

AGENTEAK: A Multi-Agent Pipeline for Terminology-Constrained Spanish–Basque Translation

Jurgi Giraud^{1,2} and Andrew Gargett²

¹School of Informatics, University of Edinburgh, Edinburgh, UK,

²School of Languages and Applied Linguistics, The Open University, Milton Keynes, UK
jgiraud2@ed.ac.uk andrew.gargett@open.ac.uk

Abstract

This paper describes the OpenU’s submission to the WMT26 Terminology Translation Task for Spanish into Basque (Track 1), covering the automotive and energy domains. We present AGENTEAK, a multi-agent system built on LANGGRAPH in which three compact open-weight models (4–8B parameters) cooperate: a reasoning model that selects and disambiguates the terminology relevant to each segment, a domain fine-tuned translation model, and a verification model that checks and repairs terminology in the draft translation. To adapt the translator, we mined in-domain Spanish–Basque bitext from Wikipedia using multilingual sentence embeddings and complemented the data with in-domain synthetic parallel paragraphs generated by an instruction-tuned Basque LLM. Domain-tagged fine-tuning improves the base model by up to 10.4 BLEU and 3.9 COMET points on our in-domain test sets. On the shared task data, supplying the pipeline with correct terminology raises terminology success rate from 0.613 to 0.757 (automotive) and from 0.601 to 0.774 (energy).

1 Introduction

Terminology control is one of the clearest cases where Machine Translation (MT) quality is judged by more than adequacy and fluency. In technical and industrial settings, a translation that renders a term inconsistently is unusable regardless of how natural it reads. The WMT terminology tasks have made this requirement explicit by pairing test documents with domain term bases and evaluating systems on whether the prescribed target terms appear in the output (Alam et al., 2021; Semenov et al., 2023).

Our submission addresses Track 1 of the WMT26 Terminology Translation Task, Spanish into Basque, in the automotive and energy domains. Basque makes the problem markedly harder than

the language pairs typically studied. It is a language isolate with agglutinative, suffix-based morphology: nouns inflect for number, definiteness and case within a single word form. Constraint mechanisms that copy a target string verbatim, such as lexically constrained decoding (Hokamp and Liu, 2017; Post and Vilar, 2018), are therefore a poor fit as enforcing the citation form produces ungrammatical output, while allowing free inflection removes the guarantee that motivated the constraint in the first place. Basque is also comparatively low-resourced, and in-domain Spanish–Basque parallel data for the automotive and energy domains is essentially unavailable off the shelf.

We tackle both problems with a modular pipeline¹, AGENTEAK,² that separates terminology reasoning from translation. Rather than a single prompted model, AGENTEAK distributes the work across three compact open-weight models orchestrated with LANGGRAPH, none larger than 8B parameters, so that the whole system is able to run on a single consumer-grade GPU. Our contributions are:

1. a domain-targeted Wikipedia bitext mining procedure for Spanish–Basque;
2. an in-domain synthetic data generation procedure using an instruction-tuned Basque LLM;
3. domain-tagged parameter-efficient fine-tuning of Llama-3.1-8B-Instruct, improving translation quality over the base model in the two domains;
4. a three-agent terminology-aware translation pipeline, and an analysis contrasting correct and randomly substituted term constraints that quantifies how far the system actually follows the terminology it is given.

¹The full pipeline is available at: <https://github.com/jurgigi/agentek-wmt26>

²From Basque *agente* ‘agent’, plural *agentek*.

2 Background

Terminology-constrained MT. Early approaches enforced terminology at decoding time, either by extending beam search to track unsatisfied constraints (Hokamp and Liu, 2017) or by allocating beam capacity dynamically to keep the cost tractable (Post and Vilar, 2018). Because hard constraints interact badly with reordering and inflection, a second line of work moved the constraint into the training data, teaching the model to use inline target-term annotations supplied at inference time (Dinu et al., 2019). Bergmanis and Pinnis (2021) showed that annotating target *lemmas* rather than surface forms is markedly more effective for morphologically rich languages, since it leaves the model free to inflect the term correctly. Our approach is a soft variant of this idea: terminology is injected into the prompt of an instruction-tuned model, and a downstream agent, rather than the decoder, is responsible for checking that it was used.

