Findings of the WMT25 Terminology Translation Task: Terminology is Useful Especially for Good MTs

Kirill Semenov University of Zurich
Xu Huang Nanjing University
Vilém Zouhar ETH Zurich
Nathaniel Berger Amazon AGI
Dawei Zhu Amazon AGI
Arturo Oncevay Independent
Pinzhen Chen University of Edinburgh & Aveni

Abstract

The WMT25 Terminology Translation Task releases new resources in high-stakes domains and investigates the capabilities of translation systems to accurately and consistently translate specialized terms. This year, we feature new domain and language coverage over previous editions, introducing two distinct tracks: (1) sentence-level translation in the information technology domain for English→German, English→Russian, and English→Spanish, and (2) document-level translation in the finance domain for English↔Traditional Chinese with a document-level one-to-many dictionary. Participants are challenged to translate texts under three modes: no terminology, proper terminology, and random terminology, allowing for a causal analysis of terminology utility. Evaluation combines overall quality, terminology accuracy, and terminology consistency. This shared task attracted broad participation, with 13 teams submitting 20 systems in Track 1 and 4 teams participating in Track 2. The results show that providing proper terminology consistently boosts both overall translation quality and term accuracy, whereas reliance on random terminology yields smaller gains. Despite the near-saturation of sentence-level benchmarks, document-level finance translation still falls short, indicating an urgent need for long-form evaluation and more robust metrics tailored to professional domains.

1 Introduction

Time flies. Since the 2023 edition of the WMT Terminology Translation Task (Semenov et al., 2023), rapid advances in machine translation (MT) and large language models (LLMs) have achieved near-human quality for general-domain translation in several languages (Kocmi et al., 2024, 2025b). Nonetheless, it remains an open question whether these powerful models or techniques can successfully address terminology translation, where the need for accurate and consistent conversion of terminologies poses extra difficulty in addition to general translation quality.

In professional fields like finance, medicine, and law, the correct use of specialized, agreed terms is critical for accuracy and clarity in communications, making terminology translation a research problem of high commercial value (Oncevay et al., 2025b). The field has seen various efforts in modeling (Hasler et al., 2018; Dinu et al., 2019), evaluation (Zouhar et al., 2020; Semenov and Bojar, 2022), and translator tool development (Vargas-Sierra, 2011; Arcan et al., 2017; Lagzdiņš et al., 2022), but recent progress seems to slow down even for high-resourced languages. For example, at the 2024 Conference on Machine Translation, only two papers were dedicated to terminology translation (Kim et al., 2024; Myung et al., 2024).

It is in this context of exciting general domain progress versus modest attention to terminology translation that we organize the WMT25 Terminology Translation Task. The task comes with two primary objectives: (1) to provide an understanding of the current landscape of terminology-aware translation, and (2) to announce and release new, human-annotated datasets to facilitate future research. We organize two tracks covering five translation directions and both sentence- and document-level translation. Particularly, the document track features large document-level one-to-many dictionaries that are more realistic in production. In terms of evaluation, we run multiple metrics that target different facets of terminology conversion, all while taking into account the general quality. Moreover, to estimate the added value (causal effect) of a provided terminology dictionary, systems are evaluated in three translation conditions: without a dictionary, with a proper terminology dictionary, and with a

★ All authors contributed considerably to the design, execution, and presentation of this work. Dawei Zhu's and Nathaniel Berger's work was done outside Amazon. Correspondence to kirill.semenov@uzh.ch. All resources are available at github.com/wmt-conference/wmt25-terminology.

Proceedings of the Tenth Workshop on Machine Translation (WMT), pages 554–576
November 8–9, 2025.

dictionary of random words.

The main findings from the WMT 2025 Terminology Translation Task across submissions are:

★ Most systems involve an LLM in some way. Top systems achieve close to perfect terminology accuracy in the sentence-level track, indicating LLM's suitability for the task, and a saturation in the sentence benchmark. We need to move to document-level benchmarking, which is also more aligned with the practical use.

★ Contrary to the expectations based on previous research, there is little trade-off between general translation ability and terminology accuracy for most participating systems.

★ Incorporating some terminologies always benefits overall translation quality; using proper terms is more useful than random terms for top-performing systems.

2 Task Description

Our task is organized into two tracks, each focused on different domains and input sizes (of both texts and dictionaries), with a unified evaluation protocol. This allows this edition of the terminology shared task to cover wider scope of fields and input types (contrary to (Alam et al., 2021), where only medical domain and sentence-level data were provided), as well as tackling the domains in a more controlled manner (contrary to (Semenov et al., 2023), where every translation direction had its own domain of texts, from NLP abstracts to web novels).

2.1 Tracks and Domains

Track 1: Sentence-level translation

- **Domain:** Information Technology (IT), SAP
- **Translation direction:** English→German, English →Russian, and English→Spanish
- **Setup:** Participants are provided with sentence segments, each with a small terminology dictionary containing only the terms present in the segment, usually 1-2 entries.

Track 2: Document-Level Translation

- **Domain:** Hong Kong finance
- **Translation direction:** English↔Chinese[1]
- **Setup:** Participants are given documents, each accompanied by a large (up to 1K entries) document-level terminology dictionary. English input documents are capped at 2K words; for Chinese→English, the Chinese input is truncated to correspond to the English output of up to 2K words.

  Terminology constraints differ by direction: for English→Chinese, terms are one-to-one mapped, same as in Track 1; for Chinese→ English, terms may be one-to-many mapped, for instance, the target can have both full entity names and acronyms. This track tests terminology accuracy and consistency in long-form translation, reflecting real-world professional needs.

2.2 Terminology Modes

To enable a causal analysis of terminology utility and translation quality, each system is requested to translate our tests under three modes separately:

- **No Terminology (No Term):** The system translates the input text without a terminology dictionary.
- **Proper Terminology (Proper Term):** The system is provided with a dictionary of domain-specific terms; relevant to the input.
- **Random Terminology (Random Term):** The system receives a dictionary of randomly selected source words and their aligned translations from the references (the random pool excludes the proper terms).

The use of Random Term mode allows for measuring if improvements in translation quality stem from incorporating terminologies or from simply seeing part of the correct translation (Zouhar, 2023; Semenov et al., 2023).

3 Data

The data for both tracks[2] is provided in the `jsonl` format, where each instance has the following entries. See Table 1 for an example.

- Text in language 1
- Text in language 2
- Real terminology mapping dictionary (Proper)
- Random word mapping dictionary (Random)
- Dummy empty dictionary (NoTerm)

3.1 Track 1: Sentence-level IT Documentation

Data source. Our sentence-level data was produced by SAP to investigate ambiguous terminology in the IT business domain (Berger et al., 2025).[3] The data originates from their online help

[1] Traditional; henceforth only "Chinese" for brevity.

[2] github.com/wmt-conference/wmt25-terminology/tree/main/ranking/references

[3] github.com/SAP/software-documentation-data-set-for-machine-translation

```
{
    "en": "Open the consumption model containing
         the measures and attributes you want to
         include in your perspective, and click
         the Perspectives tab.",
    "de": "Offnen Sie das Verbrauchsmodell mit
         den Kennzahlen und Attribute, die Sie in
          Ihre Perspektive aufnehmen möchten, un
         wechseln Sie zur Registerkarte
         Perspektiven.",
    "proper": {
        "consumption model": "Verbrauchsmodell"
    },
    "random": {
        "include": "aufnehmen",
        "want": "möchten"
    },
    "noterm": {}
}
```


Table 1: An English→German example from Track 1.

portal. Pages on the help portal are written in English, and translations are produced by post-editing machine translations to ensure proper terminology usage and adherence to corporate style guides.

Terminologies. SAP additionally maintains a one-to-many terminology dictionary across all of its production languages called SAPTerm.[4] Source and target terms were fuzzy-matched in the source and post-edited segments, with additional filtering performed to ensure post-edits made corrections of terminology usage. We make available a terminology dictionary for each segment pair, usually containing one or two entries.

The random terms were retrieved automatically by the following procedure: For every sentence pair (given input and reference translation), the pool of possible random terms is formed by all source sentence words except the terms and the stopwords (based on NLTK2024 stopword lists). Out of this pool, we sample as many words as there are in the corresponding proper dictionary. Then, we prompt ChatGPT to retrieve its translations from the reference sentence. We run an exact match search to ensure that the translation of the word is in the reference. The instruction for ChatGPT is provided in Appendix B.

Test set release. We select the English→ {German, Russian, Spanish} translation directions for the sentence track in our shared task. Per language pair, 1000 instances that contain terminology were sampled, and we split the 1000 segments into development and test sets consisting of 500 instances each. Each instance is accompanied by a terminology dictionary containing entries for that instance only. All test references are only made publicly available after the shared task submission deadline.

3.2 Track 2: Document-Level Finance

Data source. Our document track test data are sourced from the public annual reports on the official website of the Hong Kong Monetary Authority (HKMA), available in English and Traditional Chinese.[5] Each annual report contains multiple chapters, and each chapter is available to download as a standalone PDF file. In this task, we define such a chapter as a *document*. We collect all English and Traditional Chinese annual reports from 2015 to 2024. An annual report yields the same number of chapters (documents) in English and Chinese, and corresponding chapters are parallel to each other.

Document processing. We convert the chapter PDFs into markdown using MinerU (Wang et al., 2024), with table recognition, formula recognition, and optical character recognition disabled. We drop tables and formulae because they consist mostly of numbers without text, and are difficult to translate or evaluate. We then truncate each chapter to 2000 whitespace-delimited English words to keep the documents at a reasonable length for participants. Then, three authors, who are native Chinese speakers fluent in English, manually inspected and processed the markdown files. This includes truncating the Chinese side and re-aligning Chinese and English at the paragraph level in order to fix errors as a result of the automatic processing of the chapter PDFs, which are in a two-column format. As a result, each chapter (document) pair has the same number of lines in both languages.

Terminology extraction and mapping. We extract terms specific to Hong Kong finance from the source and target documents and establish a mapping via two stages. First, we prompt GPT-4.1 with a pair of source and target documents to automatically identify and align terminologies in the two languages, producing a preliminary mapping. Second, two authors independently review the generated mapping and correct it as necessary. Most revisions are removals of relatively generic named entities (e.g., Hong Kong and US dollar) that already have standard translations. The prompt used

[4] sapterm.com

[5] www.hkma.gov.hk/gb_chi/data-publications-and-research/publications/annual-report/

in this first stage is provided in Appendix A for reference. To extract a word mapping for the Random Term mode, we reuse the same technique from the sentence-level Track 1. To better approximate a real-world scenario, we merge the extracted mappings from the chapters (documents) within the same report and generate report-level mappings for both Proper and Random Term modes.

It is worth noting that the HKMA website provides a glossary for Chinese and English separately.[6] However, we did not use it because the Chinese and English lists are independently ordered, are not index-aligned, and contain different numbers of entries. Consequently, it is difficult to construct positional correspondence between entries in the two languages. This can be explored by future work.

Test set release. We release all document-level data we have prepared as the test set for this year's shared task. To avoid temporal bias and to ensure balanced representation of the translation directions, we partitioned the data so that translation direction alternates by year. Reports from odd-numbered years (2015, 2017, 2019, 2021, 2023) are used for English→Chinese tests, whereas those from even-numbered years (2016, 2018, 2020, 2022, 2024) are used for Chinese→English tests. We release each document as a single string, but paragraphs are delimited by `\n\n`, allowing participants to make their own chunking choices. With each test document, a large terminology dictionary is provided, containing mappings for all terminologies in the whole report (i.e., the dictionary is shared between all documents within the report).

4 Metrics and Evaluation

For both tracks, we run reference-based evaluation using gold translations and corresponding terminology dictionaries. We use three types of metrics in our shared task evaluation: overall quality (string match using BLEU and chrF2++; document-level MQM with LLM-as-a-judge), terminology accuracy, and terminology consistency.

This choice of metrics is motivated by two factors: (1) we wish to measure different aspects in translating terminologies, and (2) modern automated metrics can be less robust than a simple string-matching chrF, especially in domains that were not part of the metrics' training data (Lavie et al., 2025; Zouhar et al., 2024a,b).

