

GRIAL-TA participation in the WMT26 Terminology Shared Task

Gonzalo López-Sánchez^{1*} Antoni Oliver¹ Marta Cozzini¹
Mercè Vázquez¹ Patricia Morales-Hurtado¹ Sergi Alvarez-Vidal²

¹Universitat Oberta de Catalunya (UOC), Barcelona, Spain
{glopezsanch, aoliverg, mcozzini, mvazquezga, pmoraleshu}@uoc.edu

²Universitat Autònoma de Barcelona (UAB), Cerdanyola del Vallès, Spain
sergi.alvarez@uab.cat

Abstract

This paper describes the GRIAL-TA team’s submission to the WMT2026 Terminology Translation Shared Task. We participated in both tracks for the English-Polish language pair. Our core translation architecture relies on an instruction-aligned Gemma 4 model fine-tuned on highly curated, domain-specific corpora. To address the specific evaluation modes of Track 1, we constructed specialized datasets to train the model to adapt to prompts featuring no terminology, proper domain terminology, and random terminology. For Track 2, where explicit dictionaries are absent, we extracted the required glossary from the provided seed bitexts prior to translation using TBXTools, a tool for automatic terminology extraction. We merged the unique terms extracted independently by a BERT-based methodology and a linguistic part-of-speech pattern extractor, to mitigate a significant performance drop observed in the Polish data. Ultimately, while our approach maintained competitive general translation quality, the evaluations reveal that extensive fine-tuning degraded the model’s instruction-following capabilities, reducing its adherence to terminology constraints. We discuss these limitations and propose future strategies to balance domain adaptation with prompt sensitivity.

1 Introduction

With the creation of large language models (LLMs), machine translation (MT) has reached cutting-edge performance (Ataman et al., 2025); nonetheless, they have shown some struggles when it comes to multilingual machine translation (MMT) compared to traditional approaches (Zhu et al., 2024). Moreover, LLM translations of terminology-heavy texts are far from the desired outcome (Semenov et al., 2025). At the document-level, LLMs tend to be inconsistent or eliminate terminology (especially

in low-resource languages), which is detrimental for the quality of the overall translated text.

With this background in mind, the GRIAL-TA team participated in the Terminology Shared Task of WMT26, in both Track 1 (document-level MT with an explicit dictionary) and Track 2 (document-level translation with sample bitexts) for English and Polish in the fields of Medicine, and Engineering and Technology. We fine-tuned an instruction-tuned Gemma 4 model for both tracks. To extract terms from the Track 2 data we used the term extraction tool TBXTools¹, using both BERT-based and linguistic methodologies (Oliver and Vázquez, 2015; López-Sánchez et al., 2026).

2 Resources

In this section, we specify the resources used to fine-tune the translation model used in Tracks 1 and 2. We compiled a term reference list in the domains specified by the task organizers: Medicine, and Engineering and Technology. We decided to search for large resources with multiple subdomains related to the main domains. Additionally, we added bilingual corpora in order to enhance translation. All resources were obtained from reliable sources, such as institutions, international agencies, published dictionaries, and official databases. It is worth mentioning that all the resources used for this shared task, except for the European Medicines Agency and European Union’s (2026) corpus, are open access.

2.1 Corpora

We used EuroPat and EMA (Heafield et al., 2022; European Medicines Agency and European Union, 2026) with the purpose of training the translation models described in Section 2.4. We chose only English-Polish as the language pair for this task,

¹<https://github.com/grial-research-group/TBXTools>

* Corresponding author.

meaning that the unique segments mentioned in this section only apply for these two languages.

EMA and EuroPat are multilingual medicine and technical corpora, respectively. The former was constituted by documents from the European Medicines Agency and has 348,236 unique segments. The latter is the compilation of patent-specific parallel data, with 332,120 unique segments. We also created a general bilingual corpus based on NLLB² (NLLB Team et al., 2022) (74,065,994 unique segments) and Paracrawl v9³ (Bañón et al., 2020) (40,082,037 unique segments) corpora. To do so, both corpora are deduplicated, concatenated, and deduplicated again. The resulting corpus is published in Hugging Face⁴.

Since the alignment for the three corpora does not have the quality we desired (namely, segments that did not match exactly with their corresponding counterparts), we applied the rescoring algorithm⁵ proposed by Oliver and Álvarez (2023) to improve it. It uses `lid.176.bin` (Joulin et al., 2016), a language identification model based on fastText. It is trained on large multilingual datasets and supports 176 languages. As for the selection of rescored segments, we took into account language detection confidence and similarity between segments key metrics (both with a minimum score set at 0.75). After rescoring, they sized down to 162,091 segments for EMA-en-pl, 292,768 for EuroPat-en-pl, and 76,478,046 for the general corpus.

