

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

Alex Calabrese and Antonio Sabbatella



## 1 Executive Summary

Fine-tuning Large Language Models (LLMs) is essential to tailor their performance for specialized tasks. While pre-trained LLMs like GPT-4 have impressive capabilities, fine-tuning small LLMs on domain-specific datasets ensures effectiveness and efficiency in practical applications.

This project aimed to investigate the **feasibility of creating synthetic datasets for multilingual language model fine-tuning** and assessing the quality of existing datasets for this purpose. Additionally, it explored the potential of knowledge distillation techniques to enable smaller language models to imitate larger, more expensive models through fine-tuning.

The key findings included the successful creation of a multilingual and newspaper news datasets, as well as the development of data acquisition, cleaning, integration, and enrichment strategies. Comprehensive data quality evaluations were conducted, encompassing similarity analysis, sources analysis, and toxicity analysis. The results highlighted the effectiveness of the suggested approach.

## 2 Research questions

In this project, we address three key research questions related to fine-tuning Large Language Models (LLMs):

1. Starting from information publicly available online, is it possible to create a synthetic dataset for LLM fine-tuning?
2. Given the most used datasets for fine-tuning, and processing them, are they sufficient to make an open source <10b model multi-lingual?
3. Is it possible to carry out finetuning or distil knowledge so as to teach a small LLM to imitate a larger and more expensive model through fine-tuning?

### 3 Data Acquisition

In the following sections, we will illustrate the creation of **two separate datasets** during this project: one for **multilingual fine-tuning** and another for **current affairs based on newspaper news**.

#### 3.1 Data Sources

##### Multilingual fine tuning dataset (API)

**Source 1.** The **OpenOrca** dataset is an augmented FLAN (Few-shot Language Adaptation Network) dataset, primarily in English, designed for NLP tasks like language modelling and text generation. It is structured in a tabular format with data instances, fields, and splits.



**Source 2.** The **OpenHermes** is a dataset for natural language processing tasks. It is a collection of text data that has been curated and annotated for use in training machine learning models. The dataset has been created on GPT-4 generated data across the AI landscape and emphasizes an open-source approach with datasets from multiple AI industry contributors.



**Newspaper news dataset (Scraping)**

**Source 3.** Open is an Italian online newspaper founded by Enrico Mentana. It covers a wide range of topics including politics, economy, fact-checking, technology, environment, and current affairs, with multimedia content like interviews, photos, videos, and interactive graphics.



**Source 4.** Il Post is an independent Italian online newspaper that offers articles, blogs, and multimedia content on various subjects such as politics, international news, culture, and science. It was launched in April 2010 and is known for its quality journalism and insightful analysis.



**Source 5.** Agenzia Nazionale Stampa Associata (ANSA) is Italy's leading news agency and one of the top-ranking news agencies in the world. Founded in 1945, it operates as a cooperative of 23 publishers and covers news, images, and in-depth reports on significant Italian and global events across all transmission platforms.



### 3.2 Data acquisition strategies

#### Multilingual fine-tuning dataset

We created the `fetch_rows` python function to fetch data from Hugging Face APIs. The function operates by sending a GET request to the constructed API URL, which includes parameters for the offset and the number of rows to retrieve. Upon receiving a successful response, the JSON data is converted and appended to the accumulating dataset. This process continues iteratively until the desired number of rows is fetched or an error is encountered.

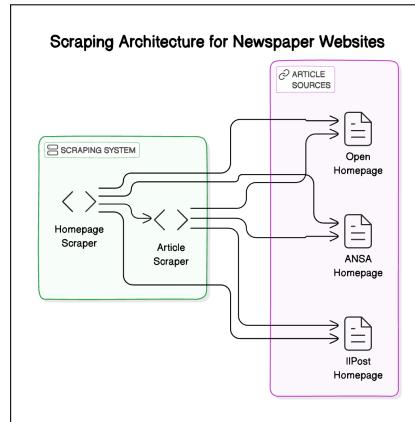
A total of 10000 rows were fetched, 5000 from OpenOrca and 5000 from Open-Hermes.

#### Newspaper news dataset

The tool employed for web data acquisition through scraping is the Python library **BeautifulSoup**. Beautiful Soup is renowned for its ability to facilitate the extraction of data from HTML and XML documents. It simplifies the process by parsing the documents and providing an intuitive interface for navigating and searching the parse tree.

The web scraping was conducted in two main stages:

1. **Homepage Scraping:** The initial phase involved scraping the homepage to collect news article titles and links. This was achieved by pinpointing the HTML elements that encapsulated the necessary data.
2. **Article Scraping:** The subsequent stage entailed scraping the content of individual articles. An Article class was crafted to systematically store the scraped data, encompassing attributes such as title, link, domain, date, subtitle, content, and tags.

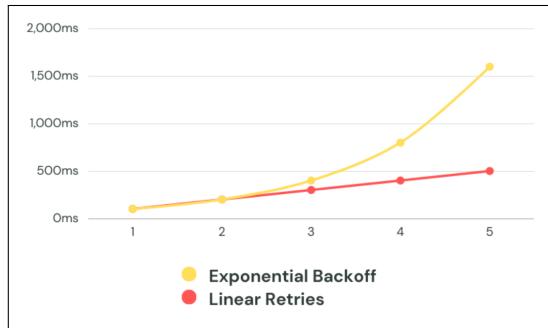


To retrieve the content of website homepages from various dates, including those in the past, the scraping process is conducted on archived versions of these pages, which are preserved and made available by the Internet Archive's Wayback Machine.

