LLaMAntino: Large Language Models per la lingua italiana

Pierpaolo Basile[1,*,†], Elio Musacchio[1], Marco Polignano[1], Lucia Siciliani[1], Giuseppe Fiameni[2] and Giovanni Semeraro[1]

[1]Università degli Studi di Bari Aldo Moro, Bari, Italy
[2]NVIDIA AI Technology Center, Milan, Italy

Abstract

I Large Language Models hanno ricevuto molta attenzione in ambito accademico grazie alle loro capacità di generalizzazione. Questi modelli sono addestrati su grandi corpus di dati scritti prevalentemente in lingua inglese e per questo sono in grado di ottenere risultati eccellenti nel rispondere ad input forniti in tale lingua. Nonostante siano stati proposti modelli addestrati con l'obiettivo di supportare più lingue, una procedura replicabile di adattamento e fine-tuning di un modello per uno specifico linguaggio non è ancora stata ben definita allo stato dell'arte. A tal fine, nasce la famiglia di modelli LLaMAntino, una collezione di modelli addestrata specificatamente per la lingua italiana. La pipeline di addestramento utilizzata per ottenere questi modelli è ben definita e descritta in dettaglio per garantirne la riproducibilità da parte della comunità scientifica.

Keywords

Large Language Models, Natural Language Processing, Language Adaptation, Instruction-tuning

1. Introduzione

Nel 2023, META ha rilasciato pubblicamente e con una licenza permissiva la seconda versione di una famiglia di Large Language Models (LLM) da loro svillupata: LLaMA 2 [1]. La disponibilità di una architettura aperta e di pesi pre-addestrati, ottenuti grazie all'utilizzo di cluster con potenti risorse di calcolo (non accessibili a chiunque), ha reso possibile ai ricercatori nell'ambito del Natural Language Processing (NLP) l'utilizzo di questi modelli. I Large Language Models sono, infatti, in grado di risolvere molti task grazie alle loro capacità di generalizzazione. La possibilità di lavorare su modelli la cui architettura è ben definita e messa liberamente a disposizione della comunità scientifica è fondamentale nello studio per il raggiungimento della Artificial General Intelligence (AGI).

Ciò nonostante, i modelli LLaMA 2 sono stati addestrati su dataset la cui lingua principale è quella inglese (89.70% del dataset di train), e senza particolari attenzioni per un addestramento multilingua, come nel caso di altri modelli ad esempio BLOOM [2].

Per i motivi appena illustrati, è evidente la necessità di sviluppare una soluzione specifica per altre lingue, in modo da poter adeguatamente supportare i ricercatori di NLP fornendo dei modelli specializzati per dei linguaggi che tendono ad avere meno risorse disponibili rispetto all'inglese. A tal fine, nasce LLaMAntino [3], una famiglia di LLMs che, a partire dai pesi pre-addestrati di LLaMA 2, sono stati ulteriormente rifiniti nei task di comprensione e generazione di testo in lingua italiana. Inoltre, i modelli sono stati sottoposti ad un´ulteriore fase di addestramento per renderli in grado di approcciarsi a disparati task di NLP, quali il dialogo o estrazione di informazioni da testo. Per questa fase, nella mentalità di utilizzare solo risorse aperte e disponibili alla comunità scientifica, abbiamo anche utilizzato dati rilasciati durante EVALITA 2023 [4], una campagna di valutazione di strumenti di NLP per la lingua italiana che promuove lo sviluppo di un benchmark di valutazione comune. Infine, seguendo quanto è stato fatto per i modelli LLaMA 2, anche i modelli della famiglia LLaMAntino sono stati rilasciati pubblicamente[1] sotto la licenza dei modelli originali.

Per riassumere, i contributi di questo lavoro sono i seguenti:

- Definizione di una pipeline di addestramento per ottenere Large Language Models specializzati in una lingua, che può essere facilmente replicata da altri ricercatori per altri linguaggi;
- Realizzazione e rilascio di una famiglia di Large Language Models addestrati per la lingua italiana utilizzando la pipeline proposta.