2.2 Terminological databases

We used the IATE (InterActive Terminology for Europe) database (Translation Centre for the Bodies of the European Union, 2018) to compile specialized terminology. This resource is used for the translation of official documents and legislation within the EU. Its website enables customized terminology extraction through structured download requests, requiring users to specify parameters such as language, domains, term types, file formats, and reliability levels. In total, we collected 22,201 terms between the three domains. Table 1 shows the parameters used to obtain them for both domains of the shared task. It is divided by languages, IATE domains, term type, reliability, evalu-

ation⁶, and total number of terms. In Section 3.4.1, the parameters used for terminology extraction are specified.

2.3 Bilingual dictionary

For the selection of random terms in the creation of the fine-tuning sets (see Section 3.2) we have used the Wikdict dictionaries⁷. In our repository, a script to convert Wikdict TEI dictionaries to tabtxt is provided. The English-Polish Wiktionary has a total of 99,394 entries.

2.4 Base models

The Gemma 4 models (El Abd et al., 2026), developed by Google DeepMind and hosted on Hugging Face, are a set of open-weights multimodal models, delivering state-of-the-art performance in several tasks and with high multilingual capabilities, covering over 140 languages.

- **google/gemma-4-12B-it**: A lightweight, instruction-tuned 12B model that directly projects audio and visual inputs into the language embedding space, removing standalone encoders to minimize latency and memory usage on standard GPUs (8–16 GB VRAM).
- **google/gemma-4-31B-it**: A dense 30.7B parameter model optimized for complex reasoning, multi-turn agentic tasks, and software engineering, featuring native video processing and adjustable visual token budgets.

We selected these models for this shared task due to their ideal balance of high-throughput efficiency, strong instruction-following capabilities, and robust multilingual pre-training across over 140 languages. Terminology-aware translation demands precise adherence to domain-specific glossaries and complex contextual constraints, making Gemma 4’s native system prompt support and configurable thinking modes particularly advantageous for controlled decoding and constrained generation. Furthermore, evaluating both the 12B and 31B variants enables a direct comparative analysis between lightweight, low-latency architectures suitable for practical edge deployments and high-capacity models capable of mastering subtle morphological inflections in terminology-heavy texts.

²<https://huggingface.co/datasets/allenai/nllb>

³<https://paracrawl.eu/>

⁴<https://huggingface.co/datasets/aoliverg/corpus-MTUOC-eng-pol>

⁵<https://github.com/mtuoc/MTUOC-PCorpus-rescorer>

⁶For further information about term type, reliability, and evaluation, see Translation Centre for the Bodies of the European Union (2026).

⁷<https://www.wikdict.com/>

	Medicine	Engineering and Technology
Languages	en, pl	en, pl
Domains	2841 health	68 INDUSTRY, 66 ENERGY, 6416 research and intellectual property, 6411 technology and technical regulations
Term type	term	term
Reliability	reliable, very reliable	reliable, very reliable
Evaluation	deprecated, obsolete, admitted, preferred, proposed, not specified	deprecated, obsolete, admitted, preferred, proposed, not specified
Total of terms	7,486	14,715

Table 1: Parameters used in the IATE request and total number of terms retrieved.

3 Methodology

3.1 Prompts

We designed a set of prompts to fine-tune the translation models under several conditions. These prompts can be found in appendix A.

3.2 Creation of the sets for fine-tuning Gemma

We developed a Python script (`createSet.py`) that, provided a parallel corpus, a general dictionary in tab separated format and a terminological database in TBX format, and specifying the names and codes of the source and target languages, can create a JSONL file that contains: the source segment, the target segment, the proper terms of both segments, and the random terms of both segments (up to three, if there are more, only the longest are selected). What follows is an example of the output:

```
{
  "en": "32 Metformin Non-clinical data for metformin reveal no special hazard for humans based on conventional studies of safety pharmacology, repeated dose toxicity, genotoxicity, carcinogenic potential, toxicity to reproduction.",
  "pl": "Metformina Dane niekliniczne dla metforminy nie wykazały specjalnych zagrożeń dla ludzi na podstawie tradycyjnych badań bezpieczeństwa farmakologicznego, toksyczności po podaniu wielokrotnym, genotoksyczności, rakotwórczości, toksycznego wpływu na rozrodczość.",
  "proper_terms": {
    "clinical data": "dane kliniczne",
    "safety pharmacology": "farmakologia bezpieczeństwa",
    "repeated-dose toxicity": "toksyczność po podaniu wielokrotnym",
    "genotoxicity": "genotoksyczność"
  },
  "random_terms": {
    "metformin": "metformina",
    "special": "specjalny",
    "safety": "bezpieczeństwo"
  }
}
```

To create these sets, we used the EMA, Europat, and the general corpus presented in Section 2.1 and the terminological databases presented in Section 2.2. As the process is slow, the task creation

of sets from the general corpus was stopped before the end. Table 2 details the size of the created sets.

Subject	Category	Total
Medical	EMEA	103,178
	General	59,160
Technical	Europat	51,644
	General	137,754

Table 2: Size of the fine-tuning datasets.