New this year, we rank submissions based on the Pareto efficiency measured by the quality metric and the terminology. We provide terminology consistency scores as an analysis, and use the document-level AutoMQM scores for a separate ranking in Track 2, as this is not fully empirically validated.

4.1 Overall Translation Quality

BLEU and chrF2++. As an indication of overall quality, we run two reference-based metrics, BLEU (Papineni et al., 2002) and chrF2++ (Popović, 2017), as implemented in sacrebleu (Post, 2018) with default settings.[7,8] In the main paper we report chrF2++ which is tokenization-insensitive, allowing for consistent evaluation of German, Russian, Spanish, English, and Traditional Chinese outputs. Specifically to run the metrics in the document translation track, we treat each entire translation and reference document as a single string (O'Brien et al., 2025).

Doc-level AutoMQM. For document-level translation quality assessment, we use an LLM-as-a-judge: LLMs are prompted to identify translation error spans and assign severity levels, from which the final score is computed. This evaluation method is well interpretable and has been shown to correlate well with human judgment (Kocmi and Federmann, 2023; Freitag et al., 2023, 2024). An extension to the document level is focus-sentence prompting (FSP), which evaluates documents sentence by sentence (Domhan and Zhu, 2025). In FSP, the judge model is provided with the full source and translation as context, along with the specific target sentence to be evaluated. To evaluate document-level translation in Track 2, we use GPT-4o and GPT-5 as judge models, applying the FSP prompt with two modifications: (1) we evaluate three consecutive sentences at a time to improve efficiency; and (2) we provide the judge model with a terminology mapping to better assess translation quality. Details of the judge prompt are provided in Appendix C. Once the model outputs the errors and their severities, we compute the final MQM score for each annual report as a weighted average over the severity levels. We define three categories of severity: minor, major, and critical, with weights

[6] E.g. www.hkma.gov.hk/eng/data-publications-and-research/guide-to-monetary-banking-and-financial-terms/

[7] BLEU|#:1|c:mixed|e:no|tok:{13a,zh}|s:exp|v:2.4.1

[8] chrF2++|#:1|c:mixed|e:yes|nc:6|nw:2|s:no|v:2.4.1

$\text{count}^{\text{src}} \leftarrow 0, \text{count}^{\text{tgt}} \leftarrow 0$
for $\text{src}_i, \text{tgt}_i, d_i \in X$ **do**
  for $\text{term}_j^{\text{src}}, \text{term}_j^{\text{tgt}} \in d_i$ **do**
    if $\text{term}_j^{\text{src}} \in \text{src}_i$ **then**
      $\text{count}^{\text{src}} \leftarrow \text{count}^{\text{src}} + 1$
      if $\text{term}_j^{\text{tgt}} \in \text{tgt}_i$ **then**
        $\text{count}^{\text{tgt}} \leftarrow \text{count}^{\text{tgt}} + 1$
if $\text{count}^{\text{src}} > 0$ **then**
  return $\text{count}^{\text{tgt}}/\text{count}^{\text{src}}$
else
  return 0

Algorithm 1: Terminology Accuracy (Track 1: sentence-level). Input X is a list of source, translation, terminology dictionary triplets $\langle(\text{src}_1, \text{tgt}_1, d_1), \ldots\rangle$.

$A \leftarrow \langle\rangle$ # accuracy for individual terms
for $\text{src}_i, \text{tgt}_i, d_i \in X$ **do**
  for $\text{term}_j^{\text{src}}, \text{Terms}_j^{\text{tgt}} \in d_i$ **do**
    if $\text{term}_j^{\text{src}} \in \text{src}_i$ **then**
      $\text{count}^{\text{src}} \leftarrow \text{src}_i.\textsc{Count}(\text{term}_j^{\text{src}})$
      $\text{count}^{\text{tgt}} \leftarrow 0$
      for $\text{term}_{j,k}^{\text{tgt}} \in \text{Terms}_j^{\text{tgt}}$ **do**
        $\text{count}^{\text{tgt}} \leftarrow \text{count}^{\text{tgt}} + \text{tgt}_i.\textsc{Count}(\text{term}_{j,k}^{\text{tgt}})$
      $A.\textsc{Append}(\textsc{Min}(\frac{\text{count}^{\text{tgt}}}{\text{count}^{\text{src}}}, 1.0))$
if $|A| \neq 0$ **then**
  return $\frac{\sum A}{|A|}$
else
  return 0

Algorithm 2: Terminology Accuracy (Track 2: document-level). Input X is a list of source, translation, terminology dictionary triplets $\langle(\text{src}_1, \text{tgt}_1, d_1), \ldots\rangle$.

of 1, 5, and 10, respectively. For example, an MT system receives an MQM score of 25 if it produces three major errors and one critical error.

4.2 Terminology Accuracy

We also evaluate how accurately translation systems can convert terms based on a given dictionary. We reckon that a source term usually occurs only once in a sentence input, but is more likely to appear multiple times in a document. Thus, we use different implementations for the two tracks as detailed below.

In the sentence track, for each source term appearing in the input text, we check if its corresponding target term appears in the translation, yielding a binary score. The accuracy is then computed as the sum of successful conversions divided by the total number of source words across all input instances. The algorithm is illustrated in Algorithm 1.

At the document level, the accuracy measure is modified to account for: (1) a source word can appear multiple times and thus a target word is expected as many times; (2) potential one-to-many mappings in a dictionary. The metric moves from a binary check to a percentage score. For each source term present in the source document, we calculate a ratio determined by the total number of appearances of all its possible target terms in the translation, divided by the total number of appearances of the source term itself. This ratio for each term is capped at 1 to avoid false positives, and the final document-level terminology accuracy is the average of all individual ratios across all source terms across all documents. The algorithm is shown in Algorithm 2.

The main difficulty of checking terminology accuracy lies in the terminology dictionary usually containing source and target entries in their stem form, but for many languages, we need to capture the inflected forms of the entries too. Hence, when we need to check whether a word is in a segment or count the number of appearances of the word, we always employ a two-pronged matching strategy. First, we run a direct surface-form match between the word and the segment. Second, to account for morphological variations, we check the lowercased lemmatized word against the lowercased lemmatized segment.[9] The final result is the higher value resulting from the two matching strategies—this applies to both binary outcomes or counts.

4.3 Terminology Consistency

We use the framework for the term consistency metric suggested by Semenov and Bojar (2022). The framework allows for automated (and more interpretable than LLM-as-a-judge) evaluation on how consistent the models are when choosing the translation of specialized terms. The modular structure of the framework allows for different levels of strictness in evaluation, so for the current shared task, we focused on two versions of the metric based on term frequency and the dictionary. As illustrated in Algorithm 3, the evaluation consists of the following steps:

- **Preprocessing:** This step requires sentence-level or paragraph-level alignments. Track 1 data already meets this; for Track 2, since the input texts have clear separation between paragraphs (double newline characters), we split the system outputs into segments accordingly.[10]

[9] github.com/stanfordnlp/stanza for lemmatization.

[10] For most systems, this simple preprocessing allowed for consistent alignment. The only exception was that for STITCH outputs, we additionally applied LaBSE embeddings (Feng et al., 2022) to align the split segments.

- **Source term selection:** At this step, we retrieved the subsets of terms present in a given segment. For Track 1, terminology dictionaries already meet the requirement. For Track 2, we filter the document-level dictionary to construct a segment-level dictionary for each segment using a substring match for Chinese and an exact match over lemmatized text for other languages.
- **Term translation alignment:** We then locate the exact part of the output that is a translation of the source term. The most effective way appeared to be few-shot prompting ChatGPT with additional post-processing, with details in Appendix E. We refer to the aligned term translations as "candidates".
- **Pseudo-reference choice:** To estimate the consistency of a system, we need to define "pseudo-references": translations against which we compare candidates. For the main analysis, we choose a frequency criterion: For each source term type, we order the candidate types by their frequency, and define the most frequent one as a pseudo-reference. Notably, this choice is insensitive to the term accuracy: the pseudo-reference may not be the best translation, but it should be used stably over the whole text. In an additional experiment in Appendix F, we also try another pseudo-reference option based on the Proper Term dictionary.
- **Evaluation:** For each term occurrence in each text segment, we check whether the observed translation candidate differs from the pseudo-reference. The final score is formalized as a multi-class accuracy: for each source term type (class), we count the percentage of the candidates matching the pseudo-reference in the submitted texts and run macro-averaging over the class percentages. As a result, we get a score within a range of 0 to 1, which shows the percentage of occurrences of the term translation that are consistent with the chosen pseudo-reference.

for $\text{src}_i, \text{tgt}_i \in X$ **do** # Source term selection & align.
  for $\text{term}_j \in \textsc{SrcTermSelect}(\text{src}_i)$ **do**
    $\text{cand}_j \leftarrow \textsc{Aligner}(\text{src}_i, \text{tgt}_i, \text{term}_j)$
    $\text{CandDict}_{\text{term}_j,\text{cand}_j} \leftarrow \text{CandDict}_{\text{term}_j,\text{cand}_j} + 1$
    $\text{AlgDict}_{i,\text{term}_j} \leftarrow \text{cand}_j$
$\text{PseudRefDict} \leftarrow \{\}$ # Pseudo-reference choice
for $\text{term}_k \in \text{CandDict}$ **do**
  $\text{PseudRefDict}_{\text{term}_k} \leftarrow \textsc{AssignPseudoRef}(\text{term}_k)$
for $\text{src}_i, \text{tgt}_i, \text{AlgDict}_i \in X$ **do** # Scoring
  for $\text{term}_j \in \text{AlgDict}_i$ **do**
    $\text{hit}_{i,j} \leftarrow \mathbb{1}[\text{term}_j = \text{PseudRefDict}_{\text{term}_j}]$
return $\sum_{k \in \text{PseudRefDict}} \frac{\sum \text{hit}_k}{|\text{hit}_k|}$ # Macro-average

Algorithm 3: Terminology Consistency (Track 1: sentence level) with pseudo-reference initialization of the most frequent terms. Input X is a list of source-translation pairs $\langle(\text{src}_1, \text{tgt}_1), \ldots\rangle$.

5 Participants and System Descriptions

This year, apart from our baseline, we see 20 systems in Track 1 and 4 systems in Track 2. Their descriptions are provided below. For an easier navigation over the variety of approaches, we label them with the main features and components of particular submissions, namely:

- models used:
  - NMT NMT model (encoder-decoder)
  - LLM LLM (decoder-only model)
  - multiple models (agents, preprocessing+postprocessing, etc.)
- training data:
  - DatAug data augmentation
  - DatCur data curation (filtering big corpora, enriching training data with annotation, etc.)
- model update techniques:
  - *FT fine-tuning, continuous pre-training, supervised fine-tuning, etc.
  - *PO various types of preference optimization: GRPO, PPO, DPO, etc.
- inference-time strategies:
  - code-switched prompts
  - ICL in-context learning, few-shot prompts, etc.
  - multi-metric decoding (using both general quality and term accuracy for sequence choice)
  - term injection (for NMT models)

o3-term-guide LLM The participant put terminology constraints in the form of explanatory statements and presumably prompted o3 from OpenAI.

DuTerm NMT LLM DatAug *FT This is a two-stage algorithm for terminology translation (Jaswal, 2025). It uses a terminology-aware NMT model fine-tuned from NLLB 3.3B (Costa-Jussà et al., 2022), and prompts GPT-4o for post-editing. To construct the NMT training data, they first extract bilingual terminology dictionaries from WMT25 dev sets, which are then supplemented with terminologies generated by the LLM. Then they use an LLM to synthesize parallel sentences

containing one or more terminologies. Specifically, the terms in both source and target sentences are bounded with special tags for identification. After filtering the training data for quality with COMET-QE and other rules, the NMT undergoes terminology-aware fine-tuning. Given the source, the NMT's translation and term pairs, they prompt GPT-4o to refine the translation for better fluency while keeping the constraints.

Erlendur LLM ICL Ingólfsdóttir et al. (2025) presented an LLM-based translation system using a pipeline approach that combines prompting with modular preprocessing and postprocessing components. In a preparatory stage, the LLM analyzes the source text to extract key terms and idioms, which are then matched with entries from bilingual dictionaries; user-provided glossaries can also be incorporated to enforce consistent terminology. After translation, additional post-processing steps may be applied. For example, a custom seq2seq grammatical error correction model is used to improve Icelandic translations. The system participated in both terminology tracks: for Track 1, it employed its standard pipeline with terminology mapping, while for Track 2, the backbone model was switched from Claude 3.5 to GPT-4.1, as the former refused to translate some test examples.

