by junior translators.

Your first task: Review the provided translations with these requirements:

- Critically evaluate each translation, noting strengths and weaknesses.
- Focus your observations on translation quality, fluency, grammar, vocabulary, and cultural appropriateness.
- After your review, reason about potential improvements and how to produce the best possible translation.
- Keep your review and reasoning succinct (under 1000 words).
- Enclose your review and reasoning within the **<think>** and **</think>** tags.

Your second task: Translate the original **{source_language}** text into fluent, natural **{target_language}**, following these guidelines:

- Complete this task only after the first task.
- Produce the best possible translation based on your previous reasoning.
- Accurately convey the meaning and nuance of the original, respecting **{target_language}** grammar, vocabulary, and cultural norms.
- Provide only the final translation as output, without explanations or comments.

Original **{source_language}** text:
{text}

Translations by junior translators:

1. Translation by the first junior translator:
   {translations[0]}

2. Translation by the second junior translator:
   {translations[1]}

3. Translation by the third junior translator:
   {translations[2]}

4. Translation by the fourth junior translator:
   {translations[3]}

5. Translation by the fifth junior translator:
   {translations[4]}

6. Translation by the sixth junior translator:
   {translations[5]}

7. Translation by the seventh junior translator:
   {translations[6]}

C.1.2 User prompt:

First, review these translations and reason about producing the best possible translation, enclosing your review in **<think>** and **</think>**.
Then, provide your improved translation of the original **{source_language}** text into **{target_language}**.

C.2 Case 2 - single-choice terms prompt (Track1)

C.2.1 System prompt:

You are a professional senior translator specializing in **{source_language}** to **{target_language}**.

You will be given an original text in **{source_language}** along with a dictionary of terms that must be translated exactly, followed by several translations into **{target_language}** produced by junior translators.

Your first task: Review the provided translations with these requirements:

- Critically evaluate each translation, noting strengths and weaknesses.

- Focus your observations on translation quality, fluency, grammar, vocabulary, and cultural appropriateness.
- Verify that all **{source_language}** terms matching keys in the dictionary below are correctly translated using the **exact {target_language}** equivalents provided.
- After your review, reason about potential improvements and how to produce the best possible translation.
- Keep your review and reasoning succinct (under 1000 words).
- Enclose your review and reasoning within the **<think>** and **</think>** tags.

Your second task: Translate the original **{source_language}** text into fluent, natural **{target_language}**, following these guidelines:

- Complete this task only after the first task.
- Produce the best possible translation based on your previous reasoning.
- Accurately convey the meaning and nuance of the original **{source_language}** text, respecting **{target_language}** grammar, vocabulary, and cultural norms.
- Replace **every** term in the **{source_language}** text found as a key in the dictionary below with its **exact {target_language}** translation from the dictionary.
- Translate all other text normally, without altering words not found in the dictionary.
- Provide only the final translation as output, without explanations or comments.

Original **{source_language}** text:
{text}

Dictionary:
{terms}

Translations by junior translators:

1. Translation by the first junior translator:
   {translations[0]}
2. Translation by the second junior translator:
   {translations[1]}
3. Translation by the third junior translator:
   {translations[2]}
4. Translation by the fourth junior translator:
   {translations[3]}
5. Translation by the fifth junior translator:
   {translations[4]}

6. Translation by the sixth junior translator:
   {translations[5]}

7. Translation by the seventh junior translator:
   {translations[6]}

C.2.2 User prompt:

First, review these translations and reason about producing the best possible translation, enclosing your review in **<think>** and **</think>**.
Then, provide your improved translation of the original **{source_language}** text into **{target_language}**.

C.3 Case 3 - multi-choice terms prompt (Track2)

C.3.1 System prompt:

You are a professional senior translator specializing in **{source_language}** to **{target_language}**.

You will be given an original text in **{source_language}** along with a dictionary of terms that must be translated exactly, followed by several translations into **{target_language}** produced by junior translators.

Your first task: Review the provided translations with these requirements:

- Critically evaluate each translation, noting strengths and weaknesses.
- Focus your observations on translation quality, fluency, grammar, vocabulary, and cultural appropriateness.
- Verify that all **{source_language}** terms matching keys in the dictionary below are correctly translated using one of the **{target_language}** alternatives listed for that term.
- After your review, reason about potential improvements and how to produce the best possible translation.
- Keep your review and reasoning succinct (under 1000 words).
- Enclose your review and reasoning within the **<think>** and **</think>** tags.

Your second task: Translate the original **{source_language}** text into fluent, natural **{target_language}**, following these guidelines:

- Complete this task only after the first task.
- Produce the best possible translation based on your previous reasoning.
- Accurately convey the meaning and nuance of the original **{source_language}** text, respecting **{target_language}** grammar, vocabulary, and cultural norms.
- For each term in the **{source_language}** text found as a key in the dictionary below, replace it with exactly one **{target_language}** translation selected from that term's list (choose the best fitting translation in context).
- Translate all other text normally, without altering words not found in the dictionary.

- Provide only the final translation as output, without explanations or comments.

Original **{source_language}** text:
{text}

Dictionary:
{terms}

Translations by junior translators:

1. Translation by the first junior translator:
   {translations[0]}

2. Translation by the second junior translator:
   {translations[1]}

3. Translation by the third junior translator:
   {translations[2]}

4. Translation by the fourth junior translator:
   {translations[3]}

5. Translation by the fifth junior translator:
   {translations[4]}

6. Translation by the sixth junior translator:
   {translations[5]}

7. Translation by the seventh junior translator:
   {translations[6]}

C.3.2 User prompt:

First, review these translations and reason about producing the best possible translation, enclosing your review in **<think>** and **</think>**.
Then, provide your improved translation of the original **{source_language}** text into **{target_language}**.