LLMs and agentic translation. LLMs are competitive with dedicated NMT systems for high-resource pairs and exploit instructions and in-context examples well (Hendy et al., 2023; Vilar et al., 2023), including glossary entries supplied at inference time (Moslem et al., 2023); targeted fine-tuning closes much of the remaining gap (Xu et al., 2024). A parallel line of work decomposes translation into a generate–critique–revise loop (Madaan et al., 2023): Feng et al. (2025) show that systematic error identification followed by targeted repair beats undirected self-correction, and Wu et al. (2024) assign distinct roles to separate model instances for literary translation. The closest precedent for our submission is IRB-MT at WMT25 (Grubišić and Korencić, 2025), which entered the terminology task with an agentic translation-revision workflow over several small open-weight LLMs, deliberately avoiding fine-tuning and selecting a final candidate with a terminology-aware quality metric. AGENTEAK differs on both counts: we fine-tune the translation agent on mined in-domain data, which we consider necessary rather than optional for a target language with no strong general-purpose translator, and we run a single bounded verification pass instead of a metric-guided search over candidates, accepting a lower ceiling for a fixed inference budget.

Basque MT and terminology. Basque NLP has been transformed by open resources such as BasqueGLUE (Urbizu et al., 2022) and Latxa, a family of Llama-based models continued-pretrained on a large Basque corpus (Etxaniz et al., 2024), which we use as our backbone. Dedicated ES–EU systems exist but are general-domain, and massively multilingual systems such as NLLB-200 (NLLB Team, 2022) cover Basque without domain adaptation or terminology control. The most direct precedent for terminology-aware Basque MT is clinical: Soto et al. (2019) report that injecting bilingual clinical terminology, combined with back-translation of monolingual health records, roughly doubled BLEU on health record templates in the absence of any in-domain parallel corpus, and the same group examined the benefits and limits of terminology-driven domain adaptation for Basque at WMT20 (Soto et al., 2020).

Parallel data mining. Where in-domain parallel data does not exist, it can be mined. Margin-based retrieval over multilingual sentence embeddings (Artetxe and Schwenk, 2019) underpins large-scale corpora extracted from Wikipedia (Schwenk et al., 2021), and LaBSE (Feng et al., 2022) provides strong cross-lingual sentence representations for over 100 languages, Basque among them. Since embedding similarity alone admits many near-parallel but unusable pairs, filtering is needed to remove problematic pairs which might introduce noise in the data. We additionally follow the observation that LLM-generated in-domain text is a useful substitute for authentic domain data when the latter is scarce (Moslem et al., 2022), a synthetic-data strategy in the tradition of back-translation (Sennrich et al., 2016).

3 Data

3.1 Domain bitext mining from Wikipedia

We built in-domain Spanish–Basque parallel data³ by mining comparable Wikipedia articles. For each domain we defined a set of Spanish root categories and traversed their subcategory graph to collect candidate articles. For the automotive domain we used *Vehículos*, *Industria automotriz*, *Automoción*, *Componentes del automóvil*, *Tecnologías del automóvil*, *Tecnologías del transporte* and *Motores*

³Datasets available on Hugging Face: <https://huggingface.co/jurgiraud/datasets> with the name tag `eseu-wmt26`

	Automotive	Energy
<i>(a) Mined bitext</i> — sentence pairs		
Candidates ($\cos. \geq 0.80$)	3,054	4,348
Retained (Bicleaner AI ≥ 0.50)	2,447	3,297
<i>(b) Synthetic</i> — paragraph pairs		
Retained (Bicleaner AI ≥ 0.50)	1,189	1,129
<i>equivalent sentences</i>	3,567	3,387

Table 1: In-domain Spanish–Basque data. The two sources are counted in different units: mined bitext is aligned at the sentence level, synthetic data at the level of three-sentence paragraphs. The italicised row converts (b) to sentences for scale only.

de combustión interna; for the energy domain, *Energía*, *Fuentes de energía*, *Tecnología energética*, *Energías renovables*, *Generación de energía eléctrica*, *Ingeniería eléctrica* and *Combustibles*.

This produced 7,788 Spanish candidate articles for the automotive domain and 3,874 for energy. Articles were linked to their Basque counterparts through Wikipedia interlanguage links, leaving 704 comparable article pairs in each domain. Both sides were sentence-split and cleaned, and every Spanish sentence was compared against every Basque sentence within the same article pair using LaBSE embeddings (Feng et al., 2022), retaining pairs with cosine similarity ≥ 0.80 ⁴. Restricting candidate generation to within-article pairs keeps the search tractable and reduces false positives relative to corpus-wide retrieval.