### Challenges

The main challenge encountered during the scraping of newspaper news sites was dealing with **rate limits** and the **potential for IP bans** from the sources.

To address this issue, we implemented a retry system that follows an **exponential backoff strategy**. This approach involved configuring a session with the *requests* library, incorporating a *Retry* object to handle connection attempts. Specifically, we set the *Retry* object to allow up to 10 retries and included a backoff factor of 0.5, which exponentially increases the wait time between retries.



**Fig. 1.** Comparison of Exponential Backoff and Linear Retry Strategies in Web Scraping Operations.

This method **effectively mitigated the risk of IP bans** and ensured more reliable and robust scraping operations.

### 3.3 Cleaning

#### Multilingual fine tuning dataset

The Openhermes dataset was already in the form of question and answer, only the conversion from json to tabular format was needed.

In the actual implementation of the project, we utilized **SlimOrca** instead of OpenOrca. **SlimOrca** is a newly curated **subset of the OpenOrca data**, designed to enhance efficiency without compromising performance. This release

includes approximately 500k GPT-4 completions, which significantly reduces the dataset size while maintaining high-quality training outputs.

In the case of SlimOrca, unfortunately, the data was not in the correct format, e.g:

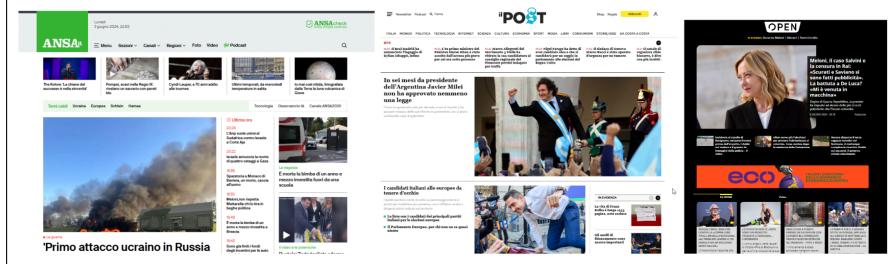
```

1 [ { "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..." }
```

So we wrote a preprocessing function specifically called ‘*df\_slimOrca\_prepoc*’ to extract the question and answer values and convert them into a pandas data frame.

### Newspaper news dataset

Generally, **BeautifulSoup extraction and small data cleaning were sufficient for most web scraping tasks**. However, in the specific case of ANSA, the process was more complex due to the need to handle inconsistent URL formats and exclude unwanted content such as video galleries.



**Fig. 2.** Source Websites for Scraped Article Links.

This required additional custom methods to clean and standardize the data. The method ‘*remove\_before\_http*’ was used to standardize URLs, and ‘*remove\_before\_ansa\_link*’ was implemented to refine ANSA-specific URLs. The ‘*detect\_videogallery*’ method filtered out video gallery URLs, ensuring only relevant articles were included. After data collection, the ‘*make\_df\_compatible*’ method was applied to adjust the DataFrame for consistency, including adding a source site column via the ‘*extract\_source\_site*’ function.

Every source class implemented the ‘*make\_df\_compatible*’ method to ensure that data from different sources could be seamlessly concatenated for further analysis.

### 3.4 Integration

#### Multilingual fine-tuning dataset

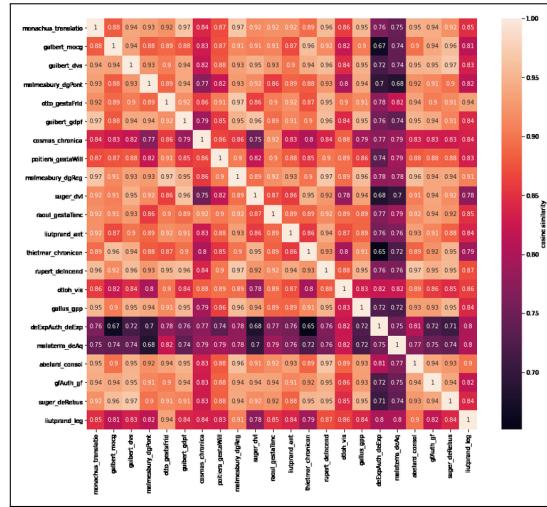
In the integration of the Multilingual fine-tuning dataset, we preserved the columns that represent the intersection of the primary datasets. Specifically, our attention was concentrated on the ‘**question**’ and ‘**answer**’ columns.

#### Newspaper news dataset

The news dataset is challenging to integrate due to its daily aggregation of articles. To address this, we used a method to evaluate the **semantic similarity of headlines**, allowing for effective linking of data.

A **similarity matrix** was constructed to compare the headlines, and based on the derived scores and a predefined similarity threshold, clusters of news sharing semantic similarity were automatically separated.

This methodology allows for the identification of news articles that cover the same topics and are prominently featured on multiple newspapers’ front pages, indicating the most significant events of the day.

