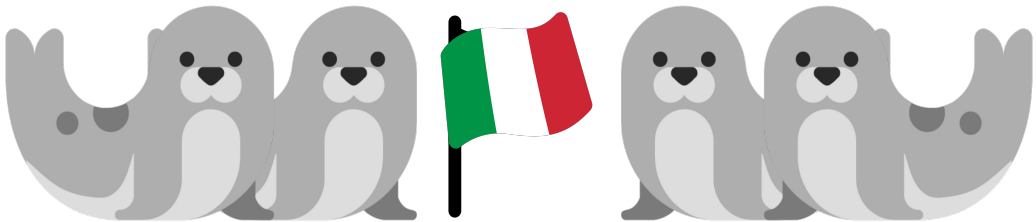
LLaMAntino

adapting LLaMA models to new languages and domains

PIERPAOLO BASILE, 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 AI RESEARCH

Large Language Models in Agriculture - CREA, Rome, 13 december 2024

The working group

SWAP researchgroup

Semantic **W**eb **A**ccess and **P**ersonalization
"Antonio Bello" research group
http://www.di.uniba.it/~swap

NVIDIA

Thanks to...

SWAP researchgroup

NVIDIA

CINECA consorzio interuniversitario

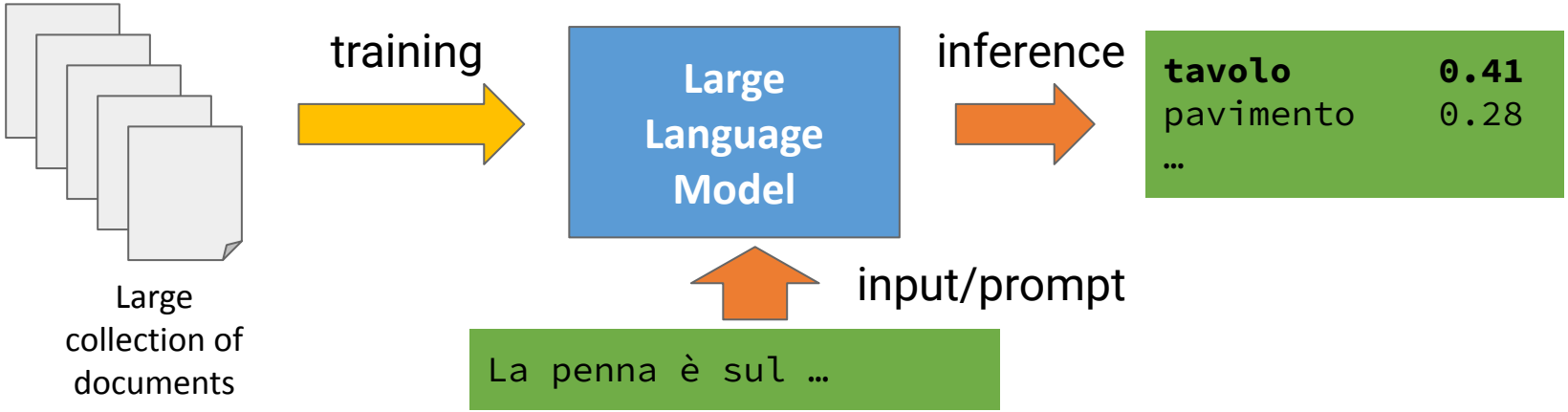
Finanziato dall'Unione europea NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

FAIR FUTURE AI RESEARCH

Large Language Models (LLMs)

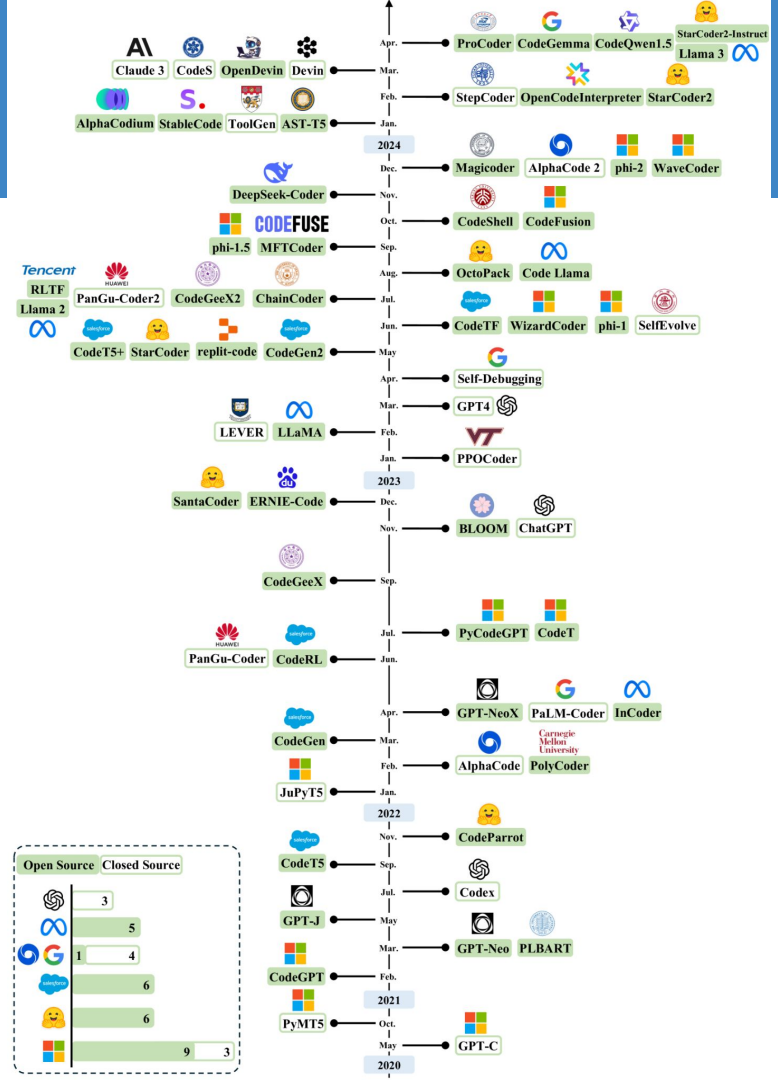


- An LLM is a **computational model** capable of **language generation** or other NLP tasks
- LLMs learns **statistical relationships** from **vast amounts of text** during a self-supervised and semi-supervised training process

What can an LLM do?

- Generation and completion of sentences
- Question answering
- Text summarization
- Machine translation
- Supervised training in specific tasks (fine-tuning)

Timeline LLMs



LLMs types

Open Weight models

- They only release the pre-trained parameters or weights
- The model can be used for inference and tuning
- The training code, original dataset, and model architecture are not provided

Open Weight LLMs

Open Source models

- They provide the complete source code and information needed to retrain the model from scratch
- This includes the model architecture code, training methodology and hyperparameters, the original training dataset

Open Source LLMs

Proprietary models

- LLMs-as-a-service
- Some can be fine-tuned with proprietary tools
- Restrictive licenses for use and modification

Proprietary LLMs

LLMs types

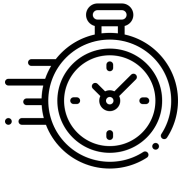
There is no "perfect" model, compromises are necessary

