

LiLa at WMT 2026 General MT Task: Exploring Beyond NMT and Learning the Limits of LLM Adaptation for Ligurian and Ladin

Mirelle Candida Bueno¹, Lucas Domingues², Samuel Frontull³, Jean Maillard⁴, Sushil Garg⁵

¹Motorola Mobility Brasil Ltda., Jaguariúna, SP, Brazil,

²Instituto Eldorado, Brazil

³University of Innsbruck, Austria,

⁴Council for Ligurian Linguistic Heritage, Italy

⁵Motorola Mobility LLC, San Jose, CA, USA

Abstract

We describe LiLa, our submission to the constrained track of the WMT 2026 General Machine Translation Task. Motivated by recent advances in large language models (LLMs) and their increasing availability as open models, we investigate the adaptation of LLMs for Ligurian and Ladin, two low-resource languages included in the shared task. We evaluate several adaptation strategies for Gemma-4, including Instruction Tuning (IT-only), continuous pre-training followed by Instruction Tuning (CPT+IT), and Low-Rank Adaptation (LoRA), and compare them with a dedicated encoder-decoder pipeline based on fine-tuned No-Language Left Behind (NLLB) models. Our experiments show that, under the evaluated conditions, fine-tuned NLLB models achieve higher automatic translation scores than the investigated LLM adaptation strategies. Based on these results, our final submission adopts a fine-tuned NLLB-based translation pipeline. We discuss the main observations from the different approaches explored during development. We release the code and models produced in these experiments through a publicly accessible GitHub repository¹.

1 Introduction

A truly multilingual translation system should not treat the world’s languages as a hierarchy in which a small number of high-resource pairs receive most of the modeling effort. The practical goal of machine translation is broader: to make information accessible across languages, including those with limited training resources and comparatively little representation in existing benchmarks. Recent multilingual models have substantially expanded the range of languages that can be addressed, but important differences in translation quality remain. The WMT26 shared task (Kocmi et al., 2026) offers an opportunity to examine this problem in a

diverse, challenging multilingual setting. This motivates an important question for low-resource languages: can general-purpose LLMs be effectively adapted to (unseen) languages with limited digital resources, and how do they compare with specialized neural machine translation (NMT) approaches when adapted to the same low-resource scenario?

Recent studies have explored LLM adaptation for multilingual and low-resource settings, including comparisons between adapted LLMs and fine-tuned encoder-decoder models (Scalvini et al., 2025). Initiatives focused on evaluating LLMs in dialectal and low-resource language scenarios (Robinson et al., 2026; Kyslyi et al., 2025; Joshi et al., 2025) have provided valuable insights into the challenges of adapting these models beyond high-resource and standardized varieties. Nevertheless, developing effective adaptation strategies for under-resourced languages remains challenging, as limited data availability, linguistic diversity, and computational constraints continue to hinder the deployment of robust multilingual LLMs.

Motivated by these developments, we investigate whether LLMs can be effectively adapted to Ligurian and Ladin, two languages of Italy included for the first time in the WMT General Machine Translation Shared Task.

Specifically, we adapt Gemma 4 (Gemma Team et al., 2026) using several strategies, including instruction tuning (Zhang et al., 2026), continuous pre-training (Ke et al., 2023) followed by instruction tuning, and parameter-efficient fine-tuning with LoRA (Hu et al., 2021). The Gemma 4 model was selected because, at the time the research was conducted, it represented the most recently released open model offering multilingual capabilities (Finkelstein et al., 2026).

Beyond translation quality, we also evaluate their ability to process and preserve common structured and formatted inputs, such as HTML documents, Markdown files, and JSON-formatted data.

¹<https://github.com/schtailmuel/wmt26-lila/>

Such capabilities could offer an alternative to traditional translation pipelines that require separate mechanisms for handling document structure and content transformation. To this end, we construct a document-level corpus containing structured and formatted examples. We compare these LLM-based adaptation strategies against established encoder-decoder translation models, using NLLB (Costa-jussà et al., 2024), which we fine-tune specifically for the English to Ligurian and English to Ladin translation directions using the same collected resources employed in our experiments. Due to hardware limitations during the training phase, the NLLB models employed in this study were restricted to a maximum of 1.3 billion parameters.

Our experiments reveal a clear practical outcome: the fine-tuned NLLB models outperform all investigated Gemma-4 variants on the considered benchmarks. Qualitative analysis of the adapted LLM outputs shows that continuous pre-training can enable the generation of fluent language-specific text in selected cases; however, instruction-tuned LLMs remain less reliable for constrained machine translation tasks, where they may hallucinate, drift from the source content, or fail to preserve required document structure.