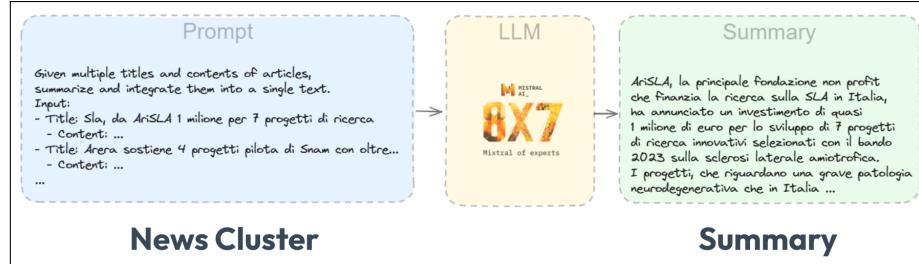


**Fig. 3.** Illustrative example

The model used for the text embedding is **Universal Sentence Encoder Multilingual** (USE), a transformer-based model that encodes text into high-dimensional vectors.

The actual **unification of news articles within the same clusters** is executed through **LLM queries to Mistral 8x7b**, incorporating the titles and bodies of the news articles into the query context.

To quantitatively discover the influence of specific news pieces and newspapers on the post-integration response, an **additional similarity assessment** is conducted between the pairs (source, integrated news).



### Challenges

To address the **limited context window of 32,768 tokens** in **Mistral 8x7b**, our solution employs a **recursive splitting mechanism**.

The '*send\_chat\_message*' function first checks if the content exceeds the token limit. If it does, the content is split into chunks. The initial chunk is preserved as the prompt. The function then divides the content chunks into two halves. Each half is sent recursively to the '*send\_chat\_message*' function to generate summaries for each part.

These summaries are then concatenated and sent again to the '*send\_chat\_message*' function to produce the final output. This **recursive splitting ensures that each request stays within the token limit while effectively summarizing large content**. Additionally, error handling and rate-limiting mechanisms are incorporated to manage HTTP errors and rate limits.

### 3.5 Data Enrichment

#### Multilingual fine-tuning dataset

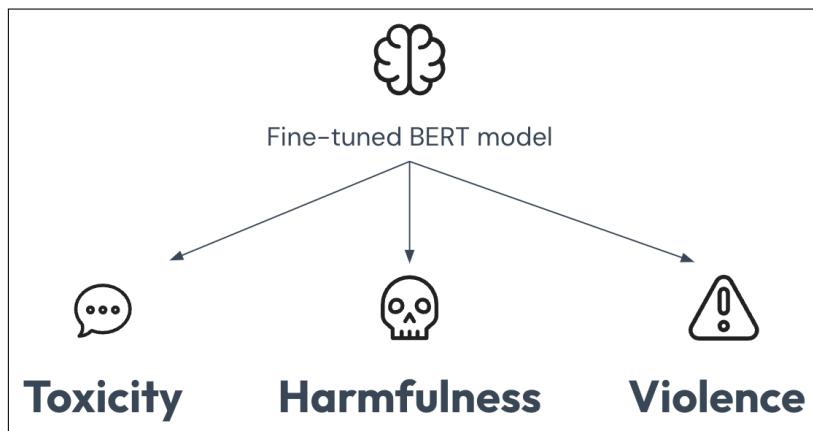
The translation system for preparing datasets for LLM finetuning translates English entries into Italian, Spanish, and French using the '*deep\_translator*' library.

The system cycles through the target languages in sequence, translating both questions and answers. After translation, it updates the Dataset with the translated content.



#### Newspaper news dataset

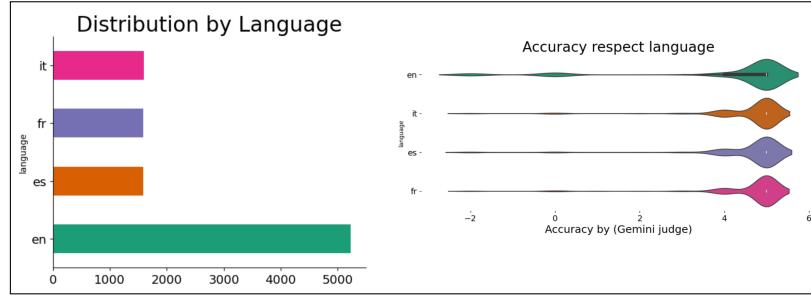
We employed a pre-trained BERT Language Model fine-tuned on a Toxic Comment Classification dataset with transformer architecture. This model allowed us to classify toxic comments and evaluate parameters related to toxicity, harmfulness, and the presence of violent content.



## 4 Data Quality

### 4.1 Multilingual fine-tuning dataset

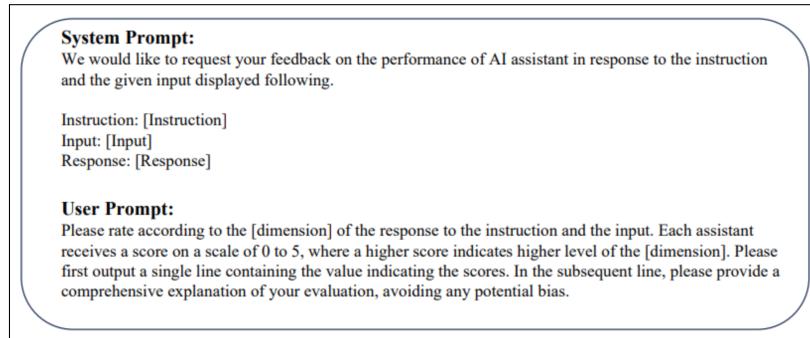
In our analysis of the multilingual fine-tuning dataset, we ensured a balanced distribution by maintaining **1,500 rows for each non-English language** (Italian, French, Spanish) and **5,000 rows for English**.