Decision Criteria

Privacy

Quality

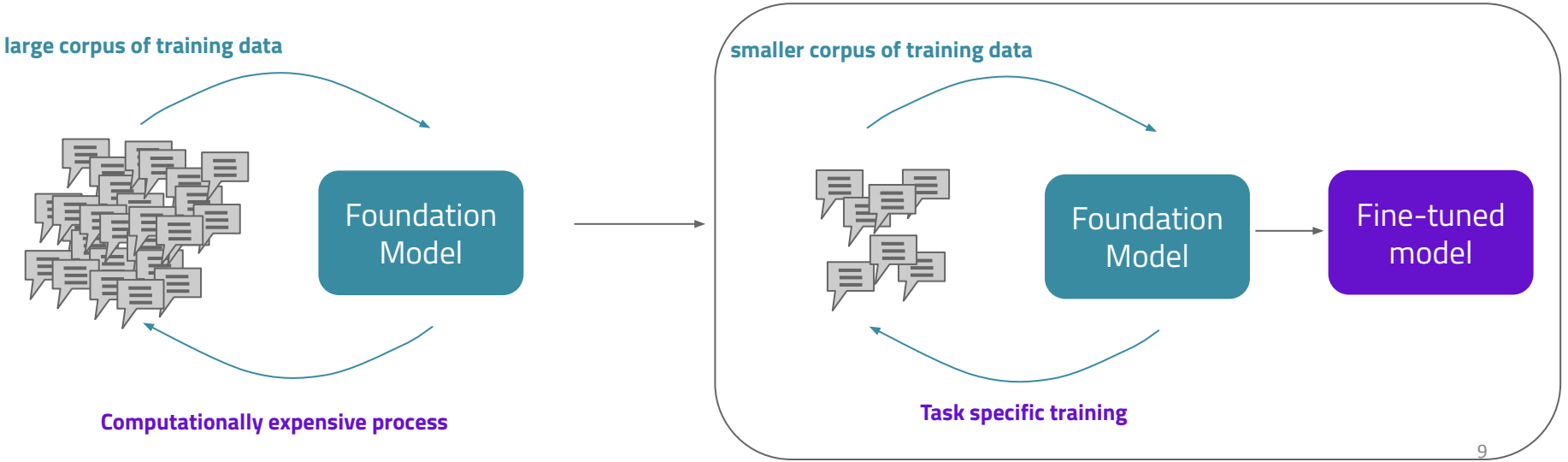
Price



Latency

Timeline LLMs

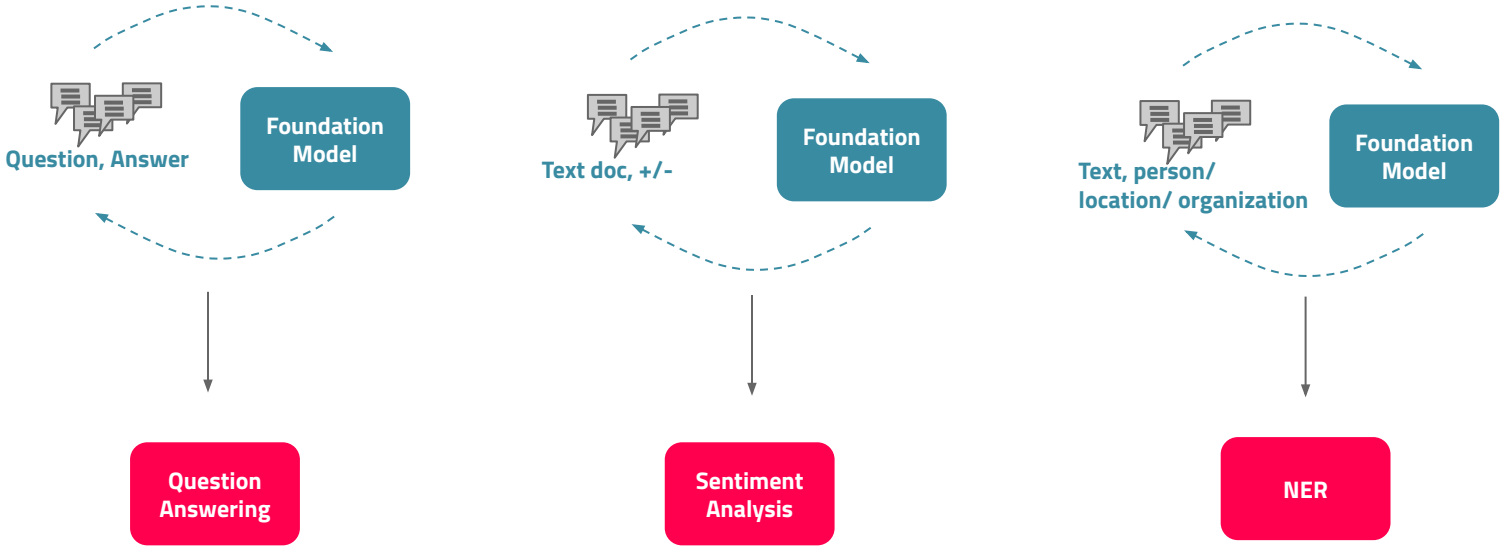
Fine-tuning: The process of further training a model on a specific task to adapt it to an application or domain



Timeline LLMs

Fine-tuning for specific tasks

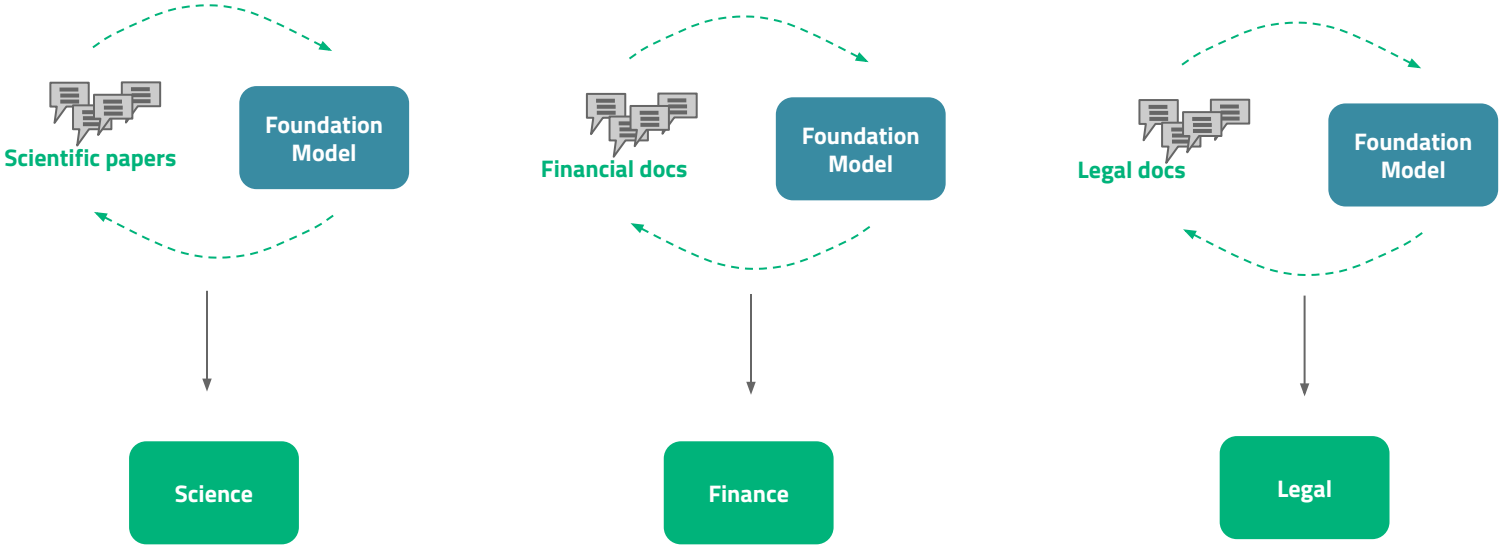
Supervised training on smaller labeled datasets



