
Creating Synthetic Datasets for LLM Fine-Tuning: Multilingual and News Domains

University of Milano-Bicocca

Our team



Alex Calabrese

Data Science Student



Antonio Sabbatella

Data Science Student



Table of contents

01

Acquisition

02

Integration

03

Enrichment

04

Quality

05

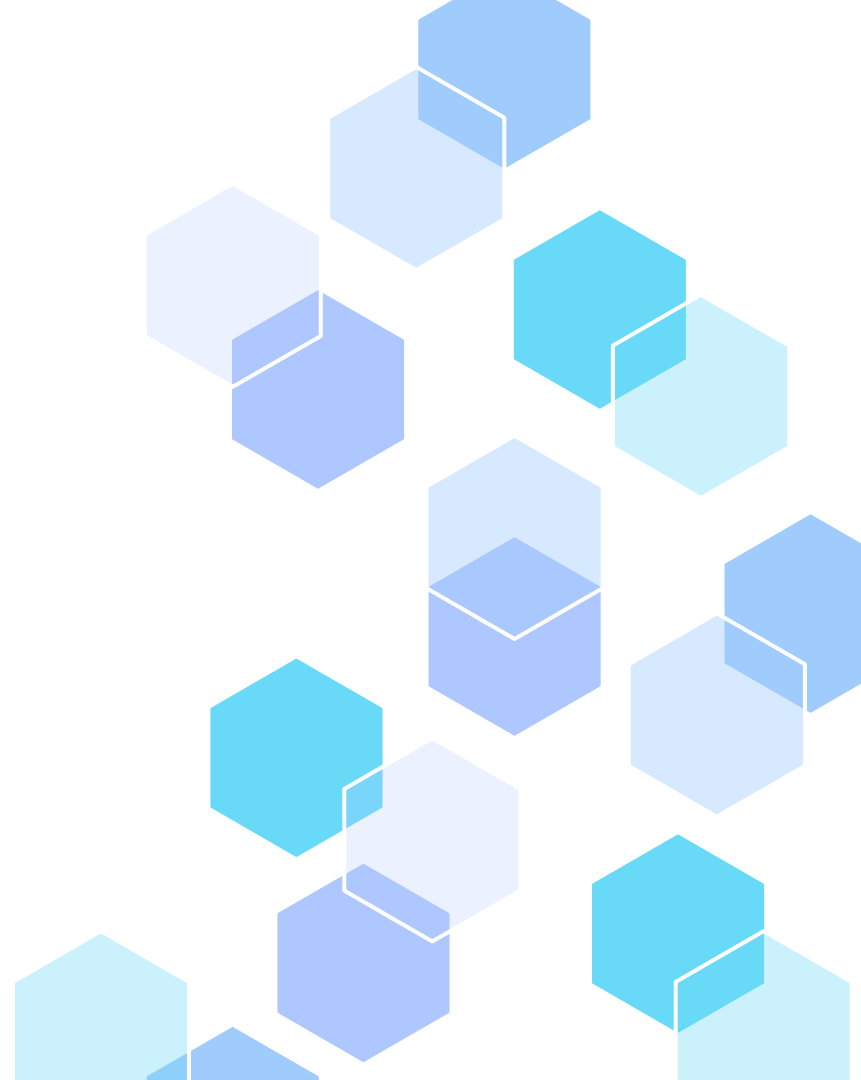
Storage

06

Conclusions

00

Research Questions






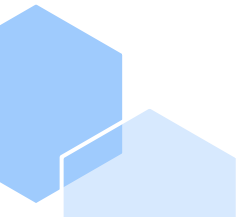
Research Question 1

Starting from information publicly available online, is it possible to create a synthetic dataset for LLM fine tuning?



Research Question 2


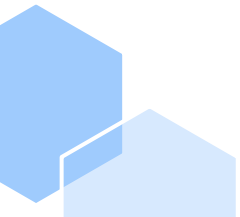
Given the most used datasets for fine tuning, processing them, are they sufficient to make an open source <10b model multi-lingual?