3.3 Fine-tuning of Gemma

We used the transformers library to fine-tune gemma4-12b-it and gemma4-31b-it for medical texts with two different settings: early stopping with patience of 3, and 2 epochs. All these variants were evaluated and compared with the base models. The best performing option was used to fine-tune a model for technical texts. All the fine-tuning scripts can be found in the repository⁸, so all the different parameters can be observed and the experiments can be reproduced.

In Table 3 we can observe the evaluation results of the base models and of the fine-tuned models. Overall, the data shows that fine-tuning enhances model performance in most metrics, with the two-epoch (FT 2 Ep.) configuration standing out as the most effective approach across both model sizes. Completing two training epochs yields improvements in standard quality metrics (BLEU, chrF, and COMET) while markedly reducing the Translation Error Rate (TER). For instance, the 12b model’s BLEU score jumps from approximately 28.5 in its base state to over 53 after two epochs, outperforming the Early Stopping (E.S.) strategy. Interestingly, while the 31b base model initially exhibits a slight

⁸<https://github.com/grial-research-group/WMT26-SharedTask-Terminology-GRIAL-TA>

Size	Model	Prompt	Terms	BLEU	chrF	TER	COMET	T.P. %
12b	Base	General	No terms	28.33	56.97	67.02	88.03	74.93
12b	Base	Medical	No terms	28.47	57.27	66.80	87.97	74.85
12b	Base	Medical	Random	28.23	56.98	66.86	87.17	73.97
12b	Base	Medical	Proper	28.82	57.60	66.41	87.83	88.27
12b	FT E.S.	General	No terms	39.33	61.89	60.86	88.66	75.37
12b	FT E.S.	Medical	No terms	39.55	62.23	58.93	88.64	75.15
12b	FT E.S.	Medical	Random	39.46	62.00	61.02	88.52	75.44
12b	FT E.S.	Medical	Proper	40.27	62.72	58.54	88.85	81.89
12b	FT 2 Ep.	General	No terms	53.62	70.95	46.41	90.20	77.20
12b	FT 2 Ep.	Medical	No terms	53.64	70.95	46.25	90.26	76.91
12b	FT 2 Ep.	Medical	Random	54.11	71.88	44.71	90.42	77.21
12b	FT 2 Ep.	Medical	Proper	53.76	70.91	46.28	90.29	79.91
31b	Base	General	No terms	30.85	59.23	65.19	88.56	75.51
31b	Base	Medical	No terms	31.10	59.52	65.27	88.62	75.51
31b	Base	Medical	Random	31.21	59.25	64.39	87.92	74.93
31b	Base	Medical	Proper	31.64	59.82	64.50	88.44	88.64
31b	FT E.S.	General	No terms	30.14	66.87	50.69	89.77	75.88
31b	FT E.S.	Medical	No terms	30.16	66.82	50.48	89.71	76.03
31b	FT E.S.	Medical	Random	30.55	66.94	50.14	89.87	75.81
31b	FT E.S.	Medical	Proper	31.03	67.93	48.84	89.86	88.50
31b	FT 2 Ep.	General	No terms	53.33	70.95	47.05	89.25	76.833
31b	FT 2 Ep.	Medical	No terms	53.87	70.90	46.87	89.28	77.05
31b	FT 2 Ep.	Medical	Random	55.63	72.46	44.65	89.60	76.69
31b	FT 2 Ep.	Medical	Proper	53.79	71.22	46.30	89.51	79.11

Table 3: Evaluation results for medical texts (FT: fine-tuned; E.S.: early stopping; 2 Ep.: 2 Epochs).

edge over its 12b counterpart, this capacity-related advantage virtually dissipates after full fine-tuning. Both models converge at remarkably similar performance levels, achieving a BLEU score of roughly 53.8 and a COMET score of around 89 to 90 when utilizing proper terminology.

Furthermore, analyzing the prompt and terminology configurations reveals nuanced model behaviors. Although the application of a domain-specific "Medical" prompt provides only marginal gains over a "General" baseline, injecting "Proper" terms dramatically boosts Terminology Precision (T.P. %) in the base models. This is clearly evidenced in the 31b Base model, where T.P. surges from 74.93% to 88.64%. However, this terminology injection effect is noticeably less pronounced in the fine-tuned (2 Epochs) models, suggesting that the extended training process already enables the models to internalize the requisite domain vocabulary robustly.