Timeline LLMs

Fine-tuning for adapting the model to a specific domain

Supervised training on smaller labeled datasets

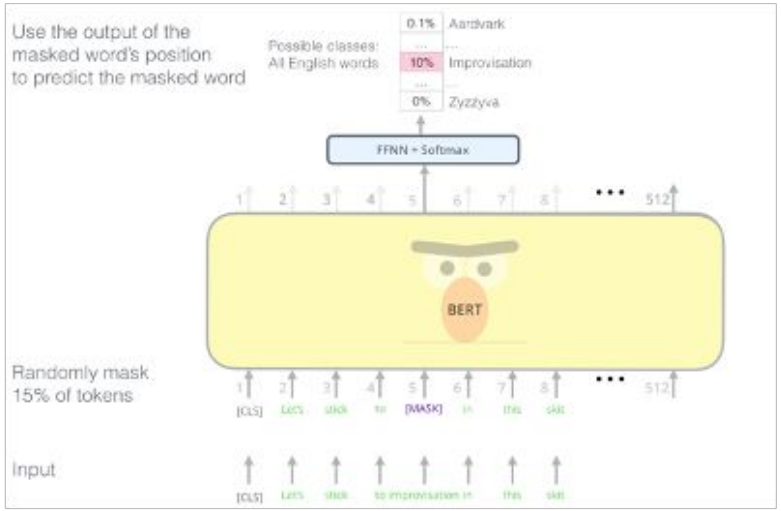


Domain-specific fine-tuned models

From the beginning (2019)

AlBERTo: The Italian Language Understanding Model

AlBERTo wants to be the first Italian language understanding model to represent a style of writing of social networks, **Twitter** in particular, written in **Italian**.



The **core** deep learning structure of BERT and AlBERTo is a **12x Transformer** Encoder, where for each input, a percentage of terms is **Masked** and then predicted for optimizing network weights in back-propagation.

* Polignano, M., Basile, P., De Gemmis, M., Semeraro, G., & Basile, V. (2019). Alberto: Italian BERT language understanding model for NLP challenging tasks based on tweets. In *Proceedings of the 6th Italian Conference on Computational Linguistics*, Bari, Italy, November 13- 15, 2019. CEUR Workshop Proceedings 2481, CEUR-WS.org, 2019.

Basic Idea - Start from Foundation Models

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

* Basile, P., Cassotti, P., Polignano, M., Siciliani, L., & Semeraro, G. (2023). On the impact of Language Adaptation for Large Language Models: A case study for the Italian language using only open resources. In *Proceedings of the 9th Italian Conference on Computational Linguistics*, Venice, Italy, November 30 - December 2, 2023. CEUR Workshop Proceedings 3596, CEUR-WS.org, 2023.

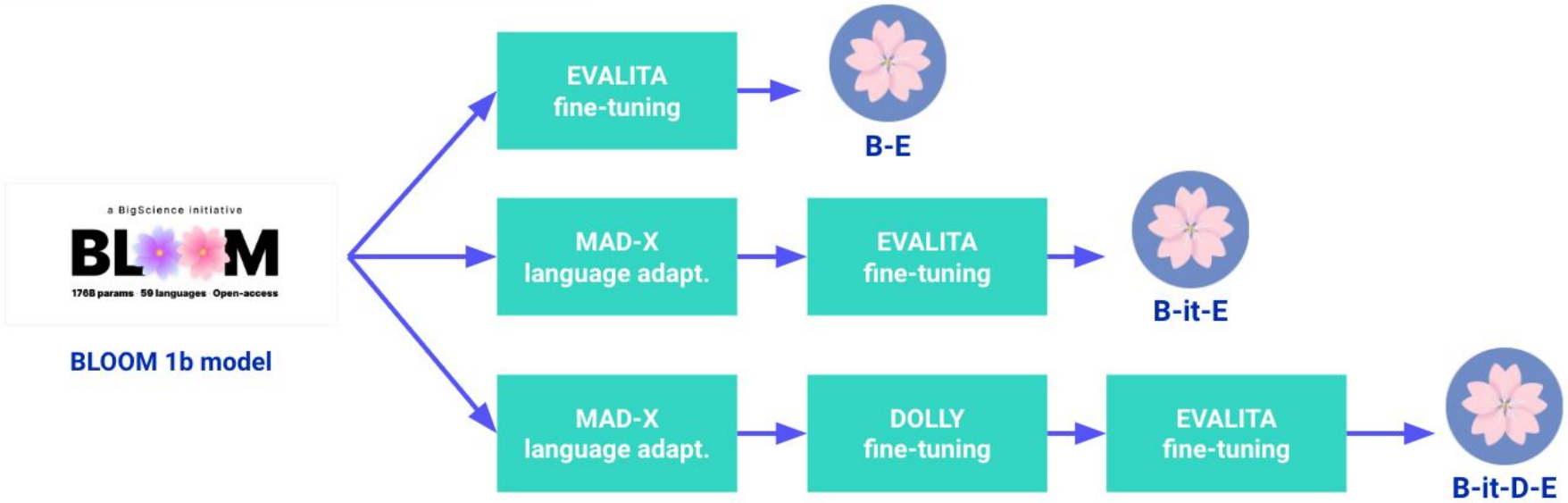
Basic Idea - Start from Foundation Models

- **Adaptation of BLOOM models to work with a new language (Italian)**, using only a limited sample size (100,000 samples)
- Exploitation of a **Language Adaptation** methodology called **MAD-X**
- Evaluation of the adapted models after a phase of instruction-based tuning on two Italian classification tasks
- **Open-science** approach using only **data created or processed using open-source tools**
- All data and models used in this work are under an open-source license

* Basile, P., Cassotti, P., Polignano, M., Siciliani, L., & Semeraro, G. (2023). On the impact of Language Adaptation for Large Language Models: A case study for the Italian language using only open resources. In *Proceedings of the 9th Italian Conference on Computational Linguistics*, Venice, Italy, November 30 - December 2, 2023. CEUR Workshop Proceedings 3596, CEUR-WS.org, 2023.

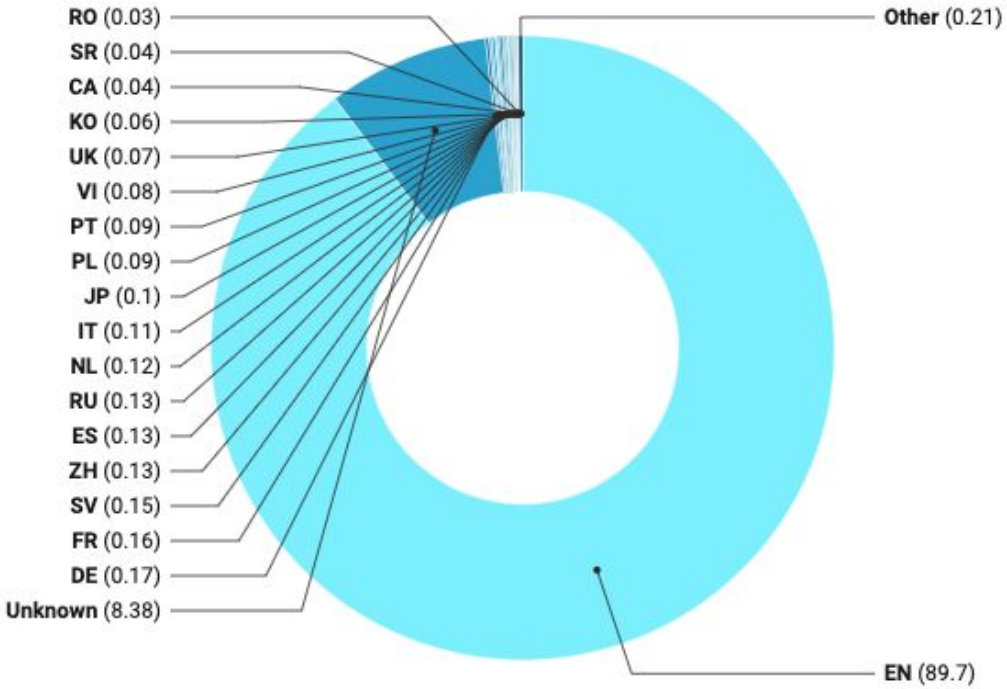
BLOOM-1b7

basilepp19/**bloom-1b7_it**



* Basile, P., Cassotti, P., Polignano, M., Siciliani, L., & Semeraro, G. (2023). On the impact of Language Adaptation for Large Language Models: A case study for the Italian language using only open resources. In *Proceedings of the 9th Italian Conference on Computational Linguistics*, Venice, Italy, November 30 - December 2, 2023. CEUR Workshop Proceedings 3596, CEUR-WS.org, 2023.

