Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

LLaMAntino: a family of open source Large Language Models for Italian based on LLaMA

Pierpaolo Basile, *Elio Musacchio, Marco Polignano, Lucia Siciliani, Giuseppe Fiameni, Giovanni Semeraro*

Spoke 6 - Work Package 6.2 / **TP 2**
Università degli Studi di Bari Aldo Moro

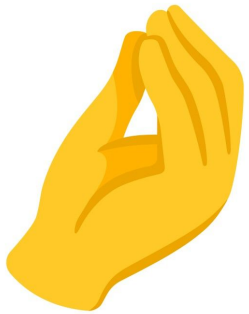
Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

What language does an LLM speak?

bigscience/**bloom** like 991

Text Generation | PyTorch | TensorBoard | Transformers | Akan | Arabic | Assamese | Bambara | Bengali | Catalan | code | English | Spanish | Basque | Fon | French | Gujarati | Hindi | Indonesian | Igbo | Kikuyu | Kannada | Ganda | Lingala | Malayalam | Marathi | Nepali | Pedi | Chichewa | Oriya | Panjabi | Portuguese | Kirundi | Kinyarwanda | Shona | Southern Sotho | Swahili | Tamil | Telugu | Tswana | Tsonga | Tumbuka | Twi | Urdu | Vietnamese | Wolof | Xhosa | Yoruba | Chinese | Zulu | arxiv:1909.08053 | arxiv:2110.02861 | arxiv:2108.12409 | bloom | feature-extraction | Eval Results | License: bigscience-bloom-rail-1.0

No Italian Language!



