VojoLe-LM the first LLM Trained on an Italian Dialect Resource-Less Language

August 26, 2025

Tommaso Sgroi

1. Introduction

About nine languages are estimated lost per year ($\approx$ one every 40 days), a rate rising toward the oft-quoted "one every 14 days" by mid-century (Simons, 2019). Even when a spoken variety has any written form, those resources may be absent. The Sorianese dialect, spoken in Soriano nel Cimino (VT) a town of less than 8k inhabitants and shrinking population is effectively resource-less and endangered. A bilingual online wordlist with Italian glosses and usage exists (della Memoria Soriano, 2025), but no substantial corpus. In this report is introduced: a new Sorianese dataset using a targeted LLM, VojoLe-LMs[1] is a fine-tune of several LLMs on it to explore preserving the dialect by encoding language knowledge in model weights with language modeling (LM) task.

Awards granted by ISCRA @ LEONARDO: We acknowledge ISCRA for awarding this project access to the LEONARDO supercomputer, owned by the EuroHPC Joint Undertaking, hosted by CINECA (Italy).

2. Related Work

Haberland et al. (Haberland et al., 2024) release a new corpus, ZenaMT, containing 7,561 parallel Ligurian-Italian sentences. This corpus spans five domains: local and international news, Ligurian literature, Genoese Ligurian linguistics concepts, traditional card game rules, and Ligurian geographic expressions. They find that a translation model augmented with ZenaMT improves a baseline by 20%, and by over 25% (BLEU); demonstrating the utility of creating data sets for MT that are tailored for local cultural contexts by target language speakers.

Alhanai et al. (Alhanai et al., 2024) built benchmarks for eight low-resource African languages and evaluated 400+ fine-tuned models. Using LLMs as annotators, high-quality dataset fine-tuning, cross-lingual transfer, and cultural adjustments produced gains: ($\approx +5.6\%$) monolingual improvement ($\approx +5.4\%$) with high-quality data, $+2.9\%$ from cross-lingual transfer, and $+3.0\%$ on culturally appropriate items.

3. Method

3.1. Dataset Creation

To construct a dataset for a resource-less language, we leveraged a dialect dictionary and feed it into a large language model (LLM) using prompt engineering techniques. The final prompt size was approximately ≈ 40k tokens

We selected the "Command A" 111B model from Cohere (Cohere & a lot of others, 2025) due to its extensive context length (up to 256k tokens) and superior empirical performance in translation tasks, particularly for short- to medium-length text. Preliminary translation quality was assessed with the help of a native dialect speaker, and several LLMs were benchmarked before finalizing this choice. For the purposes of this work, we consider the generated dataset to be of near-gold quality. Additionally, "Command A" can be deployed using only two A100 GPUs, making it feasible for local inference.

For the source material, we combined two existing datasets: "fanpage" and "ilpost" (Landro et al., 2022), originally designed for text summarization. We extracted only the summarized segments, which consist of short- to medium-length, fluent, and grammatically correct Italian textwell-suited for our task.

The translation process was carried out on the Leonardo supercomputer using the full-precision "Command A" model spitted upon four 64GB GPUs using the vLLM python library (Kwon et al., 2023). Each Italian sentence was provided to the model along with the complete dialect dictionary, including usage examples. For each sample, the model generated three translations. This process lasted approximately five days and resulted in over 300k translated

Email: Tommaso Sgroi <sgroi.1852992@studenti.uniroma1.it>.

Deep Learning and Applied AI 2025, Sapienza University of Rome, 2nd semester a.y. 2024/2025.

[1] Vojole in "Sorianese" dialect means "chestnuts", which is a traditional local food.

sentences. To avoid model's hallucinations and also having a standard translation format, it was adopted the "guided decoding" technique to generate the LLM output in JSON format.

A web interface for human evaluation was deployed[2], but no native speakers were available at the time to verify sentence correctness, so we assume that the quality of the generated data is preserved (almost gold quality).

3.2. Baseline

Since the "Sorianese" language is unknown by any LLM, the baseline is just the "perplexity" of a base pretrained completion LLM.