Based on these findings, we select the NMT approach for our final WMT 2026 submission. The submitted systems, released as LiLa², rely on NLLB-200 distilled 1.3B, combined with a format-aware inference pipeline supporting plain text, HTML, and JSON to process and preserve the structure of the heterogeneous text formats included in the WMT blind test set. The main contributions of this system description are:

- A reproducible NLLB fine-tuning pipeline for English–Ligurian and English–Ladin translation;
- A structure-aware inference pipeline that preserves HTML markup, JSON structure, placeholders, and document layout;
- An empirical comparison between fine-tuned NLLB models and several Gemma-4 adaptation strategies for two low-resource languages;

We hope that these findings contribute to further developments in the adaptation of multilingual

²Acronym for Ligurian–Ladin; in Ladin it means both “little sheep” and lilac color

LLMs for low-resource languages and encourage future research on their use as flexible alternatives to specialized translation systems.

2 Data

Ladin and Ligurian are both Romance languages and face similar challenges in terms of their representation and support in modern NLP systems. At the syntactic level, the two languages share relatively few similarities that could facilitate cross-lingual transfer for machine translation: using Ladin translations as hypotheses and Ligurian translations as references results in a BLEU score of only 2.52.³ We therefore aim to collect as much high-quality data as possible for both languages, focusing on valid, reliable, and representative resources that provide sufficient coverage to train and adapt models for each language (independently).

A key change in the WMT26 shared task is the emphasis on structured translation, requiring systems to translate content embedded in formats such as HTML, Markdown, and JSON while preserving the underlying structure and producing valid output. To address this setting, we construct two complementary types of training data: (i) a sentence-level corpus, primarily used for supervised NMT fine-tuning, and (ii) a document-level corpus for adapting LLMs to contextual and structured translation.

Both corpora include authentic and back-translated monolingual data. The sentence-level data focuses on English–target language translation pairs, whereas the document-level data provides longer contextual examples and diverse document structures, including HTML, Markdown, and JSON-like representations. The sizes and composition of the different corpora are summarized in Table 1.⁴

2.1 Ladin Corpus

Ladin comprises multiple written and standardised varieties (Frontull et al., 2025). The WMT26 shared task specifically targets the Val Badia variety. We therefore focus our corpus construction for Ladin exclusively on Val Badia Ladin and aim to use data that can be identified as belonging to this variety, rather than mixing different Ladin varieties.

We construct a multilingual corpus from publicly available parallel and monolingual resources. We

³One notable linguistic feature nevertheless connects the two languages: both are among the relatively few languages that use the letter *ë* (e with diaeresis).

⁴In the Table, we only report document-level statistics

Component	Ladin		Ligurian	
	Words	Chars	Words	Chars
Target	17.86M	94.91M	4.46M	24.93M
English	16.48M	95.85M	1.44M	8.06M
Italian	16.66M	101.19M	1.35M	8.07M
French	13.08M	77.80M	0.54M	3.11M
Spanish	12.68M	75.05M	0.51M	2.95M
German	15.91M	104.09M	–	–

Table 1: Dataset statistics by language and data component. Values are reported in millions (M).

use the Ladin Language Corpus from Motorola’s Language Revitalization Initiative⁵, the monolingual datasets^{6 7} and the parallel datasets^{8 9 10 11 12} released in Frontull and Moser (2024); Frontull and Haller-Seeber (2026). We do not use the remaining OPUS resources included in the bundle because they are not consistently restricted to the Val Badia variety.

Document-Level Corpus The document-level corpus was constructed specifically for continuous pre-training and instruction-tuning experiments with Gemma-4. Since no document-level corpus was available for Ladin, we extracted document-level texts from articles in the archive of *La Usc di Ladins*. The collected articles were then classified and filtered to retain texts in the Val Badia variety of Ladin. The resulting Ladin documents were subsequently back-translated into English, German, Italian, French, and Spanish using an internal LLM-based back-translation pipeline. To reduce lexical and semantic noise in the translation of this low-resource language, the pipeline augments the translation prompt with dictionary information from Moling (2016), providing multilingual document-level training material with longer and more coherent contextual structure. Specifically, we augmented the prompt with the dictionary definitions of words that can be unambiguously mapped to a single part of speech¹³. These definitions provide additional semantic information to guide the translation of the corresponding words.