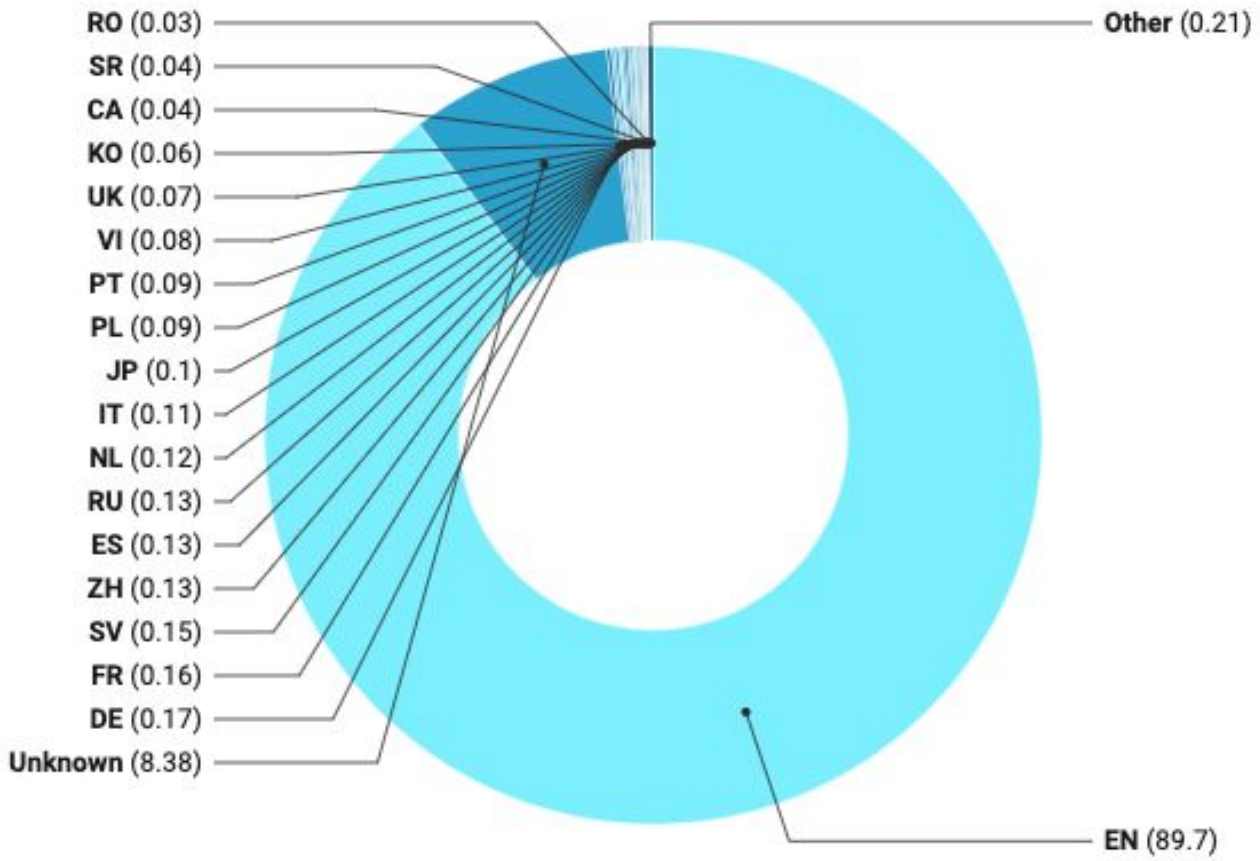
a BigScience initiative

BLOOM

176B params · 59 languages · Open-access

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

What languages does LLaMa-2 speak?



90% English pre-training data

Other languages (*German, French, Chinese, Spanish, Dutch, Italian, Japanese, Polish, Portuguese, …*)
less than 2% training data

8% training data “unknown”
(*includes programming code data*)

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

LLaMAntino

We applied the following training pipeline to the LLaMA2 models:

Language Adaptation: model training on generic data in Italian

Fine Tuning: model training on instruction data in Italian

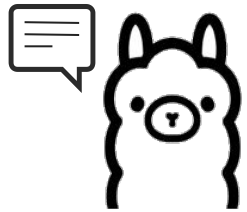
Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