4. Contribution

We make three main contributions.
Dataset We create and release the first large-scale Sorianese dataset ($\approx 300k$ sentence pairs) produced by prompting a high-context LLM with a dialect dictionary and usage examples. The dataset includes multiple variant translations per source sentence and metadata linking each example to dictionary entries and generation prompts.
Finetuned LLMs We finetune several open and proprietary base models on the Sorianese corpus and publish evaluation results (perplexity and qualitative examples). These models demonstrate that compact fine-tuning can encode dialectal knowledge and produce fluent Sorianese completions for downstream tasks.
Translation pipeline We design and validate a reproducible pipeline for bootstrapping corpora for resource-less dialects: dictionary-augmented prompting, multi-candidate generation, GPU-efficient inference (vLLM on Leonardo), and a lightweight human-in-the-loop evaluation interface. We show this pipeline enables rapid corpus creation with near-gold quality with the help of native speakers .

Each contribution is accompanied by code, prompts, and dataset.

5. Results

Experiments were run on the translated dataset from "ilpost" and "fanpage" (Landro et al., 2022), for a text completion task; the main task is to tech the LLMs a totally unknown dialect. We took into consideration the cross-entropy loss and the perplexity see Appendix A.

5.1. Experimental Setup

We fine-tuned two large language models, Mistral-7B-v0.1 (Jiang et al., 2023) and LLaMA-3-8B (Grattafiori & a lot of others, 2024), to increase model diversity and evaluate the impact of dataset quality on language modeling performance. Both models were loaded using the Unsloth framework (Daniel Han & team, 2023) with 4-bit quantization to reduce memory usage and enable efficient training on a single NVIDIA A100 GPU (64 GB VRAM). Fine-tuning was performed using a LoRA-based parameter-efficient training strategy with the configuration at Appendix B. Also, due to time issues, the dataset has been reduced by $n/3$ samples, which is the same amount of the originals datasets size; keeping only the first generation sample each, see section 3.1. Each model was tested with dropout 0 or 0.001; each fine-tune run took almost 10 hours[3]. Another notable finding concerns the token length distribution per sentence. Our analysis revealed that Sorianese consistently requires more tokens to be represented compared to standard Italian. This discrepancy arises because the tokenizer is not optimized for the dialect, leading to less efficient sub-word segmentation and a higher number of tokens per sentence. Table 3 summarizes the mean token count, its standard deviation, and the total number of tokens processed for each language–model pair. Results are reported in Appendix C.

5.2. Experimental Results

Table 1 reports the evaluation results in terms of cross-entropy loss and perplexity for both baseline (untuned) models and the corresponding fine-tuned VojoLe-LM variants. The baseline Mistral-7B and LLaMA-3.1 8B models exhibit high perplexity values (around 58–61), as expected for a dialect they were never exposed to. After fine-tuning on the Sorianese dataset, both models show a substantial reduction in perplexity: from 60.75 to 11.50 for Mistral, and from 58.27 to 12.42 for LLaMA. **This corresponds to an improvement factor of almost 6× in perplexity, indicating that the models successfully adapted to the dialect**.

Cross-entropy loss follows a similar trend, dropping from ≈ 4.1 to ≈ 2.45 for both models. The fine-tuned models thus encode a significantly more accurate representation of Sorianese while maintaining relatively stable performance on the Italian baseline, as demonstrated by the dataset split ($\approx 100k$ training and $\approx 4k$ test samples). Even with or without dropout performances were almost the same, highlighting the fact that more hyperparameter tune must be made.

The perplexity evolution during training is shown in Figure 1. Both models converge rapidly within the first few epochs, consistent with the early stopping setup (patience = 3). The final perplexity values suggest that even a relatively modest fine-tuning budget is sufficient to encode meaningful knowledge about a low-resource dialect.

[2] https://sor-eval.homeworkheroes.it/

[3] Other run and hyperparameter tuning couldn't be performed due remaining time issues with the Leonardo super-computer.