ISMT-TiU (TiUTerm-V0, TiUTerm-V1) LLM ICL The team submitted two systems, both relying on in-context learning of LLMs, with few-shot examples retrieved with BM25 from the dev set. TiUTerm-V0 is Llama-3-8B (Grattafiori et al., 2024) with 10 in-context examples; TiUTerm-V1 is XGLM-7.5B (Lin et al., 2022) with 12 in-context examples.

Barcelona Supercomputing Center (tower, salamandrata) LLM DatCur *PO Garcia Gilabert et al. (2025) submitted two models: Tower based on Llama2-7B (Touvron et al., 2023) and salamandrata based on Salamandra-7B (Gonzalez-Agirre et al., 2025). They use a novel approach of fine-tuning terminology translation using GRPO (Shao et al., 2024). Specifically, they introduce a terminology adherence reward, which penalizes outputs that do not contain the correct terminology. The training data are based on pseudo-terminology mined heuristically using named entities, noun phrases, and adverbial constructions. The reward during training is joined with a general MT quality reward using a quality estimation model.

IRB-MT (MeGuMa) LLM The submitted system, named MeGuMa, uses LLM agents with terminology-aware translation prompts, in combination with two MT metrics for per-translation-unit solution selection: MetricX (Juraska et al., 2023) and a custom approximation of terminology accuracy which uses an explicit alignment system by Steingrimsson et al. (2023) beyond surface mapping of lemmatized terms. Translation is done in two phases: translation and revision. Models used for translation are taken from three families: Gemma 3 (27B and 12B) by Team et al. (2025), Qwen3 (14B-thinking, 8B-thinking, 14B, 8B) by Yang et al. (2025), and EuroLLM (9B) by (Martins et al., 2025). In the second phase, three models are used to revise all translations: Gemma 3 27B with thinking, Gemma 3 12B with thinking, and Qwen 3 14B with thinking. While Qwen 3 supports thinking natively but not Gemma 3, all of the revision models were prompt-induced to first think and then produce the final solution. The final translation is selected from all of the generated solutions, both initial and revised. The selection criterion is an arithmetic mean of the arithmetic, geometric, and harmonic means of MetricX (Juraska et al., 2024) and terminology accuracy.

CommandA-WMT LLM DatCur *PO Kocmi et al. (2025a) submitted a post-trained version of Command A from Cohere. The data contains a mix of the languages the model was originally trained on, as well as machine translation data in new languages. The data was heavily filtered for quality. This was followed by preference tuning with a bespoke MTExpert dataset for all languages.

BIT LLM *PO Based on Qwen3-8B-Instruct, the participants used the PPO algorithm to perform reinforcement learning according to the terminology accuracy of the model's outputs without using any Dev Data.

Laniqo LLM DatCur *FT ICL Guttmann et al. (2025) use the EuroLLM-9B-Instruct model (Martins et al., 2025) as a foundation. Given an explicit dictionary, the terms in the source sentence in a prompt are substituted with the target language translations of it, creating a code-switched sentence. Additional prompt engineering analysis showed that two-shot prompts were the most efficient. The second modification was fine-tuning on augmented data from OPUS (Tiedemann

and Nygaard, 2004), where the randomly selected source text nouns and verbs were aligned with their translations, and were replaced with them in the same way as the prompts for the LLM. Fine-tuning was conducted with LoRA (Hu et al., 2022). The decoding is done with constraints: inspired by our announced metrics, the Pareto frontier between overall quality and term accuracy, they used epsilon sampling of 100 sentences, followed by multi-dimensional ranking by various QE metrics and term accuracy of a given segment. This approach was named Pareto-Optimal Decoding. The ablation experiments showed that the best scores were achieved by combining a fine-tuned model together with a modified prompt and few-shot examples.

Lingua Custodia (LC-primary, LC-2, LC-3) LLM DatCur *FT *PO Liu et al. (2025) submitted three systems in total: LC-primary, LC-2, and LC-3. They first filtered bilingual data from Common Crawl and WMT25 using LaBSE and applied an unsupervised terminology extraction approach, developed in their 2023 terminology task submission (Liu et al., 2023), to create terminology mappings. They then fine-tuned open-weight and efficient LLMs, Qwen3-4B (Yang et al., 2025) (with thinking mode disabled) and Gemma-3-4b-it (Team et al., 2025). They conducted supervised fine-tuning in the first stage and then applied GRPO in the second stage, using a sentence-level BLEU reward for overall quality and a constraint-following reward for terminology adherence.

CurTermNLLB NMT DatCur *FT Gonzalez-Gomez (2025)'s system is based on LoRA (Hu et al., 2022) fine-tuning of NLLB 200M (Costa-Jussà et al., 2022) on the consumer-grade Apple M3 with an automated pipeline for creating terminology containing data similar to the Track 1 dataset. They select data from OPUS (Tiedemann and Nygaard, 2004), specifically data from the `GNOME`, `KDE4`, and `WikiMatrix` projects. Sentence pairs from this subset of data were embedded with `all-mpnet-base-v2` in Sentence-Transformers (Reimers and Gurevych, 2019), and cosine similarity to Track 1 dev set sentence pairs was computed for filtering together with filtering based on part-of-speech. Source terminology was then aligned to target sentences to create a term dictionary; relevant dictionary entries were provided to the NLLB model as additional input.

UW-BENMT (ContextTerm) NMT DatCur Pong (2025) submitted a system named ContextTerm, a Transformer-based NMT model (roughly Transformer-base size) with terminology-aware data augmentation. The system identifies terminology constraints by selecting source–target alignments whose source words are judged most "important" by the encoder (measured via the norm of their hidden-state vectors) rather than merely low-frequency ones. Training data combined the IT-specific parallel corpora selected with Cross-Entropy Difference filtering and 30k synthetic English sentences generated using Aya-Expanse-8b (Dang et al., 2024), with inline soft constraints applied to 10% of the data.

Multitan (Systran-ft, EuroLLM-ft, MarianMT-ft) LLM NMT DatAug *FT The participants submitted three systems. The general approach was fine-tuning on in-domain data. Specifically, for Systran-ft, the authors used Systran Model Studio Lite to fine-tune Systran's baseline model with augmented in-domain data. For EuroLLM-ft, EuroLLM was updated with in-domain aligned segments and glossary by using LoRA (Hu et al., 2022). For the third system, MarianMT-ft, the team used two fine-tuning strategies: in No Term mode, using the dev set and other in-domain aligned segments; in Proper Term mode, in addition to fine-tuning the model with in-domain segments, they used a glossary for hard-forced training.

TranssionMT *FT This participant used training constraints and post-processing constraints to improve terminology translation accuracy.

STITCH LLM ICL The participants aimed at solving a recently highlighted problem of adding overly large context into prompts. The proposed method is named STITCH, which stands for Structured Terminology Integration for Translation with Context Handling. STITCH makes use of the observation that long-form documents are coarsely aligned on a paragraph-level and injects local terminology context in-flight during generation, while removing already integrated terminology information from the prompt. The approach leads to a task decomposition, allowing the model to perform document-level translation while being guided by local terminology information.

Baseline■ (GPT-4.1-nano) LLM The organizers prepared a baseline approach by querying GPT-4.1-nano (`2025-04-14`) with a long prompt containing

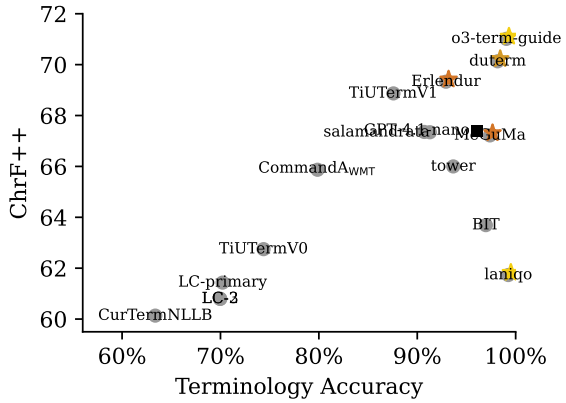

Figure 1: Tradeoff between quality (chrF++) and terminology accuracy in the proper terminology mode, averaged across three directions in Track 1. Top-performing systems are labelled as rank 1★, rank 2★, and rank 3★ according to Pareto optimality.

the input sentence or document and the entire terminology dictionary, when applicable.

6 Shared Task Results

Main results are presented in Table 2 for the sentence-level IT documentation and Table 3 for financial documents. For all three dictionary modes (Proper Term, Random Term, and No Term), the terminology accuracy is *always* measured with respect to the proper terminology dictionary. In addition to chrF++ and terminology accuracy used for system ranking, we also supply terminology consistency, which measures the consistency of the translated terminologies.

6.1 Ranking

We rank all systems in each track by a Pareto efficiency between translation quality (chrF++) and terminology accuracy in the Proper Term mode, where the systems translate real terminologies. For both Track 1 and Track 2, we average chrF++ and term accuracy across all directions for comparison.

Figure 1 visualizes the two-dimensional results for Track 1 systems. In total, there are 5 top-performing systems labelled by ★'s, namely, rank 1 (**o3-term-guide** and **laniqo**), rank 2 (**duterm**), and rank 3 (**Erlendur** and **MeGuMa**), according to Pareto optimality. In Track 2, among the 4 participants, **Erlendur** and **CommandA$_{WMT}$** are at the frontier.

6.2 Translation Quality and Terminology Accuracy

In both tracks, the quality differences between top systems are marginal, as indicated by chrF++ scores, which are usually within 1 chrF++ difference. Terminology handling exhibits sharper contrasts across the two tracks. In the sentence-level Track 1, strong systems achieve very high terminology accuracy of above 97%, implying that state-of-the-art translators can almost perfectly adhere to a few terminology constraints at the sentence level. This also implies that sentence-level terminology translation, as a rather artificial task, lacks difficulty for modern systems. By contrast, the document-level Track 2 exposes the harder challenge, as we observe that accuracy scores drop to the 70–80% range. When many terms need to be translated throughout longer contexts, systems frequently fall short.

Terminology accuracy is not uniform across translation directions. In Track 1, the best systems have similar (and very high) terminology accuracy for the three languages, but a few other systems show some divergence. Our GPT-4.1-nano baseline attains good accuracy when translating into Spanish, but not German or Russian. CommandA$_{WMT}$ and CurTermNLLB show markedly lower accuracy for Russian terminologies compared to Spanish and German. In Track 2, Erlendur delivers better quality and accuracy for English→Chinese, whereas CommandA$_{WMT}$ leads in Chinese→English.

The four systems that entered both tracks (Erlendur, MeGuMa, CommandA$_{WMT}$, and our baseline GPT-4.1-nano) enable us to compare terminology handling under different input lengths and terminology constraint loads. Given longer documents and more terminologies, we expect terminology accuracy to be lower in Track 2. However, CommandA$_{WMT}$ maintains accuracy comparable to, if not better than, its Track 1 results. Moreover, MeGuMa's accuracy is stable across the three languages in Track 1 but shows a dramatic gap of over 30 points between English→Chinese and Chinese→English in Track 2. Nonetheless, we note that the domain and translation direction also changed, which may cause the observed pattern.