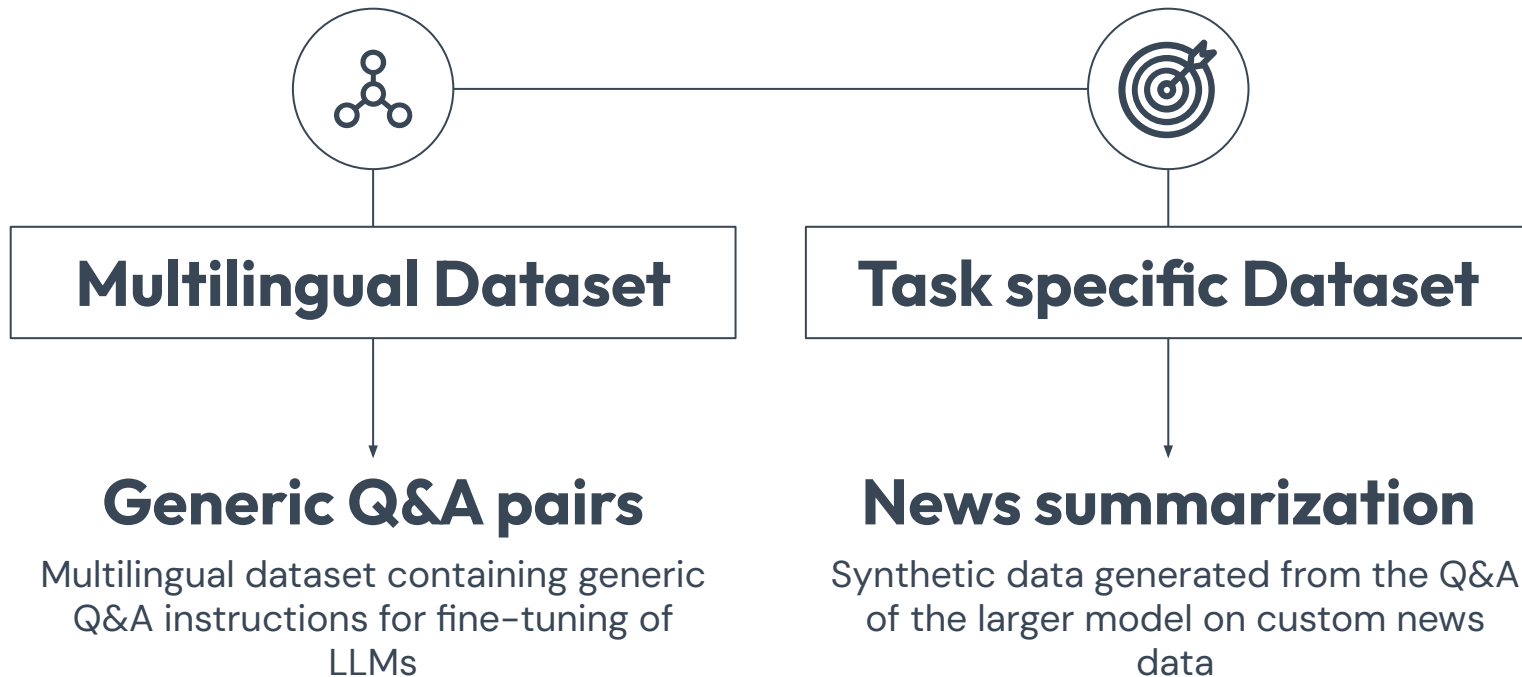


Research Question 3

Is it possible to carry out fine-tuning or distill knowledge so as to teach a small LLM to imitate a larger and more expensive model through fine tuning?



Dataset structure





01

Data

Acquisition

Sources – Multi Lingual Finetuning



Hugging Face

SlimOrca

Approximately 500k GPT-4 completions in english, designed for NLP Tasks




Hugging Face

OpenHermes

Curated and annotated text data generated using GPT-4 for NLP tasks

Sources – Multi Lingual Finetuning

Dataset Viewer	Auto-converted to Parquet	API	View in Dataset Viewer
Split (1) train · 518k rows			
Search this dataset			
conversations list · lengths			
			
[{ "from": "system", "value": "You are an AI assistant. You will be given a task. You must generate a detailed and long answer.", "weight": null }, { "from": "human", "value": "Write an article based on this \"			
[{ "from": "system", "value": "You are an AI assistant that follows instruction extremely well. Help as much as you can.", "weight": null }, { "from": "human", "value": "Answer the following question: - number is 54 -			
[{ "from": "system", "value": "You are an AI assistant. You will be given a task. You must generate a detailed and long answer.", "weight": null }, { "from": "human", "value": "Produce a long descriptive			
[{ "from": "system", "value": "You are a helpful assistant, who always provide explanation. Think like you are answering to a five year old.", "weight": null }, { "from": "human", "value": "Write a title for this			
[{ "from": "system", "value": "You are an AI assistant. User will you give you a task. Your goal is to complete the task as faithfully as you can. While performing the task think step-by-step and justify your			
[{ "from": "system", "value": "You are a helpful assistant, who always provide explanation. Think like you are answering to a five year old.", "weight": null }, { "from": "human", "value": "Read the following			
< Previous 1 2 3 ... 5,180 Next >			




SlimOrca

Dataset Viewer

Auto-converted to Parquet API View in Dataset Viewer

Split (1)
train · 243k rows

Search this dataset

instruction string · lengths	input string · lengths	output string · lengths
		
Write a Perl script that processes a log file and counts the occurrences...		'''perl #!/usr/bin/perl use strict; use warnings; # Check if a log file path is provided as command...
What can be seen once in a minute, twice in a moment, and never in a...		The letter 'M'.
Famous inventors and their inventions: Identify five well-know...		1. Thomas Edison: One of his most significant inventions is the incandescent light bulb, which...
Generate a list of 12 words that start with 'qu'.		1. Quail 2. Quarry 3. Quasar 4. Quench 5. Query 6. Quest 7. Queue 8. Quiche 9. Quick 10. Quilt 11...
Who was the first woman to win a Nobel Prize, and in which field did...		Marie Curie; Physics
A hotel chain wants to optimize its pricing strategy by implementi...		To develop a dynamic pricing model for the hotel chain, we can follow these steps: 1. Analyze...

< Previous

1 2 3 ... 2,429

Next >

OpenHermes

Sources – Newspaper news



Open

It covers a wide range of topics including politics, economy, fact-checking, technology, environment, and current affairs.



