

UniVie-HAITrans at WMT26 Terminology Translation Task: Terminology-Guided Data Filtering and Few-Shot Fine-Tuning for Lightweight LLMs

Miguel Rios

Centre for Translation Studies, University of Vienna, Austria
miguel.angel.rios.gaona@univie.ac.at

Abstract

Specialised domain translation using Large Language Models remains challenging due to terminology constraints, scarce in-domain parallel data, and the tendency of fine-tuning to degrade in-context learning performance. We present our submission for the English–Polish and Spanish–Basque translation in the Medical and Engineering & Technology domains. Our method follows three stages: i) We extract in-domain terms from a terminology database and prompt a multilingual LLM to generate contextual sentences for each term. ii) We build an efficient search index over large-scale, unconstrained web-crawled and medical parallel corpora. Using the synthetic sentences as queries, we retrieve and filter the most semantically relevant pairs to build an in-domain fine-tuning dataset. iii) We structure our training instances with a mixture of 0-shot and semantically retrieved 5-shot prompts, and fine-tune lightweight multilingual LLMs (Gemma-3-4B, and Tiny-Aya-Global) via QLoRA. Fine-tuned Gemma-3 achieves the best results on held-out data across all metrics on English–Polish (Medical) and on Spanish–Basque (Engineering & Technology), showing that few-shot fine-tuning enhances specialised domain translation quality and preserves in-context adaptation.

1 Introduction

Machine Translation (MT) based on Large Language Models (LLMs) has achieved impressive zero-shot and few-shot general-domain translation quality (Moslem et al., 2023a; Bawden and Yvon, 2023; Gaona et al., 2026). However, translating specialised domains, such as Medical guidelines or Engineering & Technology specifications, remains a major challenge. Specialised domain translation requires adherence to terminology and exact contextual information, areas where general-purpose LLMs frequently produce mistranslations or hallucinated terminology (Moslem et al., 2023b;

Gaona et al., 2023; Gaona, 2025). While supervised fine-tuning (SFT) on in-domain parallel data can adapt models to technical domains, acquiring high-quality parallel sentence pairs for specialised domains is costly and sparse. Furthermore, standard SFT often causes models to over-fit to strict direct-translation or few-shot templates, degrading their in-context learning (ICL) performance when presented with different few-shot example prompts at inference time.

Using a terminology database to filter parallel corpora is a highly effective strategy for domain adaptation and fine-tuning MT models (Lohar et al., 2022; Nayak et al., 2023; Imamura and Utiyama, 2026). On the other hand, Alves et al. (2023) examine the trade-offs between instruction fine-tuning and ICL in LLMs for MT. To address this over-specialisation, they introduce a strategy that incorporates few-shot examples directly into the fine-tuning process, effectively preserving in-context adaptation.

To overcome these challenges, we present a system for sentence-level MT that integrates synthetic sentence generation, efficient vector search with filtering over large-scale parallel corpora for domain adaptation, and few-shot fine-tuning (Alves et al., 2023) for lightweight LLMs. We evaluate our system for the English–Polish (**en-pl**) and Spanish–Basque (**es-eu**) language pairs for the Medical and Engineering & Technology domains.

We few-shot fine-tune two lightweight LLMs: Gemma-3-4B-it (Team et al., 2025) and Tiny-Aya-Global (Salamanca et al., 2026) using Quantised Low-Rank Adaptation (QLoRA) (Dettmers et al., 2023). Our contributions are:

- We provide a scalable methodology to convert terminology databases into domain extraction filters for large web corpora.
- We use a method to structure our training instances into a mix of 0-shot instructions and

5-shot in-context examples.

- We evaluate the translation quality of lightweight LLMs across across Medical and Engineering & Technology domains for both en-pl and es-eu language pairs with automatic metrics, while preserving their ICL performance.

2 Experimental Setup

We use a domain-adaptation framework for en-pl (Medical, and Engineering & Technology) and es-eu (Engineering & Technology). By leveraging a terminology database, we generate synthetic context sentences via an LLM and extract relevant parallel segments from large corpora with an efficient index. Finally, we fine-tune two lightweight LLMs on a mixed 0-shot and 5-shot dataset to preserve both instruction-following and ICL translation performance.