Ital-IA 2024: 4th National Conference on Artificial Intelligence, organized by CINI, May 29-30, 2024, Naples, Italy
*Corresponding author.
†These authors contributed equally.
pierpaolo.basile@uniba.it (P. Basile); elio.musacchio@uniba.it (E. Musacchio); marco.polignano@uniba.it (M. Polignano); lucia.siciliani@uniba.it (L. Siciliani); gfiameni@nvidia.com (G. Fiameni); giovanni.semeraro@uniba.it (G. Semeraro)
0000-0002-0545-1105 (P. Basile); 0009-0006-9670-9998 (E. Musacchio); 0000-0002-3939-0136 (M. Polignano); 0000-0002-1438-280X (L. Siciliani); 0000-0001-8687-6609 (G. Fiameni); 0000-0001-6883-1853 (G. Semeraro)

[1]https://huggingface.co/collections/swap-uniba/llamantino-models-65aa9f3357f263e3d0402346

2. Altri lavori

Per addestrare un Large Language Model in una lingua specifica, l'opzione migliore consiste nell'effettuare un addestramento "da zero´´ specifico per quella lingua. Questa operazione, tuttavia, notoriamente richiede ingenti risorse computazionali e dati di alta qualità che spesso non sono disponibili per lingue diverse dall'inglese in quanto sotto rappresentate. Ad esempio, per l'addestramento del modello a 7 miliardi di parametri LLaMA 2, sono state usate un totale di 184, 320 ore GPU [1]. Per ovviare a questo problema, sono state ideate tecniche di *Language Adaptation*, che consistono nel adattamento di un Language Model pre-addestrato su una lingua ad un'altra lingua. Questa tecnica è stata ampiamente studiata e diverse metodologie sono state descritte nello stato dell'arte, in particolare:

- **Continuare il Pre-Training**: viene continuato l'addestramento originale del modello pre-addestrato in modo da modificare i parametri usando la stessa metodologia [5]
- **Adapter**: vengono aggiunti dei layer al modello specifici per la lingua obiettivo [6]
- **Parameter Efficient Fine-Tuning** (PEFT): viene effettuato un addestramento che va a modificare solo un sottoinsieme di tutti i parametri del modello, ottenendo quindi una procedura più efficiente [7]

Oltre alla Language Adaptation, è possibile effettuare ulteriore addestramento per specifici task, questo permette ai modelli di imparare a risolvere meglio determinati compiti. A tal fine, i Large Language Models possono essere addestrati seguendo una procedura chiamata *instruction tuning* [8]. L'idea è che addestrare un modello su input formattato per seguire istruzioni permette di migliorare le sue capacità di generalizzazione. In questo modo, il modello è in grado di completare istruzioni che non ha mai visto anche senza ulteriore addestramento.

Per la lingua italiana, sono presenti in letteratura altri modelli, quali:

- **Camoscio** [9] e **Stambecco** [10], dei modelli LLaMA instruction-tuned
- **Fauno** [11], un modello Baize conversazionale [12]
- **Cerbero** [13], un modello Mistral [14]

Tutti questi modelli rilasciano pochi pesi addestrati e non superano i 13 miliardi di parametri. Inoltre, per una lingua sotto rappresentata come l'italiano, sussiste un problema nella mancanza di dataset curati e ciò comporta l'utilizzo di dataset tradotti automaticamente in italiano o ottenuti sinteticamente da altri LLMs.

3. Metodologia

La pipeline di addestramento dei modelli LLaMAntino è divisa in due step: *language adaptation*, per adattare il modello prevalentemente inglese alla lingua italiana, e *fine-tuning*, per migliorare ulteriormente le capacità del modello su specifici task. Una visualizzazione della metodologia applicata è fornita in Figura 1. In tutti i casi in cui non è stato utilizzato un dataset originariamente in lingua italiana, a causa della mancanza di risorse per il task richiesto, abbiamo utilizzato un tool di traduzione chiamato *ArgosTranslate* per effettuare una traduzione automatica[2]. Infine, in tutti i casi abbiamo utilizzato il supercomputer Leonardo che mette a disposizione nodi di calcolo con fino a 4 NVIDIA A100 64GB. Per maggiori informazioni riguardanti prompt utilizzati è possibile fare riferimento rispettivamente all´Appendice. Per quanto riguarda i parametri di addestramento, invece, sono illustrati in Tabella 1.

Step	Strategy	Epochs	Learning Rate	Batch Size
Language Adaptation	QLoRA	25,000 steps	0.0002	96
Chat Fine-Tuning	QLoRA	15,000 steps	0.0002	96
Instruction-Tuning 7b	FSDP	3	0.00002	128
Instruction-Tuning 13b	FSDP	5	0.00001	128

Table 1
Parametri usati per le fasi di addestramento

3.1. Language Adaptation

Come strategia di Language Adaptation, abbiamo continuato la fase di addestramento del modello, utilizzando come dataset *clean_mc4_it* [15], un subset di *mc4* [16] che è a sua volta un subset multilingua di *CommonCrawl*[3]. Il dataset utilizzato considera solo la parte di documenti in italiano e vengono effettuate delle operazioni di pulizia per garantirne la qualità. In particolare, vengono effettuate le tre operazioni descritte di seguito:

- Vengono rimossi documenti che contengono parole in una lista di termini italiani e inglesi chiamata "List of Dirty, Naughty, Obscene, and Otherwise Bad Words"[4];
- Vengono rimosse frasi che contengono: meno di 3 parole, una parola con più di 1000 caratteri, un simbolo di fine frase non riconosciuto o stringhe particolari che non rispecchiano il linguaggio naturale (e.g. linguaggio di programmazione, privacy policy);

[2] https://github.com/argosopentech/argos-translate
[3] https://commoncrawl.org/
[4] https://github.com/LDNOOBW/List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words

Figure 1: Pipeline di addestramento applicata per LLaMAntino

- Vengono rimossi documenti che contengono: meno di 5 frasi, meno di 500 o più di 50,000 caratteri, non sono identificati come prevalentemente in lingua italiana da LangDetect[5].

3.2. Fine-Tuning

Come strategia di fine-tuning del modello, abbiamo utilizzato due diverse procedure:

- *LLaMAntino Chat Fine-Tuning*: abbiamo ulteriormente addestrato i modelli chat per migliorare le loro capacità nel dialogo con utenti;
- *LLaMAntino Instruction-Tuning*: abbiamo ulteriormente addestrato i modelli base per migliorare le loro capacità nel risolvere task di diverso tipo.

Nel caso dei modelli chat, dopo aver adattato il modello alla lingua italiana, puntiamo anche a migliorare le loro capacità di dialogo, dal momento che gli LLM vengono spesso utilizzati come chatbot. Per fare questo, abbiamo tradotto il dataset UltraChat [17] in italiano ed ottenuto un totale di 512,837 dialoghi. Questo dataset è composto da conversazioni generate utilizzando due GPT-3 APIs che simulano rispettivamente utente e sistema. Questi dialoghi, in particolare, coprono le seguenti tematiche: *Domande riguardanti il mondo*, *Scrittura e Creatività* e *Supporto su materiali esistenti* (e.g. continuare a scrivere o riscrivere del testo).

Nel caso dei modelli base, abbiamo effettuato due procedure di instruction-tuning separate su due dataset diversi. La prima, è stata effettuata sul Dolly dataset [18] tradotto in italiano, questo consiste di 15,000 istruzioni manualmente curate per 8 diverse categorie di task, ovvero Open Question Answering, General Question Answering, Classification, Closed Question Answering, Brainstorming, Information Extraction, Summarization e Creative Writing. La seconda, è stata effettuata sui dataset di train di alcuni task di EVALITA 2023[6] [4]. EVALITA è una campagna di valutazione per modelli di Natural Language Processing specifica per la lingua italiana che raccoglie ogni anno diversi tipi di task che pertanto risultano una risorsa fondamentale per la ricerca in tale lingua. In particolare, abbiamo usato i train set dei seguenti task: ACTI, CLinkaRT, DisCoTex, EMit, GeoLing, HaSpeeDe3, HODI, LangLearn, NERMuD, PoliticIT, WiC-ITA. Per la descrizione di ciascun task si rimanda a [4].

[5] https://github.com/Mimino666/langdetect
[6] https://www.evalita.it/campaigns/evalita-2023/tasks/

4. Conclusioni

LLaMAntino è la prima famiglia di Large Language Models per l'italiano che mette a disposizione un modello a 70 miliardi di parametri e molti modelli addestrati per diversi fini. A supporto dell'open-science, i modelli vengono messi a disposizione della comunità scientifica, nella speranza che possano supportare i ricercatori italiani. Una delle limitazioni maggiori riscontrate consiste nella scarsità di dataset manualmente curati da esperti per l'italiano, il che ha reso necessario l'utilizzo di strumenti di traduzione automatica per disporre di una quantità di dati congrua per addestrare dei LLMs. In futuro, prevediamo di continuare a lavorare sui Large Language Model per la lingua italiana e di estendere il lavoro svolto in modo da prendere in considerazione tipologie di dati differenti al fine di realizzare un modello multimodale.

Acknowledgments

Riconosciamo il supporto del PNRR per il progetto FAIR - Future AI Research (PE00000013), Spoke 6 - Symbiotic AI (CUP H97G22000210007) per il programma NRRP MUR fondato da NextGenerationEU. I modelli sono stati addestrati sul supercomputer Leonardo con il supporto di CINECA-Italian Super Computing Resource Allocation, progetti di Classe C: IscrC_FineIT (HP10CCD87T), IscrC_fineNLP (HP10CT56JA) and IscrC_Pro_MRS (HP10CQO70G).