⁵<https://www.motorola.com/us/en/about/endangered-indigenous-languages-data-set>

⁶sfrontull/la-usc_valbadia_loresmt24

⁷sfrontull/south-tyrol-weather-lld

⁸sfrontull/autonomia-lld_valbadia-ita

⁹sfrontull/pinocchio-lld_valbadia-ita

¹⁰sfrontull/lld_valbadia-ita

¹¹sfrontull/stiftungsparkasse-lld_valbadia-ita

¹²sfrontull/south-tyrol-weather-lld-deu

¹³i.e., words whose dictionary entries correspond to a single possible part of speech (noun, adjective, or verb)

To further expose LLMs to realistic document-processing scenarios, subsets of the translated documents are converted into multiple representations, including HTML, $\LaTeX$, Markdown, and plain-text formats. We also create emoji-enriched variants to increase the diversity of formatting patterns and evaluate the ability of adapted models to preserve document-level structure during translation.

Sentence-Level Corpus The sentence-level corpus is used for NLLB fine-tuning. The back-translation of the document-level data did not consistently preserve sentence-level alignment, making it unsuitable for direct inclusion in the sentence-level corpus. For this corpus, we therefore relied on the *la-usc_valbadia_loresmt24* sentence-level monolingual ladin dataset¹⁴ (the largest currently available monolingual dataset for Val Badia Ladin), which we back-translated into English. To improve the quality of the generated English translations, we use a retrieval-augmented back-translation approach. For each Ladin sentence, we retrieve semantically related examples from the available authentic parallel corpus using BM25 (Robertson and Walker, 1994). We initially retrieve the top 50 candidates and apply Maximal Marginal Relevance (MMR) re-ranking to select a diverse subset of $k = 10$ demonstrations. This retrieval-augmented approach has previously been shown to be effective for Ladin, providing a reliable way to improve the quality of back-translated data (Frontull and Ströhle, 2026). For a subset of document-level data for which sentence-level alignment could be reliably maintained, we extracted the aligned sentence pairs and added them to the corpus.

2.2 Ligurian Corpus

For Ligurian, the multilingual corpus was created from a mixture of publicly available resources – ZenaMT (Haberland et al., 2024), OLDI-Seed (Dale et al., 2025), SMOL (Caswell et al., 2025) – as well as parallel and monolingual data privately held by the Council for Ligurian Linguistic Heritage¹⁵ and shared with us for the purposes of participating in this shared task.

Document-Level Corpus This portion of the corpus, consisting of Ligurian paragraphs and longer documents aligned with English, German, Italian, French and Spanish, was used for the LLM experi-

¹⁴sfrontull/la-usc_valbadia_loresmt24

¹⁵<https://conseggio-ligure.org>

ments with Gemma-4. It was made of a combination of natively multilingual document-level data and backtranslated data. Unlike for Ladin, no dictionary information was included in generation, but targeted manual post-editing was used to fix common mistranslations. Similar to Ladin, the goal was to expose LLMs to a mixture of realistic document formats. As much of the data was originally in Markdown format, this was preserved as Markdown in 70% of cases, and converted to HTML in the remaining 30% of cases.

Sentence-Level Corpus We used English-Ligurian parallel data for NLLB fine-tuning. Much of this data was natively in document format, and was segmented and sentence-aligned across languages to turn it into the sentence-level data needed for NLLB training. For the portions of the data that were backtranslated, a manual post-editing step was conducted to fix words and phrases that were observed to be commonly mistranslated.

2.3 Evaluation Benchmarks

For development and evaluation, we use publicly available multilingual benchmarks. We evaluate both Ladin and Ligurian on the FLORES+ evaluation corpus (Costa-jussà et al., 2024; Frontull et al., 2025). For Ligurian, we additionally evaluate on the BOUQuET benchmark (Andrews et al., 2025).¹⁶

3 Specialized NMT Systems

We fine-tune NLLB-200 distilled 1.3B¹⁷, a multilingual encoder-decoder model from the NLLB family (Costa-jussà et al., 2024). NLLB has demonstrated strong performance in multilingual and low-resource translation settings, making it an established reference point for evaluating adaptation approaches for low-resource languages. We fine-tune separate models for the two translation directions considered in our experiments (English–Ligurian and English–Ladin). The target languages are specified using the NLLB language tags `lij_Latn` for Ligurian and `lld_Latn` for Ladin. Models are fine-tuned using the Adafactor optimizer with an effective batch size of 64, a learning rate of 1.5×10^{-4} , a linear warmup ratio of 0.03, and a weight decay of

0.001. We train for up to 3 epochs on an NVIDIA RTX PRO 6000 Blackwell Server Edition GPU (96 GB VRAM) leveraging mixed BF16/TF32 precision, and select the best checkpoint based on validation BLEU score with a patience of three evaluations. At inference time, we decode via beam search with a beam width of 4. All NLLB experiments use a fixed random seed of 42 to ensure reproducibility.