In Table 4, we can observe the evaluation results of the 31b model in its base state compared to its fine-tuned counterpart after two training epochs (FT 2 Ep.). Mirroring the trends seen in the medical domain, fine-tuning improves lexical quality metrics across all prompt and terminology settings. Specifically, the BLEU score increases by over six points (rising from approxi-

mately 29.0 to 35.7–36.9), accompanied by a four-point gain in chrF and a notable reduction of over five points in the Translation Error Rate (TER). Interestingly, neural evaluation via COMET displays a slight divergence from the lexical metrics: COMET scores remain exceptionally stable, experiencing a minor drop from around 85.36 in the base model to roughly 84.90–85.03 in the fine-tuned model. This suggests that while fine-tuning aligns the model’s outputs much more closely with reference n -grams, its underlying semantic quality remains consistently high across both states. Regarding prompting and terminology injection, the domain-specific ‘Technical’ prompt yields negligible differences compared to the ‘General’ baseline when no terms are provided. However, the inclusion of ‘Proper’ terminology again demonstrates a clear benefit for target term preservation. In the Base model, providing proper terms drives Terminology Precision (T.P.) up from 74.49 % to 86.14 %. While this boosting effect persists in the fine-tuned model—raising T.P. from 78.06 % to 82.39 %—it is comparatively less pronounced, reinforcing the observation that fine-tuning inherently embeds domain vocabulary while slightly dampening the model’s sensitivity to external terminology prompts.

Size	Model	Prompt	Terms	BLEU	chrF	TER	COMET	T.P. %
31b	Base	General	No terms	29.50	60.13	58.69	85.38	75.16
31b	Base	Technical	No terms	29.03	59.93	59.20	85.36	74.49
31b	Base	Technical	Random	31.40	61.58	56.30	85.21	75.16
31b	Base	Technical	Proper	29.85	60.64	58.27	85.29	86.14
31b	FT 2 Ep.	General	No terms	35.70	64.21	53.65	84.94	78.58
31b	FT 2 Ep.	Technical	No terms	35.72	64.15	53.47	84.86	78.06
31b	FT 2 Ep.	Technical	Random	36.91	65.13	52.37	85.12	78.91
31b	FT 2 Ep.	Technical	Proper	35.90	64.22	53.99	85.03	82.39

Table 4: Evaluation results for technical texts (FT: fine-tuned; 2 Ep.: 2 Epochs).

Appendix B details the hardware specifications of the server, alongside the time and energy consumption of all the fine-tuning processes that were carried out.

3.4 Terminology extraction

Track 2 of the shared task required term extraction from a sample of parallel sentences. For this, we initially used the BERT methodology of the term extraction tool TBXTools (Oliver and Vázquez, 2015; López-Sánchez et al., 2026). This methodology models automatic terminology extraction as a token-level sequence labeling task using BIO tagging. A pre-trained BERT-based model is fine-tuned on an automatically annotated corpus to predict terms. We empirically evaluated several well-established models for this task. Following extensive hyperparameter tuning for each candidate, for our final implementation we selected the base variant of XLM-RoBERTa (Conneau et al., 2020), a multilingual version of RoBERTa pre-trained on 100 languages, which allowed us to seamlessly process the bilingual data. Although the larger (xlm-roberta-large) variant achieved a higher F1-score on a validation split of the data, a qualitative manual inspection of the actual predictions revealed no significant differences. Consequently, we prioritized the base model in favor of computational efficiency.

3.4.1 Resources for terminology extraction

In order to generate the training labels to be used during the fine-tuning of our terminology extraction model we needed a list of terms and a corpus for the tool to automatically annotate. For this, we compiled reference term lists using the IATE database. We downloaded lists in Polish and English for the specific domains of the task. These lists are different from those used in the dataset generation for the fine-tuning of the translation model. We aimed to maximize number of terms, so we

did not filter the download query as much as previously, as we wanted the same model to predict terms in both languages across both domains. A detailed overview of our download requests is illustrated in Table 5. Once we obtained the files, we performed some additional processing. We decided to keep only terms containing fewer than five elements and no punctuation or special characters such as parentheses, to ensure the model learned mainly to identify terms not containing them. The resulting reference lists’ sizes across domains and target languages are summarized in Table 6.

	English	Polish
Eng. & Tech.	138,905	16,227
Medicine	77,186	15,807

Table 6: Number of terms in the terminology reference lists by domain and language.

We also created the required specialized corpora using the tool corpus-from-wikipedia⁹ (Segués et al., 2026), designed for compiling specialized monolingual and parallel corpora from Wikipedia. Building on prior success with highly curated data, our objective was to evaluate whether a widely accessible platform like Wikipedia can serve as a viable source of specialized training data for our terminology extraction model. For each language, we built two separate corpora based on Wikipedia categories, one for Medicine and another one combining Engineering and Technology, the domains specified by the tasks organizers. We also performed extra filtering on this output to handle cross-lingual noise. This is particularly common in non-English Wikipedia articles, which often include English titles of books, films, or proper nouns. To address this problem we used the lid.176.bin language detector of fastText (Joulin et al., 2016). Only sen-

⁹<https://github.com/grial-research-group/corpus-from-wikipedia>

	Medicine	Engineering and Technology
Languages	en, pl	en, pl
Domains	28 SOCIAL QUESTIONS	64 PRODUCTION, TECHNOLOGY AND RESEARCH, 68 INDUSTRY
Term type	abbrev, formula, phrase, short form, term, lookup form	abbrev, formula, phrase, short form, term, lookup form
Reliability	reliable, very reliable, downgrade prior to deletion, reliability not verified, minimum reliability	reliable, very reliable, downgrade prior to deletion, reliability not verified, minimum reliability
Evaluation	deprecated, obsolete, admitted, preferred, proposed, not specified	deprecated, obsolete, admitted, preferred, proposed, not specified
Total of terms	106,856	169,470

Table 5: Parameters used in the IATE request and total number of terms retrieved for terminology extraction.

tences classified as the target language, English or Polish respectively, were retained. Table 7 summarizes the sentence counts before and after this filtering step.