LLaMA 2 same problems as before



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*)

https://slator.com/meta-warns-large-language-model-may-not-be-suitable-non-english-use/

LLaMAntino 

- *LLaMAntino* is a *family of Italian adapted LLaMA 2 models*
- The family consists of 10 different models, **4** of which are **Italian adapted versions of LLaMA 2 base models**:
  - swap-uniba/LLaMAntino-2-7b-hf-ITA
  - swap-uniba/LLaMAntino-2-13b-hf-ITA
  - swap-uniba/LLaMAntino-2-chat-7b-hf-ITA
  - swap-uniba/LLaMAntino-2-chat-13b-hf-ITA

- **Goal:** Provide Italian researchers with LLMs that show a *good understanding of the Italian language*
- Should be **further tuned** to improve their capabilities on **specific tasks** …

LLaMAntino whitepaper: https://arxiv.org/abs/2312.09993

LLaMAntino

SWAP researchgroup

Meta AI Llama 2

Llama 2 - 7B/13B

Llama 2-chat - 7B/13B

Language Adaptation *(mc4)*

LLaMAntino - 7B/13B

LLaMAntino-chat - 7B/13B

fine tuning using the **dolly** dataset: *Information extraction*
Creative writing
Open QA, …

fine tuning using **EVALITA 2023** datasets: *misogyny detection, hate-speech, …*

fine tuning using the automatic translation of **UltraChat:** *dialogue with users*

LLaMAntino-dolly 7B/13B

LLaMAntino-evalita 7B/13B

LLaMAntino-chat-UltraChat 7B/13B

LLaMAntino whitepaper: https://arxiv.org/abs/2312.09993

LLaMAntino

swap-uniba's Collections + New

- LLaMAntino-3-ANITA
- 70B Model!
- LLaMAntino Models
- Papers
- BLOOM-IT
- LLaMAntino Adapters

LLaMAntino Models

Collection of all the published LLaMAntino models

- swap-uniba/LLaMAntino-2-chat-13b-hf-UltraChat-ITA — Text Generation • Updated Jan 8 • 5.27k • 16
- swap-uniba/LLaMAntino-2-chat-7b-hf-UltraChat-ITA — Text Generation • Updated Jan 8 • 20 • 7
- swap-uniba/LLaMAntino-2-chat-13b-hf-ITA — Text Generation • Updated Dec 18, 2023 • 16 • 4
- swap-uniba/LLaMAntino-2-chat-7b-hf-ITA — Text Generation • Updated Dec 18, 2023 • 15 • 3

LLaMAntino Models & GitHub

+20K downloads!

LLaMAntino whitepaper: https://arxiv.org/abs/2312.09993

LLaMAntino

- **Techniques**
  - **Quantization** (*4-bit*)
  - **QLoRA** (*Low-Rank Adaptation*)
  - **FSDP** (*Fully Sharded Data Parallel*)
  - **Argos Translate**: open source offline translation library based on OpenMT
- **Datasets**
  - **Language Adaptation**
    - gsarti/clean_mc4_it medium split
  - **Instruction-Tuning**
    - basilepp19/dolly-15k-it
    - EVALITA 2023 tasks
  - **Chat Fine-Tuning**
    - UltraChat

LLaMAntino 70B!

swap-uniba / LLaMAntino-2-70b-hf-UltraChat-ITA ♡ like 0

Text Generation | Transformers | PyTorch | Safetensors | Italian | llama | text-generation-inferer

Model card | Files and versions | Community | Settings

LLaMAntino-2-70b-hf-UltraChat-ITA 🇮🇹 🌟

Last Update: 02/02/2024

Model description

LLaMAntino-2-70b-hf-UltraChat-ITA is a *Large Language Model (LLM)* that is an instruction-tuned version of **LLaMAntino-2-70b** (an italian-adapted **LLaMA 2 - 70B**). This model aims to provide Italian NLP researchers with an improved model for italian dialogue use cases.

70 billion parameters!