Finally, as shown in Figure 1 earlier, there is no clear quality-accuracy tradeoff for many systems; chrF++ and terminology accuracy tend to rise (or drop) together. Outliers such as MeGuMa, tower, BIT, and laniqo lean towards optimizing for termi-

System	Proper, ChrF Avg	Es	De	Ru	Proper, Acc. Avg	Es	De	Ru	Proper, Cons. Avg	Es	De	Ru	Random, ChrF Avg	Es	De	Ru	Random, Acc. Avg	Es	De	Ru	NoTerm, ChrF Avg	Es	De	Ru
o3-term-guide	71.0	75.9	71.6	65.6	99.1	99.1	99.1	99.0	87.7	86.7	86.1	90.4	68.1	72.4	69.4	62.4	49.2	50.7	52.3	44.6	63.6	69.5	64.7	56.6
duterm	70.1	76.1	70.7	63.6	98.2	98.7	98.2	97.6	87.3	86.0	86.3	89.5	66.4	72.1	67.2	59.8	46.6	48.8	48.4	42.4	61.6	67.0	62.6	55.3
Erlendur	69.3	74.8	69.9	63.3	92.9	94.4	93.2	91.2	86.7	83.8	86.3	90.0	66.4	71.6	67.6	59.8	44.4	47.1	47.1	38.9	62.6	68.1	64.0	55.6
TiUTermV1	68.9	77.1	65.7	63.8	87.6	89.4	87.3	86.1	86.7	85.7	85.9	88.5	66.8	74.2	64.4	61.8	54.6	59.2	56.7	47.9	64.4	72.4	61.9	58.9
GPT-4.1-nano■	67.4	72.4	67.4	62.3	90.7	95.2	89.0	88.0	87.5	86.3	86.3	90.0												
salamandrata	67.3	72.0	69.6	60.4	91.3	92.7	91.7	89.4	87.4	87.3	86.4	88.6	64.7	69.3	66.2	58.5	48.2	53.1	48.1	43.4	62.0	67.2	64.0	54.7
MeGuMa	67.2	72.0	67.7	61.9	97.4	97.0	96.3	98.8	88.6	86.9	88.6	90.2	64.5	70.3	64.2	59.0	46.7	53.1	46.4	40.5	58.9	65.2	59.4	52.1
tower	66.0	74.0	65.9	58.1	93.7	95.0	94.8	91.2	88.4	87.6	86.8	90.7	63.8	71.2	63.0	57.1	44.3	48.6	45.7	38.5	60.9	68.6	61.2	53.0
CommandA$_{WMT}$	65.9	70.7	67.6	59.3	79.9	81.9	86.9	70.7	86.6	84.5	87.5	87.8	63.7	68.4	65.0	57.6	45.8	49.3	48.1	40.1	60.7	65.5	62.2	54.4
BIT	63.7	69.8	62.4	58.9	97.0	96.3	98.0	96.7	87.8	86.8	86.9	89.8	65.7	67.2	66.3	63.5	80.5	47.5	97.4	96.5	66.5	69.8	66.3	63.5
TiUTermV0	62.7	69.0	61.0	58.3	74.4	75.2	71.1	76.8	86.4	85.0	85.6	88.6	61.0	68.1	59.1	55.8	49.6	54.2	49.9	44.8	60.2	68.0	57.9	54.6
laniqo	61.7	68.5	59.8	56.9	99.3	98.7	99.4	99.6	87.6	85.6	89.3	87.9	60.2	66.3	59.5	54.8	42.7	46.9	43.5	37.7	55.0	60.3	55.5	49.4
LC-primary	61.4	68.9	61.2	54.2	70.2	74.1	70.7	65.8	85.4	83.6	85.8	87.0	61.0	68.1	59.7	55.2	38.6	43.8	37.4	34.6	57.5	65.0	56.9	50.5
LC-2	60.8	67.7	61.0	53.7	70.0	73.6	70.7	65.6	85.8	85.4	85.7	86.2	60.5	67.1	59.5	54.9	38.5	43.4	37.4	34.6	56.9	64.1	56.8	49.9
LC-3	60.8	67.7	61.0	53.7	70.0	73.6	70.7	65.6	86.0	85.6	85.7	86.7	60.5	67.1	59.5	54.9	38.5	43.4	37.4	34.6	56.9	64.1	56.8	49.9
CurTermNLLB	60.1	69.1	60.3	51.0	63.4	76.5	79.0	34.6	88.0	87.5	87.6	88.8	58.8	67.4	58.0	50.8	36.1	44.1	31.7	32.6	55.6	65.6	52.8	48.4
ContexTerm	48.5	53.7	40.2	51.5	72.0	68.5	79.9	67.6	81.9	75.6	85.8	84.4	48.2	52.0	40.7	51.7	24.6	20.5	18.6	34.8	45.7	50.2	37.4	49.4
Systran-ft		71.1				44.1				88.1				71.1				44.1				71.1		
MarianMT-ft		65.6				17.5				54.1				68.9				48.8				68.9		
EuroLLM-ft		63.5				38.9				82.5				63.5				38.9				63.5		
TranssionMT				47.8				33.2				90.1				47.8				33.2				47.8

Table 2: Main results for Track 1: sentence-level IT documentation terminology-informed translation.

System	Proper, ChrF Avg	EnZh	ZhEn	Proper, Acc. Avg	EnZh	ZhEn	Proper, Cons. Avg	EnZh	ZhEn	Random, ChrF Avg	EnZh	ZhEn	Random, Acc. Avg	EnZh	ZhEn	NoTerm, ChrF Avg	EnZh	ZhEn
Erlendur	60.2	46.1	74.2	78.7	85.4	71.9	92.0	91.6	92.3	57.9	41.8	74.0	64.9	60.1	69.6	57.4	40.8	74.0
CommandA$_{WMT}$	59.6	43.6	75.5	83.6	78.9	88.3	91.5	90.1	93.0	56.7	39.8	73.7	58.8	52.1	65.4	54.9	36.9	72.9
MeGuMa	54.3	39.1	69.4	79.5	96.6	62.4	90.8	93.3	88.3	48.4	31.6	65.2	47.7	43.9	51.5	51.0	33.7	68.3
STITCH	53.4	37.5	69.3	72.8	70.9	74.8	87.4	87.2	87.6	49.9	31.1	68.8	46.9	39.5	54.4	47.5	31.8	63.1
GPT-4.1-nano■	47.9	31.6	64.1	54.7	51.6	57.9	81.9	80.3	83.5	46.5	29.1	63.9	43.8	37.6	50.0	46.1	28.6	63.7

Table 3: Main results for Track 2: document-level finance terminology-informed translation.

nology accuracy, at variable costs in chrF++. A possible explanation for that, at least for MeGuMa and laniqo, can be that the multi-metric optimization used by the authors tends to favor the terminology-specific metrics.

6.3 Terminology Consistency

Tables 2 and 3 show that, contrary to the general MT quality and success rate scores, the spreads of the consistency scores in the Proper Term mode are relatively small, ranging from 0.81 to 0.92. This shows that the models are quite stable in choosing the translations of the specific terms. The performance of the models in Track 2 is overall higher than that of Track 1: the score of 0.87 is among the highest for sentence-level translation and the lowest for document-level translation. A possible reason for that can be the contextual dependency of the generated terms: in a document-level setup, a system attends to previously generated text, and it can be more prone to copying already generated sequences, while each occurrence of a term in a sentence-level setup is translated independently. This is indirectly supported by the observation of another version of the consistency metric: with the "first-seen" pseudo-reference choice. The absolute scores in both versions of the metrics, as well as their rankings, behave in a surprisingly similar manner: the absolute scores of the "first-seen" pseudo-reference initialization are stably lower compared to the "most frequent" initialization by 0.02 on average. This suggests that the first translation of the term would tend to be the most frequent over the document.

Another observation is that, in stark contrast to the terminology accuracy, the system scores are relatively robust to different types (and presence) of explicit terminology. Yet, as was noted for the two main metrics, the difference in consistency between the proper terminology and the two other modes becomes more pronounced in the higher-scoring systems. Such a trend, however, has exceptions: while it is true for English to Russian, German (sentence level), and Chinese (document level) sentence pairs, the English to Spanish and Chinese to English outputs do not show much difference over the whole range of the systems.

Finally, we should note that if the pseudo-

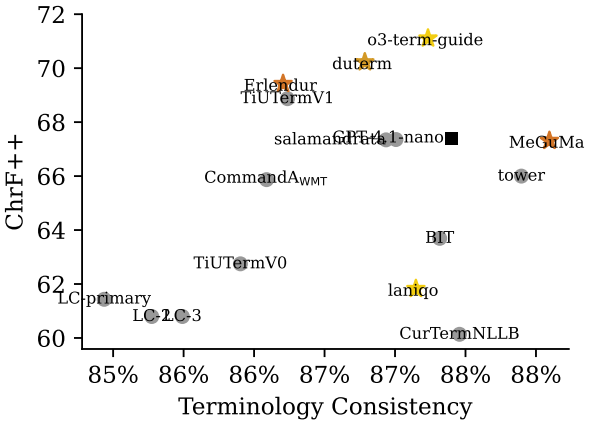

Figure 2: Relationship between quality (chrF++) and terminology consistency in the proper terminology mode, averaged across three directions in Track 1. Top-ranking systems are labelled as in Figure 1.

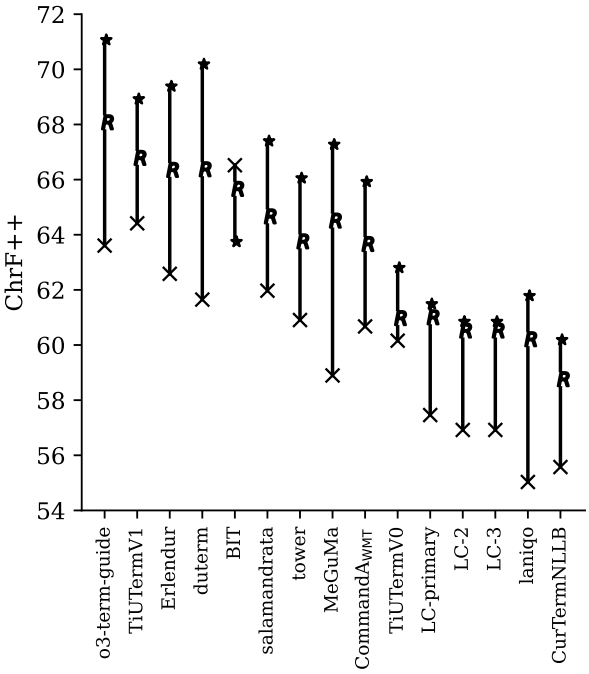

Figure 3: Effect of terminology mode on performance (measured by ChrF++) in Track 1. Legend: $\times$ denotes No Term, R denotes Random Term, and ★ denotes Proper Term.

references are initiated according to the proper terminology dictionaries, the absolute scores drop significantly (resulting in the range of 0.5 to 0.8), and the effect of the terminology mode becomes more pronounced. This is demonstrated in Appendix F, where the scores for each system in Random Term and No Term modes range between 0.2 and 0.4, while the proper terminology lies in a span of 0.5-0.8. Moreover, the difference between modes becomes more pronounced in higher-scoring systems. We conclude that the variant of the metric with dictionary-based pseudo-reference initialization may be more informative for the task of terminology translation, as it correlates with other metrics better and distinguishes between systems more clearly.

7 Analysis

Apart from reporting general translation quality, terminology accuracy, and terminology consistency, we analyze the terminology incorporation and metrics, hoping to provide insights to the community:

- A causal analysis of the impact of incorporating terminology on translation quality.
- A document-level AutoMQM using LLM-as-a-judge, all while considering a large terminology dictionary for Track 2.
- A study of the correlation between different metrics for Track 1.

7.1 Effect of Terminology Incorporation

We investigate the impact of incorporating terminologies on translation quality. Since providing a proper dictionary adds extra target-side information compared to using no dictionary, we also request participants to translate under a Random Term mode to enable a causal analysis. We plot the chrF++ scores under the three terminology modes on a vertical bar in Figure 3 for each participant in the sentence-level Track 1. This helps us easily inspect the quality gap between systems translating under different modes. It is clear that using a dictionary, either random or proper, consistently helps systems' translation quality. For top-performing systems, the benefit of using proper dictionary entries outweighs that of using a random dictionary, but for systems with a lower translation quality, there is no clear difference.

7.2 Doc-Level MQM Results

Table 4 presents the weighted MQM scores for each LLM judge, while Figures 4 and 6 visualize the distribution of error types and severities across the submitted systems in both the No Term and Proper Term settings for GPT-5 and GPT-4o, respectively.

Both LLM judges yield broadly consistent rankings among the top systems, with CommandA$_{WMT}$ and Erlendur generally outperforming MeGuMa and STITCH in terms of overall MQM scores.