LLaMAntino

LLaMA 2

7B/13B/70B

LLaMA 2 Chat



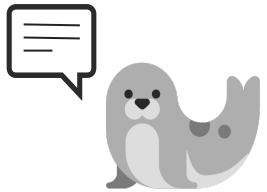
7B/13B/70B

Language Adaptation

LLaMAntino 2

7B/13B

LLaMAntino 2 Chat

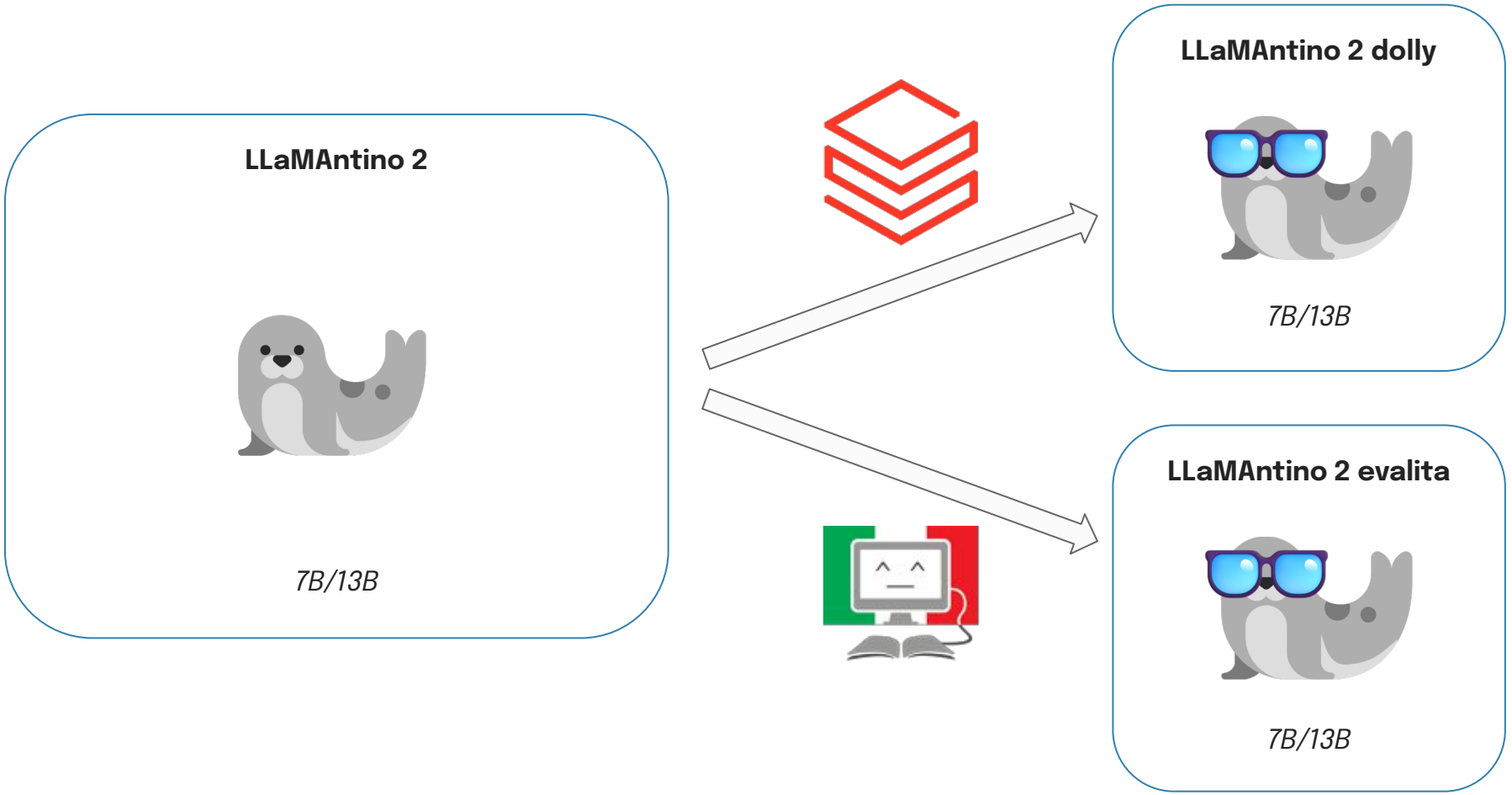


7B/13B

SWAP researchgroup

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

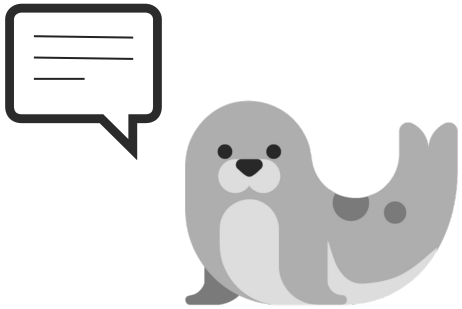
LLaMAntino: fine-tuned models



Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

LLaMAntino: chat model

LLaMAntino 2 Chat



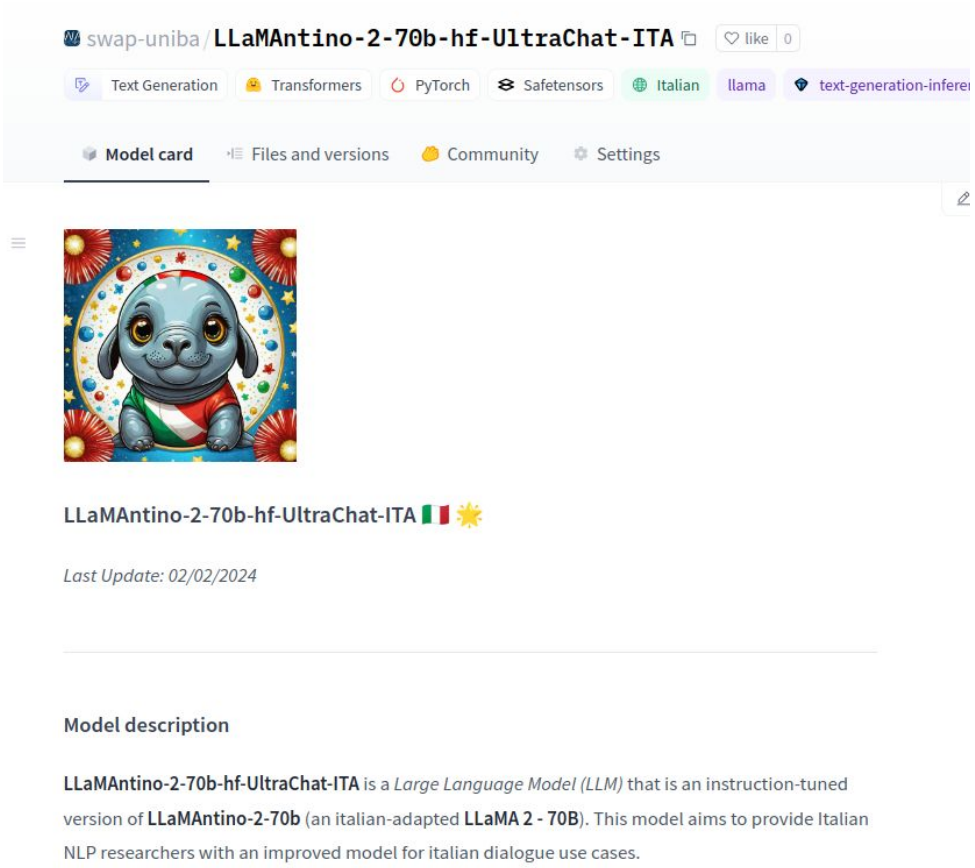
7B/13B

UltraChat

LLaMAntino 2 UltraChat

7B/13B

LLaMAntino: 70B model



Language Adaptation

+

UltraChat fine-tuning

https://huggingface.co/swap-uniba/LLaMAntino-2-70b-hf-UltraChat-ITA

LLaMAntino: computational resources

All models were trained on the **Leonardo HPC**

Language Adaptation	Fine-tuning