4 Generative LLM Adaptation

Prior literature establishes that instruction tuning frequently combined with task-specific fine-tuning or continuous pre-training (CPT) constitutes an effective paradigm for Neural Machine Translation (NMT), particularly when adapting general-purpose LLMs to low-resource target languages (Xu et al., 2023; Zhang et al., 2025a,b). Consequently, we investigated training scenarios that combine CPT with instruction tuning to evaluate prospective performance gains in translation quality. Notably, baseline performance may be constrained given that the target low-resource languages were absent from the base LLMs pre-training distribution. To assess these effects, we evaluated three experimental configurations:

1. **Zero-Shot Baseline:** Out-of-the-box evaluation of Gemma-4 without task-specific fine-tuning or parametric adaptation.
2. **Direct Instruction Tuning (IT):** Task-specific supervised instruction tuning applied directly to the base Gemma-4 architecture. For this pipeline, experiments were run with both parameter-efficient fine-tuning via low-rank adaptation (LoRA), applied to attention and feed-forward layers while keeping the base language model frozen, and full fine-tuning, which optimizes all model weights to maximize adaptation capacity.
3. **Sequential Adaptation (CPT+IT):** Supervised instruction tuning applied to Gemma-4 following an initial phase of domain-specific Continuous pre-training (CPT). The Continuous Pre-Training phase aimed to expand the Gemma-4 models proficiency in Ladin and Ligurian via a next-token prediction objective. To mitigate catastrophic forgetting, the corpus combined these target low-resource languages with selected high-resource languages.

¹⁶The Ladin subset of the BOUQuET benchmark was not publicly available during system development.

¹⁷<https://huggingface.co/facebook/nllb-200-distilled-1.3B>

4.1 Model Architecture and Optimization

The experimental framework for CPT and instruction tuning uses the google/gemma-4-E4B¹⁸ decoder-only architecture. To maximize linguistic adaptation, we adopted a Full Fine-Tuning (FT) approach. Training was executed on an NVIDIA A100 Ampere-class GPU using the Unsloth engine. We used BF16 (BFloat16) precision with FP16 fallback to ensure stability.

4.2 Continuous Pre-Training (CPT)

This section outlines the experimental framework for the CPT pipeline.

Data Organization To ensure the injection of deep linguistic understanding, we created a sentence and document level dataset with rigorous data preprocessing for continuous pre-training. The dataset integrates multi-format document-level data (HTML, $\LaTeX$, JSON, Markdown, and plain text). Synthetic sentence-level samples were removed to increase data quality, leaving the final filtered dataset with 706,330 samples and 85,233,986 words. To maintain training stability, we applied pre-compiled regular expressions to remove artifacts (e.g., reporter initials, editor credits, and signature headers). The data was partitioned using a fixed random seed (3407) to ensure a strict 99% / 1% train-test split.

Performance Tracking and Metrics Model evaluation relied on real-time monitoring and validation at 25% intervals of training progress. We tracked perplexity on the evaluation set and retained the top two checkpoints. Models were trained for two epochs, with the best checkpoint achieving a perplexity of 4.84 on the evaluation set.¹⁹ The trainer was configured with a 4096-token context window, employing sequence packing, i.e., concatenating multiple document sequences up to the maximum context length with end-of-sequence delimiters, to eliminate redundant padding tokens and maximize compute efficiency during training. Training was conducted using a learning rate of 2×10^{-4} and a batch size of 2.

4.3 Instruction Tuning (IT)

To adapt pretrained LLMs for translation in low-resource settings – specifically Ligurian and Ladin

paired with English and Italian – we implement a specialized Supervised Fine-Tuning (SFT) pipeline using a balanced sampling approach that prioritizes translation into the low-resource languages:

Dataset Curation, Probabilistic Sampling, and Deduplication

The dataset preparation pipeline utilizes a broad multilingual corpus spanning seven languages: English (eng), Ladin (lld), Ligurian (lij), Spanish (spa), Italian (ita), German (deu), and French (fra). To counteract the dominance of high-resource language pairs during optimization, data balancing is pre-computed offline using a targeted *Balanced Batch Approach*. Training batches are sampled with a fixed composition: 40% English–Ladin (eng→lld), 40% English–Ligurian (eng→lij), 10% non-English source languages to Ladin ($\{\text{deu, spa, fra, ita}\} \rightarrow \text{lld}$), and 10% non-English source languages to Ligurian ($\{\text{deu, spa, fra, ita}\} \rightarrow \text{lij}$). By heavily weighting translation directly into lld and lij, this 40/40/10/10 distribution prioritizes under-resourced target generations while preserving cross-lingual transfer from high-resource European pivot languages. All parallel pairs are mapped to a unified, zero-shot instruction prompt.²⁰

Sliding-Window Deduplication To prevent immediate response repetitions and mitigate recency bias, we execute a sliding-window deduplication pass over target sentence output hashes, removing duplicate outputs within localized context windows.