IlPost

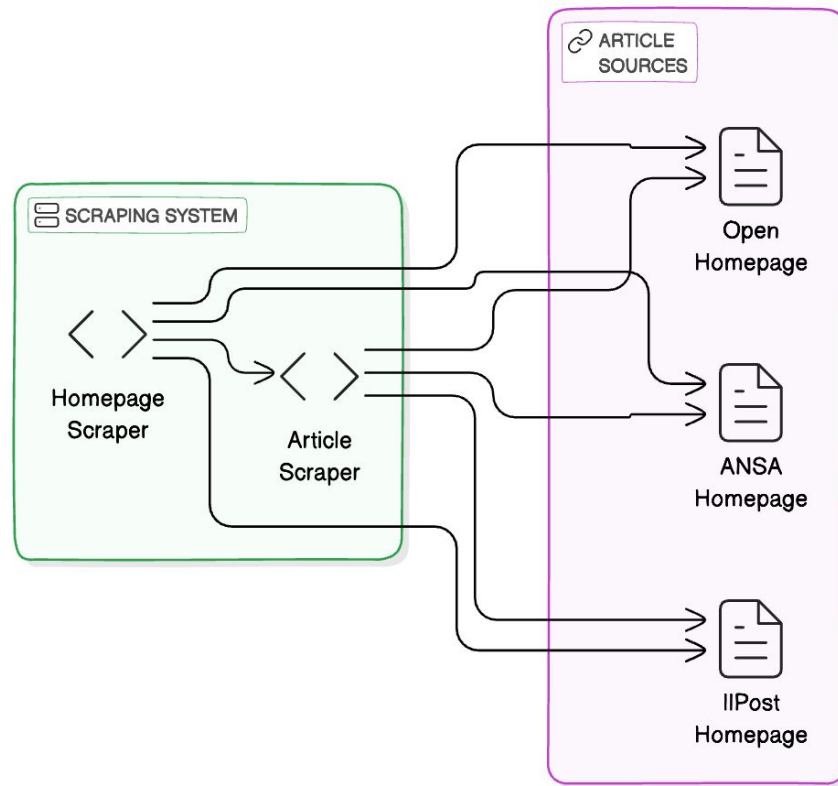
Independent Italian online newspaper that offers articles on politics, international news, culture, and science.



ANSA

Italy's leading news agency and one of the top-ranking news agencies in the world.

Scraping - Newspaper news





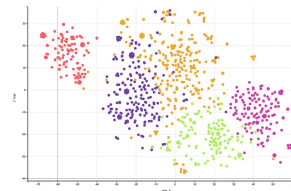
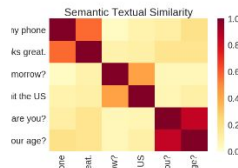
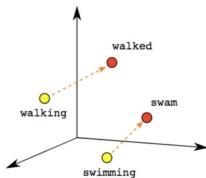
02

Data

Integration

Data Integration

Newspaper News Dataset



Embedding

Transformer-based model (USE) encode text into high-dimensional, **semantic aware**, vectors.

Similarity Matrix

Compute the **cosine similarity** among each vector pairs to find the semantically similar news.

Clustering

The *Agglomerative Hierarchical Clustering* algorithm is used for **iteratively merging the most similar news pairs**.

Data Integration

Newspaper News Dataset

Prompt

Given multiple titles and contents of articles, summarize and integrate them into a single text.

Input:

- Title: Sla, da AriSLA 1 milione per 7 progetti di ricerca
- Content: ...
- Title: Arera sostiene 4 progetti pilota di Snam con oltre...
- Content: ...
- ...

LLM



Summary

AriSLA, la principale fondazione non profit che finanzia la ricerca sulla SLA in Italia, ha annunciato un investimento di quasi 1 milione di euro per lo sviluppo di 7 progetti di ricerca innovativi selezionati con il bando 2023 sulla sclerosi laterale amiotrofica. I progetti, che riguardano una grave patologia neurodegenerativa che in Italia ...

News Cluster

Question

Summary

Answer

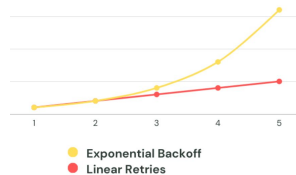
Challenges and Solutions

Dealing with Rate Limits and IP Bans



Challenge

Encountered rate limits and potential IP bans during data scraping and LLM calling.

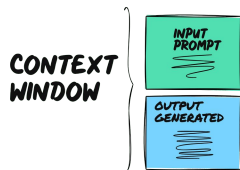


Solution

Implemented a retry system with exponential backoff strategy to mitigate IP bans

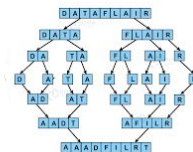
Challenges and Solutions

Large cluster exceed LLM context window



Challenge

In case of long news or numerous clusters, errors are incurred due to exceeding the maximum supported LLM context window