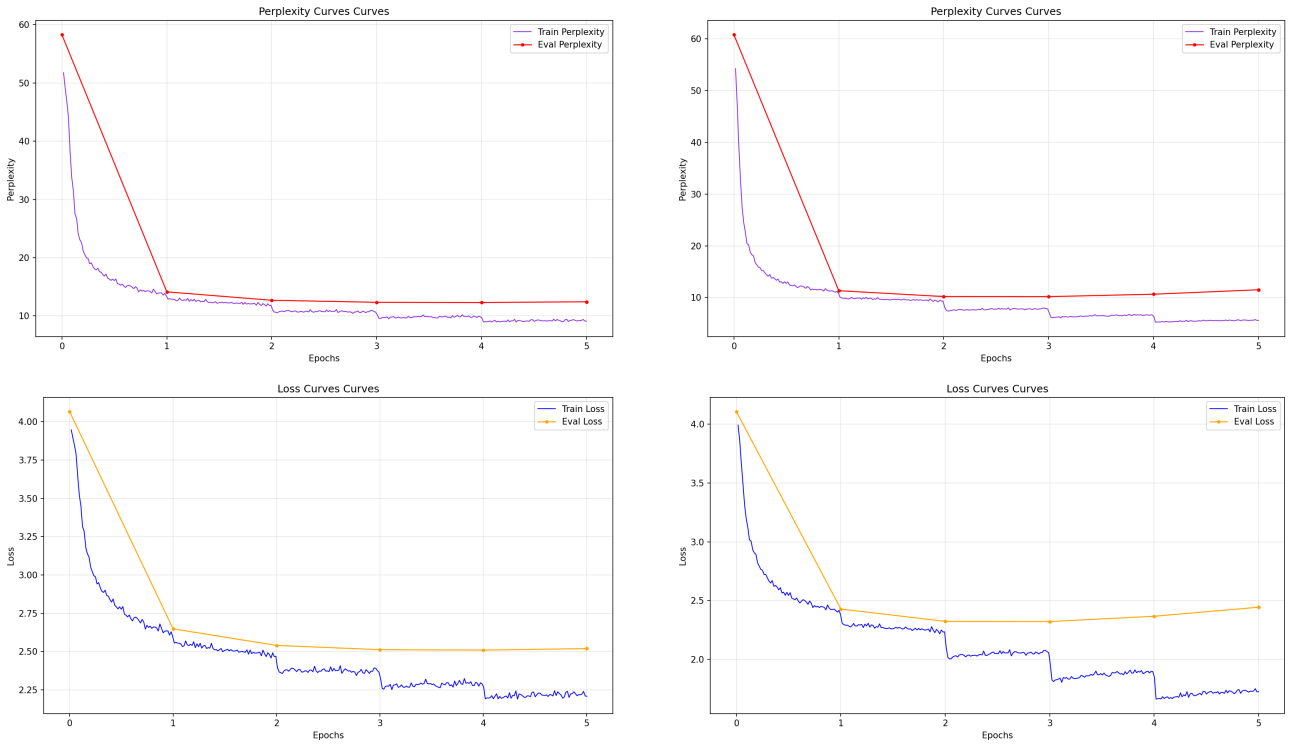

Figure 1. Left column: LLaMA-3.1 8B perplexity (top) and loss (bottom). Right column: Mistral-7B perplexity (top) and loss (bottom).

Model	Cross-Entropy Loss	Perplexity
Mistral Base	4.1068	60.7532
Mistral VojoLe-LM	**2.4427**	**11.5044**
LLaMA Base	4.0652	58.2744
LLaMA VojoLe-LM	**2.5197**	**12.4244**

Table 1. Evaluation results.

References

Alhanai, T., Kasumovic, A., Ghassemi, M., Zitzelberger, A., Lundin, J., and Chabot-Couture, G. Bridging the gap: Enhancing llm performance for low-resource african languages with new benchmarks, fine-tuning, and cultural adjustments, 2024. URL https://arxiv.org/abs/2412.12417.

Cohere, T. and a lot of others, . Command a: An enterprise-ready large language model, 2025. URL https://arxiv.org/abs/2504.00698.

Daniel Han, M. H. and team, U. Unsloth, 2023. URL http://github.com/unslothai/unsloth.

della Memoria Soriano, B. Dizionario sorianese, 2025. URL https://bancadellamemoriasoriano.weebly.com/dizionario-sorianese.html. Accessed: 2025-08-25.