Language	Domain	Before	After
English	Eng. & Tech.	253,966	247,043
	Medicine	111,889	108,700
Polish	Eng. & Tech.	24,971	23,153
	Medicine	18,638	17,435

Table 7: Number of segments in the Wikipedia corpora by language and domain before and after filtering.

3.4.2 BERT extraction

Once the corpora and reference term lists were compiled, we provided them as inputs to TBXTools to automatically annotate the data and fine-tune our model. The tool first lemmatized, using spaCy¹⁰, both the corpora and the term lists, executing the annotation process separately by language and domain. The lemmatized data were used exclusively for label generation and the resulting labels were then aligned to the original, unlemmatized corpora. The model was fine-tuned on a combined dataset of both languages and domains using the default optimization setup of the Hugging Face Trainer API, with a learning rate of 4.5×10^{-5} , a batch size of 32, a weight decay of 0.028, and a warmup ratio of 0.016 for 1.7 epochs. Hyperparameters were chosen based on the highest F1-score on a validation split of the training data.

Afterwards, we used the fine-tuned model to perform the extraction of the terms on the sample sentences given by the task organizers. The data contained 214 parallel sentences for Engineering

and Technology and 70 for Medicine. While the fine-tuned model managed to extract a high number of terms in English, we observed that in Polish it extracted a considerably smaller number of terms. To address this issue we decided to perform an additional extraction using linguistic methods.

3.4.3 Linguistic extraction

Linguistic terminology extraction is another extraction methodology supported by TBXTools. It is based on the search of certain combinations of POS tags in a text. We used the same corpora and lists of terms as used in the extraction described in section 3.4.2. The tool automatically annotated the data with part-of-speech tags, using spaCy, and retrieved the underlying POS patterns directly from the reference terminology separately by language and domain. Thus, we had four distinct lists of linguistic patterns alongside their frequency in the corpus. To reduce noise, we discarded patterns with a frequency of 2 or less. Additionally, we filtered out single-tag patterns consisting of any of the following POS tags: PUNCT, X, AUX, DET, ADP, NUM, INTJ, CONJ, and PART. We then used these lists to perform the term extraction, limiting candidate terms to unigrams, bigrams, trigrams and quadgrams. This process yielded a greater number of Polish terms. After a qualitative analysis of the results, we decided to combine the unique terms extracted by both the BERT and linguistic methodologies. Table 8 details the results of both methodologies and their combined result.

3.4.4 Terms alignment

In order for the machine translation system to use the extracted terms, they first had to be matched to their translation equivalent. Although we lemmatized all the data, it should be noted that

¹⁰<https://spacy.io/>

Methodology	Domain	English	Polish
BERT	Eng. & Tech.	830	284
	Medicine	412	69
Linguistic	Eng. & Tech.	545	338
	Medicine	163	119
Combined	Eng. & Tech.	1,056	359
	Medicine	550	181

Table 8: Extracted terms from the sample sentences by methodology, domain, and language.

we intentionally restricted the lemmatization of Polish multi-word terms with nominal modifiers, e.g. *watu rozrzuqdu*, to maintain standard citation forms and our fine-tuning data format. In these cases, the head noun was lemmatized to the nominative case (*wat*) and the modifier retained the grammatical inflection (*rozrzuqdu*). Afterwards, we generated vector embeddings using LaBSE (Feng et al., 2022), a BERT based multilingual embedding model supporting 109 languages, via the sentence-transformers (Reimers and Gurevych, 2020) library. By mapping terms into a shared vector space, LaBSE enables the computation of semantic similarity via cosine similarity. We computed pairwise cosine similarity scores across all vector pairs, which produced separate similarity matrices for the Engineering and Technology, and Medicine domains. In these matrices each cell represents the semantic similarity score between a Polish term and an English term. We subsequently applied the linear sum assignment function from the SciPy (Virtanen et al., 2020) library to each matrix to extract the optimal one-to-one alignments, therefore maximizing the total semantic similarity. After a qualitative analysis of the results, we decided on a minimum similarity threshold of 0.80 to filter out low-confidence pairs. In the end, we obtained 391 aligned pairs for Engineering and Technology along with 98 for Medicine.