The evaluation of the quality of question-answer pairs in the large-scale language model fine-tuning dataset presents a significant challenge due to the substantial volume of data involved. Manual evaluation is infeasible; therefore, we rely on methodologies proposed in the relevant literature.

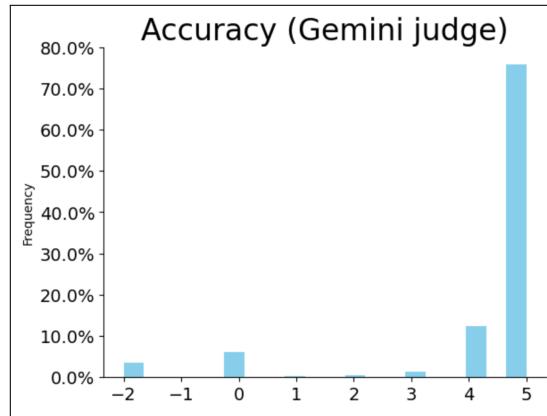
Notably, the "**ALPAGASUS: TRAINING A BETTER ALPACA WITH FEWER DATA**" by University of Maryland, Samsung Research America, University of Southern California [1] and the "**Self-Rewarding Language Models**" paper [2] by Meta and NYU propose an evaluation approach based on querying large language models (LLMs) using specifically crafted prompts. The first approach have been adopted in our solution to ensure the quality of the dataset.

The prompt template from ALPAGASUS is illustrated below, where the LLM is tasked with rating and filtering the training data based on specific criteria.



**Fig. 4.** Example of the template used to obtain, given a dataset entry, the accuracy score of the question-answer pair.

To evaluate the data quality, we analyzed **10,000 rows**, discovering that **2,414 rows had an accuracy score of less than 5**, which corresponds to **24.14%** of the dataset. Using Gemini, we found that **76% of the question-answer pairs received a perfect score of 5 out of 5**.



**Fig. 5.** Accuracy of scores using Gemini as judge.

However, entries with an accuracy score of -2 highlighted issues related to processing errors or formatting problems during the output parsing by the language model. This suggests that while the majority of the dataset is accurate, there is a significant portion that requires further attention to improve overall data quality.

Example of 0 scores:

question	answer	lang	acc	acc_explanation
What age-related phrase is represented by the anagram "over fifty"?	forty five	en	0	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.
Alsace is one of the most conservative régions of France. It is one of just two régions in metropolitan France where the conservative right won the 2004 région elections and thus controls the Alsace Regional Council. Conservative leader Nicolas Sarkozy got his best score in Alsace (over 65%) in the second round of the French presidential elections of 2007. The president of the Regional Council is Philippe Richert, a member of the Union for a Popular Movement, elected in the 2010 regional election. The frequently changing status of the région throughout history has left its mark on modern day politics in terms of a particular interest in national identity issues. Alsace is also one of the most pro-EU regions of France. It was one of the few French regions that voted 'yes' to the European Constitution in 2005. Who was elected head of the Alsace Regional Council in 2007? (If the question is unanswerable, say "unanswerable")	The question is a bit confusing, but I can help you! In 2010, Philippe Richert was elected as the president of the Alsace Regional Council. He was a member of the Union for a Popular Movement.	en	0	Explanation: The response is inaccurate and does not answer the question. The question asks who was elected head of the Alsace Regional Council in 2007, but the response talks about the election in 2010 and mentions Philippe Richert as the elected president. The response is completely irrelevant to the question and provides incorrect information.

**Fig. 6.** Table with examples of scores identified as 0.

Example of common rows with accuracy error due to translation of language syntactic tasks (correct in English but not in Italian):

question	answer	lang	acc	acc_explanation
Palindromo che significa prendere in giro o prendere in giro.	bavaglio # EN: gag	it	0	Explanation: The response "bavaglio" is not a palindromo, which means a word, phrase, or sequence that reads the same backward as forward. Therefore, the response is not accurate to the instruction and the input. The score is 0, indicating the lowest level of accuracy.
Sviluppa un elenco di 7 parole che terminano con "nave"	1. Amicizia 2. Relazione 3. Direzione 4. Borsa di studio 5. Partenariato 6. Cittadinanza 7. Apprendistato	it	0	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.

**Fig. 7.** Table with examples of Q&A pairs with scores identified as zero due to syntax errors caused by translation

#### 4.2 Newspaper news dataset

Evaluating the quality of journalistic content is a delicate task, as it is often influenced by the inherent biases of journalists and the editorial slant of their publishing outlets. To address this challenge, our approach involves **aggregating news on identical topics from diverse sources** to construct a more **balanced narrative**, thereby diminishing the impact of individual biases.