The resulting 3,054 (automotive) and 4,348 (energy) candidate sentence pairs were then further filtered using Bicleaner AI (Zaragoza-Bernabeu et al., 2022) at the recommended threshold of 0.50, discarding roughly 20% and 24% of pairs respectively. This double-filtering approach is supported by evidence that corpus quality outweighs quantity for Basque and other low-resource languages (Artetxe et al., 2022). Table 1 summarises data and filtering pipeline.

3.2 Synthetic in-domain data

To complement mined Wikipedia parallel sentences, and to increase terminological density and coverage which may not have been fully captured by encyclopaedic texts, we generated ad-

ditional parallel data⁵ with Latxa-Llama-3.1-8B-Instruct (Sainz et al., 2025). To do so, we prompted the model to produce three-sentence Basque paragraphs together with their Spanish translations, with explicit instructions to employ correct bilingual terminology for the target domain taken from the bilingual dictionary provided as part of Track 1 of the shared task. Generating short paragraphs rather than isolated sentences yields text with intra-paragraph coreference and connectives, which is closer to the document-level test data than mined sentence pairs are.

Generated pairs were passed through the same Bicleaner AI filter, leaving 1,189 automotive and 1,129 energy pairs.

3.3 Out-of-domain data and splits

Both mined and synthetic in-domain data are small, and fine-tuning on them alone can cause over-fitting to domain-specific data. We therefore combined them with EusParallel,⁶ a general-domain Spanish–Basque parallel corpus, as an out-of-domain component. All in-domain data was split into training and development portions at the article level, so that no development sentence originates from an article seen in training. The resulting datasets have been released on the Hugging Face Hub.

4 Translation Agent

Setup. We selected the translation agent against four criteria: the model should be (i) openly available, (ii) pre-trained or adapted on Basque data, (iii) instruction-tuned, and (iv) compact, below 10B parameters. Our choice fell on Latxa-Llama-3.1-8B-Instruct (Sainz et al., 2025), a Llama-3.1 model adapted to Basque on the Latxa corpus (Etxaniz et al., 2024), which meets all four criteria. We fine-tuned it⁷ with LoRA (Hu et al., 2022) on an instruction-formatted version of our data. Each training example carries an explicit domain tag ([automotive] or [energy]) in the instruction, so that a single set of adapters serves both domains and the domain can be selected at inference time. We train two configurations: (i) mined bitext plus

⁵Datasets available on Hugging Face: <https://huggingface.co/jurgiraud/datasets> with the name tag eseu-wmt26 and synthetic

⁶<https://huggingface.co/datasets/HiTZ/EusParallel>

⁷The fine-tuned model is available on Hugging Face: <https://huggingface.co/jurgiraud/latxa-eseu-wmt26-augmented>

⁴Feng et al. (2022) used a threshold of 0.60 for translation retrieval which included partial translations. We used a more stringent threshold to retain only good quality parallel sentences

out-of-domain data, and (ii) mined bitext plus out-of-domain plus synthetic data. Fine-tuning configuration and parameters are provided in Appendix A.

Evaluation. We report BLEU (Papineni et al., 2002) and TER (Snover et al., 2006) through sacreBLEU (Post, 2018), chrF2 (Popović, 2015), which is better suited to agglutinative targets because it credits partially correct word forms, and COMET (Rei et al., 2020) on our held-out in-domain development sets.

Results. Table 2 shows that domain fine-tuning improves translation quality as measured by automated metrics. Against the base model, the first configuration gains 5.4 BLEU and 3.5 COMET in the automotive domain and 9.7 BLEU and 3.3 COMET in energy, with TER falling by 12.6 and 19.7 points. Adding synthetic data gives a further consistent improvement in every metric and in both domains (+1.6 and +0.7 BLEU), confirming that terminology-conditioned generation from the same backbone supplies a signal the mined data does not. The gains are largest for energy, where more mined data was available.

5 The AGENTeak Pipeline

The submitted system is a five-stage workflow implemented with LangChain and LangGraph,⁸ of which three stages are model-driven agents and two are deterministic, as seen in Figure 1. Documents are processed chunk by chunk, with the domain tag fixed for the whole document.

Step 1: Candidate term matching (deterministic). The source chunk is matched against the term base provided with the task data. Matching is performed on the Spanish side, where morphological variation is limited to number and gender agreement, using normalised surface forms and a small set of inflectional rules. This stage is high-recall by design: it over-generates candidates, including spurious matches inside longer expressions, and leaves disambiguation to the next stage.