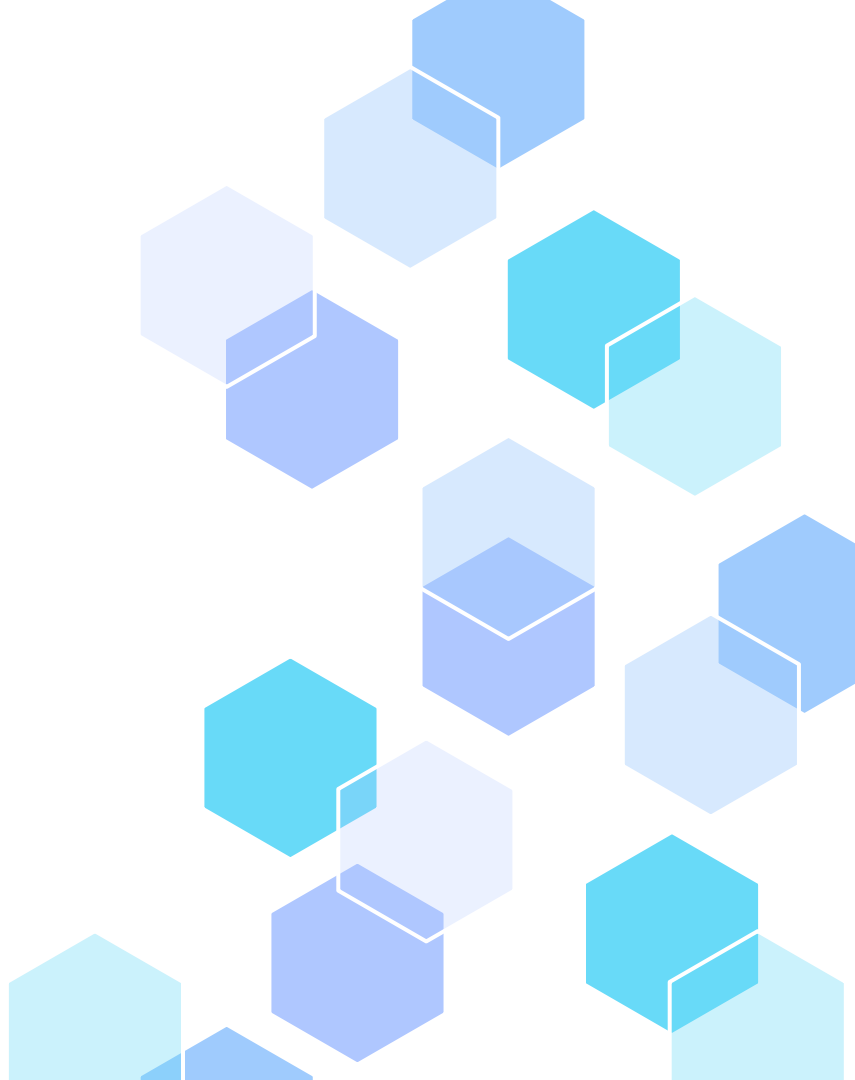


Solution

Using recursion and divide-and-conquer approach to split the initial cluster and summarize subcluster