D Model selection

D.1 Track1

The model selection for Track1 is done based on the scores achieved on the subset of DEV datasets. The subset contains 10 longest texts from each dataset.

agents	bleu	MetricX	TSR	TTQ
eurollm	0.41	**1.17**	0.44	0.65
qwen3_8b	0.29	1.36	0.56	0.73
qwen3_8b-think	0.30	1.80	0.62	0.76
qwen3_14b	0.39	1.37	0.56	0.73
qwen3_14b-think	0.34	1.23	0.55	0.72
gemma3_12b	0.39	1.61	0.64	0.77
gemma3_27b	**0.43**	1.36	**0.75**	**0.84**

Table 18: Mean scores for DEV subset, translations from English to German.

agents	bleu	MetricX	TSR	TTQ
eurollm	0.45	**3.39**	0.78	0.82
qwen3_8b	0.44	3.67	0.88	0.86
qwen3_8b-think	**0.47**	3.53	0.93	0.89
qwen3_14b	0.45	3.53	0.90	0.88
qwen3_14b-think	0.46	3.86	**0.95**	**0.90**
gemma3_12b	0.44	3.52	**0.95**	**0.90**
gemma3_27b	0.45	3.49	0.93	**0.90**

Table 19: Mean scores for DEV subset, translations from English to Spanish.

agents	bleu	MetricX	TSR	TTQ
eurollm	0.23	3.10	0.44	0.62
qwen3_8b	0.21	4.77	0.68	0.74
qwen3_8b-think	0.26	4.50	0.77	0.80
qwen3_14b	**0.28**	3.03	0.69	0.78
qwen3_14b-think	0.25	3.82	**0.85**	**0.85**
gemma3_12b	0.27	3.00	0.74	0.81
gemma3_27b	0.27	**2.66**	0.80	**0.85**

Table 20: Mean scores for Track2 subset, translations from English to Russian.

D.2 Track2

The model selection for Track2 is done based on the scores achieved on the subset of Track2 datasets. The subset contains 10 longest paragraphs from each dataset.

agents	MetricX	TSR	TTQ
eurollm	4.06	0.30	0.50
qwen3_8b	3.82	0.40	0.58
qwen3_8b-think	3.90	0.41	0.59
qwen3_14b	**3.51**	0.41	0.59
qwen3_14b-think	**3.51**	0.41	0.59
gemma3_12b	4.40	**0.45**	**0.61**
gemma3_27b	3.93	**0.45**	**0.61**

Table 21: Mean scores for Track2 subset, translations from English to Traditional Chinese.

agents	MetricX	TSR	TTQ
eurollm	3.50	0.50	0.65
qwen3_8b	3.22	0.51	0.67
qwen3_8b-think	3.26	**0.53**	**0.68**
qwen3_14b	3.34	0.52	0.67
qwen3_14b-think	**3.20**	0.52	0.67
gemma3_12b	3.31	0.51	0.66
gemma3_27b	3.43	0.52	0.67

Table 22: Mean scores for Track2 subset, translations from Traditional Chinese to English.

E Scores

E.1 Track1

E.1.1 MetricX scores

agents	noterm	proper	random	mean
trans-eurollm	0.70	**0.75**	0.77	0.74
trans-qwen3_8b	0.93	1.34	1.20	1.16
trans-qwen3_8b-think	0.92	1.34	1.27	1.18
trans-qwen3_14b	0.87	1.19	1.04	1.03
trans-qwen3_14b-think	0.86	1.31	1.15	1.11
trans-gemma3_12b	0.70	1.33	1.22	1.08
trans-gemma3_27b	0.73	1.32	1.11	1.05
revis-qwen3_14b-think	1.38	1.89	1.69	1.65
revis-gemma3_12b-think	0.71	1.32	1.12	1.05
revis-gemma3_27b-think	0.69	1.22	1.06	0.99
final	**0.36**	0.89	**0.76**	**0.67**

Table 23: MetricX scores for Track1, translations from English to German.

agents	noterm	proper	random	mean
trans-eurollm	1.83	1.90	1.91	1.88
trans-qwen3_8b	2.04	2.35	2.16	2.18
trans-qwen3_8b-think	1.85	2.31	2.32	2.16
trans-qwen3_14b	1.85	2.11	2.06	2.01
trans-qwen3_14b-think	1.84	2.27	2.25	2.12
trans-gemma3_12b	1.77	2.36	2.35	2.16
trans-gemma3_27b	1.82	2.33	2.20	2.12
revis-qwen3_14b-think	2.53	2.99	2.75	2.76
revis-gemma3_12b-think	1.97	2.30	2.24	2.17
revis-gemma3_27b-think	1.83	2.24	2.13	2.07
final	**1.24**	**1.73**	**1.63**	**1.53**

Table 24: MetricX scores for Track1, translations from English to Spanish.

agents	noterm	proper	random	mean
trans-eurollm	1.66	1.84	1.67	1.72
trans-qwen3_8b	1.77	2.01	2.13	1.97
trans-qwen3_8b-think	1.70	2.28	2.21	2.06
trans-qwen3_14b	1.61	1.96	1.86	1.81
trans-qwen3_14b-think	1.60	2.15	2.09	1.95
trans-gemma3_12b	1.42	2.21	2.11	1.91
trans-gemma3_27b	1.49	2.17	2.20	1.95
revis-qwen3_14b-think	2.25	2.74	2.63	2.54
revis-gemma3_12b-think	1.67	2.18	2.19	2.01
revis-gemma3_27b-think	1.53	2.17	1.97	1.89
final	**0.77**	**1.38**	**1.23**	**1.13**

Table 25: MetricX scores for Track1, translations from English to Russian.