Step 2: Terminology agent. The candidate list and the source chunk are passed to Qwen3-4B (Yang et al., 2025) operating with reasoning/thinking mode enabled. The agent fulfils three tasks: to discard candidates that are not genuine domain uses of the term, to recover terms the matcher

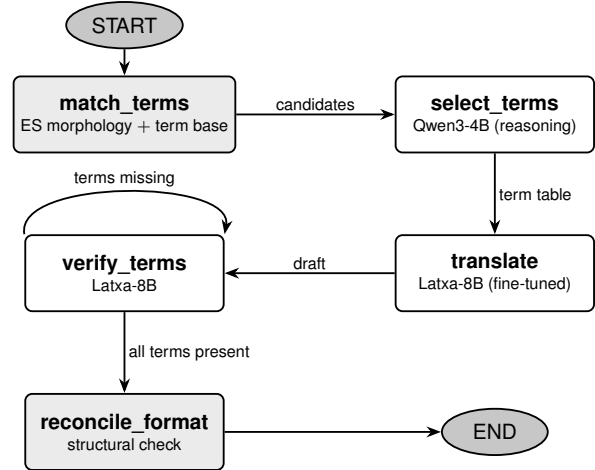


Figure 1: The AGENTeak graph. Shaded nodes are deterministic, white nodes are LLM agents. Documents are processed chunk by chunk; the state carried between nodes comprises the source chunk, the document’s domain tag, the selected term table, the current draft and the verifier’s flags.

missed, and to resolve entries with multiple admissible Basque equivalents by choosing the one appropriate to the context.

The prompt states the task explicitly and includes few-shot examples covering each of these cases. The agent returns a per-chunk table of source terms and selected target terms, which is carried forward. We use a small reasoning model here rather than the larger translator because the task is classification and selection, not generation, and because keeping this component separate means the term table can be inspected and, in principle, overridden by a human.

Step 3: Translation agent. The fine-tuned model from §4 receives the source chunk together with the selected term pairs. The prompt makes both the domain and the required terminology explicit. Crucially, the terminology is presented in citation form and the instruction states that the terms must be inflected as the Basque sentence requires; we do not ask the model to copy them verbatim. This is the soft counterpart of the lemma-annotation strategy of Bergmanis and Pinnis (2021) and is what makes the approach viable for an agglutinative target.

Step 4: Verification agent. The draft translation and the term table are passed to a verification agent, in the current pipeline Latxa-Llama-3.1-8B-Instruct, which checks, for each required term, whether an admissible inflected form is present,

⁸<https://github.com/langchain-ai/langgraph>

System	Automotive				Energy			
	BLEU↑	chrF2↑	TER↓	COMET↑	BLEU↑	chrF2↑	TER↓	COMET↑
Latxa-8B-Instruct (no fine-tuning)	20.36	56.82	82.05	80.73	17.62	58.46	87.12	82.96
+ FT (mined + OOD)	25.76	62.27	69.44	84.21	27.33	63.64	67.38	86.21
+ FT (mined + OOD + synthetic)	27.39	62.65	67.36	84.58	28.01	64.50	65.79	86.48

Table 2: Translation agent performance on the in-domain development sets. OOD denotes the out-of-domain EusParallel component. Best scores in **bold**.

and repairs the translation where it is not. The prompt includes few-shot examples of correct and incorrect realisations, with particular attention to the failure mode in which a repair strips a case ending and leaves the term grammatically stranded. Only flagged segments are rewritten, which limits the collateral edits that undirected self-refinement tends to introduce (Feng et al., 2025).

Step 5: Formatting reconciliation (deterministic). The edited translation is checked against the source for structural correspondence, including segment count, paragraph boundaries, and pass-through of symbolic or numeric-only chunks, before the final output document is assembled.

6 Results

6.1 Conditions and metrics

We evaluate under three terminology conditions: **noterm**, where the pipeline receives no terminology and the term-matching and verification stages are inactive; **proper**, where the pipeline receives the correct term base entries; and **random**, where the target side of each entry is replaced by a randomly chosen alternative. The **random** condition is a control since a system that simply ignores its terminology input would score identically across conditions, whereas a system that blindly copies would follow arbitrary terms as readily as correct ones.