03

Data Enrichment



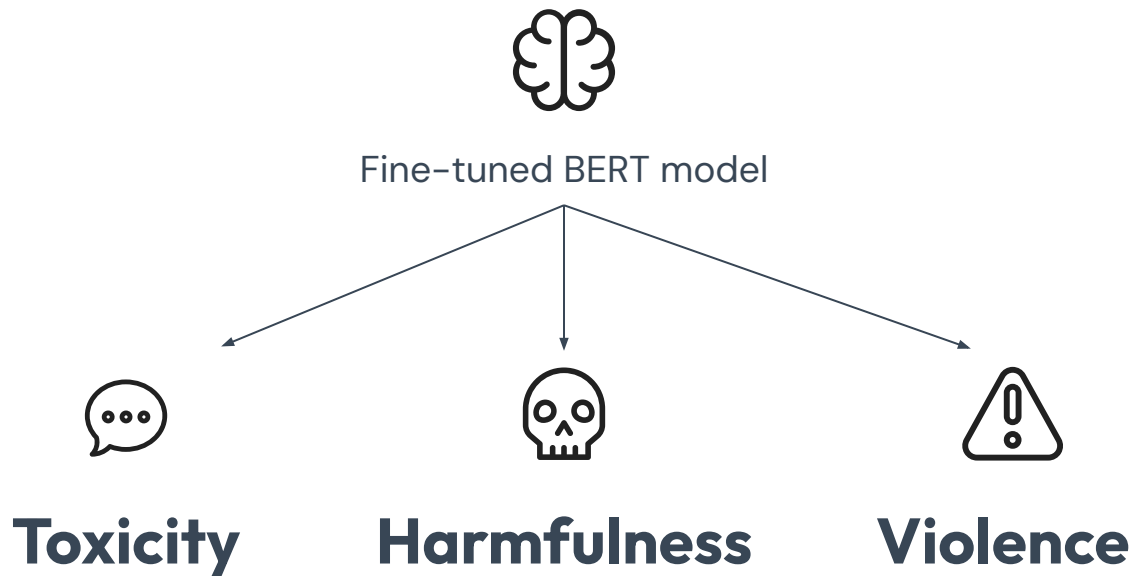
Data Enrichment

Multilingual Fine-tuning Dataset



Data Enrichment

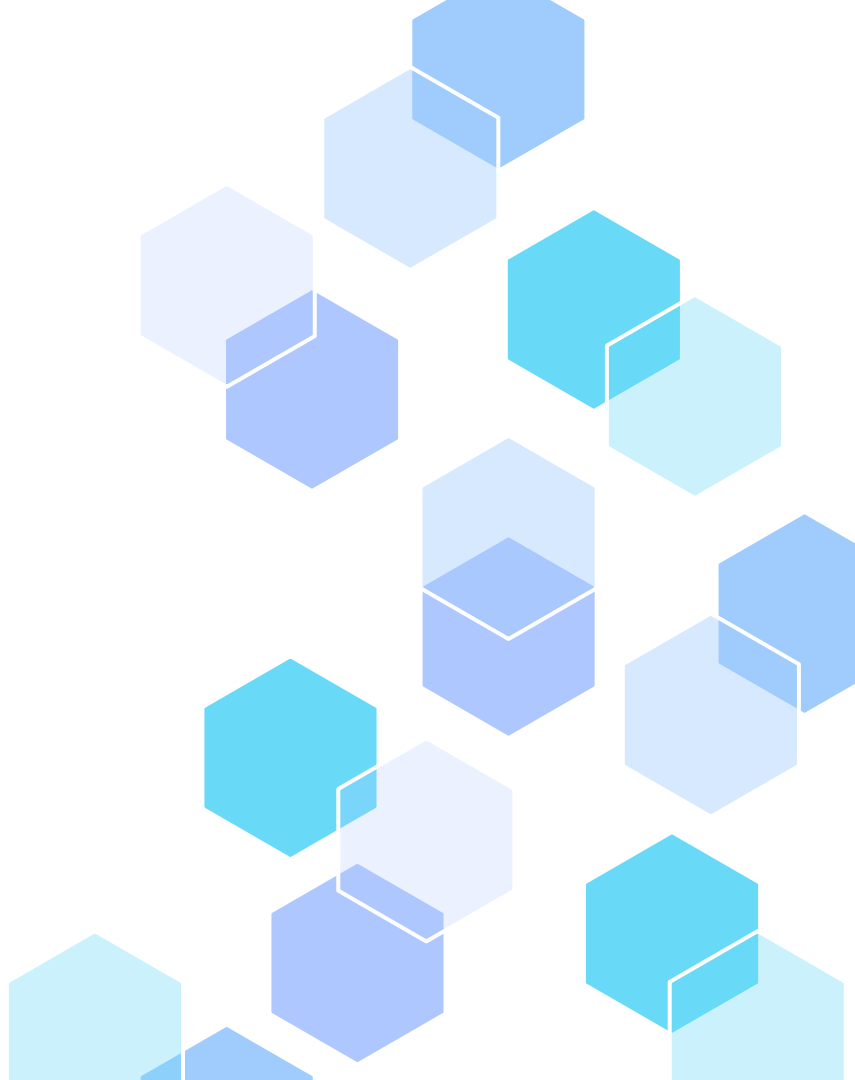
Newspaper News Dataset



04

Data

Quality

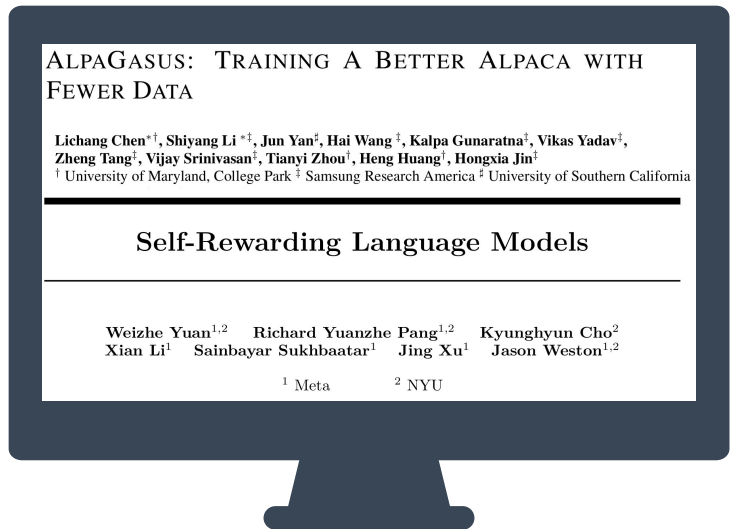


Data quality in

fine tuning

The quality evaluation of question-answer pairs presents a significant **challenge** due to the substantial **volume** of data involved.

Manual evaluation is infeasible.



Methodologies proposed in literature

Data quality in

fine tuning

The prompt ask the llm to rate the instruction, input and response with a score in [0, 5]

System Prompt:

We would like to request your feedback on the performance of AI assistant in response to the instruction and the given input displayed following.

Instruction: [Instruction]

Input: [Input]

Response: [Response]



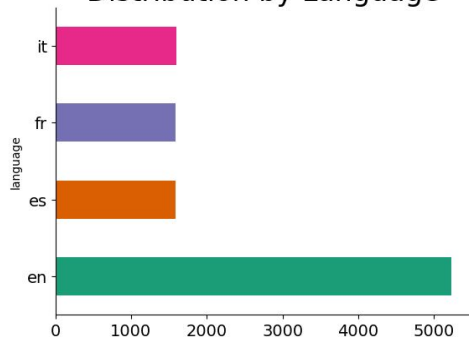
User Prompt:

Please rate according to the [dimension] of the response to the instruction and the input. Each assistant receives a score on a scale of 0 to 5, where a higher score indicates higher level of the [dimension]. Please first output a single line containing the value indicating the scores. In the subsequent line, please provide a comprehensive explanation of your evaluation, avoiding any potential bias.