2.1 Terminology

The Interactive Terminology for Europe (IATE)¹ is a terminology management system from EU institutions that covers different domains (e.g., economics, law, health, medical). We extracted terminology entries from IATE across domains associated with Medical and Engineering & Technology. Specifically, we selected terms under the following domain categories from the IATE system: *Health, Economics, Finance, Science, Production, Technology & Research, and Industry*. We only use terms with quality: 2 (minimum reliability), 3 (reliable), and 4 (very reliable) stars (human annotated quality scores). The resulting amount of terms for each domain and source language is as follows: 178,309 *Engineering & Technology* terms for English, 58,862 *Medical* terms for English, and 120,137 *Engineering & Technology* terms for Spanish.

2.2 Data

To build our domain-specific datasets, we use raw parallel pairs from two parallel sentence-level corpora: CCMatrix and EMEA.

CCMatrix A large-scale parallel corpus extracted from Common Crawl web data using multilingual sentence embeddings (LASER) (Schwenk et al., 2021). Because it covers a broad, unconstrained web-crawled texts, it serves as a candidate

pool for extracting specialised Engineering & Technology segments for both en-pl and es-eu language pairs. We use 30M segments for en-pl, and 6.5M segments for es-eu.

EMEA A domain-specific parallel corpus derived from the European Medicines Agency documents (ELG, 2020). It contains official medical, pharmacological, and clinical guidelines, providing high-quality, in-domain target sentence pairs for en-pl (Medical). The EMEA corpus contains 762,904 segments.

Before the semantic filtering and fine-tuning, we pre-processed all raw parallel corpora with standard text cleaning using the Moses² toolkit scripts to remove noisy or corrupted sentence pairs: **Length Filtering**: Sentence pairs are pruned if either the source or target sequence exceeds a maximum length of 200 tokens or falls below a minimum length of 1 token. **Length Ratio Filtering**: To eliminate misaligned sentences or alignment artefacts, we remove sentence pairs where the source-to-target and target-to-source length ratio of 1.5 times longer than the other sentence. **Deduplication**: duplicate sentence pairs across the raw datasets are removed to prevent data leakage and avoid biasing the fine-tuning process.

Following this initial cleaning step, the preprocessed corpora are passed to the filtering stage. We split a held-out sample dataset of 1,000 sentence pairs per language pair/domain combination, used strictly for evaluation. The remaining sentence pairs form the training pool.

2.3 Models

Aya Expanses To generate domain-specific contextual data, we use Aya-Expanses-32B³. This 32-billion parameter multilingual LLM is explicitly optimised for robust cross-lingual reasoning across 23 languages (Dang et al., 2024). Because of its advanced instruction-tuning and performance, we use it to generate fluent sentences from the IATE terminology, effectively bootstrapping our in-domain fine-tuning dataset.

Tiny-Aya Our first lightweight LLMs is Tiny-Aya-Global⁴. This 3.35 billion parameter multilingual model is optimised specifically for balanced, high-quality performance across non-English and

¹<https://iate.europa.eu/download-iate>

²<https://github.com/moses-smt/mosesdecoder/blob/master/scripts/training/clean-corpus-n.perl>

³CohereLabs/aya-expanses-32b

⁴CohereLabs/tiny-aya-global

low-resource language pairs. Its lightweight size and strong baseline language representation make it a good candidate for adaptation on the en-pl and es-eu language pairs.

Gemma-3 As our second lightweight LLM, we use Gemma-3-4B-it⁵. Gemma-3 is a 4 billion parameter model that integrates a 128K token context window and widespread multilingual support, spanning over 140 languages. Despite its size, it produces reasoning and strict instruction-following outputs, providing a highly competitive baseline for our few-shot fine-tuning method.