E.1.2 TSR scores

agents	proper	random	mean
trans-eurollm	0.37	0.57	0.47
trans-qwen3_8b	0.68	0.74	0.71
trans-qwen3_8b-think	0.72	0.74	0.73
trans-qwen3_14b	0.67	0.73	0.70
trans-qwen3_14b-think	0.70	0.77	0.73
trans-gemma3_12b	0.77	0.79	0.78
trans-gemma3_27b	0.78	0.77	0.78
revis-qwen3_14b-think	0.69	0.73	0.71
revis-gemma3_12b-think	0.72	0.76	0.74
revis-gemma3_27b-think	0.74	0.79	0.77
final	**0.86**	**0.87**	**0.87**

Table 26: TSR scores for Track1, translations from English to German.

agents	proper	random	mean
trans-eurollm	0.54	0.75	0.64
trans-qwen3_8b	0.78	0.81	0.79
trans-qwen3_8b-think	0.84	0.84	0.84
trans-qwen3_14b	0.78	0.81	0.79
trans-qwen3_14b-think	0.82	0.85	0.84
trans-gemma3_12b	0.83	0.85	0.84
trans-gemma3_27b	0.85	0.86	0.85
revis-qwen3_14b-think	0.80	0.80	0.80
revis-gemma3_12b-think	0.84	0.85	0.85
revis-gemma3_27b-think	0.85	0.86	0.85
final	**0.88**	**0.88**	**0.88**

Table 27: TSR scores for Track1, translations from English to Spanish.

agents	proper	random	mean
trans-eurollm	0.56	0.70	0.63
trans-qwen3_8b	0.80	0.84	0.82
trans-qwen3_8b-think	0.87	0.88	0.88
trans-qwen3_14b	0.83	0.84	0.84
trans-qwen3_14b-think	0.89	0.88	0.89
trans-gemma3_12b	0.89	0.88	0.89
trans-gemma3_27b	0.88	0.90	0.89
revis-qwen3_14b-think	0.84	0.84	0.84
revis-gemma3_12b-think	0.88	0.87	0.88
revis-gemma3_27b-think	0.88	0.87	0.88
final	**0.94**	**0.93**	**0.94**

Table 28: TSR scores for Track1, translations from English to Russian.

E.2 Track2

E.2.1 MetricX scores - English to Traditional Chinese

agents	2015	2017	2019	2021	2023	mean
trans-qwen3_8b	2.13	2.07	2.04	2.12	2.16	2.10
trans-qwen3_8b-think	2.07	2.04	1.98	2.02	2.05	2.03
trans-qwen3_14b	2.03	1.92	1.94	1.95	2.00	1.97
trans-qwen3_14b-think	1.97	1.90	1.83	1.89	1.90	1.90
trans-gemma3_12b	2.25	2.09	2.01	2.01	2.13	2.10
trans-gemma3_27b	2.05	1.92	1.89	1.99	2.01	1.97
revis-qwen3_14b-think	1.99	1.86	1.84	1.94	1.93	1.91
revis-gemma3_12b-think	1.99	1.93	1.83	1.88	1.96	1.92
revis-gemma3_27b-think	1.93	1.83	1.78	1.86	1.86	1.85
final	**1.49**	**1.40**	**1.37**	**1.41**	**1.46**	**1.43**

Table 29: MetricX scores for Track2, translations from English to Traditional Chinese with *no* terms.

agents	2015	2017	2019	2021	2023	mean
trans-qwen3_8b	2.43	2.38	2.22	2.36	2.41	2.36
trans-qwen3_8b-think	2.64	2.62	2.64	2.60	2.64	2.63
trans-qwen3_14b	**2.29**	**2.26**	**2.15**	**2.18**	**2.19**	**2.21**
trans-qwen3_14b-think	2.55	2.48	2.43	2.36	2.39	2.44
trans-gemma3_12b	2.96	2.85	2.70	2.66	2.84	2.80
trans-gemma3_27b	2.72	2.68	2.64	2.53	2.60	2.64
revis-qwen3_14b-think	2.39	2.40	2.37	2.39	2.42	2.40
revis-gemma3_12b-think	2.31	2.30	2.24	2.23	2.33	2.28
revis-gemma3_27b-think	2.46	2.36	2.41	2.29	2.35	2.38
final	2.39	2.28	2.26	2.23	2.22	2.28

Table 30: MetricX scores for Track2, translations from English to Traditional Chinese with *random* terms.

agents	2015	2017	2019	2021	2023	mean
trans-qwen3_8b	2.17	2.03	2.09	2.12	2.17	2.12
trans-qwen3_8b-think	2.19	2.04	2.09	2.09	2.14	2.11
trans-qwen3_14b	2.05	1.95	1.99	1.98	2.05	2.01
trans-qwen3_14b-think	2.05	1.90	1.92	1.97	2.02	1.97
trans-gemma3_12b	2.36	2.10	2.11	2.17	2.22	2.19
trans-gemma3_27b	2.18	1.98	2.01	2.04	2.10	2.06
revis-qwen3_14b-think	2.11	1.91	1.98	1.94	2.07	2.00
revis-gemma3_12b-think	2.09	1.93	2.04	2.00	2.01	2.01
revis-gemma3_27b-think	2.05	1.85	1.94	1.97	2.00	1.96
final	**1.88**	**1.70**	**1.73**	**1.77**	**1.84**	**1.78**

Table 31: `MetricX` scores for `Track2`, translations from English to Traditional Chinese with *proper* terms.