Training parameters During training, we used a sequence length of 1024 tokens for the google/gemma-4-E4B direct instruction tuning via full fine-tuning, and 2048 tokens for both the CPT + IT and Direct Instruction Tuning via LoRA models. For the LoRA experiments, we set the rank (r) and alpha (α) parameters to 16. The learning rate was set to 2×10^{-5} for full fine-tuning and 2×10^{-4} for LoRA. All instruction tuning experiments were trained for 3 epochs.

Performance Tracking and Metrics Validation was run on a held-out validation set of 2,000 samples and performed at every quarter of an epoch. Checkpoint selection was based exclusively on performance on this validation set. After model selection was complete, the selected checkpoint was

¹⁸<https://huggingface.co/google/gemma-4-E4B>

¹⁹This model checkpoint will be made publicly available on Hugging Face.

²⁰We format all inputs using the template: "Translate from {SRC_LANG} to {TGT_LANG}: {SRC_TEXT}", where language names are expanded in full English.

Model Variant	Ladin (lld)		Ligurian (lij)			
	FLORES+		FLORES+		BOUQuET	
	BLEU ↑	chrF++ ↑	BLEU ↑	chrF++ ↑	BLEU ↑	chrF++ ↑
Gemma-4-E4B-IT (Zero-shot)	3.37	31.02	4.01	33.60	3.30	26.52
Gemma-4-E4B-IT (Direct IT - Full FT)	21.38	48.77	25.63	54.60	23.47	48.61
Gemma-4-E4B-IT (Direct IT - LoRA)	23.21	50.40	27.15	55.35	24.83	49.89
Gemma-4-E4B (CPT + Full FT IT)	2.52	25.01	2.95	27.84	1.47	19.14
NLLB 1.3B (Full Fine-Tuning)	31.41	55.24	32.50	57.40	32.62	53.98

Table 2: System performance comparison across Gemma-4 adaptation strategies and NLLB fine-tuning setups on the FLORES+ and BOUQuET datasets.

evaluated once on the FLORES+ and BOUQuET test sets, which were not used for checkpoint selection or hyperparameter tuning. Translation quality was measured using sacreBLEU and chrF++ for the evaluated English-to-Ladin and English-to-Ligurian directions. During generation, we enabled sampling and generated one candidate translation per input using a temperature of 0.7, top-p of 0.9, and a repetition penalty of 1.1.

5 Results and Analysis

Table 2 presents a performance comparison across our primary fine-tuned NLLB model and several exploratory Gemma-4 adaptation strategies on the FLORES+ and BOUQuET test set. All evaluation metrics were computed using standard sacreBLEU settings via the Hugging Face evaluate library.

Across the exploratory setups, zero-shot google/gemma-4-E4B-IT²¹ struggled significantly on both low-resource language pairs, confirming that the model lacks sufficient proficiency in both Ladin and Ligurian. Fine-tuning substantially improved performance, with LoRA consistently outperforming full fine-tuning and emerging as the strongest LLM adaptation strategy. Interestingly, combining CPT with full instruction fine-tuning obtained lower automatic scores than the zero-shot instruction-tuned checkpoint across all three benchmarks, with the results suggesting a strong tendency for the model to fallback to the task of next token prediction instead of instruction following

Despite the competitive results obtained with LLM adaptation, the fully fine-tuned NLLB 1.3B system achieved consistently higher automatic scores than the evaluated Gemma-4 variants across language pairs, datasets, and metrics.

5.1 Translating the blind test set

We applied the fine-tuned NLLB model using a structure-aware inference pipeline rather than directly translating each document as a single sequence, as illustrated in Figure 1.²² This was necessary because the WMT 2026 blind test set documents can be substantially longer than the maximum sequence length used during training. We therefore introduced a dynamic segmentation and reconstruction procedure implemented through a custom TranslationPlanner. The planner decomposes each document into translatable and non-translatable spans while preserving the original document structure. For natural-language content, sentence boundaries are detected using the rule-based pysbd segmenter (Sadvilkar and Neumann, 2020). Line breaks, blank segments, and leading and trailing whitespace are retained as non-translated spans, while only linguistically meaningful segments containing alphabetic characters are passed to the translation model. This allows long documents to be processed as smaller translation units while enabling the generated segments to be reconstructed into the original document layout.

Format-Aware Parsing The inference procedure dynamically selects a processing strategy based on the instructions associated with each WMT instance. Three document types are handled separately: plain text, HTML-formatted content, and JSON data. For HTML documents, the input is first split around HTML tags using a regular-expression-based parser. HTML tags are treated as immutable structural elements and are therefore never passed to the translation model. Only the text between tags is sentence-segmented and translated. After generation, the translated text segments are recombined with the original tags in their original or-

²¹<https://huggingface.co/google/gemma-4-E4B-it>

²²The system would otherwise break most of the markup or structured data.