2.4 Few-Shot Fine-Tuning

Our method follows *three stages*:

Synthetic Seed Generation We extract terms for English and Spanish from the IATE terminology database, filtered specifically for the target domains (Medical, Engineering & Technology). Because the raw terminology lacks contextual information, or examples. We generate synthetic context sentences for each term. Using Aya-Expanse, we prompt the system to produce one natural, domain-appropriate sentence for each term. The synthetic data generation prompt is defined in Section A. For example, with the English Medical domain term "drug injection" the LLM generates: "*The patient required a drug injection to alleviate the severe pain in their lower back.*".

Semantic Embedding Filtering We compute vector representations for all synthetic sentences and the source-side corpus using a dense text encoder with a multilingual model⁶ from sentence-transformers (embedding size, $d = 384$). To efficiently mine domain-relevant samples from our large parallel corpora, we use a scalable vector index based on Facebook AI Similarity Search (FAISS) (Douze et al., 2025).

To overcome the memory size and search speed related to brute-force comparisons over millions of sentences, we index the corpus embeddings using an Inverted File with Product Quantisation (IVFPQ) index optimised for $L2$ Euclidean distance. The embedding space is partitioned into N Voronoi cells via K -means clustering. During retrieval, the search is restricted only to the inverted lists corresponding to the nearest centroid clusters,

this drastically narrows the search space. To minimise memory, each d -dimensional embedding vector is split into M sub-vectors. Each sub-vector is independently quantised into a cluster centroid via K -means, converting high-dimensional floating-point vectors into lightweight codebooks. The index hyper-parameters depend on the size of the pool corpus. The values of the FAISS index hyper-parameters for each language pair are defined in Section C.

For each seed synthetic domain sentence, we query the IVFPQ index to retrieve the nearest neighbour corpus source sentence within $L2$ distance. Corpus sentence pairs falling within the *top-1* are extracted and deduplicated to build the final fine-tuning dataset.

Few-Shot Example Retrieval To retain both 0-shot performance and few-shot adaptability, we format the data into a mixed structure. Following (Alves et al., 2023), we build a hybrid fine-tuning dataset where instances randomly alternate between **0-shot** (direct translation instructions) and **5-shot** configurations (instructions with 5 example pairs). The 5 examples are retrieved from the remaining pool of segments using the previous FAISS index, ensuring that the in-context prompts are domain-relevant.

For Track 2 of WMT26, the organisers do not provide a terminology database. Instead, they supply two sets of texts. *Seed Bitexts*: Sentence and paragraph pairs designed for systems to extract and learn relevant domain terminology. *Source Documents*: Texts containing the terminology established in the seed. We use an in-context learning strategy. The 0-shot baseline incorporates no target terms (bitexts) for prompting, whereas the 5-shot approach incorporates semantically retrieved seed bitexts into the prompt. The few-shot fine-tuning prompts are defined in Section B.

Sampling instances from both formats during fine-tuning ensures the model learns to output clean translations without over-generating, while remaining responsive to few-shot prompts. The final fine-tuning data split contains a total of 269,152 segments: 129,173 (Engineering & Technology), 48,834 (Medical) segments for en-pl, and 91,145 (Engineering & Technology) segments for es-eu.

We fine-tune two lightweight LLMs to evaluate translation quality: **Gemma-3**, and **Tiny-Aya**.

⁵[google/gemma-3-4b-it](https://huggingface.co/google/gemma-3-4b-it)

⁶[sentence-transformers/
paraphrase-multilingual-MiniLM-L12-v2](https://pypi.org/project/sentence-transformers-paraphrase-multilingual-MiniLM-L12-v2/)

Model	BLEU	chrF	COMET	BLEU	chrF	COMET
0-shot		Medical		Engineering & technology		
Tiny-Aya	21.8	51.8	0.8578*	23.2	56.7	0.8730*
Tiny-Aya FFT	24.1*	60.5*	0.8189	29.0*	62.1*	0.8442
Gemma-3	23.4	53.6	0.8622	26.5	58.8	0.8806
Gemma-3 FFT	47.7*	69.2*	0.9017*	43.1*	67.4*	0.8825
5-shot		Medical		Engineering & technology		
Tiny-Aya	30.9	58.5	0.8753*	30.1	60.8	0.8814*
Tiny-Aya FFT	32.1	65.3*	0.8518	31.0	65.6*	0.8198
Gemma-3	34.8	61.2	0.8853	33.9	63.3	0.8928
Gemma-3 FFT	53.1*	72.7*	0.9106*	50.9*	71.8*	0.8949