A multi-stage filtering process is employed:

1. **Intrinsic Filter:** We prioritize content that is featured on the homepage, under the assumption that it represents the most significant news curated by the editorial team.
2. **Relevance Filter:** We further refine the selection by including only those news items that are deemed relevant across multiple publications, ensuring a broader consensus on their importance.

By adopting these methodologies, we aim to ensure the quality and reliability of both the language model fine-tuning dataset and the news dataset, which are crucial components of our research.

#### Evaluation Methodology and Metrics

To comprehensively assess the aggregated news data from ANSA, IlPost, and Open sources, a robust evaluation framework was employed, encompassing various metrics. This section delineates the evaluation structure and key findings derived from the provided dataset.

#### Data Description

The data includes various columns such as title, link, domain, date, subtitle, content, tags, and source\_site. The quantity of articles featured on the homepages of each news website varies significantly. This discrepancy is due to the differing editorial choices of each source. Upon examination, it is evident that **ANSA consistently hosts a larger volume of articles compared to the other sources under consideration**. For the initial three days, the dataset exhibits varying article counts across the news sources:

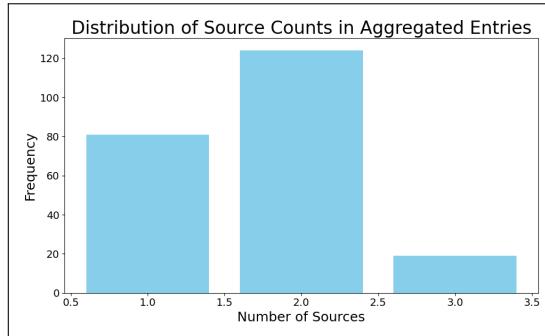
Source Site	Number of Articles
ANSA	652
Il Post	59
Open	147

**Fig. 8.** Table shows the cardinality of items scraped from the three sites during a 3-day period.

The total number of articles across all sources amounts to 858. After aggregating the articles using semantic similarity-based clustering, the total count was reduced to 224, yielding an aggregation score of **26.11%**. This score represents the percentage reduction between the scraped and aggregated articles, computed as:

$$\text{Aggregation Score} = \left( \frac{\text{Number of Aggregated Articles}}{\text{Total Number of Articles}} \right) \times 100$$

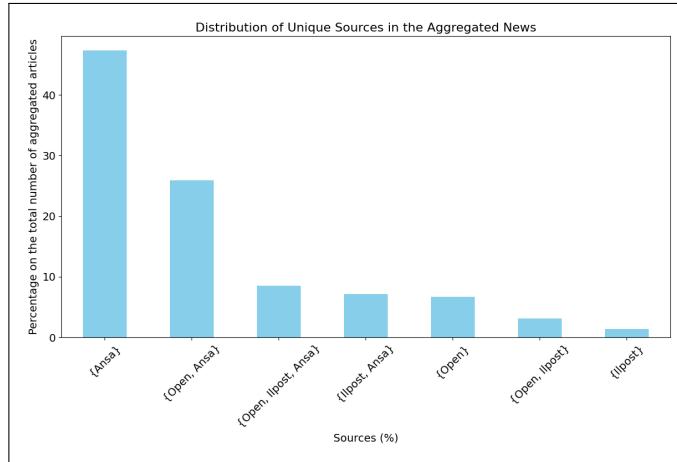
The majority of the clusters integrated news from multiple sources, as visualized by the distribution of source counts in the aggregated entries.



**Fig. 9.** Total number of different websites per cluster.

The bar chart below visualizes the distribution of unique sources in the aggregated news:

To gain deeper insights into the similarity scores, we computed two key metrics for each row: the **mean similarity** and the **similarity difference**.

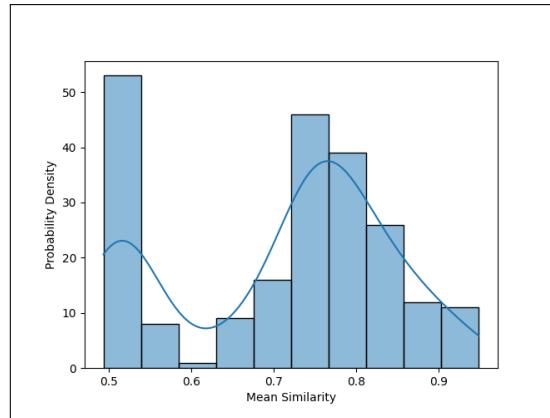


**Fig. 10.** Distribution of the number of sites by cluster, indicates the origin of the articles aggregated in the clusters.

#### 4.3 Similarity Analysis

##### Mean Similarity

This was calculated by taking the average of the cleaned similarity scores within each row and the generated summary.



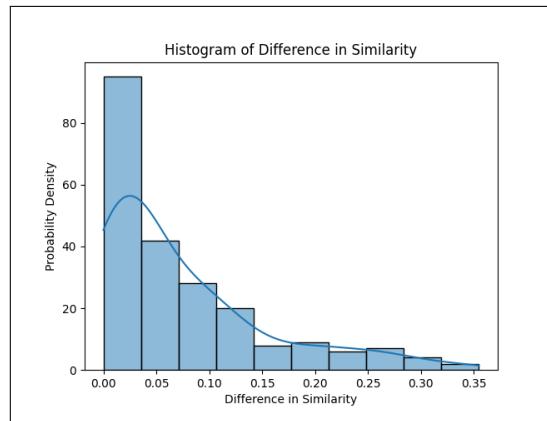
**Fig. 11.** Average similarity among items in the same cluster and the aggregated news.