Language Adaptation

+

UltraChat fine-tuning

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

LLaMAntino

SWAP researchgroup

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**

LLaMAntino - Anita

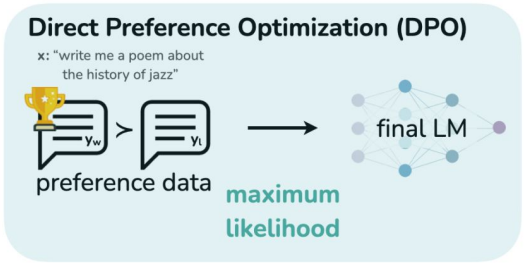
SWAP researchgroup

Supervised fine-tuned + DPO using several English datasets

Stanford Alpaca

Databricks' Dolly 2.0

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

LLaMAntino

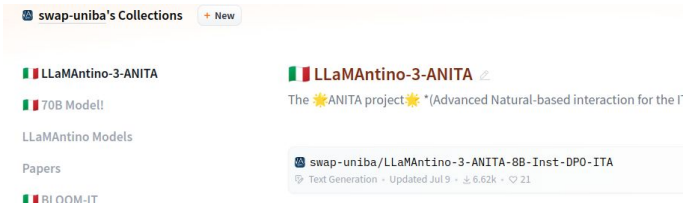
SWAP researchgroup

LLaMAntino: Llama-3-ANITA 8B

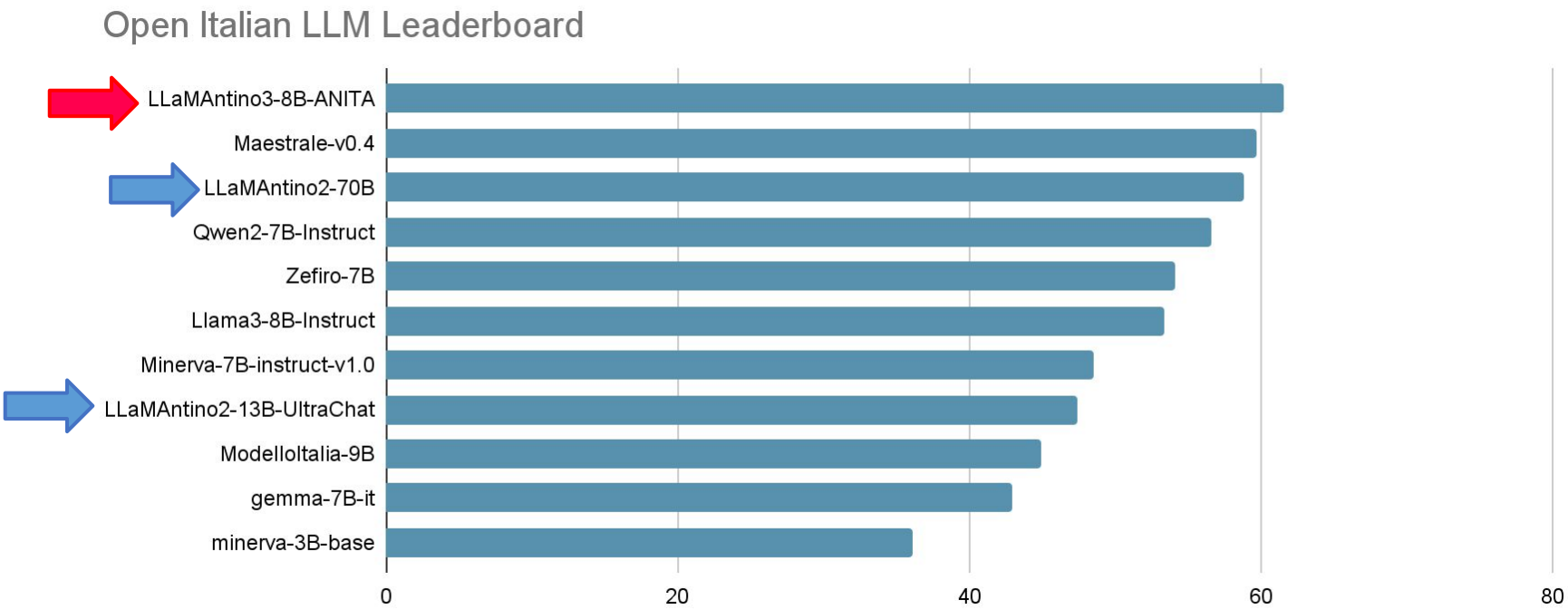
Il progetto 💥ANITA💥 *(Advanced Natural-based interaction for the ITAlian language)* vuole fornire ai ricercatori italiani un modello LLM migliorato per i casi d'uso della lingua italiana 🇮🇹.

https://chat.llamantino.it/

SCAN ME



LLaMAntino



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

Adapting a Large Language Model to the Legal Domain: A Case Study in Italian

Finanziato dall'Unione europea NextGenerationEU

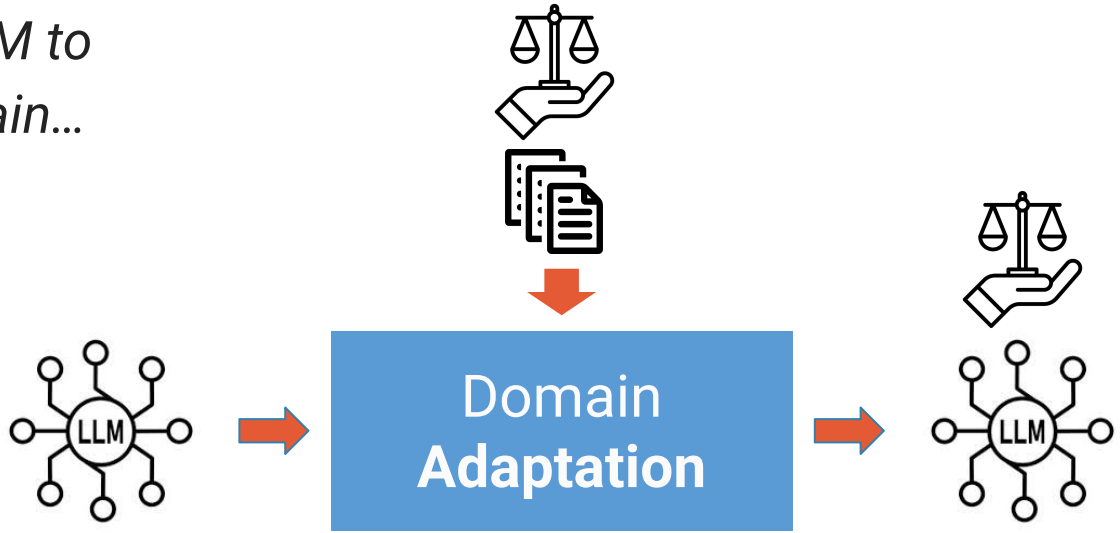
Ministero dell'Università e della Ricerca

Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA

FAIR FUTURE AI RESEARCH

Idea

Adapting an open LLM to the Italian legal domain…



Training data

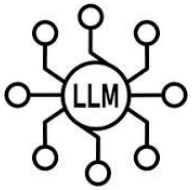
Dedicated Scraper

Cleaned corpus of ~**400k** documents

- We build a custom crawler for the Normattiva website
- The crawler considers the page layout and structure
- We use **Selenium + Beautiful soap**

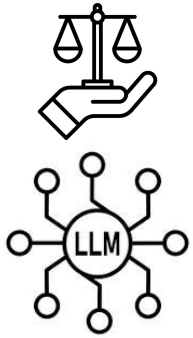
Crawler: https://github.com/FValerio96/NormattivaCrawling/
Corpus: https://huggingface.co/datasets/swap-uniba/normattiva-dump

Adaptation Strategy



Meta LLaMA-3 8b

> **continuous pre-training** using domain-specific corpora
> **PEFT + LoRA** (r=16, alpha=32, all linear layers, batch = 16)
> **unsloth** + **one RTX A6000** with 48GB



Evaluation

Goal: test the model to complete sentences from the legal corpus

>1,000 sentences

> Two different prompt lengths: 20 and 40

> **Metrics:**

>> BLEU

>> ROUGE

>> BERTscore

>> Perplexity

N-grams level

Semantic level

Model uncertainty

Results

Prompt length	20			40		
Model	Llama3.1	Llama3.1-NA-100k	Llama3.1-NA	Llama3.1	Llama3.1-NA-100k	Llama3.1-NA
BLEU-1	.132	.142	.152	.170	.193	.182
BLEU-2	.071	.102	.110	.111	.148	.142
Rouge-1	.313	.436	.448	.402	.509	.521
Rouge-2	.163	.294	.304	.266	.384	.397
Rouge-L	.254	.377	.387	.342	.451	.462
BERTscore	.705	.772	.777	.744	.796	.802
Perplexity	2.180	1.557	1.513	2.760	1.586	1.539

BLEU, Rouge and BERTscore increase according to number of documents, while perplexity decreases

The model adapted on 100k documents has similar performance to the one trained on the whole dataset!

Case study in Public Administration

Retrieval-Augmented Generation (RAG)