	Weighted MQM score					
	GPT-4o			GPT-5		
System	**Avg**	**EnZh**	**ZhEn**	**Avg**	**EnZh**	**ZhEn**
Proper						
Erlendur	75.7	75.4	76.0	29.2	37.5	19.5
CommandA$_{WMT}$	81.2	77.6	84.8	37.7	41.0	34.5
STITCH	174.4	185.6	163.4	57.2	65.8	48.7
MeGuMa	166.8	151.2	182.2	85.1	69.5	100.5
Random						
Erlendur	76.0	81.2	71.0	35.0	49.4	20.8
CommandA$_{WMT}$	85.2	88.9	81.5	56.1	67.7	44.7
STITCH	90.5	89.6	91.3	74.4	83.3	65.8
MeGuMa	217.4	160.2	279.1	186.5	164.4	208.2
NoTerm						
Erlendur	74.0	74.6	73.5	33.1	44.4	22.1
CommandA$_{WMT}$	84.8	86.2	83.5	56.0	62.2	49.9
STITCH	104.9	110.4	99.5	90.4	103.3	77.8
MeGuMa	133.1	125.1	141.0	116.0	117.3	114.7

Table 4: Weighted MQM scores (lower is better, sorted ascending by GPT-5 Proper Avg), averaged over all, EnZh, and ZhEn documents in Track 2. Detailed counts for different error severities are presented in Appendix D.

Lower scores for the leading systems indicate fewer and less severe errors. However, while the overall patterns are similar, the judges diverge in the final ranking of the lower-performing systems in the Proper mode. Notably, GPT-5 tends to be more conservative, flagging fewer errors overall, whereas GPT-4o is stricter in its error identification.

Examining the error type distribution, Figure 4 (GPT-5) shows that most errors are classified as minor or major, with critical errors being relatively rare. The most frequent error types across all severity levels are accuracy, mistranslation, and terminology. The Proper terminology mode consistently reduces the number of terminology-related errors compared to the No Term mode, confirming the utility of providing domain-specific dictionaries.

This trend, however, is not consistently observed with GPT-4o (see Appendix Figure 6). Manual inspection revealed that GPT-4o occasionally produces false positives for terminology mismatch errors, sometimes flagging even exact matches as errors. As a result, we place greater reliance on the GPT-5 results for these outcomes. For future shared tasks, it may be beneficial to combine automated MQM with targeted manual review, or to further refine judge prompts to better accommodate acceptable variation in terminology use.

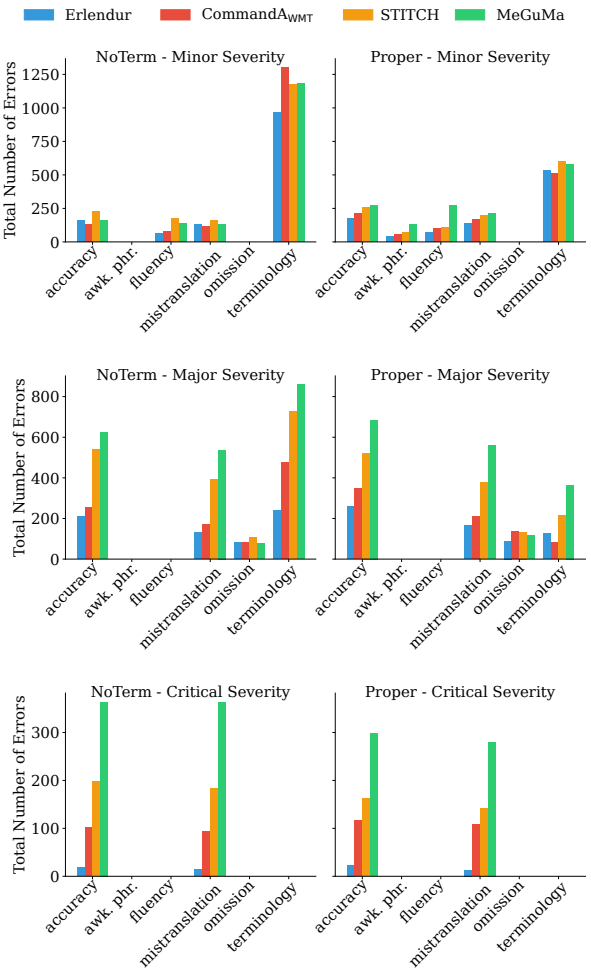

Figure 4: Distribution of error types and severities in No Term and Proper settings using GPT-5 as a judge. Total error counts indicate the number of errors each system made on the complete Task 2 test sets, comprising 10 annual reports across two translation directions. Error types with fewer than 100 occurrences across all severity levels and systems are omitted for clarity. "awk. phr." denotes awkward phrasing.

7.3 Ranking Correlation

One of the reasons for the slow progress in the field of terminology-aware translation is the lack of clarity on the best evaluation protocol(s) due to it being a multifaceted problem. We suggested several types of metrics—general quality, accuracy, and consistency—which differ significantly in the intended focus on the hypothesis and in their implementation. In this section, we analyze how differently or similarly those metrics rank the participating systems alone. This will give insights into the optimal choice of metric(s) in the future.

The comparison of the system ranking by different metrics was conducted using Kendall's τ (Kendall, 1938). We ran the correlation study only

	BLEU	ChrF++	Cons (Freq)	Cons (Dict)
ChrF++	0.81*	–		
Cons (Freq)	0.26	0.39*	–	
Cons (Dict)	0.20	0.36*	0.54*	–
Term Acc	0.22	0.39*	0.49*	0.96*

Figure 5: Kendall's τ correlation between various metrics used in the analysis, Track 1. Comparisons with $p < 0.05$ are marked with *.

for Track 1, since for Track 2, there are only 4 data points, so the τ scores will never surpass the thresholds of statistical significance. We visualize the correlation scores in Figure 5.

We can see two areas of high correlation. The first is well-known, between BLEU and ChrF++. Moreover, terminology accuracy and terminology consistency with dictionary-based pseudo-references display a near-perfect correlation. The reasons for this are clear: both metrics rely heavily on terminology dictionaries; therefore, they are aimed at finding and grouping the proper term occurrences in the system outputs. A surface-level (although lemmatized) match of translated terms is a good approximation of a more tailored system of term alignment between the input and output sequences. Therefore, considering the computational cost of running terminology consistency, future research could rely more on terminology accuracy, as it does not require any external alignment method.

The rest of the metrics show considerably weak positive correlation, in descending order: different types of consistency against term accuracy; and chrF++ against consistency and term accuracy. Although the correlation between the general quality and term-specific metrics is statistically significant, we still conclude that, as shown in Figures 1 and 2, the three types of metrics show different trends. For example, there is a 10 chrF++ gap between the best and second best system in terminology accuracy (Figure 1); top-ranking systems by chrF++ and terminology accuracy do not achieve the best terminology consistency. Therefore, using term-specific metrics is an important aspect of the evaluation of terminology-aware machine translation.

8 Related Work

Past shared tasks. Alam et al. (2021) introduced the first WMT shared task on MT using terminologies, focusing on the medical domain (including COVID-19 terminology) across five language pairs: English to French, Chinese, Russian, and Korean, as well as Czech to German. This pioneering effort established the foundation for systematic evaluation of terminology translation quality and consistency, with terminologies mined semi-automatically from parallel corpora. Building on this, Semenov et al. (2023) organized the second iteration in 2023, which expanded the range of domains (apart from medical texts, it included CL abstracts and web novels), while narrowing down the scope of translation directions: Chinese↔English, English↔Czech, and German↔English. Similar to the previous edition, their terminologies were mined semi-automatically, and they extended this line of work by contrasting random and proper terminologies. Their findings revealed that while incorporating terminology dictionaries led to improvements in translation quality, incorporating equivalent amounts of information from reference translations yielded similar results, challenging the prevailing assumption about terminologies being the crux of meaning in translation. Complementarily, Conia et al. (2025) organized the SemEval-2025 Task 2 on Entity-Aware Machine Translation, which focused on translating text containing complex named entities such as culture-specific titles, location names, and food names across 10 language pairs, introducing the XC-Translate benchmark with over 50K manually-translated sentences with entities that can deviate significantly from word-to-word translations.

Terminology translation test release. To the best of our knowledge, this shared task is among the few that release a high-quality terminology for translation in high-stakes domains such as IT and finance, with the exception of past shared tasks and a contemporary work (Oncevay et al., 2025a).

9 Conclusions

We now conclude the third iteration of the WMT Terminology Translation Task. In comparison to the 2021 and 2023 editions, this time we featured both sentence and document translation tracks with brand new data and domains. The former track ensured continuity, while the latter approximated real-life use cases better. We introduced an LLM-based document-level AutoMQM and used Pareto optimality to rank participants, but we kept the three

inference modes from 2023 for a causal analysis.

We attracted more than 20 submissions, three times more than the previous edition. The overwhelming majority used LLM-based solutions with different types of training techniques. This goes in line with a general trend in the machine translation field towards LLM-based solutions highlighted by Kocmi et al. (2024). Top-scoring systems in the sentence-level track reached good overall translation quality and nearly perfect term accuracy; the document track remains a more challenging task with respect to both metrics. The term consistency, on the contrary, shows a more stable behavior in both tracks, with overall higher scores for document-level MT. In terms of the inference modes, better systems benefit more from proper terminologies, while lower-scoring systems are less sensitive to dictionaries. Finally, we see high correlations between term-based metrics, but not between them and the overall quality, which highlights the necessity to keep at least one terminology-specific metric for this task.

Outlook. The lessons from the shared task also hint at the possible directions for its future iterations:

- Data: continue with document-level terminology translation evaluation
- Metrics: investigate suitable ranking measures and the trade-off between informativeness and computational costs of term-oriented metrics.
- Human evaluation: run human judgment on terminology translations and analyse its correlation with automatic scores. This, to our knowledge, has not been explored before.
- Language: extend the task to more, especially lower-resourced languages, while preventing contamination.

We are open to collaborations, and we especially welcome resources that can be used towards test sets or human evaluation. Stay tuned!

Limitations

The sentence-level test sets have been used in line with their original translation directions; for the document track, we are unsure of the original translation direction, so one of the two directions has the potential problem of translating translated/post-edited text back to its original language.

In terms of evaluation, while we have used several best metrics we can design, there could be some room for considerations and improvements: 1) document-level AutoMQM, especially with terminologies, has not been validated against human judgment; 2) although our terminology match runs lowercasing and lemmatization before string matching, it may not capture all occurrences of an intended word; and 3) certain correlation exists between metrics, e.g. surface string match and terminology match, so they are not fully orthogonal.

Finally, we used a quality-terminology tradeoff to rank participating systems, but as LLMs are more often deployed in practice, cost-effectiveness has become another important aspect.

Acknowledgments

We thank all participants for their submissions.

The document-level finance data used in this shared task is derived from the publicly available annual reports on the Hong Kong Monetary Authority (HKMA)'s website. We acknowledge HKMA as the source and owner of the reports, and we are grateful for the availability of these materials for research purposes.

Pinzhen Chen is supported by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee [grant number 10052546].

Vilém Zouhar gratefully acknowledges the support of the Google PhD Fellowship.

References

Md Mahfuz Ibn Alam, Ivana Kvapilíková, Antonios Anastasopoulos, Laurent Besacier, Georgiana Dinu, Marcello Federico, Matthias Gallé, Kweonwoo Jung, Philipp Koehn, and Vassilina Nikoulina. 2021. Findings of the WMT shared task on machine translation using terminologies. In *Proceedings of the Sixth Conference on Machine Translation*.

Mihael Arcan, Marco Turchi, Sara Tonelli, and Paul Buitelaar. 2017. Leveraging bilingual terminology to improve machine translation in a CAT environment. *Natural Language Engineering*, 23(5):763–788.

Nathaniel Berger, Johannes Eschbach-Dymanus, Miriam Exel, Matthias Huck, and Stefan Riezler. 2025. Learning to translate ambiguous terminology by preference optimization on post-edits. *arXiv preprint arXiv:2507.03580*.

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

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, and others. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *arXiv preprint arXiv:2412.04261*.

Georgiana Dinu, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. 2019. Training neural machine translation to apply terminology constraints. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Tobias Domhan and Dawei Zhu. 2025. Same evaluation, more tokens: On the effect of input length for machine translation evaluation using large language models. *arXiv preprint arXiv:2505.01761*.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, and others. 2024. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*.

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, and others. 2023. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation*.