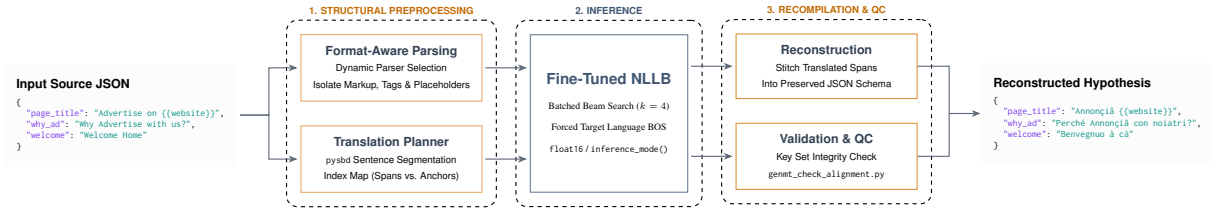


Figure 1: End-to-end format-aware machine translation pipeline preserving structural anchors and dynamic markup.

der. For JSON examples, we employ a custom streaming JSON parser that tracks the syntactic state of the JSON document using a context stack. The parser distinguishes object keys from values and selectively extracts string values that occur in translatable positions. JSON keys, brackets, commas, colons, numeric values, and other structural elements are preserved without modification. Extracted string values are subsequently processed using the same sentence-level translation planner as ordinary text. The translated values are then serialized back into the original JSON structure. Template placeholders are also explicitly protected during preprocessing. Strings matching the pattern `{{ . . . }}` are isolated from the surrounding natural-language text before sentence segmentation. The placeholders are consequently not exposed to the translation model and are inserted unchanged during document reconstruction. This prevents variable names and template expressions from being translated, dropped, or otherwise modified.

Translation and Document Reconstruction After structural preprocessing, the resulting translatable segments are collected and processed in batches of 32 examples. Each batch is tokenized with padding and truncation, using a maximum input length of 512 tokens. The fine-tuned NLLB model generates up to 256 new tokens per segment. We use beam search with four beams (`num_beams = 4`) and explicitly force the beginning-of-sequence token corresponding to the desired target language. Following generation, the translated segments are reconstructed using the original translation plan. The reconstruction process combines generated translations with all non-translated spans, including whitespace, line breaks, HTML markup, JSON syntax, and protected placeholders. Thus, the inference system produces complete document-level hypotheses while restricting neural generation to linguistically relevant segments. This design allows documents exceeding the model’s input context to be processed without

sacrificing their original structure.

Post-Inference Validation and Quality Control

We performed multiple levels of validation after generation to ensure that the resulting WMT hypotheses were both linguistically generated and structurally aligned with their corresponding source documents. First, local validation checks verify that the number of output documents matches the number of input documents for the selected target language, that document identifiers are unique, and that all source fields remain unchanged. The generated hypothesis is added as a separate field without altering the original WMT metadata.

For JSON documents, the validation procedure compares the source and hypothesis representations to verify that the JSON root type is preserved and that the number of elements remains consistent. For dictionary-like structures, the set of object keys is additionally checked to ensure that keys have not been translated or removed. Placeholder sets are compared between the source and hypothesis to verify that all template variables are preserved. For HTML documents, structural consistency is verified by comparing the number of paragraph tags and the complete ordered sequence of HTML tags between the source and generated hypothesis. These checks ensure that translation does not introduce, remove, or reorder markup.

Finally, the generated output is passed through the WMT `genmt_check_alignment.py`²³ checker. The checker is executed against both the original WMT input and the generated translation file, and the inference pipeline requires a successful return code before the output is accepted. Consequently, the complete test-time procedure combines fine-tuned NLLB generation, sentence-level segmentation, format-aware parsing, batched beam-search decoding, structural reconstruction, and automated alignment validation.

²³https://github.com/wmt-conference/wmt-collect-translations/blob/main/genmt_check_alignment.py

6 Conclusion

We presented LiLa, our submission to the WMT 2026 General MT shared task for Ligurian and Ladin. The final system combines fine-tuned NLLB models with a deterministic, structure-aware inference pipeline. In our experiments, the encoder-decoder NLLB system achieved higher automatic evaluation scores than the evaluated decoder-only Gemma-4 adaptation strategies. Specifically, continuous pre-training followed by Instruction Tuning (CPT + IT) obtained lower BLEU and chrF++ scores than both direct instruction tuning and the zero-shot instruction-tuned checkpoint. Because the configurations differed in initialization, data composition, optimization, and training budget, this result should not be interpreted as evidence that continued pretraining is generally detrimental. Limited instruction tuning after CPT, catastrophic forgetting, and the aggressive full-parameter CPT learning rate are possible explanations that require controlled ablations. Future work will explore hybrid approaches that combine the deterministic processing and translation-focused architecture of encoder-decoder MT systems with the broader contextual generation capabilities of LLMs for low-resource regional languages.