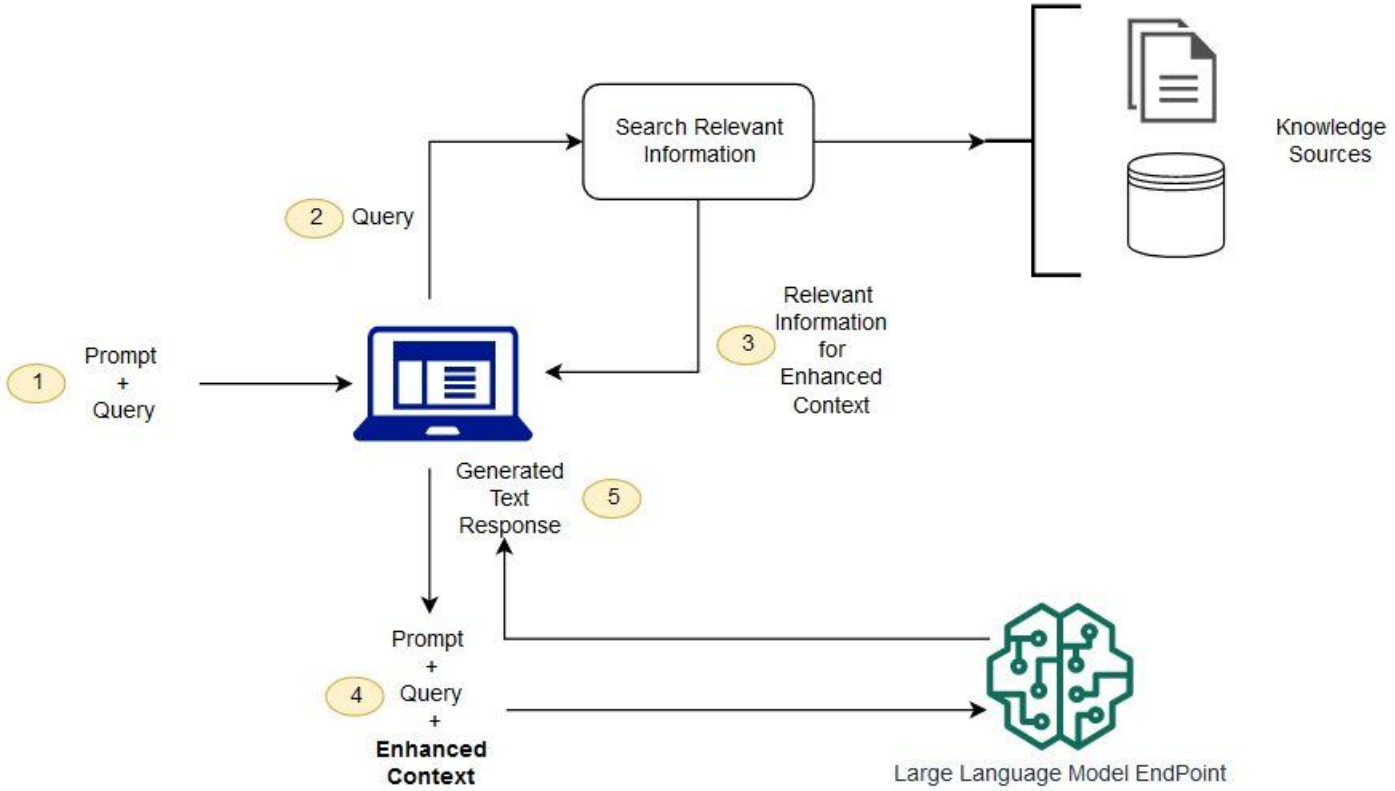
Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

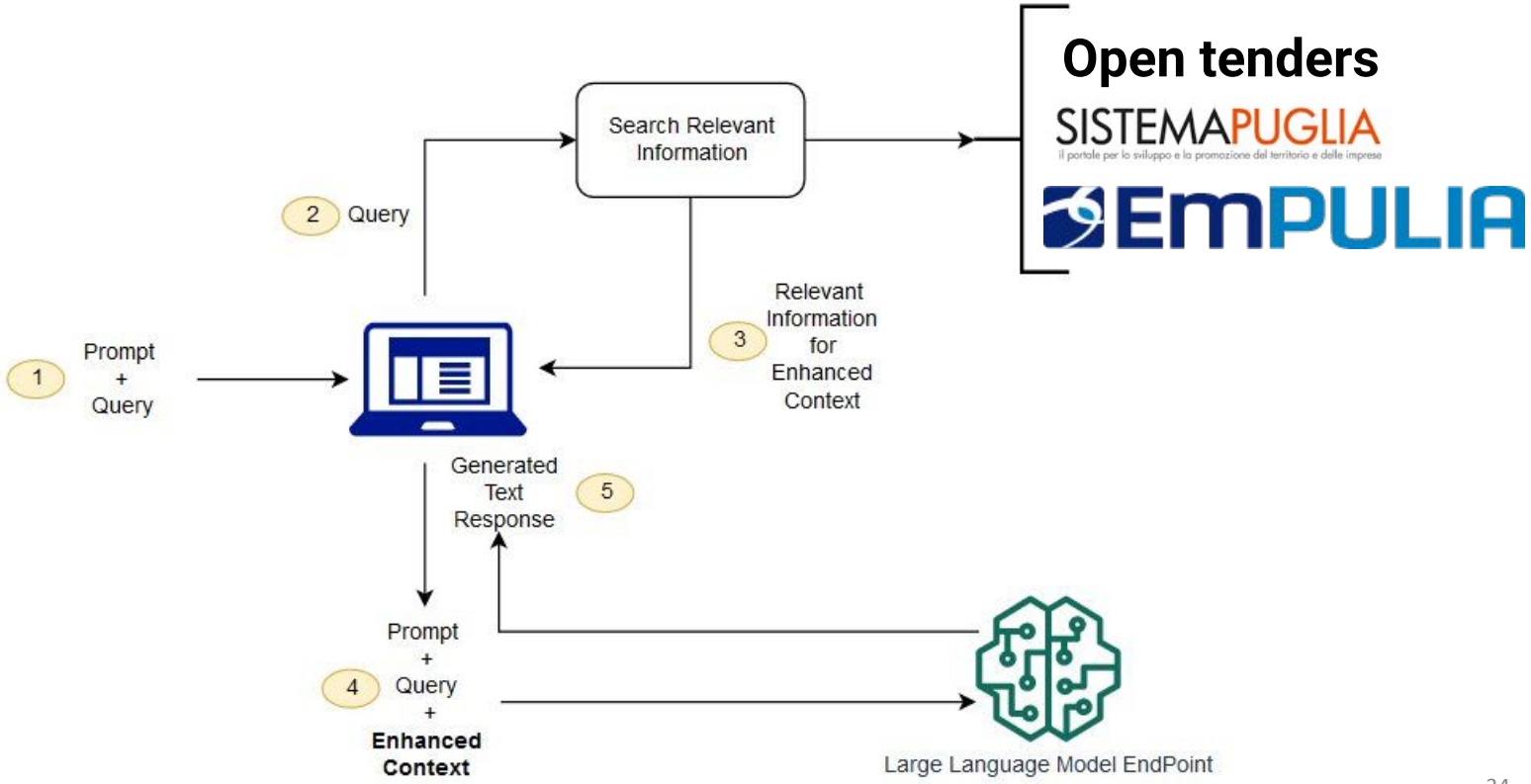
Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