Javier Garcia Gilabert, Carlos Escolano, Xixian Liao, and Maite Melero. 2025. Terminology-Constrained Translation from Monolingual Data using GRPO. In *Proceedings of the Tenth Conference on Machine Translation*.

Aitor Gonzalez-Agirre, Marc Pàmies, Joan Llop, Irene Baucells, Severino Da Dalt, Daniel Tamayo, José Javier Saiz, Ferran Espuña, Jaume Prats, Javier Aula-Blasco, and others. 2025. Salamandra technical report. *arXiv preprint arXiv:2502.08489*.

Mariano Gonzalez-Gomez. 2025. CurTermNLLB: Automatic Data Curation and Terminology-Aware Fine-Tuning of NLLB-600M. In *Proceedings of the Tenth Conference on Machine Translation*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and others. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Kamil Guttmann, Adrian Charkiewicz, Zofia Rostek, Mikołaj Pokrywka, and Artur Nowakowski. 2025. Laniqo at WMT25 Terminology Translation Task: A Multi-Objective Reranking Strategy for Terminology-Aware Translation via Pareto-Optimal Decoding. In *Proceedings of the Tenth Conference on Machine Translation*.

Eva Hasler, Adrià de Gispert, Gonzalo Iglesias, and Bill Byrne. 2018. Neural machine translation decoding with terminology constraints. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank adaptation of large language models. In *International Conference on Learning Representations*.

Svanhvít Lilja Ingólfsdóttir, Haukur Páll Jónsson, Kári Steinn Aðalsteinsson, Róbert Fjölnir Birkisson, Sveinbjörn Þórðarson, and Þorvaldur Páll Helgason. 2025. Miðeind at WMT25 General Machine Translation Task. In *Proceedings of the Tenth Conference on Machine Translation*.

Akshat Jaswal. 2025. It Takes Two: A Dual Stage Approach for Terminology-Aware Translation. In *Proceedings of the Tenth Conference on Machine Translation*.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. MetricX-23: The Google submission to the WMT 2023 metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation*.

Maurice G. Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30:81–93.

Sejoon Kim, Mingi Sung, Jeonghwan Lee, Hyunkuk Lim, and Jorge Gimenez Perez. 2024. Efficient terminology integration for LLM-based translation in specialized domains. In *Proceedings of the Ninth Conference on Machine Translation*.

Tom Kocmi, Arkady Arkhangorodsky, Alexandre Berard, Phil Blunsom, Samuel Cahyawijaya, Théo Dehaze, Marzieh Fadaee, Nicholas Frosst, Matthias Galle, Aidan Gomez, and others. 2025a. Command-a-translate: Raising the bar of machine translation with difficulty filtering. In *Proceedings of the Tenth Conference on Machine Translation*.

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, and others. 2025b. Findings of the WMT25 general machine translation shared task: Time to stop evaluating on easy test sets. In *Proceedings of the Tenth Conference on Machine Translation*.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, and others. 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*.

Tom Kocmi and Christian Federmann. 2023. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation*.

Andis Lagzdiņš, Uldis Siliņš, Toms Bergmanis, Mārcis Pinnis, Artūrs Vasiļevskis, and Andrejs Vasiljevs. 2022. Open terminology management and sharing toolkit for federation of terminology databases. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*.

Alon Lavie, Greg Hanneman, Sweta Agrawal, Kanojia Diptesh, Chi-Kiu Lo, Vilém Zouhar, Frederic Blain, Chrysoula Zerva, Eleftherios Avramidis, Sourabh Deoghare, and others. 2025. Findings of the WMT25 shared task on automated translation evaluation systems: Linguistic diversity is challenging and references still help. In *Proceedings of the Tenth Conference on Machine Translation*.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, and others. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.

Jingshu Liu, Mariam Nakhlé, Gaëtan Caillout, and Raheel Qadar. 2023. Lingua custodia's participation at the WMT 2023 terminology shared task. In *Proceedings of the Eighth Conference on Machine Translation*.

Jingshu Liu, Mariam Nakhlé, Gaëtan Caillaut, and Raheel Qader. 2025. Lingua Custodia's participation at the WMT 2025 Terminology shared task. In *Proceedings of the Tenth Conference on Machine Translation*.

Pedro Henrique Martins, João Alves, Patrick Fernandes, Nuno M Guerreiro, Ricardo Rei, Amin Farajian, Mateusz Klimaszewski, Duarte M Alves, José Pombal, Nicolas Boizard, and others. 2025. EuroLLM-9B: Technical report. *arXiv preprint arXiv:2506.04079*.

Jiyoon Myung, Jihyeon Park, Jungki Son, Kyungro Lee, and Joohyung Han. 2024. Efficient technical term translation: A knowledge distillation approach for parenthetical terminology translation. In *Proceedings of the Ninth Conference on Machine Translation*.

Dayyán O'Brien, Bhavitvya Malik, Ona de Gibert, Pinzhen Chen, Barry Haddow, and Jörg Tiedemann. 2025. DocHPLT: A massively multilingual document-level translation dataset. In *Proceedings of the Tenth Conference on Machine Translation*.

Arturo Oncevay, Elena Kochkina, Keshav Ramani, Toyin Aguda, Simerjot Kaur, and Charese Smiley. 2025a. Translating domain-specific terminology in typologically-diverse languages: A study in tax and financial education. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*.

Arturo Oncevay, Charese Smiley, and Xiaomo Liu. 2025b. 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).

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*.

Benjamin Pong. 2025. UW-BENMT at WMT25 Terminology Translation Task: Contextually Selected Pseudo-Terminology Constraints for Terminology-Aware Neural Machine Translation in the IT Domain. In *Proceedings of the Tenth Conference on Machine Translation*.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.

Kirill Semenov and Ondřej Bojar. 2022. Automated evaluation metric for terminology consistency in MT. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*.

Kirill Semenov, Vilém Zouhar, Tom Kocmi, Dongdong Zhang, Wangchunshu Zhou, and Yuchen Eleanor Jiang. 2023. Findings of the WMT 2023 shared task on machine translation with terminologies. In *Proceedings of the Eighth Conference on Machine Translation*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

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*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Jörg Tiedemann and Lars Nygaard. 2004. The OPUS corpus - parallel and free: `http://logos.uio.no/opus`. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Chelo Vargas-Sierra. 2011. Translation-oriented terminology management and ICTs: Present and future. *Interdisciplinarity and languages: Current Issues in Research, Teaching, Professional Applications and ICT*, pages 45–64.

Bin Wang, Chao Xu, Xiaomeng Zhao, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Rui Xu, Kaiwen Liu, Yuan Qu, Fukai Shang, and others. 2024. MinerU: An open-source solution for precise document content extraction. *arXiv preprint arXiv:2409.18839*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Vilém Zouhar, Pinzhen Chen, Tsz Kin Lam, Nikita Moghe, and Barry Haddow. 2024a. Pitfalls and outlooks in using COMET. In *Proceedings of the Ninth Conference on Machine Translation*.

Vilém Zouhar, Shuoyang Ding, Anna Currey, Tatyana Badeka, Jenyuan Wang, and Brian Thompson. 2024b. Fine-tuned machine translation metrics struggle in unseen domains. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*.

Vilém Zouhar, Tereza Vojtěchová, and Ondřej Bojar. 2020. WMT20 document-level markable error exploration. In *Proceedings of the Fifth Conference on Machine Translation*.

Vilém Zouhar. 2023. Machine translation that peeks at the reference. Report.

A Automatic Terminology Extraction

We prompt GPT-4.1 (`gpt-4.1-2025-04-14`) to automatically extract terminology from the source text and align it with the terminology used in the reference translation. Our prompt is shown in Table 5. Extraction and mapping are performed at the document level. The mappings of documents from the same annual report are then merged into a single terminology mapping for the entire report. In the data release, each source document is accompanied by this report-level terminology mapping, simulating realistic scenarios in which terminology mappings are predefined for a specific domain or task. In such cases, the predefined terms may or may not appear in the source document requiring translation.

B Automatic Random Terminology Alignment

The random terms were aligned with the help of GPT-4o. First, we randomly sample the words from the sentence that are not in the set of the proper terms, then we prompt GPT-4o with the text demonstrated in Table 7. To avoid hallucinations, we post-check the target sentence on whether it contains a highlighted word.

C Focus-Segments Prompting (FSP)

Focus-Sentence Prompting (FSP) was originally proposed by Domhan and Zhu (2025) as a method for using LLMs as judges in long-form translation evaluation while mitigating length bias, which is the tendency of LLMs to underreport errors when evaluating an entire long translation in a single pass. Their approach evaluates one sentence at a time while still providing the entire source and target documents as context. Although effective, the original FSP is costly because it requires many inference calls. To reduce this cost, we introduce Focus-Segments Prompting, in which a segment of three sentences is evaluated at once. This modification reduces the computational cost of FSP by approximately a factor of three. In our meta-evaluation on the WMT'24 Metrics Shared Task data, Focus-Segments Prompting performed comparably to the original FSP.

Another modification we introduced is adapting FSP to better suit our terminology-focused task. We consider accurate terminology translation a key quality dimension that the LLM judge should evaluate. However, even an LLM judge may not always know the correct translations of certain terms. To address this, we provide the judge with a ground-truth terminology mapping for reference. Recall that our original mapping was report-level and included many terms that might not appear in the segment under evaluation. To avoid unnecessary distraction for the judge, we tailor the mapping so that it only retains terms present in the source segment. Furthermore, the judge is explicitly instructed to evaluate terminology usage. Note that providing the mapping means our metric is not entirely reference-free and may correlate more with other terminology-focused metrics. To study this effect, we also tested a standard FSP prompt without access to the terminology mapping across all submissions and settings. We found that the system rankings remained unchanged with the standard FSP prompt, suggesting that the inclusion of the terminology mapping primarily improves the interpretability and focus of the evaluation without fundamentally altering its outcomes.

Our terminology-aware FSP prompt is presented in Table 6.

D MQM Error Count

The output of the MQM judges using the FSP prompt is a list of errors. Each error is assigned a severity of minor, major, or critical. Table 8 reports the number and severity of errors produced by each submitted system, averaged across all documents in Track 2.

E Automatic Term Alignment in the Output Texts

The initial edition of the consistency metric (Semenov and Bojar, 2022) suggested that for term translation, specialized word alignment methods would be used. However, our preliminary analysis shows that both popular solutions, FastAlign by Dyer et al. (2013) and AwesomeAlign by Dou and Neubig (2021) show a lack of robustness with respect to morphological variation of the words, as well as casing and punctuation. Therefore, we used GPT-4o to retrieve the aligned terms from the system outputs. Our experiments showed that few-shot prompting was helpful for the quality of the term retrieval; therefore, we used 20-shot prompts. An example of the alignment prompt can be found in Table 9. For Track 2, we first split the documents into smaller paragraphs and retrieved the subsets of the terms for each segment, i.e., applied the same

```
TASK:
You are an expert linguist and terminologist.

Your job is to:
1. Analyze the source document and identify all domain-specific terminology and key terms (e.g. technical terms, product names,
named entities, etc.).
2. Find the corresponding translations in the translated document.
3. Output the result as a Python dictionary in the format:
{
    "source_term_1": "translated_term_1",
    "source_term_2": "translated_term_2",
    ...
}

RULES:
- Both source and translated documents are in Markdown format and may include image paths (e.g. ![image](path/to/image.png)) or
links. Ignore such elements.
- Only extract relevant terminology – avoid common words, function words, and markdown/control elements.
- If a translation is ambiguous or missing, set the value to null.
- Follow Python dict syntax strictly.
- Do NOT include explanations or extra text – only output the Python dictionary.

{{'-'*40}}
SOURCE DOCUMENT:
{{ source_document }}
{{'-'*40}}
TRANSLATED DOCUMENT:
{{ translated_document }}
{{'-'*40}}

OUTPUT:
(Please provide only the Python dictionary below)
```

Table 5: Prompt used for automatic terminology extraction in Track 2.

preprocessing schema as described in Section 4.3. To avoid hallucinations, every output is compared to a system output (by simple substring search).