4 Participation in Track 1: Document-Level Translation with Explicit Dictionary

For both tracks of the task, we fully fine-tuned gemma4-31B-it for two epochs. We trained two separate models: one specialized in medical texts and another in mechanical engineering, as described in Section 3.3. As required by the task, we evaluated three distinct translation setups based on the provided resources: (1) given no terminol-

ogy, (2) given random terminology, and (3) given proper terminology.

Initially, we attempted full-document translation; however, this resulted in incomplete or truncated translations. Adjusting generation parameters to generate longer outputs did not resolve the issue. This limitation was likely due to the fact that the fine-tuning process was conducted at the segment level. Consequently, we adopted a segment-by-segment translation approach after segmenting each document. When incorporating terms, only the terms explicitly present in the given segment were included in the prompt. Segments without any associated terms were translated using the standard domain-specific prompt. This strategy successfully allowed us to retrieve complete translations for all documents.

4.1 Results

Our system’s performance demonstrates competitive surface-level translation quality, though it struggles with strict terminology enforcement. As shown in Table 9, our system achieved a chrF++ score of 71.0, closely aligned with the average system score (72.2). However, the neural metrics scores indicate a wider gap in semantic quality compared to the average.

The most significant holdup was term success. Under the *proper* dictionary condition, our system achieved only a 57.3% term success rate, significantly below the 85.5% system average. Despite this, the ablation across different prompt conditions somewhat validates our fine-tuning, since injecting the correct terminology improves term precision, as detailed in Table 10.

5 Participation in Track 2: Document-Level Translation with Sample Bitexts

For this track, the task organizers provided a group of sample sentences, that could be used by our system in different ways to extract the relevant terminological information, and a document to translate, from which the sample sentences were retrieved. We divided this task in two different processes:

- Extraction of relevant terms in both languages and subsequent alignment, described in detail in Section 3.4.
- Using these extracted terms to carry out the same procedure as done for Track 1, described in Section 4.

System	chrF++		Term success (lemma)		COMET-22		XCOMET-XXL		MetricX-24 XL (lower better)	
	Avg	enpl	Avg	enpl	Avg	enpl	Avg	enpl	Avg	enpl
gold	100.0	100.0	96.4	93.6	96.4	96.3	93.2	96.5	2.04	2.26
Avg. system	72.2	74.2	85.5	85.2	85.6	87.0	79.0	83.0	4.93	4.94
GRIAL-TA	71.0	71.0	57.3	57.3	76.9	76.9	64.9	64.9	8.39	8.39

Table 9: Results of track 1 (explicit dictionary, proper mode): document chrF++, lemmatized exclusive term success (%) and neural metrics on aligned segments, averaged over directions and per direction. Gold is the reference scored against itself. Data sourced from task organizers. The remaining systems were averaged into one single row (Avg. system) for conciseness.

System	Proper			Random			Noterm		
	Term	chrF++	COMET	Term	chrF++	COMET	Term	chrF++	COMET
gold	96.4	100.0	96.4	–	–	–	–	–	–
Avg. system	85.5	72.2	85.6	56.6	69.5	84.0	59.0	70.6	85.1
GRIAL-TA	57.3	71.0	76.9	40.0	69.3	75.3	38.8	69.2	75.6

Table 10: Results of track 1: lemmatized term success (%), document chrF++ and COMET-22 under the released dictionary, a random dictionary and no dictionary. Random and noterm are scored against the proper dictionary. Data sourced from task organizers. The remaining systems were averaged into one single row (Avg. system) for conciseness.

System	chrF++		Term success (lemma)		COMET-22		XCOMET-XXL		MetricX-24 XL (lower better)	
	Avg	enpl	Avg	enpl	Avg	enpl	Avg	enpl	Avg	enpl
gold	100.0	100.0	93.4	94.9	95.2	96.3	91.4	96.4	1.88	2.37
Avg. system	73.8	76.0	74.8	73.1	85.8	87.5	81.0	84.3	4.53	4.79
GRIAL-TA	62.2	62.2	32.1	32.1	64.4	64.4	43.8	43.8	13.12	13.12

Table 11: Results of track 2 (sample bitexts, sample mode): document chrF++, lemmatized exclusive term success (%) and neural metrics on aligned segments, averaged over directions and per direction. Gold is the reference scored against itself. Data sourced from task organizers. The remaining systems were averaged into one single row (Avg. system) for conciseness.

System	Sample			Noterm		
	Term	chrF++	COMET	Term	chrF++	COMET
gold	93.4	100.0	95.2	–	–	–
Avg. system	74.8	73.8	85.8	60.2	70.8	85.1
GRIAL-TA	32.1	62.2	64.4	43.3	68.1	76.6

Table 12: Results of track 2: lemmatized term success (%), document chrF++ and COMET-22 under with and without the sample bitexts. Random and noterm are scored against the proper dictionary. Data sourced from task organizers. The remaining systems were averaged into one single row (Avg. system) for conciseness.