E.2.2 `MetricX` scores - Traditional Chinese to English - Bug

agents	2016	2018	2020	2022	2024	mean
trans-qwen3_8b	1.69	1.75	1.60	1.62	1.62	1.65
trans-qwen3_8b-think	1.64	1.66	1.61	1.59	1.62	1.62
trans-qwen3_14b	1.61	1.58	1.54	1.62	1.56	1.58
trans-qwen3_14b-think	1.59	1.57	1.51	1.57	1.52	1.55
trans-gemma3_12b	1.77	1.69	1.76	1.66	1.69	1.72
trans-gemma3_27b	1.66	1.66	1.62	1.62	1.65	1.64
revis-qwen3_14b-think	1.68	1.66	1.75	1.77	1.86	1.75
revis-gemma3_12b-think	1.63	1.66	1.65	1.67	1.70	1.66
revis-gemma3_27b-think	1.65	1.66	1.58	1.63	1.58	1.62
final	**1.29**	**1.26**	**1.18**	**1.22**	**1.20**	**1.23**

Table 32: `MetricX` scores for `Track2`, translations from Traditional Chinese to English with *random* terms.

agents	2016	2018	2020	2022	2024	mean
trans-qwen3_8b	1.76	1.74	1.60	1.68	1.69	1.69
trans-qwen3_8b-think	1.71	1.72	1.63	1.63	1.68	1.67
trans-qwen3_14b	1.62	1.63	1.58	1.63	1.68	1.63
trans-qwen3_14b-think	1.65	1.61	1.55	1.61	1.63	1.61
trans-gemma3_12b	1.75	1.76	1.65	1.72	1.75	1.73
trans-gemma3_27b	1.74	1.73	1.60	1.68	1.76	1.70
revis-qwen3_14b-think	1.83	1.71	1.80	1.98	1.92	1.85
revis-gemma3_12b-think	1.74	1.70	1.65	1.77	1.74	1.72
revis-gemma3_27b-think	1.72	1.66	1.62	1.64	1.68	1.66
final	**1.33**	**1.29**	**1.23**	**1.28**	**1.28**	**1.28**

Table 33: `MetricX` scores for `Track2`, translations from Traditional Chinese to English with *proper* terms.

E.2.3 `MetricX` scores - Traditional Chinese to English - Correct

agents	2016	2018	2020	2022	2024	mean
trans-qwen3_8b	1.69	1.69	1.57	1.59	1.60	1.63
trans-qwen3_8b-think	1.67	1.65	1.59	1.60	1.62	1.63
trans-qwen3_14b	1.57	1.57	1.52	1.62	1.56	1.57
trans-qwen3_14b-think	1.56	1.51	1.46	1.51	1.52	1.51
trans-gemma3_12b	1.59	1.54	1.51	1.47	1.51	1.52
trans-gemma3_27b	1.63	1.61	1.54	1.57	1.56	1.58
revis-qwen3_14b-think	2.19	2.13	2.08	2.07	2.07	2.11
revis-gemma3_12b-think	1.68	1.70	1.64	1.69	1.75	1.69
revis-gemma3_27b-think	1.67	1.62	1.59	1.61	1.60	1.62
final	**1.22**	**1.17**	**1.12**	**1.14**	**1.14**	**1.16**

Table 34: `MetricX` scores for `Track2`, translations from Traditional Chinese to English with *no* terms.

agents	2016	2018	2020	2022	2024	mean
trans-qwen3_8b	1.75	1.76	1.72	1.76	1.79	1.76
trans-qwen3_8b-think	1.77	1.96	1.86	1.73	1.85	1.83
trans-qwen3_14b	1.70	1.71	1.63	1.72	1.67	1.69
trans-qwen3_14b-think	1.74	1.79	1.75	1.67	1.70	1.73
trans-gemma3_12b	2.63	2.72	2.65	2.41	2.64	2.61
trans-gemma3_27b	2.09	2.37	2.10	1.96	2.03	2.11
revis-qwen3_14b-think	1.87	2.02	2.00	1.97	2.11	1.99
revis-gemma3_12b-think	1.87	2.09	1.92	1.84	1.92	1.93
revis-gemma3_27b-think	1.99	2.00	2.08	1.95	1.96	2.00
final	**1.54**	**1.58**	**1.53**	**1.46**	**1.52**	**1.53**

Table 35: `MetricX` scores for `Track2`, for translations from Traditional Chinese to English, with random terms.

agents	2016	2018	2020	2022	2024	mean
trans-qwen3_8b	1.83	1.80	1.72	1.85	1.87	1.81
trans-qwen3_8b-think	1.91	1.88	1.83	1.94	1.90	1.89
trans-qwen3_14b	1.82	1.78	1.68	1.80	1.83	1.78
trans-qwen3_14b-think	1.88	1.86	1.72	1.84	1.85	1.83
trans-gemma3_12b	2.08	2.01	2.01	2.24	2.05	2.08
trans-gemma3_27b	2.10	2.05	1.94	2.10	2.02	2.04
revis-qwen3_14b-think	2.04	1.99	1.97	2.04	2.13	2.03
revis-gemma3_12b-think	1.99	1.92	1.90	2.01	1.93	1.95
revis-gemma3_27b-think	2.05	1.95	1.84	1.97	1.91	1.95
final	**1.59**	**1.57**	**1.47**	**1.58**	**1.53**	**1.55**

Table 36: `MetricX` scores for `Track2`, for translations from Traditional Chinese to English, with proper terms.