Table 1: Evaluation of base LLMs and few-shot fine-tuned models (FFM) on the en-pl language pair for the Medical and Engineering & Technology domains in 0-shot and 5-shot setups. Bold denotes top results, and asterisks (*) denote statistical significance ($p < 0.05$) for comparisons between the baseline and FFT models.

Model	BLEU	chrF	COMET
0-shot	Engineering & technology		
Tiny-Aya	11.8	51.4	0.7893
Tiny-Aya FFT	35.4	71.6	0.8222*
Gemma-3	7.0	43.6	0.6733
Gemma-3 FFT	66.4*	82.9*	0.9012*
5-shot	Engineering & technology		
Tiny-Aya	25.1	57.2	0.8216
Tiny-Aya FFT	40.9*	75.0*	0.8390*
Gemma-3	24.8	55.3	0.7570
Gemma-3 FFT	71.6*	85.7*	0.9108*

Table 2: Evaluation of base LLMs and few-shot fine-tuned models (FFM) on the es-eu language pair for the Engineering & Technology domain in 0-shot and 5-shot configurations. Bold denotes top results, and asterisks (*) denote statistical significance ($p < 0.05$) for comparisons between the baseline and FFT models.

Both models are fine-tuned on the mixed few-shot dataset using QLoRA (Dettmers et al., 2023). We apply adapters to the query, key, value, and projection layers to optimise parameter efficiency while maintaining baseline model performance. The values for QLoRA and SFT hyper-parameters are defined in Section D.

WMT26 Submission Our submission is the fine-tuned Gemma-3 model for the WMT Track 2 terminology task, covering English–Polish (Medical, Engineering & Technology) and Spanish–Basque (Engineering & Technology). We process both the supplementary few-shot candidate pool and the target documents at the sentence-level using our system. For Track 2 document-level translation, we submit MT outputs with both 0-shot (*no termi-*

nology) and 5-shot (*bitext*) setups.

To replicate our experiments, we are releasing our scripts on GitHub⁷.

3 Results

For automatic translation evaluation on the held-out test set, we report BLEU (Papineni et al., 2002; Post, 2018), chrF (Popović, 2015), and COMET (Rei et al., 2020). For COMET, we specifically use the `Unbabel/wmt22-comet-da` model. We evaluate the base instruction-tuned LLMs as baselines and compare their performance against the few-shot fine-tuned models (FFM) across both 0-shot and 5-shot configurations.

Table 1 shows the performance of the baseline lightweight LLMs (Tiny-Aya and Gemma-3) compared to the FFT models on the en-pl language pair. The asterisks denote statistical significance ($p < 0.05$) between the baseline and the FFM models for automatic metrics based on bootstrap resampling (Post, 2018).

The results for the en-pl language pair show that fine-tuning with our few-shot adaptation approach produces significant improvements across models and domains, with Gemma-3 FFT outperforming in all the setups. The FFT models outperform the baselines in almost all scenarios. Gemma-3 FFT with 0-shot achieves a high improvement in BLEU in the Medical and Engineering & Technology domains. Similar improvements are observed in chrF.

Our few-shot training strategy preserves and enhances ICL performance. Across all models, moving from 0-shot to 5-shot evaluation produces

⁷https://github.com/HAITrans-lab/UniVie-HAITrans_WMT26

Model	chrF		Term		COMET		XCOMET		MetricX	
	en-pl	es-eu	en-pl	es-eu	en-pl	es-eu	en-pl	es-eu	en-pl	es-eu
CUNI-UFAL	81.3	76.9	90.4	88.5	91.8	87.0	91.6	79.9	3.33	3.78
Gemma-3 FFT	73.0	67.5	62.1	53.5	85.3	78.9	77.4	59.4	5.71	7.46

Table 3: **WMT26 Track 2** (sample bitexts, sample mode): document chrF++, lemmatized exclusive term success (%) and neural metrics on aligned segments.

