SiciLLaMa: a fine-tuned version of LLaMa for the Sicilian language

January 29, 2025

Matteo Mortella

Abstract

Large Language Models (LLMs) showed excellent performance in text understanding and generation tasks. However, this is not true for low-resource languages like Sicilian due to a lack of training data. This study introduces SiciLLaMa, a fine-tuned version of LLaMA 3 (8B), specifically adapted for Sicilian. Using LoRA (Low-Rank Adaptation), models were fine-tuned on a dataset sourced from Wikipedia and books. SiciLLaMa, the best-performing one, is capable of answering questions in Sicilian, but further research is needed to deploy it in real scenarios. Code and other details are available on Github.

1. Introduction

In recent years, Large Language Models (LLMs) have made remarkable progress in understanding and generating text in widely spoken languages. However, they struggle to achieve satisfactory results in low-resource languages (Huang et al., 2023), languages for which little data is available, and that risk to vanish in the near future (Wurm, 2001; Ostler, 1999). One of these languages is Sicilian, spoken in Sicily. It has been chosen as the target language for this study.

Relying on tailored LLMs for low-resource languages has demonstrated better results (Chang et al., 2023). Specifically, fine-tuning has shown a significant performance improvement for the text generation task (Khade et al., 2024), especially using parameter-efficient techniques like LoRa (Low-Rank Adaptation) (Hu et al., 2021).

Following the results described before, this study present SiciLLaMA, a fine-tuned version of LLama for the Sicilian language. Text data for fine-tuning were collected from Wikipedia and books; they were pre-processed, and several train and test sets were created. The selected base model was a 4-bit quantized version of Meta's Llama 3.1 8b instruction-tuned model(Unsloth & Meta, 2024); it is optimized for Unsloth (Han et al., 2023), a framework that grants fast and memory-efficient fine-tuning. The Fine-tuning phase has been conducted with different configurations and for each of them, a model version has been produced. Then, the models were evaluated with perplexity and compared with the baseline. In the end, both quantitative and qualitative results have been gathered.

2. Related work

Recently, LLMs have gained attention for their notable performance across various tasks and domains, such as question answering, text and code generation, and analysis. (Colombo et al., 2024; Thirunavukarasu et al., 2023; Jiang et al., 2024) However, most studies consider only high-resource languages such as English. Languages with limited data are overlooked, and research shows poor performance compared to the same experiments done with data from high-resource languages (Hasan et al., 2024). Low-resource languages constitute a valuable legacy for mankind that reflects centuries of cultural and historical development. Also, they serve as base for fields like anthropology, literature, and linguistics. LLMs can provide real aid to the goal of preserving these languages (Zhong et al., 2024).

Studies showed that LLMs fine-tuned on low-resource languages have fair results on tasks like machine translation (Lankford et al., 2023; Lu et al., 2025), especially using parameter-efficient fine-tuning methods like LoRa (Joshi et al., 2024). Regarding Italian languages, Sicilian, a language used in Sicily island, is considered endangered (Moseley, 2010).

While there are specific models for Italian like Camoscio, DanteLLM, and Fauno (Santilli & Rodolà, 2023; Bacciu et al., 2024; 2023), there are no tailored models for Italian dialect languages. The only available study presents an analysis that measures GPT4 (OpenAI, 2024) potential, for Sicilian exhibiting good overall understanding capabilities, but unsatisfactory capacities on text production (Lilli, 2023).

Email: Matteo Mortella <mortella.2058262@studenti.uniroma1.it>.

Deep Learning and Applied AI 2024, Sapienza University of Rome, 2nd semester a.y. 2023/2024.

3. Method

3.1. Original Dataset

Sicilian data was collected from two sources: Wikipedia pages and books.

Wikipedia content was selected because it is various in the topics and it is written in modern Sicilian. The Sicilian Wikipedia dump has been extracted using the WikiExtractor code (Attardi, 2015). Wikipedia data is composed of 41,518 pages with an average number of 47.29 words per page.