FAIR
FUTURE AI RESEARCH

Retrieval-Augmented Generation

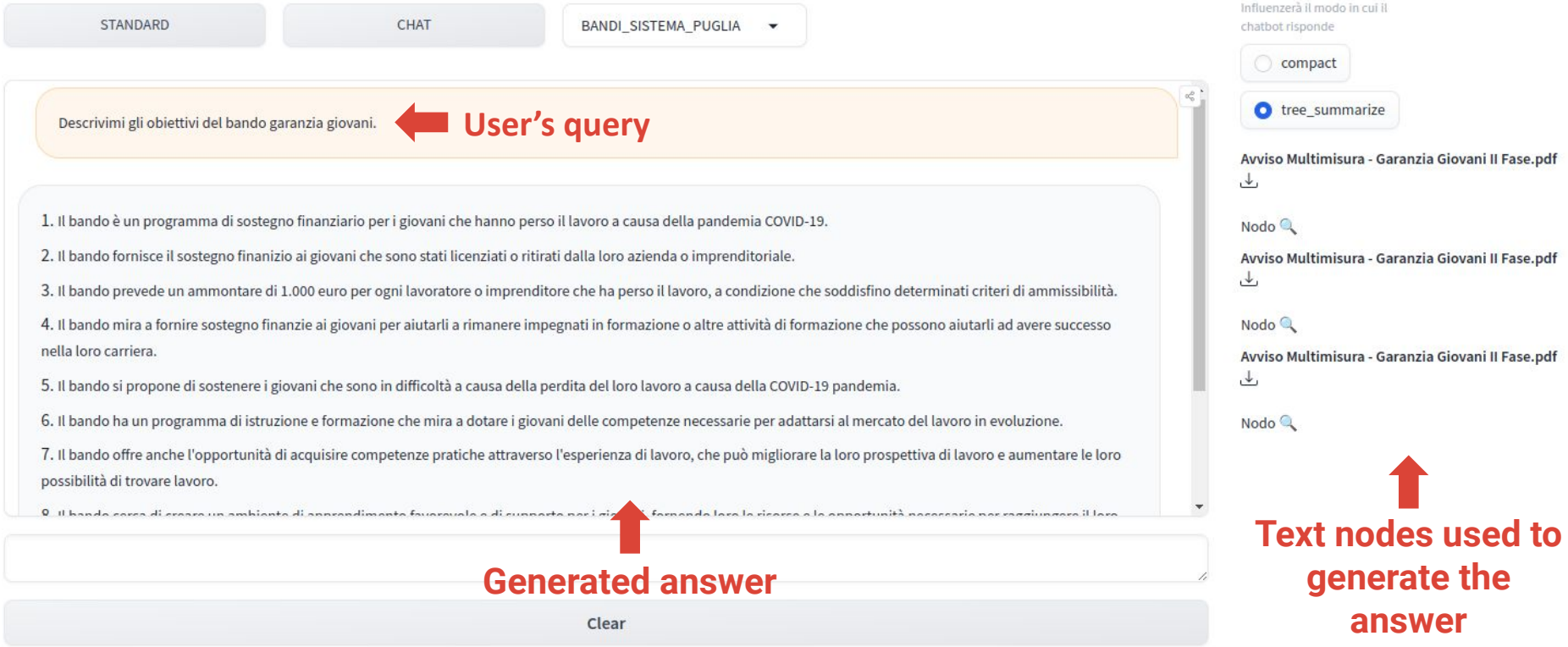


Retrieval-Augmented Generation



Retrieval-Augmented Generation

SWAP researchgroup



LLaVA-NDiNO: Empowering LLMs with Multimodality for the Italian Language

Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

FAIR
FUTURE AI RESEARCH

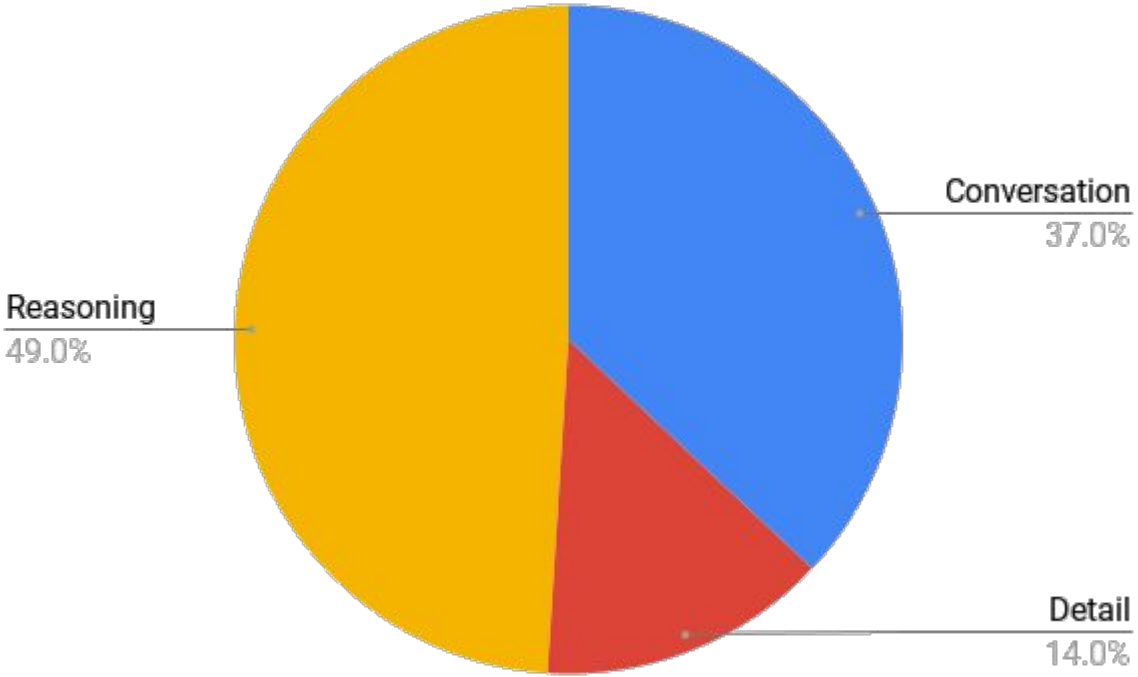
LLMs and LVLMs

After the success of LLMs (LLaMA, Mistral, …), **LVLM**s were developed

These models employ an adaptation strategy to make decoder-only text LLMs understand vision-language inputs

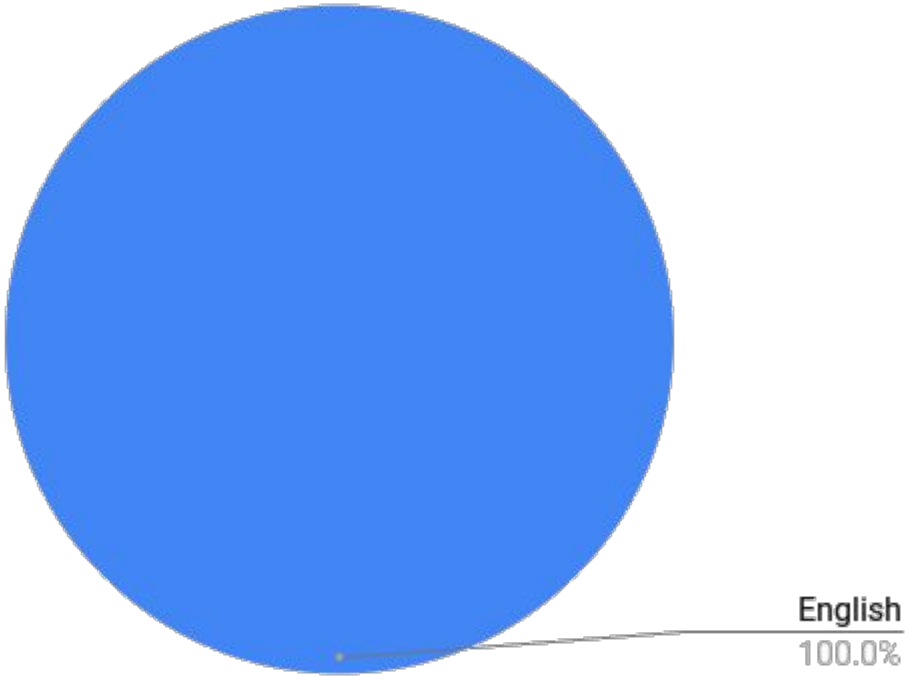
The problem

The first LLaVA model was instruction-tuned using a mixture of 158k instances covering three different meta-categories