Model	0-shot (no term)			5-shot (bitexts)		
	Term	chrF	COMET	Term	chrF	COMET
CUNI-UFAL	73.7	73.4	87.5	89.2	78.1	88.7
Gemma-3 FFT	40.4	64.9	79.1	57.8	70.3	82.1

Table 4: **WMT26 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.

noticeable score improvements. Gemma-3 FFT achieves the highest performance in the 5-shot configuration, reaching top scores on BLEU, chrF, and COMET. While Tiny-Aya FFT achieves consistent improvements in BLEU and chrF over its baseline, its COMET scores show a decrease after FFT (0-shot Medical). In contrast, Gemma-3 FFT scales well, improving simultaneously across BLEU, chrF, and COMET (Medical), indicating that larger model capacity combined with in-domain few-shot examples, plays a role in translation quality.

Table 2 shows the performance of base LLMs against the FFT models on es-eu language pairs within the Engineering & Technology domain. The results for es-eu show improvements following the few-shot fine-tuning, highlighting the necessity of domain and language adaptation for low-resource languages like Basque. The baseline models struggle significantly in the 0-shot setting.

This shows that base lightweight LLMs lack sufficient pre-training data for Basque without ICL guidance or targeted adaptation. In the 0-shot setup, Gemma-3 FFT shows an increase in BLEU, followed by chrF and COMET. ICL provides further improvements when combined with fine-tuning. Moving from 0-shot to 5-shot evaluation improves all models, Gemma-3 FFT achieves the overall highest scores of BLEU, chrF, and COMET. Fine-tuning for es-eu produces statistically significant improvements across all three metrics for both FFT LLMs.

WMT26 Results The organisers evaluated all submissions using chrF++, lemmatised **Term** Success percentage, COMET, XCOMET, and MetricX (Charkiewicz et al., 2026). We compare our

system’s performance with the runner-up submission (CUNI-UFAL) (Charkiewicz et al., 2026). Table 3 shows the comparison between our model and CUNI-UFAL, while Table 4 details the performance across 0-shot and 5-shot configurations. The fine-tuning few-shot approach shows consistent gains over the baseline, however it does not surpass the top-performing CUNI-UFAL system.

4 Conclusions and Future Work

In this work, we presented a system for domain adaptation in specialised MT using lightweight LLMs, without compromising their ICL performance. We evaluate across en-pl and es-eu for specialised domains (Medical and Engineering & Technology). Our findings indicate: The few-shot fine-tuning method (Gemma-3 FFT) shows statistically significant improvements over base LLM baselines across automatic metrics. The FFT approach effectively preserves and enhances ICL. Evaluating models in a 5-shot configuration consistently produces the top performance across all metrics.

Building on these findings, we identify promising directions for future work: Studying lexical retrieval-based (e.g., BM25, fuzzy-match) selection for filtering and few-shot prompting could further improve the interaction between the fine-tuned adapter parameters and prompt context. While automatic metrics show high translation quality performance, conducting human evaluations (e.g., MQM framework) will help assess terminology consistency, and potential hallucinations in the fine-tuned LLMs (Lommel et al., 2013; Gaona et al., 2023; Kocmi et al., 2024).

Limitations

Relying on the synthetic generated sentences for filtering, potentially perpetuates or amplifies underlying biases of the LLM while failing to capture complex in-domain examples. In-house computing resources force trade-offs in LLMs size, restricting overall MT capacity during training and testing. The system’s terminology database (IATE) is heavily based on EU institutions, potentially leading to coverage gaps and lower precision when translating non-EU text.