Grattafiori, A. and a lot of others. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Haberland, C. R., Maillard, J., and Lusito, S. Italian-Ligurian machine translation in its cultural context. In Melero, M., Sakti, S., and Soria, C. (eds.), *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pp. 168–176, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.sigul-1.21/.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., Gonzalez, J. E., Zhang, H., and Stoica, I. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems*

Principles, 2023. URL https://github.com/vllm-project/vllm.

Landro, N., Gallo, I., La Grassa, R., and Federici, E. Two new datasets for italian-language abstractive text summarization. *Information*, 13(5), 2022. ISSN 2078-2489. doi: 10.3390/info13050228. URL https://www.mdpi.com/2078-2489/13/5/228.

Simons, G. Two centuries of spreading language loss. *Proceedings of the Linguistic Society of America*, 4:27, 03 2019. doi: 10.3765/plsa.v4i1.4532.

Appendix

A. Loss functions

Cross-entropy loss for a sequence of tokens $x_{1..T}$ with model probabilities $p(x_t|x_{<t})$ is

$$\mathcal{L}_{\text{CE}} = -\frac{1}{T}\sum_{t=1}^{T}\log p(x_t \mid x_{<t}).$$

Perplexity is defined from the average negative log-likelihood (cross-entropy) as

$$\text{PPL} = \exp\big(\mathcal{L}_{\text{CE}}\big) = \exp\left(-\frac{1}{T}\sum_{t=1}^{T}\log p(x_t \mid x_{<t})\right).$$

When reporting dataset-level metrics across N sequences, we average token counts:

$$\mathcal{L}_{\text{CE, dataset}} = -\frac{1}{\sum_{i=1}^{N} T_i}\sum_{i=1}^{N}\sum_{t=1}^{T_i}\log p(x_{i,t} \mid x_{i,<t}),$$

$$\text{PPL}_{\text{dataset}} = \exp\big(\mathcal{L}_{\text{CE, dataset}}\big).$$

B. Hyperparameters

LoRA rank	16
LoRA α	32
LoRA dropout	0 or 0.001
rsLoRA	True
Optimizer	AdamW (8-bit)
Learning rate	2×10^{-5} (cosine decay)
Embedding learning rate	1×10^{-6}
Batch size	32
Warmup ratio	0.1
Epochs	up to 10 (early stopping after 3)
Weight decay	0.0

Table 2. Training hyperparameters.

C. Token Length Distribution

Lang	Model	Mean	StdDev	Total
Sorianese	Mistral-7B	83.84	36.55	10,762,744
Italian	Mistral-7B	73.28	32.38	9,406,935
Sorianese	Llama-3.1	77.00	33.50	9,883,627
Italian	Llama-3.1	66.84	29.50	8,579,514

Table 3. Token length distribution per sentence across models and languages. Sorianese consistently requires more tokens due to suboptimal tokenizer segmentation.

D. Failed Experiments and Thoughts

The first an main idea, was about using ChatGPT to convert the full dataset, but it was very error prone and tendance to hallucinate; also the estimated amount on money for this task was over the $1k$ euros (even more with Command A API). A lot of datasets were found, but no one was fitting the interests; it was necessary a dataset with medium and/or short sentences. Tatoeba dataset was an option, but sentences were too short. To speedup inference were thought a pipeline which consisted calculating the similarity between a word embedding in the dialect dictionary Italian gloss and the one in the dataset sentence text; then build a custom prompt using just with the necessary dialect dictionary entries, saving a lot of tokens. This method, unfortunately, had poor performance. It was also tried with sentences but no progress was made. The last try was using the stem of the words and try a match, but in many cases there weren't a 1:1 association between dialect and Italian words; also it was degrading performances. Was observed that more context of usage, even out of context dictionary entries helps the model to perform better. As last chance, thanks to Professor Rodolà, i came across CINECA ISCRA C project, asking a try for the Leonardo Booster super computer. The whole project was carried by Leonardo, with all problems that an HPC, and distributed computing, can have (it took almost 3 weeks of out-of-office/workplace free time to be setup and ready). In the end, this project took a long time to be done, and will be necessary even more to make it research valuable.
Hope you enjoyed the paper :)