We noticed that, while being able to correctly identify the part of the sentence containing a term translation, GPT-4o tends to return an overly long string (for example, if the ground truth term correspondence for English-Spanish sentence is "predefined"-"predeterminado", GPT, given a sentence "Es el valor predeterminado." would return the phrase "**valor** predeterminado" (lit. "predefined **value**". To overcome this, we used the following post-processing schema: each GPT output is compared against the reference term translation. If the number of words in the aligned term is more than it is in the reference translation, we run AwesomeAlign (Dou and Neubig, 2021) on the sentence pair and retrieve the word mappings of each word. Then, we check if the words selected by GPT (lemmatized) indeed correspond to the (lemmatized) source sentence tokens. If not, we cut these words out and leave only the part that corresponds to the exact term translation.

F Consistency Scores with Dictionary-Defined Pseudo-References

Figures 7 and 8 show the term consistency scores of the submitted systems with respect to pseudoreference initialization based on terminology dictionaries. We see that, firstly, the difference between the proper terminology mode, on one side, and random terminology and no terminology modes, on the other side, is significantly larger than in case of most frequent pseudoreference initialization. We also observe the increasing deltas between the proper terminology mode and two other modes in the best scoring systems.

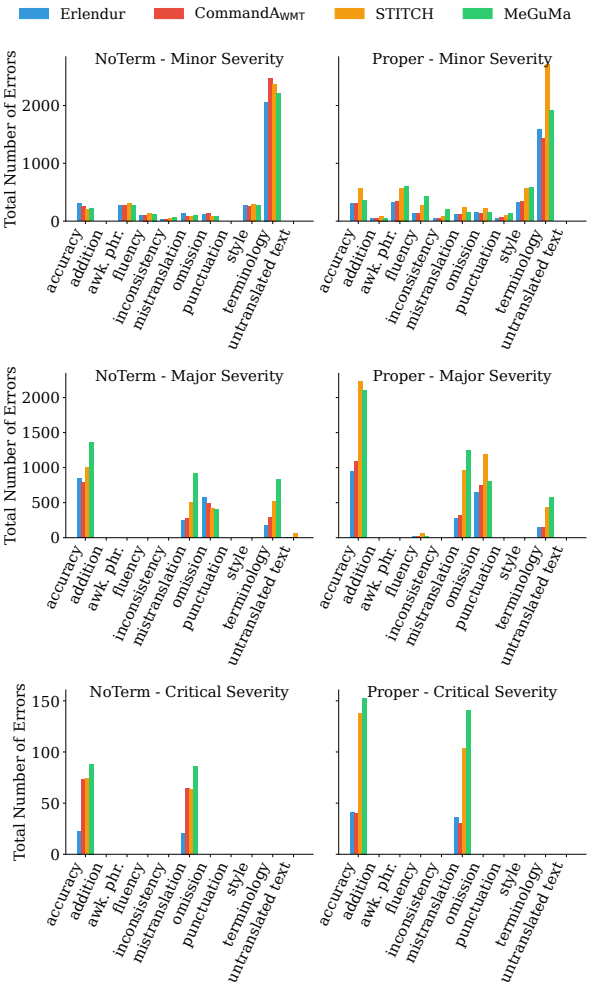

Figure 6: Distribution of error types and severities in No Term and Proper settings using GPT-4o as a judge. Total error counts indicate the number of errors each system made on the complete Task 2 test sets, comprising 10 annual reports across two translation directions. Error types with fewer than 50 occurrences across all severity levels and systems are omitted for clarity. "awk. phr." denotes awkward phrasing.

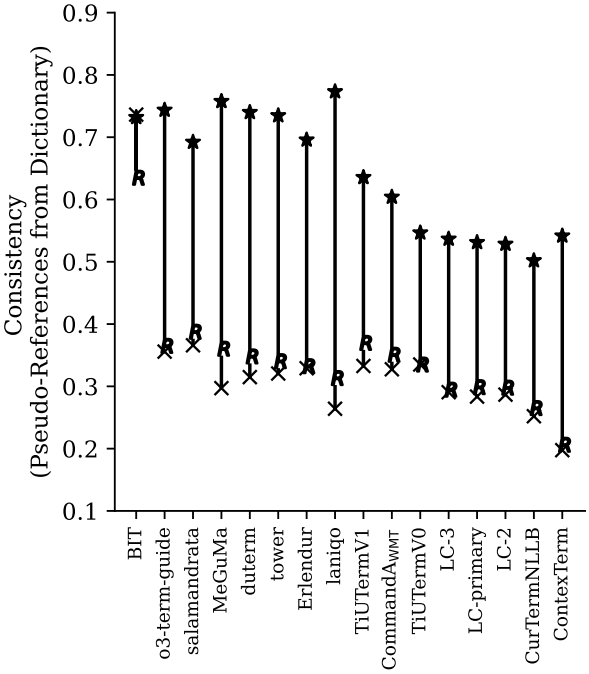

Figure 7: Effect of terminology mode on performance (measured by consistency score with dictionary-defined pseudo-references); Track 1. Legend: × denotes No Term, R denotes Random Term, and ★ denotes Proper Term.

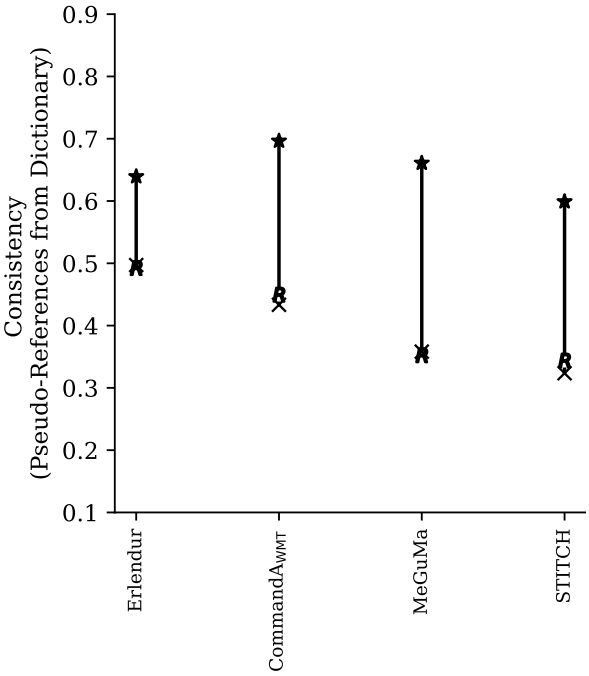

Figure 8: Effect of terminology mode on performance (measured by consistency score with dictionary-defined pseudo-references); Track 2. The legend is identical to that of the Figure 7.

```
You are an annotator for the quality of machine translation. Your task is to identify errors and assess the quality of the
translation using MQM. Based on the source text (in <source></source> tags) and the machine translation (surrounded by
<translation></translation> tags), identify error types in the translation and classify them.

The categories of errors are:
- accuracy (addition, mistranslation, omission, untranslated text, wrong language)
- fluency (character encoding, grammar, inconsistency, punctuation, register, spelling)
- style (awkward phrasing)
- terminology (see subcategories below)
- other

Each error, including omissions or untranslated content, is classified as one of three categories:
- Critical: Errors that make the text incomprehensible or misleading.
- Major: Errors that disrupt flow or distort meaning, but the text is still understandable.
- Minor: Errors that do not affect comprehension but are grammatically, stylistically, or formally incorrect.

The source text must be fully covered, and any omissions should be annotated as errors.
If the error is an omission (missing translation), set "error_span": "" and describe the missing content in "explanation".
Only include spans with errors; exclude correct text.

You will be given a full document and its translations. Only score the sentence in <target_segment></target_segment>, but use the
rest of the document for context. Consistency issues may be flagged per segment, even if similar issues are repeated in other
segments.

Terminology use (if <terminology> is provided; it will be a dictionary and may be empty `{}`):
- Treat <terminology> as a useful reference, not an absolute rule. Prefer its entries when appropriate, but do not penalize
natural, domain-correct alternatives.
- Ignore very minor variations such as capitalization, plural/singular, or the presence/absence of articles (``a,'' ``the'')
unless they clearly change the meaning and/or cause translation errors.
- If an entry in <terminology> seems implausible or clearly incorrect in context, do not enforce it.
- Terminology error subcategories:
  • terminology_mismatch: a correct entry exists in <terminology> for the context, but the translation uses a meaningfully
  different wording that diverges from the provided or established term.
  • terminology_omitted: a source term with a required translation is left untranslated.


Please respond in JSON following this schema:
{
  "type": "object",
  "properties": {
    "errors": {
      "type": "array",
      "items": {
        "type": "object",
        "properties": {
          "error_span": {
            "type": "string",
           "description": "The portion of the translation containing the error. If the error is an omission, use an empty string."
          },
          "explanation": {
            "type": "string",
            "description": "A brief explanation of the error and its impact; reference document context and/or <terminology> only
            if necessary."
          },
          "error_category": {
            "type": "string",
            "enum": ["accuracy", "fluency", "style", "terminology", "other"],
            "description": "The main category of the error."
          },
          "error_type": {
            "type": "string",
            "description": "The specific type of error (e.g., omission, mistranslation, punctuation, terminology_mismatch,
            terminology_omitted)."
          },
          "severity": {
            "type": "string",
            "enum": ["critical", "major", "minor"],
            "description": "Critical: incomprehensible or misleading. Major: distorts meaning but is understandable. Minor: does
            not affect comprehension but is incorrect."
          }
        },
        "required": ["error_span", "explanation", "error_category", "error_type", "severity"]
      }
    },
    "quality_score": {
      "type": "integer",
      "description": "Overall quality score of the translation for the target segment. Use any integer between 0 and 100.
      Guidance: 0 = No meaning preserved, nearly all information is lost. 33 = Some meaning preserved but significant parts are
      missing or garbled; hard to follow. 66 = Most meaning preserved, with only minor grammar/fluency issues; understandable
      overall. 100 = Perfect meaning and grammar, fluent and natural."
    }
  },
  "required": ["errors", "quality_score"]
}

Please score the following input: <input>
<source_language>{{ src_lang }}</source_language>
<source>{{ src }}</source>
<target_language>{{ tgt_lang }}</target_language>
<translation>{{ output_seq }}</translation>
<target_segment>{{ target_segment }}</target_segment>
<terminology>{{ terminology_dict }}</terminology>
</input>

Output requirements:
- Respond with valid JSON only (no text before or after the JSON).
- Produce strings as plain text without Markdown formatting.
```

Table 6: The FSP prompt used to identify MQM errors in the translation.

```
You are a professional translator from {{ source_language }} to {{ target_language }}. You need to help the user with finding the
corresponding words in the {{ target_language }} sentence translated from {{ source_language }}.

Below you are given the {{ source_language }} sentence and its translation, and a list of words to which you need to find the
corresponding words.

You need to return the words and their translations in the form of the Python dictionary, where keys are {{ source_language }}
words and values are their translations, for example, {{ 4-shot example dictionary }}.
DO NOT PRETTIFY THE DICTIONARY, return the raw dictionary in one string.

Source sentence: {{ source sentence }}
Target sentence: {{ target sentence }}
Words that need correspondence: {{ randomly selected words }}
Dictionary of correspondences:
```

Table 7: Prompt for automatically mapping randomly selected words from the source text to the target text.