Limitations Our analysis of the continuous pre-training model is limited: beyond automatic benchmark scores, we only performed qualitative inspection on a small set of examples. The claims about its fluency and failure modes should therefore be interpreted as exploratory rather than as a comprehensive human evaluation. Additionally, we utilized the same dataset for both the CPT and IT phases, which may have limited the model’s ability to effectively disentangle general domain adaptation from instruction-following behaviors. In particular, the model showed limited ability to adhere to the instruction-following task after IT, instead tending to revert to the objective learned during CPT. One possible explanation is that the instruction-tuning stage was simply insufficiently long for the model to acquire the desired instruction-following behavior; given our limited computational budget, we were unable to determine whether this behavior would have been mitigated with additional instruction-tuning steps.

Furthermore, while Gemma was trained on a structure-aware, document-level translation task, the NLLB model was trained on non-structure-aware, sentence-level translation. A notable lim-

itation of this study is the absence of an experiment training the LLM on the same simplified task as NLLB to establish a direct performance baseline for Gemma. Finally, our experimental scope was heavily constrained by limited computational resources; restrictions on GPU availability and overall training time prevented us from performing exhaustive hyperparameter optimization or extending the training schedule to fully mitigate the challenges observed in the CPT + IT pipelines.

Acknowledgments

The authors would like to thank Motorola Mobility, Lenovo Foundation, and Eldorado Research Institute for their generous support and contributions to this research. We extend our sincere appreciation to Juliana Rebellato and Paul Videsott for initiating this collaboration and to Georg Moser, for supporting it. We also thank the organizers of the WMT General MT Shared Task for coordinating the shared task and fostering a collaborative research environment that enables advances in multilingual language technologies and the reviewers for their valuable comments and questions that helped to improve the presentation.