Books were selected among resources indicated by Cademia Siciliana. The books, available for research on Google Books, are from the XIX century. Their TXT files were manually inspected: samples were reviewed by Sicilian-speaking people to confirm an accetable form of Sicilian. 11 books were chosen, with an average length of 22,482 words and a full length ranging from 1,733 to 79,918 words.

3.2. Data preparation

Data was processed to remove unwanted content (URL, markup language, non-standard characters) and to keep only alphanumeric characters and punctuation. Then, a single dataset of text samples was produced. The maximum length for a sample to limit context, is 512 words; exceeding samples were split at full stops or at \n.

3.3. Final dataset: train and test

The final dataset was split into train (85% of samples) and test set. The script consents to generate three train-test splits: ”full” original data, ”partial” with a reduced percentage of uni-grams, bi-grams, and tri-grams, and ”reduced” without uni-grams and bi-grams and a reduced amount of tri-grams. The aim of these divisions is to observe how the presence of small N-grams impact on the performance given the limited amount of data.

3.4. Fine-tuning

LoRa technique has been used, and different models were fine-tuned, at least one for each selected R parameter corresponding to the rank of the trainable parameters matrices (32, 64, 128). The base-line model is llama-3-8b-Instruct-bnb-4bit because it had the best trade-off between computational power required and performance on Unsloth, and the instruction model was recommended as a better choice with less data in the documentation. Also, Unsloth allows fine-tuning with limited GPU, and that was crucial because the free 15 GB GPU subscription on Google Colab was used to fine-tune the models. Also, two configurations of LoRa target modules were picked to generate different model versions: the aim was to see the impact of including embeddings and output probability modules in training.

4. Results

The evaluation was conducted both with quantitative and qualitative results, full findings are available on Github. The chosen measure to compare the models' performance is perplexity: a lower perplexity indicates better understanding and less uncertainty in text generation.

Table 1 shows quantitative results:

Baseline: llama-3-8b-Instruct-bnb-4bit

noEmbed: used the default Unsloth target modules

Embed: extends default target modules with ”embed_tokens”, ”lm_head”

Model name	Perplexity	Test set
r32_reduced_noEmbed	**21.95**	**Reduced**
r128_reduced_noEmbed	22.00	Reduced
r64_reduced_noEmbed	22.03	Reduced
Baseline	33.73	Reduced
Baseline	34.95	Partial
Baseline	37.21	Full

Table 1. Perplexity scores for top-3 and bottom-3 models, ordered by perplexity

The table shows that the default target modules are better to chose, the R parameter makes no difference (representational power is enough for R=32 given the limited number of samples), and excluding uni-grams and bi-grams is the best approach. Qualitative question-answer pairs were collected both for the best and the baseline models. SiciLLaMa, the best model, showed a correct understanding of the language (the answers were related to the question), and a good structure of the phrase. However, even though the text were generated always in Sicilian, some of the generated words were imperfect and the information unreliable, probably due to a limited learning on Sicilian vocabulary and word forms. The baseline model replied in English and Italian, always with correct content.

5. Discussion and conclusions

In this study, SiciLLaMA, a fine-tuned llama model for the Sicilian language was introduced. Results showed that there was an improvement compared to the baseline model in terms of perplexity, and that a decent text generation capability has been reached (completely absent in the baseline), the produced answers are understandable by Sicilian-speaking people, but the reliability of the produced information is not sufficient. Further research on different baseline models, parameter exploration and data augmentation are required to obtain better LLMs for the Sicilian language. The code and other resources are available on Github.

References

Attardi, G. Wikiextractor. https://github.com/attardi/wikiextractor, 2015.

Bacciu, A., Trappolini, G., Santilli, A., Rodolà, E., and Silvestri, F. Fauno: The italian large language model that will leave you senza parole!, 2023. URL https://arxiv.org/abs/2306.14457.

Bacciu, A., Campagnano, C., Trappolini, G., and Silvestri, F. Dantellm: Let‘s push italian llm research forward! In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 4343–4355, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.388/.