5.1 Results

The results of this track highlight the challenges faced in the term extraction from the sample bitexts. As detailed in Table 11, our system achieved scores below the average system. Terminology success, shown in Table 12, reached 32.1%, significantly below of the 74.8% average system baseline.

The most revealing insight from this track is the performance inversion. The *noterm* setting, where the model translated without the help of the extracted glossary, actually outperformed the *sample* setting accross all metrics. This drop in performance when using the glossary correlates with the difficulties detailed in Section 3.4.

6 Conclusions and future work

In this work, we presented the GRIAL-TA submission to the WMT26 Terminology Shared Task for the English-Polish language pair across both Track 1 and 2. A major challenge of our approach was constructing a specialized training corpus to fine-tune Gemma 4, explicitly teaching the model to follow terminology-constrained prompts structured for the task. Although the algorithm developed for corpus creation—which extracts random terms from a dictionary and proper terms from a terminology database—leaves room for improvement (particularly regarding morphological inflection handling in highly inflectional languages like Polish), our experimental results demonstrate the suitability of this strategy.

However, our approach faced certain limitations. First, due to time constraints inherent to the Shared Task schedule, our experiments were strictly limited to two model sizes of a single architecture, preventing a broader comparative analysis across other open-weight large language models. Second, because our fine-tuning dataset was compiled exclusively at the segment level, the fine-tuned models exhibited document truncation issues when attempting zero-shot full-document translation, necessitating a segment-by-segment decoding approach during inference.

In addition, the shared task’s results suggest that the performance gap compared to the average system primarily stems from the model’s reduced sensitivity to prompt constraints. As observed during internal evaluations, the fine-tuning degraded the model’s instruction-following capabilities, making it less receptive to the provided glossary of terms—a significant limitation, given that this was a funda-

mental requirement of the task.

Regarding terminology extraction, the BERT-based methodology proved highly effective for English but significantly underperformed in Polish. This imbalance in the quantity of the predicted terms negatively impacted the subsequent alignment phase, although we managed to partially mitigate it using linguistic methods. While multilingual models offer strong baseline cross-lingual transfer, their ability to extract highly specialized terminology was noticeably degraded in the Polish data. This degradation may be attributed to the ‘curse of multilinguality’ described in [Conneau et al. \(2020\)](#). Under this phenomenon, model capacity is diluted across languages, that is, while low-resource languages benefit from shared representations, moderate-to-high resource languages, like Polish, can be hindered. In addition, because English outweighs Polish in XLM-RoBERTa’s pre-training data, it is probable that the model simply lacked sufficient exposure to specialized Polish terminology.

In future work we plan to investigate whether dedicated monolingual BERT-based models mitigate these issues, and further analyze the impact of the quality of the data, especially in languages other than English. Likewise, regarding the training phase of the translation model, we plan to refine the dataset creation algorithm to better account for complex target-language morphology. Regarding prompt sensitivity, we intend to explore alternative fine-tuning strategies to ensure the model retains its ability to adhere to external glossaries. Furthermore, we aim to expand the generation of these enriched datasets to paragraph and document levels, as well as to additional language pairs and specialized domains.

Acknowledgments

This work was supported by the Industrial Doctorates Plan from the Department of Research and Universities of the Government of Catalonia (reference number: 2025 DI 00112).

This work has been supported by project "TaMTAS: Terminology-Aware Machine Translation for Accessible Science" - PCI2025-167063-2, funded by MICIU/AEI/10.13039/501100011033 and by the European Union (CHIST-ERA).

This work is partially supported by the OpenEU Alliance, with funding from the European Union (EUROPEAN EDUCATION AND CULTURE EX-

ECUTIVE AGENCY - EACEA) under grant agreement No 101177241.