4-bit quantization, QLoRA, SFTTrainer	Fully-Sharded Data Parallel (FSDP)
3 nodes for a total of **12 GPUs A100 64GB**	**2 nodes** for a total of **8 GPUs A100 64GB**
LoRA parameters: attention dimension (64), scaling parameter (16), dropout (0.1). Single GPU batch size (8). Steps (25K) Text length of (1024)	Single GPU batch size (16). Epochs (3 for 7B, 5 for 13B). Text length (1024)
~100.000 Leonardo hours	**~50.000 Leonardo hours**

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

LLaMAntino ANITA (Llama-3)

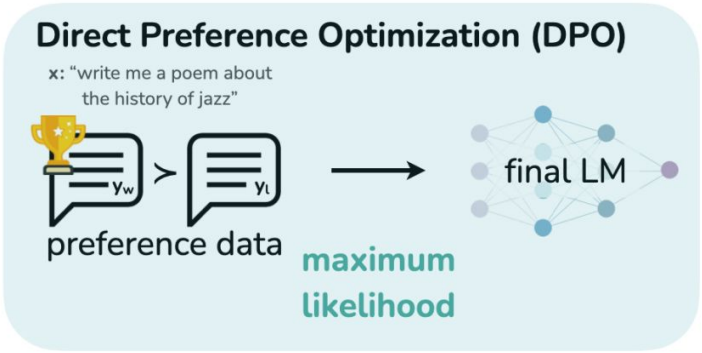
Supervised fine-tuned + DPO using several English datasets

Stanford Alpaca

Databricks' Dolly 2.0

WizardLM

ORPO-DPO-mix-40k v1.2



Language Adaptation

https://huggingface.co/swap-uniba/LLaMAntino-3-ANITA-8B-Inst-DPO-ITA
https://arxiv.org/pdf/2405.07101