References

[1] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al., Llama 2: Open foundation and fine-tuned chat models, arXiv preprint arXiv:2307.09288 (2023).

[2] T. Le Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon,

M. Gallé, et al., Bloom: A 176b-parameter open-access multilingual language model (2022).
[3] P. Basile, E. Musacchio, M. Polignano, L. Siciliani, G. Fiameni, G. Semeraro, Llamantino: Llama 2 models for effective text generation in italian language, 2023. arXiv:2312.09993.
[4] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, Evalita 2023: Overview of the 8th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR. org, Parma, Italy, 2023.
[5] E. C. Chau, L. H. Lin, N. A. Smith, Parsing with multilingual bert, a small corpus, and a small treebank, arXiv preprint arXiv:2009.14124 (2020).
[6] R. Wang, D. Tang, N. Duan, Z. Wei, X. Huang, G. Cao, D. Jiang, M. Zhou, et al., K-adapter: Infusing knowledge into pre-trained models with adapters, arXiv preprint arXiv:2002.01808 (2020).
[7] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Lora: Low-rank adaptation of large language models, arXiv preprint arXiv:2106.09685 (2021).
[8] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, et al., Scaling instruction-finetuned language models, arXiv preprint arXiv:2210.11416 (2022).
[9] A. Santilli, E. Rodolà, Camoscio: an italian instruction-tuned llama, 2023. arXiv:2307.16456.
[10] Michael, Stambecco: Italian instruction-following llama model, https://github.com/mchl-labs/stambecco, 2023.
[11] A. Bacciu, G. Trappolini, A. Santilli, E. Rodolà, F. Silvestri, Fauno: The italian large language model that will leave you senza parole!, arXiv preprint arXiv:2306.14457 (2023).
[12] C. Xu, D. Guo, N. Duan, J. McAuley, Baize: An open-source chat model with parameter-efficient tuning on self-chat data, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 6268–6278. URL: https://aclanthology.org/2023.emnlp-main.385. doi:10.18653/v1/2023.emnlp-main.385.
[13] F. A. Galatolo, M. G. Cimino, Cerbero-7b: A leap forward in language-specific llms through enhanced chat corpus generation and evaluation, arXiv preprint arXiv:2311.15698 (2023).
[14] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, et al., Mistral 7b, arXiv preprint arXiv:2310.06825 (2023).
[15] G. Sarti, M. Nissim, It5: Large-scale text-to-text pretraining for italian language understanding and generation, ArXiv preprint 2203.03759 (2022). URL: https://arxiv.org/abs/2203.03759.
[16] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, C. Raffel, mT5: A massively multilingual pre-trained text-to-text transformer, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 483–498. URL: https://aclanthology.org/2021.naacl-main.41. doi:10.18653/v1/2021.naacl-main.41.
[17] N. Ding, Y. Chen, B. Xu, Y. Qin, Z. Zheng, S. Hu, Z. Liu, M. Sun, B. Zhou, Enhancing chat language models by scaling high-quality instructional conversations, arXiv preprint arXiv:2305.14233 (2023).
[18] M. Conover, M. Hayes, A. Mathur, J. Xie, J. Wan, S. Shah, A. Ghodsi, P. Wendell, M. Zaharia, R. Xin, Free dolly: Introducing the world's first truly open instruction-tuned llm, 2023. URL: https://tinyurl.com/58fkx3hu.

A. Prompt Format

Listing 1: Chat template

```
<s>[INST] <<SYS>>
Sei un assistente disponibile, rispettoso
     e onesto. Rispondi sempre nel modo
    piú utile possibile, pur essendo
    sicuro. Le risposte non devono
    includere contenuti dannosi, non
    etici, razzisti, sessisti, tossici,
    pericolosi o illegali. Assicurati che
     le tue risposte siano socialmente
    imparziali e positive. Se una domanda
     non ha senso o non è coerente con i
    fatti, spiegane il motivo invece di
    rispondere in modo non corretto. Se
    non conosci la risposta a una domanda
    , non condividere informazioni false.
</SYS>

{{ user_msg_1 }} [/INST] {{
    model_answer_1 }}</s>
<s>[INST] {{ user_msg_2 }} [/INST] {{
    model_answer_2 }}</s>

...

<s>[INST] {{ user_msg_N }} [/INST] {{
    model_answer_N }}</s>
```

Listing 2: Instruction-tuning template

```
Di seguito è riportata un'istruzione che
    descrive un'attivitá, abbinata ad un
    input che fornisce ulteriore
    informazione. Scrivi una risposta che
     soddisfi adeguatamente la richiesta.


### Istruzione:
{instruction}

### Input:
{input}

### Risposta:
{response}
```