We measure overall quality with the reference-free metric CometKiwi (Rei et al., 2022), and terminology usage with terminology success rate (TSR). The term base entries matched on the source side of a segment define its required target terms, and TSR is the proportion of those realised in the hypothesis, micro-averaged over all (segment, term) pairs. Since required terms almost never surface in citation form (§1), exact string matching is unusable. We instead match stems with a left word-boundary anchor and unconstrained suffixation, adapting the scorer released with the WMT25 task (Semenov

Terminology given	Automotive	Energy
None	0.7598	0.8097
Random	0.7618	0.8125
Proper	0.7662	0.8138

Table 3: Reference-free translation quality (CometKiwi) of the full pipeline under the three terminology conditions.

et al., 2025). Appendix B gives the procedure in full.

6.2 Overall translation quality

Table 3 reports CometKiwi. Quality is ordered **proper** > **random** > **noterm** in both domains, but the spread is small: 0.0064 for automotive and 0.0041 for energy. Adding a constraint-and-repair loop on top of a fine-tuned translator does not degrade general translation quality, which is the usual price of hard constrained decoding, and even the **random** terminology does not push quality below the unconstrained baseline as the verification agent integrates the imposed term into a well-formed sentence rather than pasting it in. The absolute gap between domains (≈ 0.05 CometKiwi) is much larger than any effect of terminology, and mirrors the development-set gap in §4.

6.3 Terminology success rate

Table 4 crosses the terminology supplied to the pipeline with the term set against which success is measured. Three observations stand out.

First, the pipeline uses the terminology it is given. Supplying proper terms raises proper-term success from 0.6128 to 0.7570 in the automotive domain (+14.4 points, +23.5% relative) and from 0.6013 to 0.7736 in energy (+17.2 points, +28.7% relative). The same holds in the control: supplying random terms raises random-term success from 0.6314 to 0.6896 and from 0.6721 to 0.7434.

Secondly, the interference between conditions is asymmetric and mild. Feeding random terms costs little proper-term accuracy (−1.2 points automo-

Domain	Terms given	Evaluated against	
		Proper	Random
Automotive	None	0.6128	0.6314
	Random	0.6013	0.6896
	Proper	0.7570	0.6315
Energy	None	0.6013	0.6721
	Random	0.6025	0.7434
	Proper	0.7736	0.6902

Table 4: Terminology success rate. Rows indicate the terminology supplied to the pipeline; columns indicate the term set against which success is measured. Diagonal cells (in bold) are the compliance figures of interest.

tive, +0.1 energy, i.e. within noise), meaning the system largely falls back on its own preferences. Conversely, feeding proper terms leaves random-term success at or slightly above baseline (0.6315 and 0.6902), which we attribute to stem overlap between the true and substituted entries in a fair number of cases; this also indicates that the absolute success-rate values carry some measurement noise from stem-level matching.

Finally, the **noterm** baselines are already high (0.60–0.67). The fine-tuned translator produces a majority of the required terms unprompted, a direct return on the in-domain mined and synthetic data of §3: terminology injection is refining an already domain-adapted model rather than compensating for a generic one.

7 Official Shared Task Results

Table 5 reports the official Track 1 scores released by the task organisers (Charkiewicz et al., 2026). AGENTEAK was entered for Spanish–Basque only, and all figures below are for that direction, pooled over the automotive and energy domains. We show our own row alongside the reference translations and three contrasting submissions; the full ranking over all systems and both language directions is available from the leaderboard.

Among the nineteen systems submitted for Spanish–Basque in Track 1, AGENTEAK places 13th on term success under both the surface and the lemmatised variant, and between 10th and 11th on the quality metrics (chrF++, COMET-22, CometKiwi-22, XCOMET-XXL and both MetricX-24 variants). Terminology compliance is thus the weaker side of our submission relative to the field, which is consistent with the ceiling we anticipated in §2 for soft, prompt-level constraint injection with a single bounded verification pass: the sys-

tems above us on term success cluster tightly at 89–90% surface success, a band that a fixed one-shot repair stage does not reach.

Two further observations follow from the official figures. First, the organisers’ term success is computed with their own source-side matcher and their own lemmatiser, and is therefore *not* comparable with the stem-level TSR of §6; the official figure is the higher of the two, which suggests our deterministic matcher (Step 1) marks more source occurrences as requiring a target term than the official scorer does, so our own numbers are measured against a stricter denominator. Second, the gap between surface and lemmatised term success is 10.3 points for AGENTEAK, against a median of approximately 7.5 points across the nineteen es–eu submissions. This is direct external evidence for the argument of §8: our failures are disproportionately failures of *form* rather than of lexical selection, with the prescribed term present but realised in a way that surface matching rejects.