E.2.4 `MetricX` scores - Traditional Chinese to English - Bug Assessment

We are unable to adequately assess the expected `MetricX` scores and the drop in the final results of our submission as a result of the bug. For this assessment, we need access to the translation reference texts that are unavailable. However, as we can observe in Sections E.1.1, E.2.1, E.2.2 and E.2.3, `MetricX` scores are increasing (i.e. translation quality drops) for random and proper term cases compared to the

no-term case. The intuition behind this trend is that forcing terminology reduces the general translation quality because translators need to balance between two different objectives: 1) general translation and 2) terminology constraint. Based on this intuition, we can approximate the drop in the final MetricX scores as a missed increase in MetricX scores (i.e. missed drop in general translation quality) as a result of less constrained terminology.

agents	2016	2018	2020	2022	2024	mean
final-bug	1.29	1.26	1.18	1.22	1.20	1.23
final-correct	1.54	1.58	1.53	1.46	1.52	1.53
final-diff	0.25	0.32	0.35	0.24	0.32	0.30

Table 37: Estimation of the final score drop due to the bug, MetricX scores for Track2, translations from Traditional Chinese to English with *random* terms.

agents	2016	2018	2020	2022	2024	mean
final-bug	1.33	1.29	1.23	1.28	1.28	1.28
final-correct	1.59	1.57	1.47	1.58	1.53	1.55
final-diff	0.26	0.28	0.24	0.30	0.25	0.27

Table 38: Estimation of the final score drop due to the bug, MetricX scores for Track2, translations from Traditional Chinese to English with *proper* terms.

E.2.5 TSR scores - English to Traditional Chinese

agents	2015	2017	2019	2021	2023	mean
trans-qwen3_8b	0.56	0.47	0.50	0.53	0.58	0.53
trans-qwen3_8b-think	0.55	0.51	0.55	0.58	0.58	0.55
trans-qwen3_14b	0.53	0.42	0.49	0.47	0.52	0.49
trans-qwen3_14b-think	0.53	0.48	0.51	0.50	0.51	0.51
trans-gemma3_12b	0.61	0.54	0.58	0.56	0.60	0.58
trans-gemma3_27b	0.60	0.54	0.60	0.59	0.65	0.60
revis-qwen3_14b-think	0.53	0.41	0.50	0.48	0.54	0.49
revis-gemma3_12b-think	0.55	0.44	0.53	0.51	0.57	0.52
revis-gemma3_27b-think	0.58	0.47	0.56	0.52	0.59	0.55
final	**0.75**	**0.70**	**0.72**	**0.73**	**0.75**	**0.73**

Table 39: TSR scores for Track2, for translations from English to Traditional Chinese, with *random* terms, Track2.

agents	2015	2017	2019	2021	2023	mean
trans-qwen3_8b	0.65	0.64	0.64	0.66	0.67	0.65
trans-qwen3_8b-think	0.66	0.65	0.66	0.67	0.68	0.66
trans-qwen3_14b	0.66	0.64	0.64	0.66	0.67	0.66
trans-qwen3_14b-think	0.66	0.65	0.66	0.68	0.68	0.67
trans-gemma3_12b	0.70	0.69	0.68	0.72	0.71	0.70
trans-gemma3_27b	0.70	0.69	0.69	0.72	0.71	0.70
revis-qwen3_14b-think	0.67	0.66	0.67	0.68	0.68	0.67
revis-gemma3_12b-think	0.67	0.67	0.67	0.69	0.68	0.68
revis-gemma3_27b-think	0.68	0.68	0.67	0.69	0.69	0.68
final	**0.78**	**0.77**	**0.78**	**0.81**	**0.79**	**0.79**

Table 40: TSR scores for Track2, for translations from English to Traditional Chinese, with *proper* terms, Track2.

E.2.6 TSR scores - Traditional Chinese to English - Bug

agents	2016	2018	2020	2022	2024	mean
trans-qwen3_8b	0.56	0.90	0.90	0.90	0.86	0.82
trans-qwen3_8b-think	0.55	0.90	0.90	0.90	0.86	0.82
trans-qwen3_14b	0.54	0.90	0.90	0.90	0.86	0.82
trans-qwen3_14b-think	0.55	0.90	0.90	0.90	0.86	0.82
trans-gemma3_12b	0.55	0.89	0.90	0.90	0.86	0.82
trans-gemma3_27b	0.55	0.89	0.90	0.91	0.86	0.82
revis-qwen3_14b-think	0.55	0.89	0.90	0.90	0.85	0.82
revis-gemma3_12b-think	0.55	0.89	0.90	0.90	0.85	0.82
revis-gemma3_27b-think	0.55	0.89	0.89	0.90	0.86	0.82
final	**0.56**	**0.91**	**0.90**	**0.91**	**0.87**	**0.83**

Table 41: TSR scores for Track2, for translations from Traditional Chinese to English, with *random* terms, Track2.

agents	2016	2018	2020	2022	2024	mean
trans-qwen3_8b	0.90	0.88	0.91	0.89	0.90	0.89
trans-qwen3_8b-think	0.91	0.88	0.91	0.89	0.89	0.90
trans-qwen3_14b	0.90	0.89	0.91	0.90	0.90	0.90
trans-qwen3_14b-think	0.91	0.90	0.92	0.91	0.90	0.91
trans-gemma3_12b	0.91	0.87	0.91	0.88	0.89	0.89
trans-gemma3_27b	0.91	0.88	0.90	0.88	0.89	0.89
revis-qwen3_14b-think	0.91	0.90	0.91	0.89	0.90	0.90
revis-gemma3_12b-think	0.90	0.87	0.90	0.90	0.89	0.89
revis-gemma3_27b-think	0.91	0.88	0.91	0.90	0.90	0.90
final	**0.92**	**0.91**	**0.92**	**0.92**	**0.91**	**0.92**

Table 42: TSR scores for Track2, for translations from Traditional Chinese to English, with *proper* terms, Track2.