Alpagasus prompt template

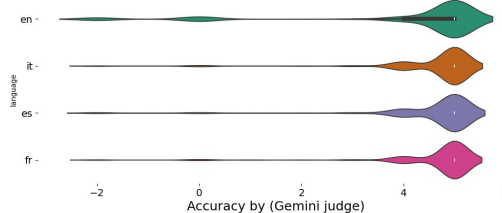
Data Quality

Distribution by Language



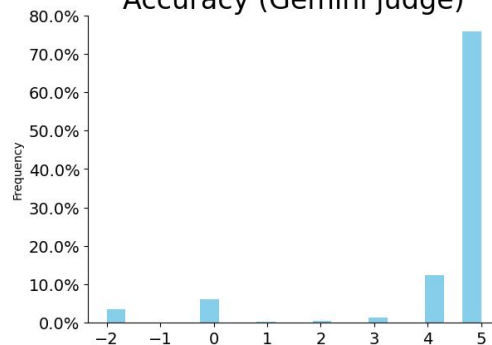
**Language
distribution**

Accuracy respect language



**Differences in
languages**

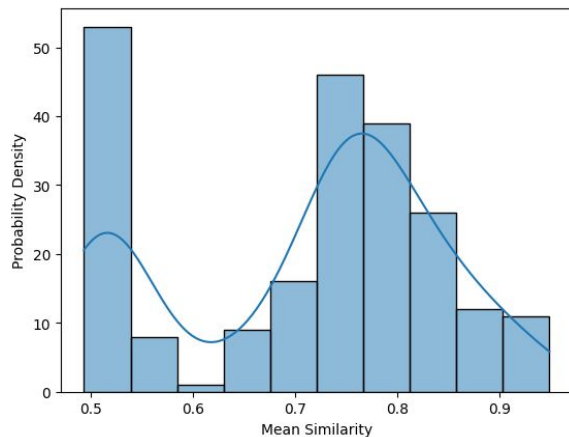
Accuracy (Gemini judge)



Overall score

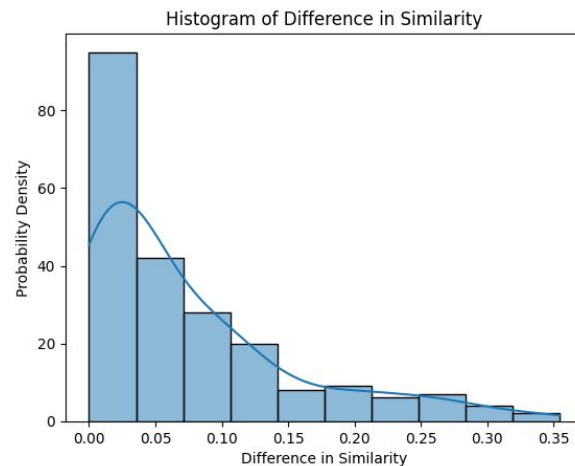
Data Quality

Post aggregation analysis



Average Similarity

Similarity between the summary and original new sources. This score can be used as criterion for filtering *hallucination*.



Variability

This metric represent the of variability similarity scores in each cluster.

Example of 0 score types

Tasks	Response	Score
Sviluppa un elenco di 7 parole che terminano con "nave"	<ol style="list-style-type: none">1. Amicizia2. Relazione3. Direzione4. Borsa di studio5. Partenariato6. Cittadinanza7. Apprendistato	Explanation: The response is completely inaccurate. The instruction is to develop a list of 7 words that end with "nave", but the response does not contain any words that end with "nave". Instead, it provides a list of words that are unrelated to the instruction.
What age-related phrase is represented by the anagram "over fifty"?	forty five	Explanation: The response "forty five" is not an age-related phrase represented by the anagram "over fifty". The correct answer should be "fifty over". The response is inaccurate and does not match the instruction and the input.



05

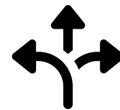
Storage

Storage: mongoDB



NoSQL

NoSQL databases scale horizontally, handling large data and traffic. Ideal for high-traffic applications and big data workloads.



Flexibility

NoSQL databases offer flexible schema designs, allowing for easy adaptation to changing data structures.

Storage: mongoDB

```
▼ 44 : Object
  _id : ObjectId('665f82c2d9fab9d7ed114006')
  date : "14 dicembre 2023, 15:50"
  tags : null
▼ sources : Array (2)
  ▼ 0: Object
    title : "Al Maxxi nasce la Collezione di design contemporaneo"
    link : "https://www.ansa.it/ansacom/notizie/cultura/maxxi-2/2023/12/13/al-maxx..."
    date : "14 dicembre 2023, 15:50"
    content : "Dalla grande struttura per piscina, in scala 1:1, al Sanatorio progett..."
    tags : null
    source_site : "Ansa"
    id : 44
  ▶ 1: Object
▶ similarity : Array (2)
  question : "Considerati più titoli e contenuti degli articoli, riassumili e integr..."
  answer : "Il Maxxi presenta quattro mostre, tra cui la prima omaggio ufficiale a..."
  toxicity : 0.0003700370143633336
  severe_toxicity : 0.006508501712232828
  obscene : 0.043308693915605545
  identity_attack : 0.008164607919752598
  insult : 0.03998107090592384
  threat : 0.002338578924536705
  sexual_explicit : 0.0013928256230428815
```