	MQM Judge: GPT-4o									MQM Judge: GPT-5								
	# Minor			# Major			# Critical			# Minor			# Major			# Critical		
System	**Avg**	**EnZh**	**ZhEn**	**Avg**	**EnZh**	**ZhEn**	**Avg**	**EnZh**	**ZhEn**	**Avg**	**EnZh**	**ZhEn**	**Avg**	**EnZh**	**ZhEn**	**Avg**	**EnZh**	**ZhEn**
Proper																		
Erlendur	21.1	19.0	23.1	10.2	10.9	9.5	0.4	0.2	0.6	7.9	7.9	8.0	3.8	5.2	2.1	0.2	0.3	0.1
CommandA$_{WMT}$	19.8	20.9	18.8	11.6	10.8	12.3	0.4	0.3	0.4	7.8	10.8	5.0	3.9	5.4	2.4	1.1	0.3	1.8
STITCH	37.1	37.4	36.8	25.0	28.1	21.9	1.2	0.8	1.7	9.2	10.3	8.2	6.7	8.6	4.8	1.5	1.3	1.7
MeGuMa	29.6	28.3	30.8	24.7	22.2	27.1	1.4	1.2	1.6	10.8	9.2	12.3	9.5	8.3	10.7	2.7	1.9	3.5
Random																		
Erlendur	24.7	26.0	23.4	10.0	10.8	9.1	0.1	0.1	0.2	11.1	13.9	8.3	4.5	6.7	2.4	0.1	0.2	0.1
CommandA$_{WMT}$	26.4	27.8	25.1	10.6	11.2	10.0	0.6	0.5	0.6	13.0	16.0	10.0	6.8	9.1	4.5	0.9	0.6	1.2
STITCH	25.7	25.2	26.1	11.5	11.8	11.2	0.7	0.5	0.9	14.2	14.6	13.9	9.2	11.2	7.3	1.4	1.3	1.5
MeGuMa	27.4	31.2	23.2	22.3	24.0	20.4	7.9	0.9	15.4	14.9	15.4	14.3	18.9	20.8	17.0	7.7	4.5	10.9
NoTerm																		
Erlendur	24.5	25.8	23.3	9.5	9.5	9.5	0.2	0.1	0.3	11.0	13.3	8.8	4.1	5.7	2.5	0.2	0.2	0.1
CommandA$_{WMT}$	27.9	31.2	24.5	10.1	10.4	9.8	0.7	0.3	1.0	13.8	17.4	10.3	6.6	8.4	4.8	0.9	0.3	1.5
STITCH	27.2	30.4	24.1	14.2	15.3	13.1	0.7	0.4	1.0	14.7	16.4	13.1	11.4	14.1	8.9	1.9	1.6	2.0
MeGuMa	25.4	26.9	24.0	20.0	18.9	21.0	0.8	0.4	1.2	13.7	13.8	13.7	13.4	14.6	12.2	3.5	3.0	4.0

Table 8: Mean number of errors at each severity level (lower is better). Systems are sorted in ascending order, consistent with Table 4. Results are shown overall across 111 documents (Avg) and separately for the EnZh and ZhEn subsets.

```
You are a professional {{ source_language }}-{{ target_language }} translator, teaching the students the course on technical
translation. You are checking a student's translation of a sentence that contains a technical term. You are given an {{
source_language }} term (it can be a word or an expression), a source {{ source_language }} sentence containing this term (it may
be cased differently or contain additional punctuation), and a student's {{ target_language }} translation. You need to find how
the student has translated the term in question in {{ target_language }}, and return only that term.

Important: do not change the translated term anyhow, copy it straight from the sentence! For example, keep the casing and the
grammar form of the translated term as is.

When completing the task, follow the examples below:

        {{ source_language }} sentence: {{ sentence in source language }}
        {{ source_language }} term: {{ source language term }}
        {{ target_language }} translation: {{ reference translation }}
        Translated term: {{ reference term translation }}

                .
                .
                .

        {{ source_language }} sentence: {{ sentence in source language }}
        {{ source_language }} term: {{ source language term }}
        {{ target_language }} translation: {{ reference translation }}
        Translated term:
```

Table 9: Prompt used for automatic terminology alignment. Only one shot of 20 examples was shown explicitly.

System	Proper, ChrF				Proper, Acc.				Proper, Cons.				Random, ChrF				Random, Acc.				Random, Cons.				NoTerm, ChrF				NoTerm, Acc.				NoTerm, Cons.			
	Avg	Es	De	Ru	Avg	Es	De	Ru	Avg	Es	De	Ru	Avg	Es	De	Ru	Avg	Es	De	Ru	Avg	Es	De	Ru	Avg	Es	De	Ru	Avg	Es	De	Ru	Avg	Es	De	Ru
o3-term-guide	71.0	75.9	71.6	65.6	99.1	99.1	99.1	99.0	87.7	86.7	86.1	90.4	68.1	72.4	69.4	62.4	49.2	50.7	52.3	44.6	88.3	89.1	87.1	88.5	63.6	69.5	64.7	56.6	44.4	46.9	47.5	38.9	89.5	88.8	88.3	91.3
duterm	70.1	76.1	70.7	63.6	98.2	98.7	98.2	97.6	87.3	86.0	86.3	89.5	66.4	72.1	67.2	59.8	46.6	48.8	48.4	42.4	86.6	88.7	84.9	86.3	61.6	67.0	62.6	55.3	42.9	46.9	42.5	39.1	86.9	86.7	86.8	87.0
Erlendur	69.3	74.8	69.9	63.3	92.9	94.4	93.2	91.2	86.7	83.8	86.3	90.0	66.4	71.6	67.6	59.8	44.4	47.1	47.1	38.9	86.2	86.5	84.5	87.5	62.6	68.1	64.0	55.6	42.3	44.9	42.5	39.5	87.1	87.2	86.0	88.0
TiUTermV1	68.9	77.1	65.7	63.8	87.6	89.4	87.3	86.1	86.7	85.7	85.9	88.5	66.8	74.2	64.4	61.8	54.6	59.2	56.7	47.9	85.1	86.4	84.0	84.9	64.4	72.4	61.9	58.9	52.1	54.6	54.1	47.7	85.2	87.9	84.4	83.2
GPT-4.1-nano■	67.4	72.4	67.4	62.3	90.7	95.2	89.0	88.0	87.5	86.3	86.3	90.0																								
salamandrata	67.3	72.0	69.6	60.4	91.3	92.7	91.7	89.4	87.4	87.3	86.4	88.6	64.7	69.3	66.2	58.5	48.2	53.1	48.1	43.4	87.4	87.9	86.3	88.2	62.0	67.2	64.0	54.7	44.4	49.9	42.7	40.7	87.9	88.5	86.9	88.2
MeGuMa	67.2	72.0	67.7	61.9	97.4	97.0	96.3	98.8	88.6	86.9	88.6	90.2	64.5	70.3	64.2	59.0	46.7	53.1	46.4	40.5	87.1	88.4	84.7	88.1	58.9	65.2	59.4	52.1	40.1	46.9	38.5	34.8	86.3	85.6	85.3	87.9
tower	66.0	74.0	65.9	58.1	93.7	95.0	94.8	91.2	88.4	87.6	86.8	90.7	63.8	71.2	63.0	57.1	44.3	48.6	45.7	38.5	87.4	87.7	85.9	88.5	60.9	68.6	61.2	53.0	40.9	46.7	39.6	36.3	88.3	87.9	87.0	90.0
CommandA$_{WMT}$	65.9	70.7	67.6	59.3	79.9	81.9	86.9	70.7	86.6	84.5	87.5	87.8	63.7	68.4	65.0	57.6	45.8	49.3	48.1	40.1	88.3	87.5	86.2	91.3	60.7	65.5	62.2	54.4	43.0	47.3	42.7	38.9	87.7	85.5	86.4	91.3
BIT	63.7	69.8	62.4	58.9	97.0	96.3	98.0	96.7	87.8	86.8	86.9	89.8	65.7	67.2	66.3	63.5	80.5	47.5	97.4	96.5	87.9	87.6	86.8	89.3	66.5	69.8	66.3	63.5	96.7	96.3	97.4	96.5	87.9	87.4	86.9	89.3
TiUTermV0	62.7	69.0	61.0	58.3	74.4	75.2	71.1	76.8	86.4	85.0	85.6	88.6	61.0	68.1	59.1	55.8	49.6	54.2	49.9	44.8	84.9	85.1	84.7	84.9	60.2	68.0	57.9	54.6	49.1	53.6	49.4	44.2	85.2	85.4	84.9	85.4
laniqo	61.7	68.5	59.8	56.9	99.3	98.7	99.4	99.6	87.6	85.6	89.3	87.9	60.2	66.3	59.5	54.8	42.7	46.9	43.5	37.7	82.3	82.9	82.8	81.4	55.0	60.3	55.5	49.4	36.9	41.5	35.2	34.0	82.2	81.0	83.2	82.4
LC-primary	61.4	68.9	61.2	54.2	70.2	74.1	70.7	65.8	85.4	83.6	85.8	87.0	61.0	68.1	59.7	55.2	38.6	43.8	37.4	34.6	85.4	85.8	83.1	87.2	57.5	65.0	56.9	50.5	36.5	41.2	35.5	32.8	84.7	85.3	84.2	84.6
LC-2	60.8	67.7	61.0	53.7	70.0	73.6	70.7	65.6	85.8	85.4	85.7	86.2	60.5	67.1	59.5	54.9	38.5	43.4	37.4	34.6	85.7	86.5	83.7	86.9	56.9	64.1	56.8	49.9	36.3	40.8	35.5	32.6	85.0	85.8	84.3	85.0
LC-3	60.8	67.7	61.0	53.7	70.0	73.6	70.7	65.6	86.0	85.6	85.7	86.7	60.5	67.1	59.5	54.9	38.5	43.4	37.4	34.6	84.9	85.0	83.2	86.5	56.9	64.1	56.8	49.9	36.3	40.8	35.5	32.6	85.3	85.7	84.4	85.7
CurTermNLLB	60.1	69.1	60.3	51.0	63.4	76.5	79.0	34.6	88.0	87.5	87.6	88.8	58.8	67.4	58.0	50.8	36.1	44.1	31.7	32.6	84.1	85.3	82.0	84.9	55.6	65.6	52.8	48.4	34.2	41.7	27.1	33.8	85.7	86.1	84.7	86.2
ContexTerm	48.5	53.7	40.2	51.5	72.0	68.5	79.9	67.6	81.9	75.6	85.8	84.4	48.2	52.0	40.7	51.7	24.6	20.5	18.6	34.8	80.0	75.0	78.3	86.7	45.7	50.2	37.4	49.4	22.4	18.6	13.8	34.8	79.2	72.3	80.8	84.5
Systran-ft		71.1				44.1				88.1				71.1				44.1				88.6				71.1				44.1				88.2		
MarianMT-ft		65.6				17.5				54.1				68.9				48.8				85.1				68.9				48.8				86.4		
EuroLLM-ft		63.5				38.9				82.5				63.5				38.9				83.1				63.5				38.9				82.8		
TranssionMT				47.8				33.2				90.1				47.8				33.2				88.3				47.8				33.2				88.4

Table 10: Extended results for Track 1: sentence-level IT documentation terminology-informed translation. See Table 2 for a subset.

System	Proper, ChrF			Proper, Acc.			Proper, Cons.			Random, ChrF			Random, Acc.			Random, Cons.			NoTerm, ChrF			NoTerm, Acc.			NoTerm, Cons.		
	Avg	EnZh	ZhEn	Avg	EnZh	ZhEn	Avg	EnZh	ZhEn	Avg	EnZh	ZhEn	Avg	EnZh	ZhEn	Avg	EnZh	ZhEn	Avg	EnZh	ZhEn	Avg	EnZh	ZhEn	Avg	EnZh	ZhEn
Erlendur	60.2	46.1	74.2	78.7	85.4	71.9	92.0	91.6	92.3	57.9	41.8	74.0	64.9	60.1	69.6	90.6	89.2	91.9	57.4	40.8	74.0	65.0	60.3	69.6	90.7	89.2	92.1
CommandA$_{WMT}$	59.6	43.6	75.5	83.6	78.9	88.3	91.5	90.1	93.0	56.7	39.8	73.7	58.8	52.1	65.4	90.6	89.1	92.2	54.9	36.9	72.9	56.6	49.1	64.1	90.6	89.0	92.1
MeGuMa	54.3	39.1	69.4	79.5	96.6	62.4	90.8	93.3	88.3	48.4	31.6	65.2	47.7	43.9	51.5	85.8	84.1	87.4	51.0	33.7	68.3	48.3	44.6	51.9	85.1	83.2	87.0
STITCH	53.4	37.5	69.3	72.8	70.9	74.8	87.4	87.2	87.6	49.9	31.1	68.8	46.9	39.5	54.4	84.3	76.9	91.7	47.5	31.8	63.1	44.8	41.2	48.5	84.9	82.7	87.1
GPT-4.1-nano■	47.9	31.6	64.1	54.7	51.6	57.9	81.9	80.3	83.5	46.5	29.1	63.9	43.8	37.6	50.0				46.1	28.6	63.7	43.5	37.2	49.8			

Table 11: Extended results for Track 2: document-level finance terminology-informed translation. See Table 3 for a subset.