E.2.7 TSR scores - Traditional Chinese to English - Correct

agents	2016	2018	2020	2022	2024	mean
trans-qwen3_8b	0.35	0.43	0.44	0.44	0.43	0.42
trans-qwen3_8b-think	0.36	0.43	0.44	0.44	0.42	0.42
trans-qwen3_14b	0.34	0.43	0.44	0.44	0.42	0.42
trans-qwen3_14b-think	0.35	0.43	0.44	0.44	0.42	0.42
trans-gemma3_12b	0.35	0.43	0.44	0.45	0.44	0.42
trans-gemma3_27b	0.36	0.44	0.45	0.45	0.43	0.43
revis-qwen3_14b-think	0.34	0.42	0.43	0.43	0.41	0.41
revis-gemma3_12b-think	0.36	0.42	0.43	0.44	0.42	0.41
revis-gemma3_27b-think	0.35	0.44	0.45	0.45	0.43	0.42
final	**0.39**	**0.48**	**0.48**	**0.48**	**0.46**	**0.46**

Table 43: TSR scores for Track2, for translations from Traditional Chinese to English, with *random* terms.

agents	2016	2018	2020	2022	2024	mean
trans-qwen3_8b	0.66	0.70	0.73	0.76	0.71	0.71
trans-qwen3_8b-think	0.66	0.70	0.73	0.75	0.71	0.71
trans-qwen3_14b	0.66	0.71	0.73	0.77	0.72	0.72
trans-qwen3_14b-think	0.67	0.72	0.73	0.77	0.72	0.72
trans-gemma3_12b	0.66	0.69	0.72	0.75	0.70	0.70
trans-gemma3_27b	0.66	0.70	0.72	0.76	0.72	0.71
revis-qwen3_14b-think	0.67	0.71	0.72	0.76	0.72	0.72
revis-gemma3_12b-think	0.66	0.68	0.71	0.76	0.71	0.70
revis-gemma3_27b-think	0.66	0.70	0.73	0.76	0.72	0.71
final	**0.69**	**0.74**	**0.75**	**0.80**	**0.75**	**0.75**

Table 44: TSR scores for Track2, for translations from Traditional Chinese to English, with *proper* terms.

E.2.8 TSR scores - Traditional Chinese to English - Bug Assessment

Here, we provide an assessment of the expected TSR scores and the drop in the final results of our submission as a result of the bug. The assessment is calculated with full correct terms on the original submitted translations that contain the bug.

agents	2016	2018	2020	2022	2024	mean
trans-qwen3_8b	0.34	0.41	0.41	0.42	0.40	0.40
trans-qwen3_8b-think	0.33	0.42	0.41	0.42	0.40	0.40
trans-qwen3_14b	0.33	0.41	0.41	0.42	0.40	0.39
trans-qwen3_14b-think	0.33	0.41	0.41	0.42	0.40	0.39
trans-gemma3_12b	0.33	0.40	0.41	0.42	0.39	0.39
trans-gemma3_27b	0.33	0.40	0.41	0.42	0.40	0.39
revis-qwen3_14b-think	0.33	0.40	0.40	0.41	0.39	0.39
revis-gemma3_12b-think	0.33	0.40	0.40	0.41	0.39	0.38
revis-gemma3_27b-think	0.33	0.41	0.40	0.42	0.40	0.39
final	**0.36**	**0.45**	**0.44**	**0.45**	**0.43**	**0.43**
final-delta (bug - correct)	**-0.03**	**-0.03**	**-0.04**	**-0.03**	**-0.03**	**-0.03**

Table 45: Assessment of the final score drop due to the bug, TSR scores for Track2, for translations from Traditional Chinese to English, with *random* terms.

agents	2016	2018	2020	2022	2024	mean
trans-qwen3_8b	0.53	0.53	0.54	0.61	0.55	0.55
trans-qwen3_8b-think	0.54	0.54	0.56	0.61	0.56	0.56
trans-qwen3_14b	0.56	0.59	0.57	0.65	0.59	0.59
trans-qwen3_14b-think	0.56	0.59	0.59	0.65	0.59	0.60
trans-gemma3_12b	0.52	0.53	0.54	0.60	0.54	0.55
trans-gemma3_27b	0.55	0.56	0.56	0.64	0.58	0.58
revis-qwen3_14b-think	0.57	0.58	0.58	0.64	0.59	0.59
revis-gemma3_12b-think	0.55	0.55	0.56	0.63	0.56	0.57
revis-gemma3_27b-think	0.55	0.56	0.57	0.64	0.57	0.58
final	**0.61**	**0.63**	**0.64**	**0.69**	**0.64**	**0.64**
final-delta (bug - correct)	**-0.08**	**-0.11**	**-0.11**	**-0.11**	**-0.11**	**-0.11**

Table 46: Assessment of the final score drop due to the bug, TSR scores for Track2, for translations from Traditional Chinese to English, with *proper* terms.