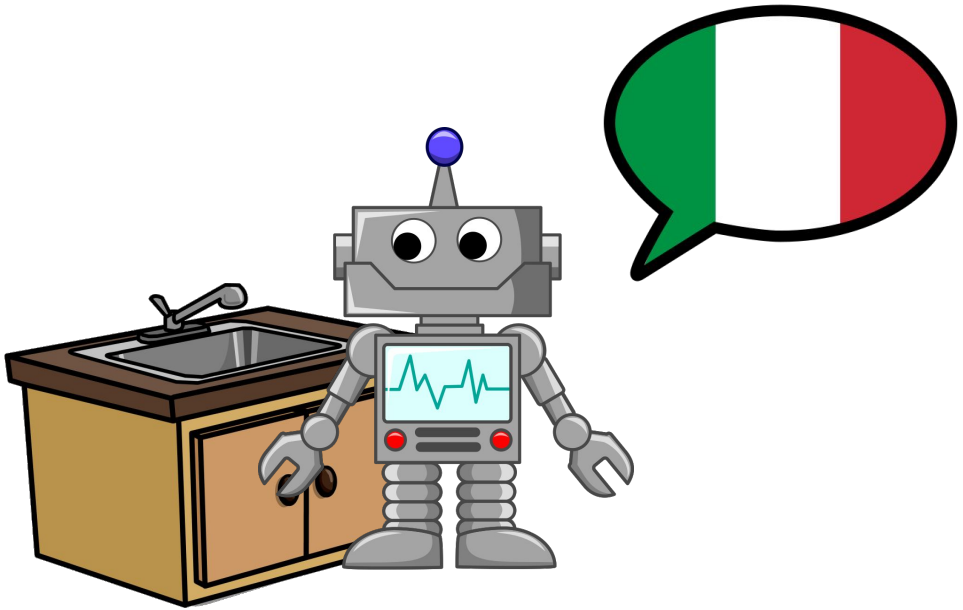


The problem

All of the instances were English-language vision-text data :(



LLaVA-NDiNO to the rescue: overview



We developed and released the first family of LVLMs trained on **Italian vision-language instances** only

The model family name is **LLaVA-NDiNO**

LLaVA-NDiNO to the rescue: base models

LLM:

- **LLaMA 3 8B base**

Vision Encoder:

- **CLIP ViT large-patch14-336**

We use **LLaVA-NeXT** as multimodal adaptation strategy and training framework

LLaVA-NDiNO to the rescue: models

We release three different model categories:

- **Adapt**
- **Short Instruction-Tuning**
- **Long Instruction-Tuning**

We test several training settings using the trained checkpoints

(e.g. performing the short instruction-tuning step for both the adapted and non-adapted model)

You can download the models from the collection!

LLaVA-NDiNO to the rescue: data

SWAP researchgroup

Adapt

Short IT

Long IT

Conversation

All datasets are either natively in the Italian language or translated using **MADLAD 3B**

Example

Quante persone ci sono in questa immagine? Rispondi brevemente.

LLaVA-NDiNO PT + IT:
1.

LLaVA-NDiNO PT + IT + LONG-IT:
C’è una persona in questa immagine.

Example

Cosa c'è di strano in questa immagine?

LLaVA-NDiNO PT + IT:
Un uomo è seduto su una sedia a rotelle che lava i panni.

LLaVA-NDiNO PT + IT + LONG-IT:
L'immagine è strana perché mostra un uomo che asciuga
le camicie mentre è in piedi sulla parte superiore di un camion giallo, che è un modo insolito e non
convenzionale per asciugare le camicie.

Evaluation of LLMs (for Italian)

Finanziato dall'Unione europea NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA

FAIR FUTURE AI RESEARCH

BACKGROUND

There are two possible settings to evaluate LLMs' linguistic abilities and knowledge

Open-ended Tasks

In una gara di papere di gomma, la papera vincente ha percorso 300 metri in 10 minuti.
Qual era la velocità media della papera?

Closed-ended Tasks

In una gara di papere di gomma, la papera vincente ha percorso 300 metri in 10 minuti.
Qual era la velocità media della papera? ***1.*** *0,5 m/s* ***2.*** *1,0 m/s*

Is closed-ended evaluation of language-adapted models optimal?

IDEA

- Several benchmarks (MMLU, ARC, …) are designed as **closed-ended tasks**
- We redesign the closed-ended tasks as open-ended ones

In una gara di papere di gomma, la papera vincente ha percorso 300 metri in 10 minuti.
Qual era la velocità media della papera?
1. 0,5 m/s
2. 1,0 m/s

Scegli l'opzione corretta.

Expected answer: 1

In una gara di papere di gomma, la papera vincente ha percorso 300 metri in 10 minuti.
Qual era la velocità media della papera?

Genera la risposta.

Expected answer: 0,5 m/s

In una gara di papere di gomma, la papera vincente ha percorso 300 metri in 10 minuti.
Qual era la velocità media della papera?
1. 0,5 m/s
2. 1,0 m/s

Genera l'opzione corretta.

Expected answer: 0,5 m/s

- This strategy allows to test the generative abilities of the model, especially in languages different from English

CALAMITA

Challenge the Abilities of LAnguage Models in ITAlian

Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

FAIR
FUTURE AI RESEARCH

CALAMITA

Evaluation of LLMs on languages other than English

Possible strategies

Translate existing benchmarks for English

Repurpose existing benchmarks for target language

Create new data echoing existing benchmarks for English

Start from scratch entirely

CALAMITA

Evaluation of LLMs on languages other than English

Possible strategies

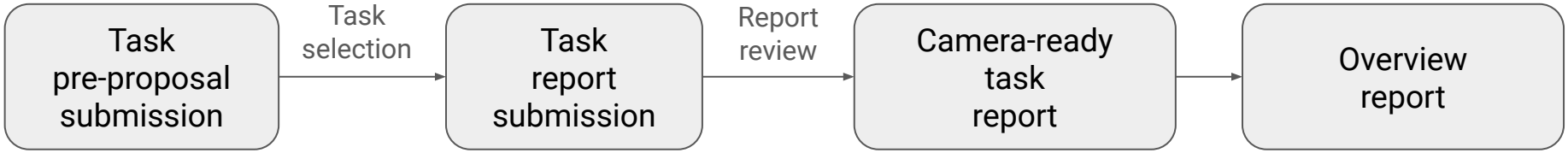
~~Translate existing benchmarks for English~~

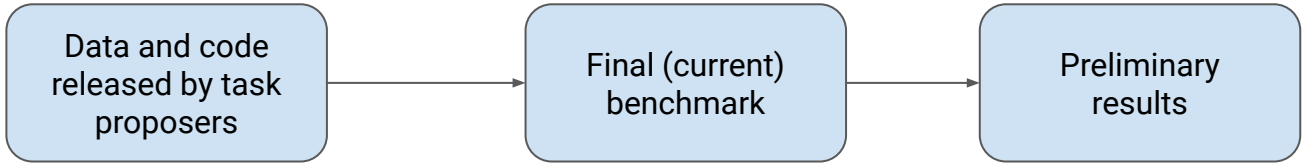
Repurpose existing benchmarks for target language

Create new data echoing existing benchmarks for English

Start from scratch entirely

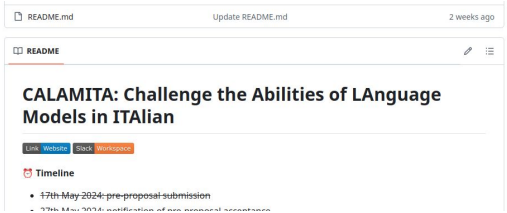
CALAMITA: a collaborative effort





AILC Associazione Italiana di Linguistica Computazionale

- Data and code integrated in a new fork of **evaluation-harness**
- Public datasets on Hugging Face


README.md Update README.md 2 weeks ago

README

CALAMITA: Challenge the Abilities of LAnguage Models in ITAlian

Link Website Slack Workspace

Timeline

- ~~17th May 2024: pre-proposal submission~~

Code: https://github.com/CALAMITA-AILC/calamita2024
Leaderboard: https://calamita-ailc.github.io/calamita2024/

SWAP researchgroup

THANK YOU FOR YOUR ATTENTION!