References

- Pierre Andrews, Mikel Artetxe, Mariano Coria Meglioli, Marta R. Costa-jussà, Joe Chuang, David Dale, Mark Duppenthaler, Nathaniel Paul Ekberg, Cynthia Gao, Daniel Edward Licht, Jean Maillard, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Eduardo Sánchez, Ioannis Tsiamas, Arina Turkatenco, Albert Ventayol-Boada, and Shireen Yates. 2025. [BOUQuET : dataset, benchmark and open initiative for universal quality evaluation in translation](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27515–27535, Suzhou, China. Association for Computational Linguistics.
- Isaac Caswell, Elizabeth Nielsen, Jiaming Luo, Colin Cherry, Geza Kovacs, Hadar Shemtov, Partha Talukdar, Dinesh Tewari, Baba Mamadi Diane, Djibrila Diane, Solo Farabado Cissé, Koulako Moussa Doumbouya, Edoardo Ferrante, Alessandro Guasoni, Christopher Homan, Mamadou K. Keita, Sudhamoy DebBarma, Ali Kuzhuget, David Anugraha, and 5 others. 2025. [SMOL: Professionally translated parallel data for 115 under-represented languages](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 1103–1123, Suzhou, China. Association for Computational Linguistics.
- Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, 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, John Hoffman, and 20 others. 2024. [Scaling neural machine translation to 200 languages](#). *Nature*, 630(8018):841–846.
- David Dale, Laurie Burchell, Jean Maillard, Idris Abdulmumin, Antonios Anastasopoulos, Isaac Caswell, and Philipp Koehn. 2025. [Findings of the WMT 2025 shared task of the open language data initiative](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 495–502, Suzhou, China. Association for Computational Linguistics.
- Mara Finkelstein, Isaac Caswell, Tobias Domhan, Jan-Thorsten Peter, Juraj Juraska, Parker Riley, Daniel Deutsch, Geza Kovacs, Cole Dilanni, Colin Cherry, Eleftheria Briakou, Elizabeth Nielsen, Jiaming Luo, Kat Black, Ryan Mullins, Sweta Agrawal, Wenda Xu, Erin Kats, Stephane Jaskiewicz, and 2 others. 2026. [TranslateGemma technical report](#). *Preprint*, arXiv:2601.09012.
- Samuel Frontull and Simon Haller-Seeber. 2026. [Align-Fix: A tool for parallel corpora augmentation and refinement](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 215–224, Rabat, Morocco. Association for Computational Linguistics.
- Samuel Frontull and Georg Moser. 2024. [Rule-based, neural and LLM back-translation: Comparative insights from a variant of Ladin](#). In *Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024)*, pages 128–138, Bangkok, Thailand. Association for Computational Linguistics.
- Samuel Frontull, Thomas Ströhle, Carlo Zoli, Werner Pescosta, Ulrike Frenademez, Matteo Ruggeri, Daria Valentin, Karin Comploj, Gabriel Perathoner, Silvia Liotto, and Paolo Anvidalfare. 2025. [Bringing Ladin to FLORES+](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 1061–1071, Suzhou, China. Association for Computational Linguistics.
- Samuel Frontull and Thomas Ströhle. 2026. [Every word presented in context: Syntactic coverage as objective for low-resource machine translation with large language models](#). In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 8824–8837, Palma, Mallorca, Spain. European Language Resources Association (ELRA).
- Gemma Team et al. 2026. [Gemma 4 technical report](#). *Preprint*, arXiv:2607.02770.
- Christopher R. Haberland, Jean Maillard, and Stefano Lusito. 2024. [Italian-Ligurian machine translation in its cultural context](#). In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 168–176, Torino, Italia. ELRA and ICCL.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. [Lora: Low-rank adaptation of large language models](#). *Preprint*, arXiv:2106.09685.
- Raviraj Joshi, Kanishk Singla, Anusha Kamath, Raunak Kalani, Rakesh Paul, Utkarsh Vaidya, Sanjay Singh Chauhan, Niranjan Wartikar, and Eileen Long. 2025. [Adapting multilingual LLMs to low-resource languages using continued pre-training and synthetic corpus: A case study for Hindi LLMs](#). In *Proceedings of the First Workshop on Natural Language Processing for Indo-Aryan and Dravidian Languages*, pages 50–57, Abu Dhabi. Association for Computational Linguistics.
- Zixuan Ke, Yijia Shao, Haowei Lin, Tatsuya Konishi, Gyuhak Kim, and Bing Liu. 2023. [Continual pre-training of language models](#). *Preprint*, arXiv:2302.03241.
- Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Sergey Dukanov, Anton Dvorkovich, Mark Fishel, Markus Freitag, Samuel Frontull, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Stan Kharevich, Philipp Koehn, Zheng Li, Jean Maillard, Christof Monz, Alexander Murauski, and 12 others. 2026. Findings of the WMT26 general machine translation shared task: Contrastive dynamic human evaluation at scale. In *To appear in Proceedings of the Eleventh Conference on Machine Translation*. Association for Computational Linguistics.
- Roman Kyslyi, Yuliia Maksymiuk, and Ihor Pysmennyi. 2025. [Vuyko Mistral: Adapting LLMs for low-resource dialectal translation](#). In *Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025)*, pages 86–95, Vienna, Austria (online). Association for Computational Linguistics.
- Sara Moling, editor. 2016. *Dizionario Italiano-Ladino Val Badia / Dizionar Ladin Val Badia-Talian*. Istitut Ladin “Micurà de Rü”, San Martin de Tor (BZ).
- S. E. Robertson and S. Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’94*, page 232241, Berlin, Heidelberg. Springer-Verlag.
- Nathaniel R. Robinson, Shahd Abdelmoneim, Anjali Kantharuban, Otba Alsoul, Salima Lamsiyah, Kelly Marchisio, and Kenton Murray. 2026. [AMIYA shared task: Arabic modeling in your accent at VarDial 2026](#). In *Proceedings of the 13th Workshop on NLP for Similar Languages, Varieties and Dialects*,

pages 1–17, Rabat, Morocco. Association for Computational Linguistics.

Nipun Sadvilkar and Mark Neumann. 2020. [PySBD: Pragmatic sentence boundary disambiguation](#). In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, pages 110–114, Online. Association for Computational Linguistics.

Barbara Scalvini, Iben Nyholm Debess, Annika Simonsen, and Hafsteinn Einarsson. 2025. [Rethinking low-resource MT: the surprising effectiveness of fine-tuned multilingual models in the LLM age](#). In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 609–621, Tallinn, Estonia. University of Tartu Library.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2023. [A paradigm shift in machine translation: Boosting translation performance of large language models](#). *Preprint*, arXiv:2309.11674.

Bolin Zhang, Jiahao Wang, Qianlong Du, Jiajun Zhang, Zhiying Tu, and Dianhui Chu. 2025a. [A survey on data selection for llm instruction tuning](#). *J. Artif. Int. Res.*, 83.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Guoyin Wang, and Fei Wu. 2026. [Instruction tuning for large language models: A survey](#). *ACM Comput. Surv.*, 58(7).

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. 2025b. [Instruction tuning for large language models: A survey](#). *Preprint*, arXiv:2308.10792.