Example MongoDB Query

```
# Get questions with all similarity scores above a threshold
similarity_threshold = 0.7
query = {"similarity": {"$elemMatch": {"$gt": similarity_threshold}}}
cursor = collection.find(query)
print(f"Number of rows with all similarity scores above {similarity_threshold}: {len(list(cursor))}.")
# output: Number of rows with all similarity scores above 0.7: 162.
```

LLM fine-tuning



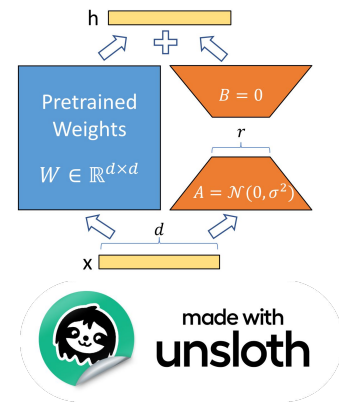
Llama 3 8B

Open model from **Meta**, pretrained and instruction-tuned generative text model.



Gemma 2B

Lightweight open model from **Google**, designed for natural language processing tasks.

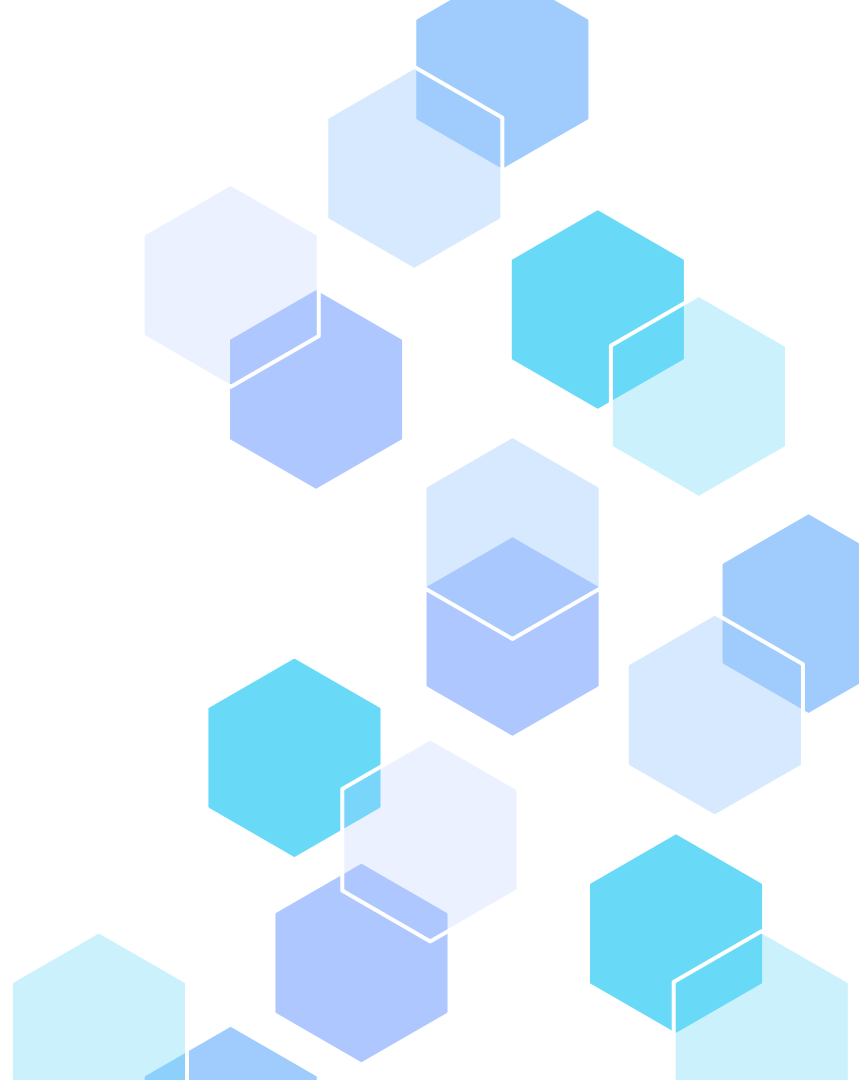


PEFT QLoRA

Quantized Low-Rank Adaptation efficiently compress the model and update only the adapters

06

Conclusions



Research Question 1

Starting from information publicly available online, is it possible to create a synthetic dataset for LLM fine tuning?

	question	answer	language	accuracy	acc_explanation
1188	Mia figlia allora adolescente e io andammo in ...	Il nuovo tavolo era situato nel corridoio che ...	it	5	\nExplanation:\n\nThe response accurately answer...
3531	Suggérez trois projets de décoration intérieur...	1. Vase à bouteille de vin :\nMatériel : Boute...	fr	5	\nExplanation:\n\nThe response accurately follow...
1392	È vero che Porto Rico fa parte degli Stati Uni...	Sì, in base al testo fornito, è vero che Porto...	it	5	\nExplanation:\n\nThe response accurately addres...
904	Provide a list of 8 words that have 'ther' som...	1. Together\n2. Thermometer\n3. Thermostat\n4....	en	5	\nExplanation:\n\nThe AI assistant's response is...
601	An automotive manufacturer is planning to laun...	To determine the optimal number and location o...	en	5	\nExplanation:\n\nThe response accurately addres...
...
2402	Write a Python script to monitor an Azure Stor...	To monitor an Azure Storage account for new fi...	es	5	\n\nThe response is highly accurate and comprehe...