The plot reveals two prominent peaks of average similarity, one at the lower end (0.5) and the other forming a bell-shaped curve around 0.75. This observation

could serve as an additional criterion for **further filtering to enhance the quality of the fine-tuning dataset.**

### Similarity Difference

This metric signifies the range of similarity scores, computed as the difference between the maximum and minimum values within each row's similarity array (the similarity is calculated between each news text and the cluster summary). It underscores the variability in similarity scores, indicating the extent to which individual sources influence the summary and can also be utilized to present the user with an ordered list of sources based on their similarity to the summary.



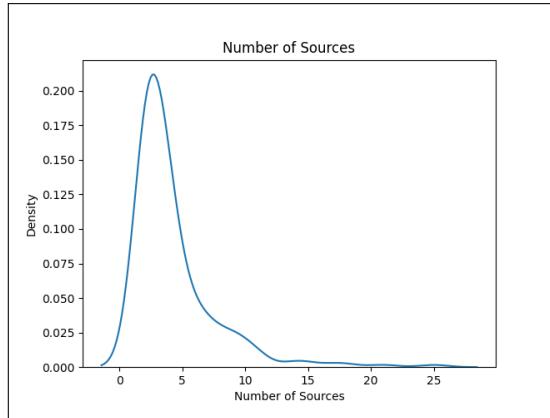
**Fig. 12.** Mean difference in mean similarities between aggregate news and items in the cluster from which it originates.

As illustrated in the graph, the difference is generally minimal, suggesting that all news articles contribute to the creation of the output summary.

#### 4.4 Sources Analysis

##### Number of Sources

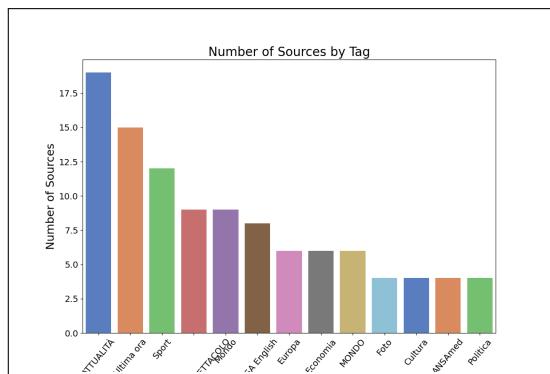
This plot shows how the number of sources varied across the dataset and helped identify the most frequent source counts.



**Fig. 13.** Distribution of the number of news items for each cluster.

##### Number of Sources by Tag

This bar plot visualizes the count of sources associated with these popular tags (appear more than three times), providing insights into which tags were most prevalent and their associated source counts.

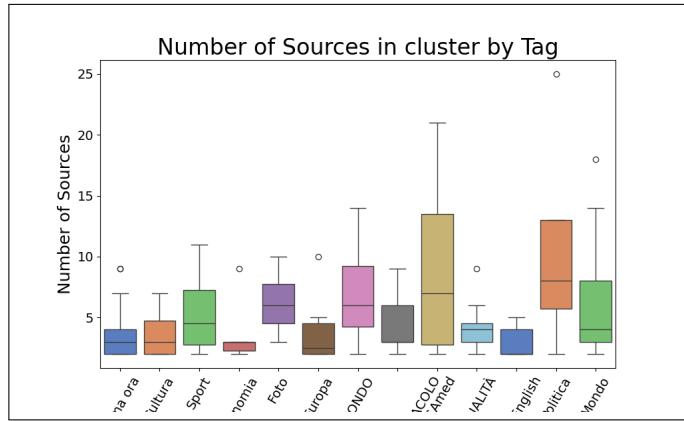


**Fig. 14.** Distribution of the number of news items per regrouped by tag.

The graph reveals that the majority of the news is tagged as ‘*Attualità*’ (actuality), followed by ‘*ultima ora*’, ‘*sport*’, and ‘*Culture and Entertainment*’. As a further improvement, semantic similarity could be used to merge similar tags such as, for example, ‘*Culture and Entertainment*’ and ‘*Culture*’. Different tags that had equal lower and upper cases (exact match) were merged, an example was “*SPORT*” from ansa.it and the tag from the other sources: ‘*sport*’.