E.3 Metric-guided agent selection

E.3.1 Track1

agents	noterm	proper	random	total
trans-eurollm	10.4%	6.2%	12.6%	9.7%
trans-qwen3_8b	4.6%	7.0%	4.8%	5.4%
trans-qwen3_8b-think	5.2%	5.2%	5.2%	5.2%
trans-qwen3_14b	6.4%	5.0%	7.4%	6.2%
trans-qwen3_14b-think	3.6%	5.4%	5.6%	4.8%
trans-gemma3_12b	8.4%	9.2%	8.2%	8.6%
trans-gemma3_27b	7.2%	7.4%	5.8%	6.8%
revis-qwen3_14b-think	12.0%	12.4%	10.2%	11.5%
revis-gemma3_12b-think	**23.5%**	**24.6%**	**26.4%**	**24.8%**
revis-gemma3_27b-think	18.6%	17.5%	13.8%	16.6%

Table 47: The frequency of agents selected for the final solution, for translations from English to German, Track1.

agents	noterm	proper	random	total
trans-eurollm	12.2%	13.2%	11.7%	12.4%
trans-qwen3_8b	3.4%	6.4%	5.6%	5.1%
trans-qwen3_8b-think	8.2%	5.0%	5.8%	6.3%
trans-qwen3_14b	6.2%	8.7%	6.0%	7.0%
trans-qwen3_14b-think	4.0%	5.0%	3.2%	4.0%
trans-gemma3_12b	9.6%	4.3%	5.6%	6.5%
trans-gemma3_27b	8.0%	7.0%	7.0%	7.3%
revis-qwen3_14b-think	13.0%	7.6%	9.4%	10.0%
revis-gemma3_12b-think	17.2%	**27.6%**	**31.2%**	**25.3%**
revis-gemma3_27b-think	**18.2%**	15.0%	14.4%	15.8%

Table 48: The frequency of agents selected for the final solution, for translations from English to Spanish, Track1.

agents	noterm	proper	random	total
trans-eurollm	11.6%	9.6%	11.7%	11.0%
trans-qwen3_8b	4.3%	5.2%	5.2%	4.9%
trans-qwen3_8b-think	5.4%	6.2%	7.0%	6.2%
trans-qwen3_14b	7.2%	6.0%	5.0%	6.0%
trans-qwen3_14b-think	3.2%	4.6%	4.2%	4.0%
trans-gemma3_12b	10.6%	8.6%	10.2%	9.8%
trans-gemma3_27b	7.4%	7.0%	7.8%	7.4%
revis-qwen3_14b-think	11.0%	10.8%	10.2%	10.6%
revis-gemma3_12b-think	**21.4%**	**27.8%**	**25.2%**	**24.8%**
revis-gemma3_27b-think	17.8%	14.2%	13.4%	15.1%

Table 49: The frequency of agents selected for the final solution, for translations from English to Russian, Track1.

E.3.2 Track2

agents	2015	2017	2019	2021	2023	total
trans-qwen3_8b	11.0%	6.4%	6.9%	7.4%	6.0%	7.4%
trans-qwen3_8b-think	7.8%	4.8%	5.1%	6.1%	6.6%	6.1%
trans-qwen3_14b	6.1%	6.7%	6.3%	7.9%	6.8%	6.8%
trans-qwen3_14b-think	7.5%	7.1%	10.1%	9.0%	9.4%	8.7%
trans-gemma3_12b	7.2%	11.1%	12.1%	12.0%	13.8%	11.6%
trans-gemma3_27b	9.3%	11.8%	13.5%	8.8%	10.1%	10.6%
revis-qwen3_14b-think	12.2%	14.6%	14.2%	12.3%	10.5%	12.6%
revis-gemma3_12b-think	17.7%	**19.0%**	**16.2%**	17.3%	16.6%	17.3%
revis-gemma3_27b-think	**20.6%**	18.0%	15.1%	**18.7%**	**19.7%**	**18.4%**

Table 50: The frequency of agents selected for the final solution, for translations from English to Traditional Chinese, with no terms, Track2.

agents	2015	2017	2019	2021	2023	total
trans-qwen3_8b	7.2%	7.6%	6.0%	4.7%	5.3%	6.0%
trans-qwen3_8b-think	5.2%	6.9%	5.6%	6.5%	7.3%	6.4%
trans-qwen3_14b	6.4%	6.7%	8.3%	6.9%	5.9%	6.8%
trans-qwen3_14b-think	8.7%	6.2%	9.2%	10.0%	6.6%	8.2%
trans-gemma3_12b	13.1%	15.5%	12.4%	15.0%	15.5%	14.4%
trans-gemma3_27b	12.8%	9.9%	12.4%	9.7%	14.0%	11.7%
revis-qwen3_14b-think	9.6%	11.3%	11.2%	13.4%	12.1%	11.7%
revis-gemma3_12b-think	**20.6%**	**19.4%**	**19.4%**	**20.1%**	**20.5%**	**20.0%**
revis-gemma3_27b-think	16.0%	16.0%	15.1%	13.2%	12.3%	14.3%

Table 51: The frequency of agents selected for the final solution, for translations from English to Traditional Chinese, with *proper* terms, Track2.

agents	2015	2017	2019	2021	2023	total
trans-qwen3_8b	9.0%	11.6%	9.7%	7.4%	6.4%	8.6%
trans-qwen3_8b-think	6.1%	5.5%	8.5%	7.4%	8.3%	7.3%
trans-qwen3_14b	9.3%	6.2%	6.5%	6.3%	7.3%	7.0%
trans-qwen3_14b-think	5.2%	7.1%	8.5%	7.4%	6.6%	7.1%
trans-gemma3_12b	15.4%	**17.6%**	**15.1%**	**18.0%**	13.1%	15.8%
trans-gemma3_27b	15.7%	16.0%	14.8%	14.8%	15.7%	15.4%
revis-qwen3_14b-think	8.1%	7.6%	9.2%	6.7%	9.6%	8.2%
revis-gemma3_12b-think	**17.4%**	12.7%	14.8%	16.8%	**18.2%**	**16.1%**
revis-gemma3_27b-think	13.4%	15.3%	12.4%	14.8%	14.4%	14.1%

Table 52: The frequency of agents selected for the final solution, for translations from English to Traditional Chinese, with *random* terms, Track2.