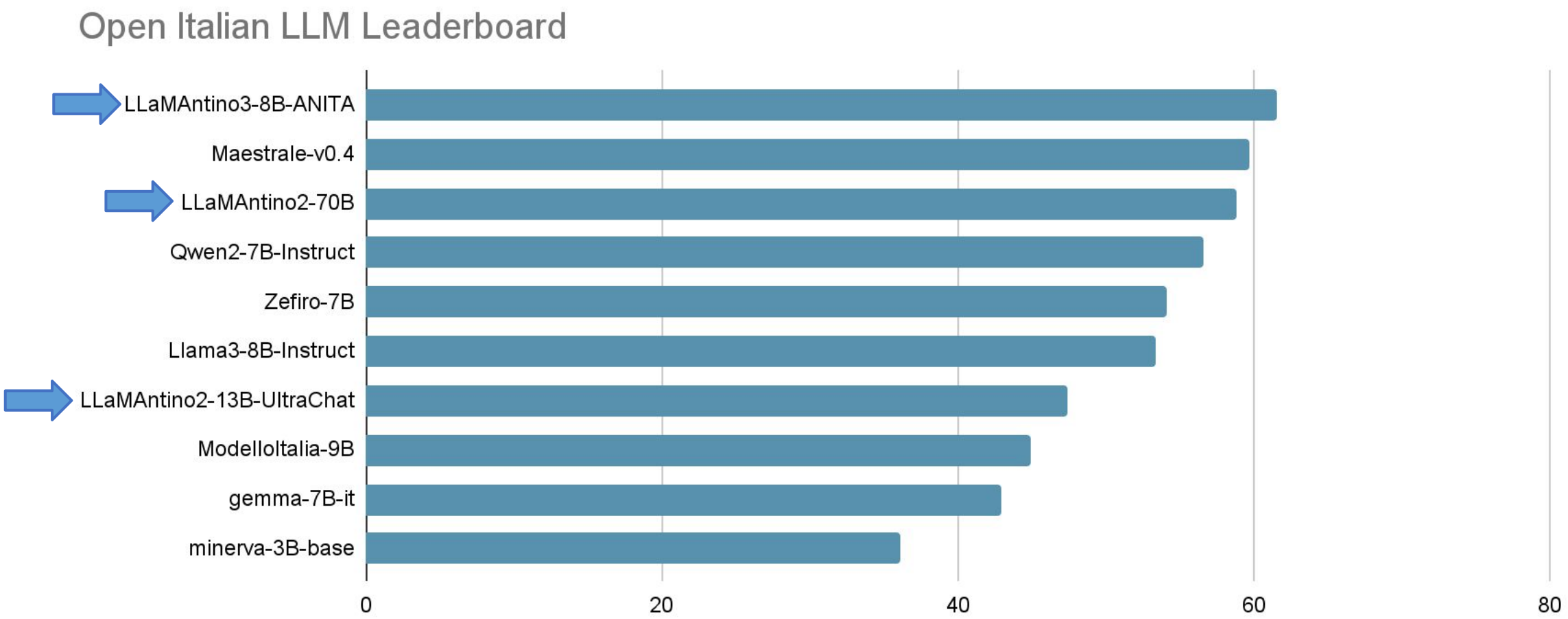
Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

LLaMAntino: Llama-3-ANITA 8B

Il progetto 💥ANITA💥 *(Advanced Natural-based interaction for the ITAlian language)* vuole fornire ai ricercatori italiani un modello LLM migliorato per i casi d'uso della lingua italiana 🇮🇹.

SCAN ME

https://chat.llamantino.it/



https://huggingface.co/spaces/FinancialSupport/open_ita_llm_leaderboard

- A family of LLMs for the Italian language
- A set of basic models that can be fine-tuned according to specific downstream tasks in Italian
- A powerful model for building conversational agents
- A replicable and scalable pipeline that can be applied to several LLMs

Ongoing works…

- Development of application in specific domains: PA, Health, Security, …
- Multimodal extension using LLaVA
- Extend the pipeline to other LLMs (*Mistral, Mixtral, Llama 3.1, Phi, …*)
- *CALAMITA…*

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

CALAMITA - Challenge the Abilities of LAnguage Models in ITAlian

- A collaborative effort to develop a dynamic and growing benchmark for evaluating LLMs' capabilities in Italian
  - **short-term:** building the benchmark through a series of **challenges collaboratively construed** by the research community
  - **long-term:** a suite of tasks in the form of a benchmark which can be accessed through a **shared platform and a live leaderboard**
- The first call expired on May 17th
  - **22 benchmark proposal!!!**
  - the 1st version of the benchmark will be presented in December at CLiC-it 2024
  - the benchmark will be included in the Language Model Evaluation Harness

https://clic2024.ilc.cnr.it/calamita/

AILC Associazione Italiana di Linguistica Computazionale

Finanziato dall'Unione europea
NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

THANK YOU FOR YOUR ATTENTION!