Chang, T. A., Arnett, C., Tu, Z., and Bergen, B. K. When is multilinguality a curse? language modeling for 250 high- and low-resource languages, 2023. URL https://arxiv.org/abs/2311.09205.

Colombo, P., Pires, T. P., Boudiaf, M., Culver, D., Melo, R., Corro, C., Martins, A. F. T., Esposito, F., Raposo, V. L., Morgado, S., and Desa, M. Saullm-7b: A pioneering large language model for law, 2024. URL https://arxiv.org/abs/2403.03883.

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

Hasan, M. A., Tarannum, P., Dey, K., Razzak, I., and Naseem, U. Do large language models speak all languages equally? a comparative study in low-resource settings, 2024. URL https://arxiv.org/abs/2408.02237.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

Huang, H., Tang, T., Zhang, D., Zhao, W. X., Song, T., Xia, Y., and Wei, F. Not all languages are created equal in llms: Improving multilingual capability by cross-lingual-thought prompting, 2023. URL https://arxiv.org/abs/2305.07004.

Jiang, J., Wang, F., Shen, J., Kim, S., and Kim, S. A survey on large language models for code generation, 2024. URL https://arxiv.org/abs/2406.00515.

Joshi, S., Khan, M. S., Dafe, A., Singh, K., Zope, V., and Jhamtani, T. Fine Tuning LLMs for Low Resource Languages . In *2024 5th International Conference on Image Processing and Capsule Networks (ICIPCN)*, pp. 511–519. IEEE Computer Society, 2024. doi: 10.1109/ICIPCN63822.2024.00090. URL https://doi.ieeecomputersociety.org/10.1109/ICIPCN63822.2024.00090.

Khade, O., Jagdale, S., Phaltankar, A., Takalikar, G., and Joshi, R. Challenges in adapting multilingual llms to low-resource languages using lora peft tuning, 2024. URL https://arxiv.org/abs/2411.18571.

Lankford, S., Afli, H., and Way, A. adaptmllm: Fine-tuning multilingual language models on low-resource languages with integrated llm playgrounds. *Information*, 14(12):638, November 2023. ISSN 2078-2489. doi: 10.3390/info14120638. URL http://dx.doi.org/10.3390/info14120638.

Lilli, S. Chatgpt-4 and italian dialects: Assessing linguistic competence. *Umanistica Digitale*, 7(16): 235–263, Jan. 2023. doi: 10.6092/issn.2532-8816/18221. URL https://umanisticadigitale.unibo.it/article/view/18221.

Lu, K., Yang, Y., Yang, F., Dong, R., Ma, B., Aihemaiti, A., Atawulla, A., Wang, L., and Zhou, X. Low-resource language expansion and translation capacity enhancement for llm: A study on the uyghur. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 8360–8373. Association for Computational Linguistics, January 2025. URL https://aclanthology.org/2025.coling-main.559/.

Moseley, C. *Atlas of the World's Languages in Danger*. Unesco, 2010.

OpenAI. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Ostler, R. Disappearing languages. *The Futurist*, 33(7):16, 1999.

Santilli, A. and Rodolà, E. Camoscio: an italian instruction-tuned llama, 2023. URL https://arxiv.org/abs/2307.16456.

Thirunavukarasu, A., Ting, D., and Elangovan, K. Large language models in medicine, 2023. URL https://doi.org/10.1038/s41591-023-02448-8.

Unsloth and Meta. Meta-llama-3.1-8b-instruct-bnb-4bit, 2024. URL https://huggingface.co/unsloth/Meta-Llama-3.1-8B-Instruct-bnb-4bit. Accessed: 2025-01-19.

Wurm, S. A. *Atlas of the World's Languages in Danger of Disappearing*. Unesco, 2001.

Zhong, T., Yang, Z., Liu, Z., Zhang, R., Liu, Y., Sun, H., Pan, Y., Li, Y., Zhou, Y., Jiang, H., Chen, J., and Liu, T. Opportunities and challenges of large language models for low-resource languages in humanities research, 2024. URL https://arxiv.org/abs/2412.04497.