References

- Duygu Ataman, Alexandra Birch, Nizar Habash, Marcello Federico, Philipp Koehn, and Kyunghyun Cho. 2025. [Machine Translation in the Era of Large Language Models: A Survey of Historical and Emerging Problems](#). *Information*, 16(9):723.
- Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. [ParaCrawl: Web-scale acquisition of parallel corpora](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.
- Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, Mayank Chaturvedi, Aditya Chawla, Victor Cotruta, Alice Coucke, Phil Culliton, Robert Dadashi, Lucas Dixon, Mohamed Elhawaty, Utku Evci, Clément Farabet, and 303 others. 2026. [Gemma 4 Technical Report](#). *Preprint*, arXiv:2607.02770.
- European Medicines Agency and European Union. 2026. [European Medicines Agency \(EMA\)](#).
- Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Ariavazhagan, 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, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.
- Kenneth Heafield, Elaine Farrow, Jelmer van der Linde, Gema Ramírez-Sánchez, and Dion Wiggins. 2022. [The EuroPat corpus: A parallel corpus of European patent data](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 732–740, Marseille, France. European Language Resources Association.
- Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. [Bag of tricks for efficient text classification](#). *Preprint*, arXiv:1607.01759.
- Gonzalo López-Sánchez, Patricia Morales-Hurtado, Mercè Vázquez, Albert Morales-Moreno, and Silvia Rodríguez Vázquez. 2026. [Adapting TBXTools to automatic terminology extraction with BERT](#). In *Proceedings of the DEfinition and Term Extraction CHallenge (DETECH 2026)*, volume 4215, Zadar, Croatia. CEUR Workshop Proceedings.
- NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Hefernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. [No language left behind: Scaling human-centered machine translation](#). *Preprint*, arXiv:2207.04672.
- Antoni Oliver and Sergi Álvarez. 2023. [Filtering and rescore the CCMATRIX corpus for Neural Machine Translation training](#). In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 39–45, Tampere, Finland. European Association for Machine Translation.
- Antoni Oliver and Mercè Vázquez. 2015. [TBXTools: A Free, Fast and Flexible Tool for Automatic Terminology Extraction](#). In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 473–479, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.
- Nils Reimers and Iryna Gurevych. 2020. [Making monolingual sentence embeddings multilingual using knowledge distillation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- Gemma Segué, Gonzalo López-Sánchez, and Antoni Oliver. 2026. Monolingual and Parallel Specialized Corpora from Wikipedia. In *Proceedings of the International Conference of the Spanish Society for Natural Language Processing (SEPLN)*, León, Spain. CEUR.org. Forthcoming.
- Kirill Semenov, Xu Huang, Vilém Zouhar, Nathaniel Berger, Dawei Zhu, Arturo Oncevay, and Pinzhen Chen. 2025. [Findings of the WMT25 Terminology Translation Task: Terminology is Useful Especially for Good MTs](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 554–576, Suzhou, China. Association for Computational Linguistics.
- Translation Centre for the Bodies of the European Union. 2018. [IATE](#).
- Translation Centre for the Bodies of the European Union. 2026. [Term Level](#).
- Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay

Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, and 16 others. 2020. [SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python](#). *Nature Methods*, 17:261–272.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. [Multilingual Machine Translation with Large Language Models: Empirical Results and Analysis](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.

A Prompts

General: In this prompt no subject is declared.

```
SYSTEM_PROMPT = "You are a professional English-Polish translator."
USER_PROMPT_TEMPLATE = (
    "Translate the following text from English to Polish, providing only the translation and nothing else:\n"
    "\"\"\"{SEGMENT_TO_TRANSLATE}\"\"\""
)
```

Medical/Technical: In this prompt the subject is declared, either medical or technical:

```
SYSTEM_PROMPT = "You are a professional medical/technical English-Polish translator."
USER_PROMPT_TEMPLATE = (
    "Translate the following text from English to Polish, providing only the translation and nothing else:\n"
    "\"\"\"{SEGMENT_TO_TRANSLATE}\"\"\""
)
```

Medical/Technical random terms: In this prompt the subject is declared, either medical or technical, and the random terms are provided:

```
SYSTEM_PROMPT = "You are a professional medical/technical English-Polish translator."
USER_PROMPT_TEMPLATE = (
    "Use the following glossary:\n{RANDOM_TERMS}\n\n"
    "Translate the following text into Polish. Provide only the translation and nothing else:\n\"\"\"{SEGMENT_TO_TRANSLATE}\"\"\""
)
```

Medical/Technical proper terms: In this prompt no the subject is declared, either medical or technical, and the proper terms are provided:

```
SYSTEM_PROMPT = "You are a professional medical/technical English-Polish translator."
USER_PROMPT_TEMPLATE = (
    "Use the following glossary:\n{PROPER_TERMS}\n\n"
    "Translate the following text into Polish. Provide only the translation and nothing else:\n\"\"\"{SEGMENT_TO_TRANSLATE}\"\"\""
)
```

B Fine-tuning hardware, time, and energy consumption

All fine-tuning experiments were conducted on a high-performance compute node featuring four NVIDIA RTX PRO 6000 Blackwell Edition GPUs (96 GB VRAM each, providing a cumulative 384 GB VRAM). The system was powered by dual AMD EPYC 9135 16-Core processors (32 physical cores in total) and 384 GB of system RAM, running on Ubuntu 22.04.5 LTS.

B.1 Gemma4-12b-it early stopping

Total Time: 0h 35m 23s
Total GPU Consumption: 0.7027 kWh

B.2 Gemma4-12b-it Medical 2 epochs

Total Time: 8h 56m 18s
Total GPU Consumption: 11.4567 kWh

B.3 Gemma4-31b-it Medical early stopping

Total Time: 11h 16m 36s
Total GPU Consumption: 15.1518 kWh

B.4 Gemma4-31b-it 2 Medical epochs

Total Time: 18h 47m 19s
Total GPU Consumption: 24.8788 kWh

B.5 Gemma4-31b-it Technical 2 epochs

Total Time: 24h 0m 9s
Total GPU Consumption: 39.6976 kWh