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

Ivan Grubišić* and Damir Korenčić*
Division of Electronics, Ruđer Bošković Institute, Zagreb, Croatia
* Equal contribution

INTRODUCTION

When translating texts from specialized technical domains such as medicine, finance, or law, it is important to **translate technical terms accurately and consistently**. 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 machine translation (MT), they are rarely evaluated on specialized domains that require strict adherence to terminology. The goal of the WMT25 Terminology Translation Task is to determine **how well do the modern MT systems tackle this challenge**.

TASK

WMT25 Terminology Task datasets are divided into Track1 and Track2, consisting of texts from the information technology and financial domains, respectively. Track1 datasets contain paragraph-level texts and cover **en-de, en-es, and en-ru language pairs**. Track2 datasets contain long documents 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 predefined terminology on system performance. For the same reason, an additional "no terminology" setup is included in the task.

OBJECTIVES

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.

METHOD

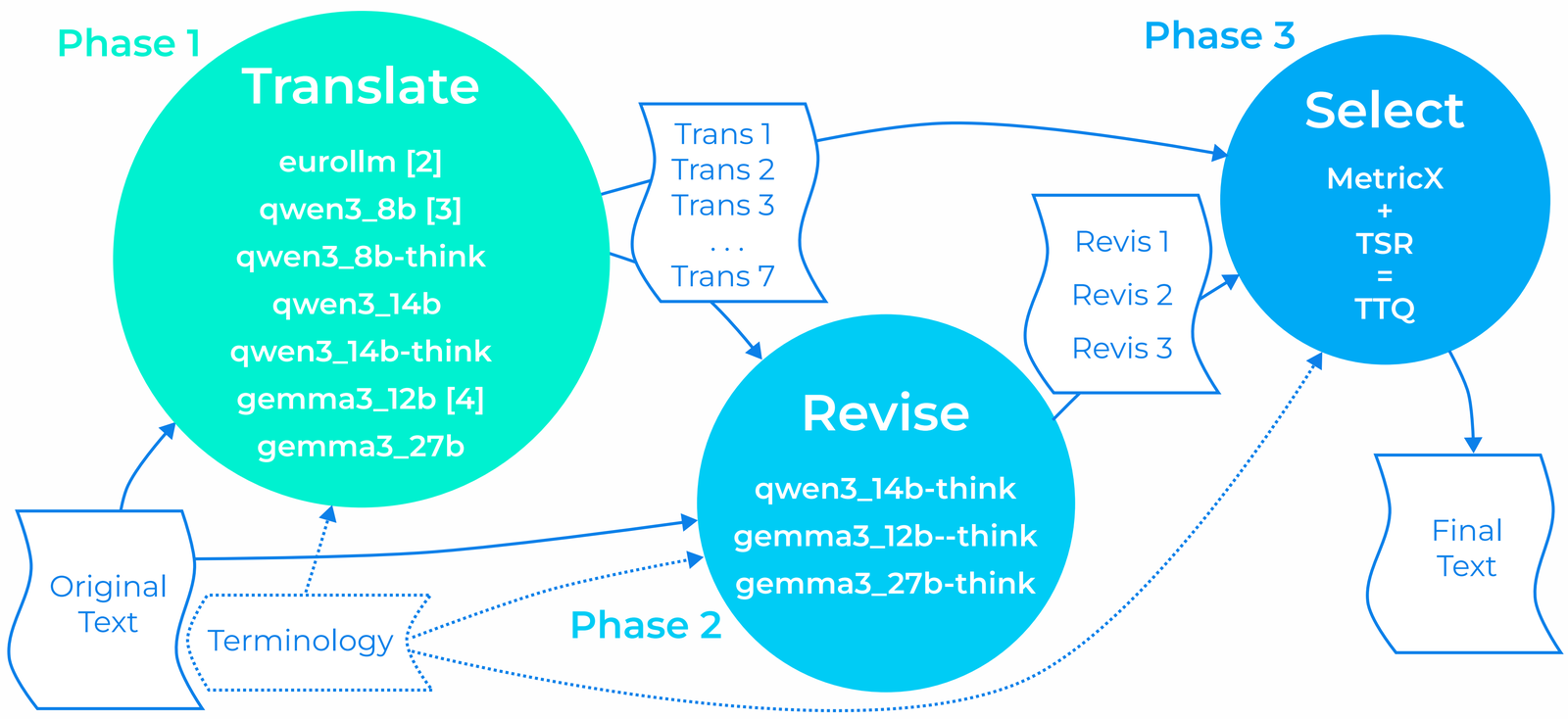
Our translation **system produces an output in three phases**:

1) individual translator LLMs generate initial translations,

2) reviser LLMs generate improved translations from the initial ones,

3) all of the candidate translations are pooled, and the best one is selected based on a custom metric.

Terminology Translation Quality (TTQ) metric combines two other metrics: MetricX [1], for evaluating general translation quality, and custom Terminology Success Rate (TSR).



RESULTS

The system produces high-quality translations with low MetricX and high TSR scores.

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 future work we plan to improve our system with a more granular agentic workflow that incorporates additional specialized roles like pre-editor and post-editor and expand the evaluation to more language pairs with qualitative analysis of the outputs.

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.

REFERENCES

[1] Juraska et al., 2024. Metricx-24: The google submission to the wmt 2024 metrics shared task (http://arxiv.org/abs/2506.04079)
[2] Martins et al., 2025. EuroLLM-9B: Technical Report (https://arxiv.org/abs/2506.04079)
[3] Yang et al., 2025. Qwen3 Technical Report (http://arxiv.org/abs/2505.09388)
[4] Team et al., 2025. Gemma 3 technical report. (http://arxiv.org/abs/2503.19786)