References

- Duarte M. Alves, Nuno M. Guerreiro, João Alves, José Pombal, Ricardo Rei, José G. C. de Souza, Pierre Colombo, and André F. T. Martins. 2023. [Steering large language models for machine translation with finetuning and in-context learning](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11127–11148, Singapore. Association for Computational Linguistics.
- Rachel Bawden and François Yvon. 2023. [Investigating the Translation Performance of a Large Multilingual Language Model: the Case of BLOOM](#). In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 157–170, Tampere, Finland. European Association for Machine Translation.
- Adrian Charkiewicz, Pinzhen Chen, Thierry Etchegoyhen, Harritxu Gete Ugarte, Kamil Gutmman, Xu Huang, David Ponce, Artur Nowakowski, Frédéric Odebert, Arturo Oncevay, Kirill Semenov, Dawei Zhu, and Vilém Zouhar. 2026. Findings of the WMT26 terminology translation task. In *To appear in Proceedings of the Eleventh Conference on Machine Translation*. Association for Computational Linguistics.
- John Dang, Shivalika Singh, Daniel D’souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, and 26 others. 2024. [Aya expand: Combining research breakthroughs for a new multilingual frontier](#). *Preprint*, arXiv:2412.04261.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. [Qlora: Efficient finetuning of quantized llms](#). *Preprint*, arXiv:2305.14314.
- Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2025. [The faiss library](#). *Preprint*, arXiv:2401.08281.
- ELG ELG. 2020. [ELG - Bilingual corpus made out of PDF documents from the European Medicines Agency, \(EMA\)](#), <https://www.ema.europa.eu>, (February 2020).
- Miguel Angel Rios Gaona. 2025. [Instruction-tuned large language models for machine translation in the medical domain](#). In *Proceedings of Machine Translation Summit XX: Volume 1*, pages 162–172, Geneva, Switzerland. European Association for Machine Translation.
- Miguel Angel Rios Gaona, Raluca-Maria Chereji, Alina Secara, and Dragos Ciobanu. 2023. [Quality Analysis of Multilingual Neural Machine Translation Systems and Reference Test Translations for the English-Romanian language pair in the Medical Domain](#). In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 355–364, Tampere, Finland. European Association for Machine Translation.
- Miguel Angel Rios Gaona, Claudia Pliesseis, Dragos Ioan Ciobanu, and Alina Secara. 2026. [Fuzzy Matching and Sentence Embeddings for Few-shot Machine Translation with Large Language Models](#). *Proceedings of the 26th Annual Conference of the European Association for Machine Translation*, 1:567–579.
- Kenji Imamura and Masao Utiyama. 2026. [Text filter based on automatically acquired vocabularies for multilingual machine translation](#). In *Proceedings for the Ninth Workshop on Technologies for Machine Translation of Low Resource Languages (LoResMT 2026)*, pages 37–48, Rabat, Morocco. Association for Computational Linguistics.
- Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024. [Error Span Annotation: A Balanced Approach for Human Evaluation of Machine Translation](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 1440–1453, Miami, Florida, USA. Association for Computational Linguistics.
- Pintu Lohar, Sinead Madden, Edmond O’Connor, Maja Popovic, and Tanya Habruseva. 2022. [Building machine translation system for software product descriptions using domain-specific sub-corpora extraction](#). In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 1–13, Orlando, USA. Association for Machine Translation in the Americas.
- Arle Richard Lommel, Aljoscha Burchardt, and Hans Uszkoreit. 2013. [Multidimensional quality metrics: a flexible system for assessing translation quality](#). In *Proceedings of Translating and the Computer 35*, London, UK. Aslib.
- Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023a. [Adaptive Machine Translation with Large Language Models](#). In *Proceedings*

of the 24th Annual Conference of the European Association for Machine Translation, pages 227–237, Tampere, Finland. European Association for Machine Translation.

Yasmin Moslem, Gianfranco Romani, Mahdi Molaei, John D. Kelleher, Rejwanul Haque, and Andy Way. 2023b. [Domain Terminology Integration into Machine Translation: Leveraging Large Language Models](#). In *Proceedings of the Eighth Conference on Machine Translation*, pages 902–911, Singapore. Association for Computational Linguistics.