#### Number of Sources news in each cluster by Tag

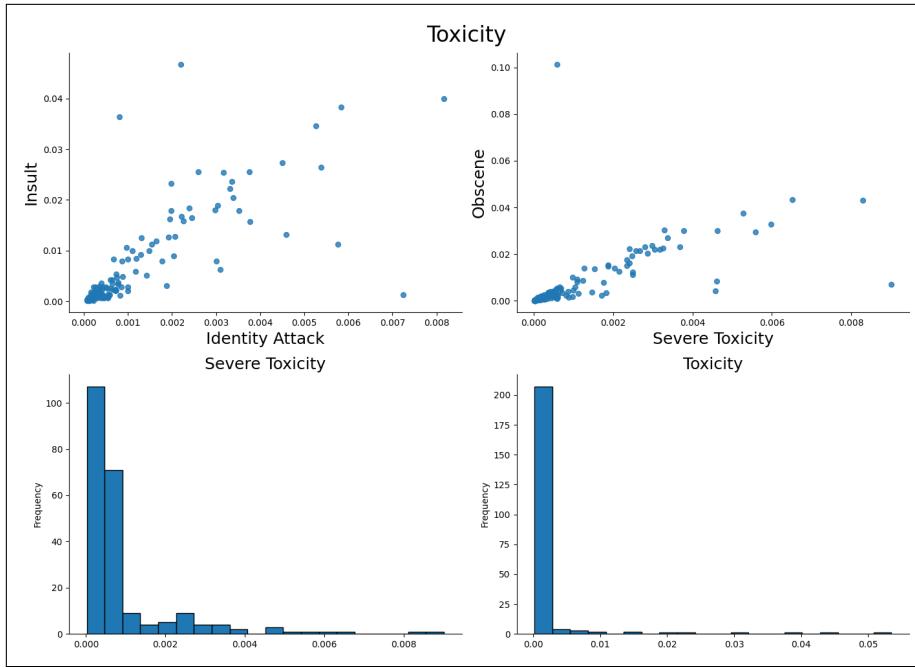
Complementing the bar plot, the boxplot offers a summary of the distribution of the number of sources for each popular tag. It highlights the median, quartiles, and any potential outliers, giving a detailed view of the data spread.



**Fig. 15.** Boxplot of the number of news items in the clusters aggregated by tag.

The boxplot graph indicates that the tags “*Mondo*”, “*sport*”, and “*politica*” exhibit the highest variability in the number of news articles in each cluster. The outlier in this case is ANSAmed, a tag related to the Ansa Mediterranean agency, which could potentially be disregarded or removed from the tag list.

## Toxicity analysis



**Fig. 16.** Grid representing at the top row the scatterplots of the correlation between toxicity metrics and in the second row the values of severe toxicity and toxicity.

Histograms indicated a concentration of values towards lower toxicity levels, suggesting the overall dataset exhibits relatively **low toxicity**. The presence of any offensive content necessitates continued monitoring and mitigation efforts, particularly for sensitive fields like digital communication, following these results, there is no apparent toxic content in the generated news.

## 5 Storage

In our project, **MongoDB** serves as the primary database for storing news articles and the multilingual finetuning dataset. Using the **Python pymongo library**, we have implemented a *DatabaseHandler* class to facilitate efficient database operations.

This class connects to the MongoDB server using a connection string, manages collections and performs various data handling tasks. Key functionalities include deleting all documents within a collection and cleaning data for insertion.

The use of MongoDB's flexible schema design is particularly advantageous for handling the varied and complex data structures involved in our application.

Thanks to MongoDB's flexible NoSQL structure, we could store news within clusters as **nested sub-objects**, as a news story only exists within its corresponding cluster in our project.

After storing the multilingual dataset and final news dataset in MongoDB, various queries can be performed to retrieve specific data subsets. For instance, you can use the following queries:

### Multilingual fine-tuning dataset

```
● ● ● Query
# Get all the rows with it language
query = {"language": "it"}
cursor = collection.find(query)
print(f"Number of rows with {query['language']} language: {len(list(cursor))}.")
# output: Number of rows with it language: 1595.
```

```
● ● ● Query
# Get all the rows with 4 accuracy
query = {"accuracy": 4}
cursor = collection.find(query)
print(f"Number of rows with {query['accuracy']} accuracy: {len(list(cursor))}.")
# output: Number of rows with 4 accuracy: 1250.
```

## 5.1 Newspaper news dataset

```

● ● ● Query

# Get all questions with a specific tag
tag_to_find = "CULTURA & SPETTACOLO"
query = {"tags": tag_to_find}
cursor = collection.find(query)
print(f"Number of rows with the tag \'{tag_to_find}\': {len(list(cursor))}.")
# output: Number of rows with the tag "CULTURA & SPETTACOLO": 9.

# 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.

# Get the question with the highest toxicity
result = collection.aggregate([
    {"$group": {"_id": None, "max_toxicity": {"$max": "$toxicity"}}}
])
print(f"Max toxicity: {list(result)[0]['max_toxicity']}.")

# Get the average toxicity
result = collection.aggregate([
    {"$group": {"_id": None, "avg_toxicity": {"$avg": "$toxicity"}}}
])
print(f"Avg toxicity: {list(result)[0]['avg_toxicity']}.")

# output: Max toxicity: 0.05349026247859001.
# output: Avg toxicity: 0.001879764144112934.

```

## 6 Conclusions

### 6.1 Research questions

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

Yes, it is possible to create a synthetic dataset for LLM fine-tuning by leveraging publicly available online resources, such as web scraping, and text generation models, to generate high-quality, multilingual data that can be used to fine-tune LLMs.

Publicly available datasets, while abundant in English, present a challenge for fine-tuning multilingual models due to the **scarcity of multilingual resources**. **Literal translation approaches introduce errors**, particularly in question-answering tasks that depend on syntax and context. **The integration of LLMs within the model architecture is crucial to mitigate these errors** and ensure accurate performance across diverse languages.