8 Discussion

Morphology is the residual error source. The gap between what the terminology agent selects and what appears in the final output is largely morphological rather than lexical. In our development work the dominant failure mode was not a missing term but a term realised in citation form where the syntax required an inflected one, or conversely a repair that stripped a case ending. It also suggests that terminology success rate, computed over stems, systematically overstates usable accuracy for Basque considering that a term counted as present may still be ungrammatical in context. A task-appropriate metric for agglutinative targets would need to check morphosyntactic well-formedness, not only presence.

Why separate the agents. Ablating the pipeline into a single prompted model would be simpler, but the separation buys two things. The term table produced at Step 2 is an inspectable intermediate artefact, which matters in professional settings where terminology decisions are auditable. And the division of labour lets each component be sized to its task: a 4B reasoning model suffices for selection, while the 8B fine-tuned model is reserved for generation. The entire system uses no model above 8B parameters and requires no proprietary API.

System	chrF++		Term success		COMET		MetricX-24 XL	
	doc↑	para↑	surface↑	lemma↑	COMET-22↑	Kiwi-22↑	ref↓	QE↓
<i>gold</i> (references)	100.0	100.0	98.7	99.1	96.5	75.4	1.82	3.47
COZY (best overall)	75.3	70.1	89.5	96.4	88.8	75.0	3.24	3.77
Agentic-OPUS	73.5	67.5	89.9	96.0	87.6	75.1	3.53	3.84
SalamandraTA	73.3	67.5	86.0	95.2	87.0	74.6	3.66	4.01
MASC-Dec	72.7	67.1	82.1	86.8	84.5	72.2	4.61	4.71
AGENTEAK (ours)	68.8	61.3	78.4	88.7	84.6	72.5	4.71	4.67

Table 5: Official WMT26 Track 1 results for Spanish–Basque (Charkiewicz et al., 2026), pooled over both domains. XCOMET-XXL is omitted for space (75.7 for AGENTEAK, 83.2 for the best system). *gold* is the reference translation, not a participant, and bounds each metric. Comparison systems are shown for calibration only and are not an exhaustive ranking: nineteen systems were submitted for this direction.

9 Conclusion

We described AGENTEAK, the OpenU’s submission to the WMT26 Terminology Translation Task for Spanish into Basque. The system combines targeted in-domain data creation (Wikipedia bitext mining and prompted synthetic generation) with a three-agent pipeline that separates terminology selection, translation and verification across compact open-weight models. Domain-tagged fine-tuning improves the backbone by up to 10.4 BLEU, and terminology injection raises terminology success rate by 14–17 points at negligible cost to reference-free quality. Our analysis with randomised terminology shows that the system follows correct constraints considerably more readily than arbitrary ones, characterising both the robustness and the limits of soft, prompt-level terminology control for a morphologically rich low-resource target language. In the official evaluation (Charkiewicz et al., 2026), AGENTEAK placed mid-field among the nineteen Spanish–Basque submissions to Track 1 (§7), with the surface/lemma term-success gap confirming that its residual terminology errors are predominantly morphological.

Limitations

We report no human evaluation, and CometKiwi has not been validated specifically for Basque, so absolute values should be treated with caution, further evidenced by Falcão et al. (2024) who find that COMET-family metrics transfer unevenly to Spanish–Basque. Terminology success rate is computed by stem-level matching, which both over-counts (crediting morphologically ill-formed realisations) and potentially under-counts (missing legitimate allomorphy).

It is also worth noting that while the translator agent was fine-tuned on in-domain data and trans-

lation instruction, the verification agent was not, and the pipeline relies on its instruct-based nature and on the few-shot examples provided at inference to define the task and output required. A natural follow-up would be to fine tune a verification agent specifically on the task of verifying and editing Basque terminology alongside proper declension, and to evaluate this agent.