agents	2016	2018	2020	2022	2024	total
trans-qwen3_8b	7.0%	8.6%	8.5%	8.6%	9.8%	8.6%
trans-qwen3_8b-think	8.6%	5.0%	7.2%	7.5%	7.2%	7.1%
trans-qwen3_14b	9.4%	8.4%	10.9%	8.6%	9.4%	9.4%
trans-qwen3_14b-think	10.0%	10.3%	9.3%	9.6%	8.7%	9.5%
trans-gemma3_12b	13.2%	13.7%	**15.5%**	**17.3%**	**15.3%**	15.2%
trans-gemma3_27b	9.1%	9.3%	9.3%	9.4%	11.0%	9.7%
revis-qwen3_14b-think	15.1%	**18.2%**	14.6%	16.3%	**15.3%**	**15.8%**
revis-gemma3_12b-think	**17.2%**	13.7%	13.4%	13.6%	11.2%	13.5%
revis-gemma3_27b-think	10.0%	12.5%	10.9%	8.6%	11.7%	10.7%

Table 53: The frequency of agents selected for the final solution, for translations from Traditional Chinese to English, with *no* terms, Track2.

agents	2016	2018	2020	2022	2024	total
trans-qwen3_8b	11.0%	8.8%	8.9%	8.4%	11.3%	9.7%
trans-qwen3_8b-think	7.0%	5.2%	5.0%	6.7%	6.3%	6.0%
trans-qwen3_14b	10.2%	11.5%	13.6%	10.7%	10.5%	11.3%
trans-qwen3_14b-think	9.1%	8.1%	6.8%	10.0%	8.7%	8.5%
trans-gemma3_12b	10.5%	10.8%	9.7%	9.6%	9.1%	9.8%
trans-gemma3_27b	8.9%	10.3%	11.6%	10.4%	11.0%	10.5%
revis-qwen3_14b-think	15.1%	**19.4%**	**19.2%**	**17.7%**	**15.6%**	**17.4%**
revis-gemma3_12b-think	**15.6%**	16.1%	13.6%	15.7%	13.6%	14.8%
revis-gemma3_27b-think	12.1%	9.3%	11.1%	10.4%	13.4%	11.3%

Table 54: The frequency of agents selected for the final solution, for translations from Traditional Chinese to English, with *proper* terms, Track2.

agents	2016	2018	2020	2022	2024	total
trans-qwen3_8b	7.2%	6.9%	7.6%	8.2%	9.3%	8.0%
trans-qwen3_8b-think	6.7%	8.1%	7.0%	7.7%	7.0%	7.3%
trans-qwen3_14b	8.1%	10.5%	9.5%	9.8%	14.3%	10.7%
trans-qwen3_14b-think	7.8%	10.0%	6.4%	8.4%	7.5%	8.0%
trans-gemma3_12b	14.0%	10.5%	12.8%	11.9%	13.6%	12.6%
trans-gemma3_27b	11.6%	12.7%	11.1%	13.1%	9.3%	11.4%
revis-qwen3_14b-think	**16.4%**	**17.3%**	**18.5%**	**14.4%**	**14.3%**	**16.0%**
revis-gemma3_12b-think	16.2%	11.0%	12.2%	12.9%	11.5%	12.6%
revis-gemma3_27b-think	11.6%	12.5%	14.6%	13.2%	12.9%	13.0%

Table 55: The frequency of agents selected for the final solution, for translations from Traditional Chinese to English, with *random* terms, Track2.

E.4 Final results - Track2- Bug Correction

Here we compare the Track2 scores of the submitted MeGuMa system and the scores of the error-corrected version. The error occurred when projecting document-level terminology map to the paragraph level. This operation, performed because our system operates on the paragraph level, used whitespace-delimiters for term matching. This approach, suitable for European languages, is not correct for Chinese texts. For a more detailed description, see the end of Subsection 2.

The corrected system was scored using the evaluation code from the official repository.[5] The corrected show a large increase (34 points) of the terminology success rate Term-Acc (labeled as "Proper, Acc." by the organizers). This is expected, since the erroneous term matching caused the loss of predefined term translations fed to the system. For the random terminology, there is a smaller increase of 2 points. Translation accuracies, in terms of ChrF2++, increase by approximately 3.5 points.

In order to ensure transparency, our repository[6] contains detailed code-level description of the bug, the outputs of the corrected system, and the instructions how to run our system.

System	Bleu4	ChrF	Proper, Acc.	Random, Acc.
MeGuMa [submitted]	32.96	69.41	62.43	86.76
MeGuMa [debugged]	39.07	72.73	96.62	87.66
Difference	6.11	3.32	34.19	0.90

Table 56: The final scores of MeGuMa system, before and after correcting the data preparation bug. Translations from Traditional Chinese to English, with *proper* terms, Track2.

System	Bleu4	ChrF	Proper, Acc.	Random, Acc.
MeGuMa [submitted]	23.88	65.21	51.55	86.44
MeGuMa [debugged]	31.38	69.07	53.68	92.58
Difference	7.50	3.86	2.13	6.14

Table 57: The final scores of MeGuMa system, before and after correcting the data preparation bug. Translations from Traditional Chinese to English, with *random* terms, Track2.

System	Bleu4	ChrF	Proper, Acc.	Random, Acc.
MeGuMa [submitted]	30.83	68.31	51.91	85.91
MeGuMa [debugged]	30.83	68.31	51.91	85.91
Difference	0.00	0.00	0.00	0.00

Table 58: The final scores of MeGuMa system, before and after correcting the data preparation bug. Translations from Traditional Chinese to English, with *noterm* terms, Track2.

[5] https://github.com/wmt-conference/wmt25-terminology/
[6] https://github.com/igrubi/irb-mt-wmt2025