Prashanth Nayak, John Kelleher, Rejwanul Haque, and Andy Way. 2023. [Instance-based domain adaptation for improving terminology translation](#). In *Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track*, pages 222–234, Macau SAR, China. Asia-Pacific Association for Machine Translation.

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*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. [chrF: character n-gram F-score for automatic MT evaluation](#). In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. [A call for clarity in reporting BLEU scores](#). In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. [COMET: A neural framework for MT evaluation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Alejandro R. Salamañca, Diana Abagyan, Daniel D’souza, Ammar Khairi, David Mora, Saurabh Dash, Viraat Aryabumi, Sara Rajaei, Mehrnaz Mofakhami, Ananya Sahu, Thomas Euyang, Brittawnya Prince, Madeline Smith, Hangyu Lin, Acyr Locatelli, Sara Hooker, Tom Kocmi, Aidan Gomez, Ivan Zhang, and 7 others. 2026. [Tiny aya: Bridging scale and multi-lingual depth](#). *Preprint*, arXiv:2603.11510.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021. [CCMatrix: Mining billions of high-quality parallel sentences on the web](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6490–6500, Online. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. [Gemma 3 technical report](#). *Preprint*, arXiv:2503.19786.

A Synthetic Seed Generation Prompt

The synthetic seed data generation prompt is as follows:

Produce one sentence in {src} using the following term from the {domain-str} domain: "{term}"

with *src* source language (*English* or *Spanish*), *domain-str* domain name (*health* or *engineering and technology*), and *term* source test segment.

B Few-shot Fine-tuning Prompts

The 0-shot and 5-shot prompts are as follows:

0-shot prompt:

Translate the {src} source text to {tgt} in the {domain-str} domain.
Rules:
Output strictly the {tgt} translation.
Do NOT repeat the {src} text.
Do NOT include labels like "{src}:" or "{tgt}:".
Do NOT add quotation marks, explanations, or any extra text.
Now translate the following:
{src}: {src-segment}
{tgt}:

5-shot prompt:

Translate the {src} source text to {tgt} in the {domain-str} domain.
Rules:
Output strictly the {tgt} translation.
Do NOT repeat the {src} text.
Do NOT include labels like "{src}:" or "{tgt}:".
Do NOT add quotation marks, explanations, or any extra text.
Examples from the {domain-str} domain:
{src}: {example-segment}
{tgt}: {example-segment}
...
{src}: {example-segment}
{tgt}: {example-segment}
Now translate the following:
{src}: {src-segment}
{tgt}:

with *src* source language (*English* or *Spanish*), *tgt* target language (*Polish* or *Basque*), *domain-str* domain name (*health* or *engineering and technology*), *example-segment* source or target retrieved examples, and *src-segment* source test segment.

C FAISS Index Hyper-parameters

Instead of searching through all dataset vectors, FAISS uses *IndexFlatL2* to store and search the coarse cluster centroids that partition your vector space into k -means clusters. The *IndexFlatL2* uses Euclidean distance (L2) to perform vector searches. When the index is trained FAISS runs k -means clustering on the a sample of the training data to learn N global centroids/clusters. The FAISS index hyper-parameters are as follows:

Language pair	Training size	N clusters	M sub-quantisers	$nbits$
en-pl	20M	30000	64	8
es-eu	4M	1000	16	8

Table 5: FAISS index hyper-parameters with a *IndexFlatL2* coarse-quantiser and embedding dimensionality of $d = 384$.

D QLoRA and SFT Hyper-parameters

The hyper-parameter values for both LLMs are as follows:

Hyper-parameter	Value
r	32
α	64
Dropout	0.1
Target modules	q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj
Epoch	2
Max sequence length	1024
Batch size	2
Gradient accumulation	4
Learning rate	$2e - 4$
Warm-up steps	0.03
Scheduler type	cosine

Table 6: FFT hyper-parameter values. The upper section contains the 4-bit QLoRA hyper-parameters, and the lower section contains the overall SFT.