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 (WMT)*, pages 652–663.
- Mikel Artetxe, Itziar Aldabe, Rodrigo Agerri, Olatz Perez-de Viñaspre, and Aitor Soroa. 2022. [Does corpus quality really matter for low-resource languages?](#) In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7383–7390.
- Mikel Artetxe and Holger Schwenk. 2019. [Margin-based parallel corpus mining with multilingual sentence embeddings](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3197–3203.
- Toms Bergmanis and Mārcis Pinnis. 2021. [Facilitating terminology translation with target lemma annotations](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3105–3111.
- Adrian Charkiewicz, Pinzhen Chen, Thierry Etchegoyhen, Harritxu Gete Ugarte, Kamil Guttmann, Xu Huang, David Ponce, Artur Nowakowski, Frédéric Odermatt, 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.
- 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*, pages 3063–3068.
- Julen Etxaniz, Oscar Sainz, Naiara Perez, Itziar Aldabe, German Rigau, Eneko Agirre, Aitor Ormazabal, Mikel Artetxe, and Aitor Soroa. 2024. [Latxa: An open language model and evaluation suite for Basque](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 14952–14972.
- Júlia Falcão, Claudia Borg, Nora Aranberri, and Kurt Abela. 2024. [COMET for low-resource machine translation evaluation: A case study of English-Maltese and Spanish-Basque](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3553–3565.
- 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*, pages 878–891.
- Zhaopeng Feng, Yan Zhang, Hao Li, Bei Wu, Jiayu Liao, Wenqiang Liu, Jun Lang, Yang Feng, Jian Wu, and Zuozhu Liu. 2025. [TEaR: Improving LLM-based machine translation with systematic self-refinement](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3922–3938.
- Ivan Grubišić and Damir Korencic. 2025. [IRB-MT at WMT25 terminology translation task: Metric-guided multi-agent approach](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 1302–1334, Suzhou, China. Association for Computational Linguistics.
- Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. [How good are GPT models at machine translation? a comprehensive evaluation](#). *arXiv preprint arXiv:2302.09210*.
- Chris Hokamp and Qun Liu. 2017. [Lexically constrained decoding for sequence generation using grid beam search](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1535–1546.
- 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 (ICLR)*.
- Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. [Self-refine: Iterative refinement with self-feedback](#). In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2022. [Domain-specific text generation for machine translation](#). In *Proceedings of the 15th Biennial Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 14–30.
- Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023. [Adaptive machine translation with large language models](#). In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 227–237.
- NLLB Team. 2022. [No language left behind: Scaling human-centered machine translation](#). *arXiv preprint arXiv:2207.04672*.
- 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.
- 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.
- 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.
- Matt Post and David Vilar. 2018. [Fast lexically constrained decoding with dynamic beam allocation for neural machine translation](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1314–1324.
- 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.
- Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C. Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. [COMETkiwi: IST-unbabel 2022 submission for the quality estimation shared task](#). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645.
- Oscar Sainz, Naiara Perez, Julen Etxaniz, Joseba Fernandez de Landa, Itziar Aldabe, Iker García-Ferrero,

- Aimar Zabala, Ekhi Azurmendi, German Rigau, Eneko Agirre, Mikel Artetxe, and Aitor Soroa. 2025. [Instructing large language models for low-resource languages: A systematic study for Basque](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29136–29160, Suzhou, China. Association for Computational Linguistics.
- Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. [WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1351–1361.
- 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.
- 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 (WMT)*, pages 663–671.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. [Improving neural machine translation models with monolingual data](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 86–96.
- Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. [A study of translation edit rate with targeted human annotation](#). In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231.
- Xabier Soto, Olatz Perez-de Viñaspre, Gorka Labaka, and Maite Oronoz. 2019. [Neural machine translation of clinical texts between long distance languages](#). *Journal of the American Medical Informatics Association*, 26(12):1478–1487.
- Xabier Soto, Olatz Perez-de Viñaspre, Gorka Labaka, and Maite Oronoz. 2020. [Ixamed’s submission description for WMT20 biomedical shared task: Benefits and limitations of using terminologies for domain adaptation](#). In *Proceedings of the Fifth Conference on Machine Translation*, pages 875–880.
- Gorka Urbizu, Iñaki San Vicente, Xabier Saralegi, Rodrigo Agerri, and Aitor Soroa. 2022. [BasqueGLUE: A natural language understanding benchmark for Basque](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1603–1612.
- David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. [Prompting PaLM for translation: Assessing strategies and performance](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 15406–15427.
- Minghao Wu, Yulin Yuan, Gholamreza Haffari, and Longyue Wang. 2024. [\(perhaps\) beyond human translation: Harnessing multi-agent collaboration for translating ultra-long literary texts](#). *arXiv preprint arXiv:2405.11804*.
- Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024. [A paradigm shift in machine translation: Boosting translation performance of large language models](#). In *The Twelfth International Conference on Learning Representations (ICLR)*.
- An Yang et al. 2025. [Qwen3 technical report](#). *arXiv preprint arXiv:2505.09388*.
- Jaume Zaragoza-Bernabeu, Gema Ramírez-Sánchez, Marta Bañón, and Sergio Ortiz Rojas. 2022. [bi-cleaner AI: Bicleaner goes neural](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 824–831.