Research Question 2

Given the most used datasets for fine tuning, processing them, are they sufficient to make an open source <10b model multi-lingual?

Question	Lang	Answer (Gemma 2b, LLaMa 3-8b)
<p>Scrivi una frase che parli di un ristorante con tutte le seguenti proprietà: nome = Strada, eatType = ristorante, cibo = cinese, fascia di prezzo = economico, valutazione del cliente = 5 su 5, adatto alle famiglie = sì, vicino = Rainbow Vegetarian Café</p> <p>Translation: "Write a sentence about a restaurant with all of the following properties: name = Street, eatType = restaurant, food = Chinese, price range = cheap, customer rating = 5 out of 5, family-friendly = yes, nearby = Rainbow Vegetarian Café"</p>	it	<p>Ristorante Strada: Un'esperienza culinaria cinese a prezzi accessibili. Il ristorante "Strada" offre un'autentica cucina cinese a prezzi convenienti, situato vicino al "Rainbow Vegetarian Café". Con un punteggio di 5 su 5 e un'atmosfera accogliente, è perfetto per le famiglie che cercano un pasto delizioso e conveniente</p> <p>Translation: "Strada Restaurant: An affordable Chinese dining experience. The "Strada" restaurant offers authentic Chinese cuisine at affordable prices, located next to the "Rainbow Vegetarian Café." With a score of 5 out of 5 and a cozy atmosphere, it is perfect for families looking for a delicious and affordable meal."</p> <p>Strada è un delizioso ristorante cinese economico dove il cibo è valutato 5 su 5 dai clienti e che è adatto alle famiglie, situato vicino al Rainbow Vegetarian Café.</p> <p>Translation: "Strada is a delicious inexpensive Chinese restaurant where the food is rated 5 out of 5 by customers and is family-friendly, located next to Rainbow Vegetarian Café."</p>

<p>Mi esposa y yo fuimos a ver Elizabethtown anoche (27/10/2005) y quedamos terriblemente decepcionados. De hecho, entramos esperando ver una buena película. NO esperábamos nada grandioso, pero imaginamos algunas escenas lindas y algunas líneas divertidas de Paula Dean (normalmente presentadora de Food Network). Es difícil subestimar exactamente cuán equivocados estábamos (de hecho, consideré tomar un descanso para ir al baño), ir a jugar uno o dos videojuegos y luego regresar.\n\nSeleccione entre los siguientes.\n* Al hablante le encanta Elizabethtown.\n*El orador recomienda Elizabethtown.\n*El orador detesta Elizabethtown.\n* Ninguna de las opciones anteriores.\nP: ¿Qué piensa el orador sobre Elizabethtown?</p> <p>Translation: "My wife and I went to see Elizabethtown last night (10/27/2005) and were terribly disappointed. In fact, we went in expecting to see a good movie. We were NOT expecting anything great, but imagined some cute scenes and some funny lines from Paula Dean (normally a Food Network host). It's hard to understate exactly how wrong we were (in fact, I considered taking a bathroom break),</p>	es	<p>El orador está profundamente decepcionado con Elizabethtown. A pesar de entrar con expectativas moderadas, la experiencia del orador con la película fue significativamente peor de lo que imaginaban. La película no cumplió con sus expectativas ni ofreció las escenas o líneas divertidas que esperaban.</p> <p>Translation: "The speaker is deeply disappointed with Elizabethtown. Despite going in with moderate expectations, the speaker's experience with the film was significantly worse than they imagined. The film did not live up to their expectations nor did it deliver the funny scenes or lines they expected."</p> <p>El orador detesta Elizabethtown.</p> <p>Translation: "The speaker detests Elizabethtown."</p>
--	----	--

Research Question 3

Is it possible to carry out fine-tuning or distill knowledge so as to teach a small LLM to imitate a larger and more expensive model through fine tuning?

Answer	Response Summary
<p>Original Question:</p> <p>- Considerati più titoli e contenuti degli articoli, riassumili e integrali in un unico testo. \n Input: Title: Vertici Ue, l'euro un successo ma ora per noi nuove sfide \n Content: BRUXELLES, 30 DIC - "Nel corso degli anni ci sono state sfide enormi, compresi dubbi sul futuro dell'euro stesso. Ma ogni volta abbiamo trovato le risposte giuste, come il sistema armonizzato di vigilanza e risoluzione delle crisi bancarie o il Meccanismo europeo di stabilità. Oggi, il sostegno alla moneta unica da parte dei cittadini dell'area dell'euro è vicino a livelli record ...</p> <p>Total word: 879</p>	<p>Output Summary fine-tuned Gemma 2B:</p> <p>I vertici delle istituzioni Ue hanno scritto un messaggio congiunto agli europei, in cui descrivono le sfide che i Paesi Ue devono affrontare insieme. Questi includono tensioni geopolitiche crescenti, come la guerra della Russia contro l'Ucraina, che richiede decisioni collettive coraggiose, e un'accelerazione della crisi climatica che possiamo risolvere solo insieme. Inoltre, i vertici hanno sottolineato che le questioni come la difesa, la transizione verde e la transizione digitale sono diventate urgenti e di interesse comune. Per ...</p> <p>Total word: 201. Mistral total word (reference model): 249</p>

Thanks!

Do you have any questions?