A Fine-Tuning Setup

Prompt. The system prompt (Figure 2) is written in Spanish and carries the domain as an explicit `Ámbito:` line. It is wrapped in the Llama-3.1 chat template that Latxa inherits (Figure 3). Prompt tokens are masked and the loss is computed on the Basque completion only. The same system prompt, with the same `Ámbito:` value, is used at inference inside the pipeline.

Model and adapter.

- bf16, no quantisation; SDPA attention; gradient checkpointing.
- LoRA $r = 16$, $\alpha = 32$, dropout 0.1, bias none.
- Targets: q, k, v, o, gate, up, down.
- 41.9M trainable parameters

Optimisation.

- Max length 1024, packing disabled, seed 42.
- LR $1e-4$, cosine schedule, warmup ratio 0.03, weight decay 0.01.
- Batch 4 per device $\times$ 4 accumulation steps (effective 16).
- Evaluate and checkpoint every 100 steps; best combined dev loss restored;

Eres un traductor experto del castellano al euskera (euskara batua), {OPENING}. Ámbito: {TAG}. Traduce con precisión y naturalidad, respetando la declinación y la sintaxis propias del euskera, {REGISTER} Conserva exactamente todas las cifras, unidades y nombres propios tal como aparecen en el original, así como el formato del texto de partida (mayúsculas, puntuación, saltos de línea y etiquetas). No añadas explicaciones, comentarios ni el texto de partida. Responde únicamente con la traducción.	
Slot	Value
<i>automotive / energy</i>	
{OPENING}	especializado en el ámbito de la automoción de la energía
{TAG}	automoción energía
{REGISTER}	y empleando la terminología asentada de ese ámbito.
<i>general</i>	
{OPENING}	capaz de trabajar con textos de cualquier ámbito
{TAG}	general
{REGISTER}	y empleando un registro estándar.

Figure 2: System prompt, with the three domain-conditioned slots.

< begin_of_text >< start_header_id >system < end_header_id > {SYSTEM PROMPT}< eot_id > < start_header_id >user< end_header_id > {SPANISH SOURCE}< eot_id > < start_header_id >assistant< end_header_id > {BASQUE TRANSLATION}< eot_id >	
---	--

Figure 3: Rendered training example. Everything above the rule is the masked prompt; the loss is computed only on the completion below it.

B Computing Terminology Success Rate

Figure 4 summarises the procedure. Required terms are determined by matching the term base against the *source* segment with the same deterministic matcher used at Step 1 of the pipeline (§5), so the scorer and the system see identical source-side matches; repeated occurrences of an entry within a segment count once. Each required target term is lowercased and reduced to the invariant prefix its inflected forms share, and counts as realised if that stem occurs in the hypothesis preceded by a word boundary and followed by anything. TSR is the ratio of realised to required terms, pooled across the test set and reported per domain and per condition.

Our scorer adapts the one released with the WMT25 task (Semenov et al., 2025); the stem reduction and boundary anchoring are our addi-

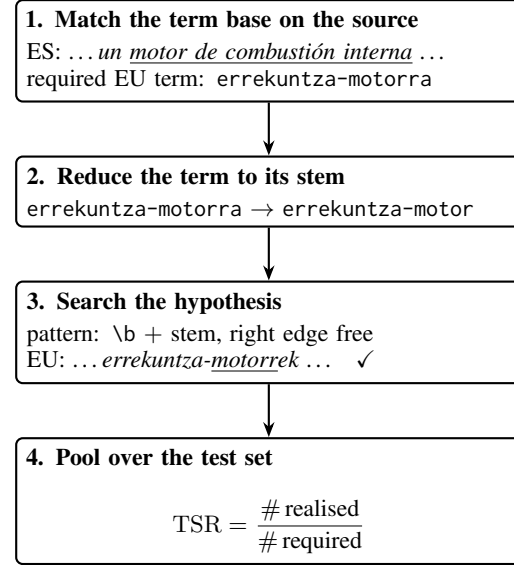


Figure 4: Terminology success rate, computed for one required term. The left word boundary stops a stem being credited inside an unrelated word.

tions, so our figures are *not* directly comparable with those of the unmodified official scorer. Two biases follow from the design. The measure overcounts, because a stem match certifies presence but not grammaticality since a term left morphologically stranded still scores as a success (§8). It is also noisy, because proper and random entries sometimes share a stem, which inflates success rates measured against the term set that was not supplied.