Example rows:

	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:\nThe response accurately answer...
3531	Suggérez trois projets de décoration intérieur...	1. Vase à bouteille de vin.\nMatiériel : Boute...	fr	5	\nExplanation:\nThe response accurately follow...
1392	È vero che Porto Rico fa parte degli Stati Uni...	Si, in base al testo fornito, è vero che Porto...	it	5	\nExplanation:\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:\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:\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	\nExplanation:\n\nThe response is highly accurate and compreh...

**Fig. 17.** Example rows from the final fine-tuning dataset (filtered by score = 5)

It is important to consider that the quality of the synthetic dataset generated and used for tuning is directly dependent on several factors associated with the chosen model:

- **Model quality:** A more robust and accurate LLM will inherently generate higher-quality synthetic data.
- **Context length:** The length of the context considered by the model during training can have a significant impact on the richness and complexity of the synthetic data.
- **Cost:** The use of larger LLMs generally results in higher expense.

The models used to generate synthetic data were chosen primarily for their speed, medium cost, and the possibility of high context (32k for Mistral), which is critical for the aggregation task that tends to use large news texts as input.

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

From the empirical analyses conducted by analyzing the outputs of the models in the different languages (also visible in the table below) the fine-tuning with the dataset obtained for only one epoch (using the Qlora algorithm) is **sufficient to allow Gemma and Llama to answer the questions in all languages** (including Italian (it), Spanish (es) and French (fr)).

Question	Lang	Answer (Gemma 2b, LLaMa 3-8b)
<p>Scriv 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><i>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"</i></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><i>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."</i></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><i>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."</i></p>

**Fig. 18.** Gemma & Llama answer in IT

Mi esposo 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. nSelecteone 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. n P: ¿Qué piensa el orador sobre Elizabethtown?  <i>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 underestimate exactly how wrong we were (in fact, I considered taking a bathroom break, going to play a video game or two, and then coming back). Select from the following. The speaker loves Elizabethtown. The speaker recommends Elizabethtown. The speaker dislikes Elizabethtown. None of the above."</i>	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><i>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."</i></p> <p>El orador detesta Elizabethtown.</p> <p><i>Translation: "The speaker detests Elizabethtown."</i></p>
--	----	--

**Fig. 19.** Gemma & Llama answer in ES

<p>Dans une gare, quatre amis déposent leurs bagages dans des casiers séparés avant de partir faire du tourisme. Adam met sa valise dans le casier A, Beth met la sienne dans le casier B, Carl met la sienne dans le casier C et Dana met la sienne dans le casier D. Plus tard, un agent de sécurité déplace les valises en raison d'un problème de maintenance, plaçant la valise d'Adam dans le casier D, la valise de Beth dans le casier A, la valise de Carl dans le casier B et la valise de Dana dans le casier C. Lorsque les amis reviendront récupérer leurs bagages, où chacun cherchera-t-il initialement sa valise ?</p> <p><i>Translation: "At a train station, four friends deposit their luggage in separate lockers before setting off on a sightseeing trip. Adam puts his suitcase in locker A, Beth puts hers in locker B, Carl puts his in locker C and Dana puts hers in locker D. Later, a security guard moves the suitcases due to a maintenance problem, placing Adam's suitcase in locker D, Beth's suitcase in locker A, Carl's suitcase in locker B and Dana's suitcase in locker C. When the friends return to collect their luggage, where will everyone initially look for their suitcases?"</i></p>	<p>fr</p>	<p>Chacun cherchera sa valise dans le casier où il l'a initialement placée, ignorant la permutation effectuée par l'agent de sécurité. Adam cherchera sa valise dans le casier A, Beth cherchera sa valise dans le casier B, Carl cherchera sa valise dans le casier C et Dana cherchera sa valise dans le casier D.</p> <p><i>Translation: "Each person will look for their suitcase in the locker where they originally placed it, ignoring the permutation made by the security guard. Adam will look for his suitcase in locker A, Beth will look for her suitcase in locker B, Carl will look for his suitcase in locker C and Dana will look for her suitcase in locker D."</i></p> <p>Adam cherchera d'abord sa valise dans le casier D, puisqu'il l'a déposée dans le casier A. Beth cherchera d'abord sa valise dans le casier A, puisqu'elle l'a déposée dans le casier B. Carl cherchera d'abord sa valise dans le casier B, puisqu'il l'a déposée dans le casier C. Dana cherchera d'abord sa valise dans le casier C, puisqu'elle l'a déposée dans le casier D.</p> <p><i>Translation: "Adam will first look for his suitcase in locker D, since he left it in locker A. Beth will first look for her suitcase in locker A, since she deposited it in locker B. Carl will first look for his suitcase in locker B, since he deposited it in locker C. Dana will first look for her suitcase in locker C, since she deposited it in locker D."</i></p>
--	-----------	--

Fig. 20. Gemma &amp; Llama answer in FR

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

Yes, it is possible as shown with **Gemma 2B model that has demonstrated the potential to effectively capture and replicate the capabilities of larger, resource-intensive models through finetuning**. By finetuning Gemma 2B on data from a powerful model like Gemini and Mistral8x7b (GPT-3.5 level performance), the smaller model can be loaded with much of the knowledge and performance of its larger counterpart.

Example:

Original Question:

- Considerati più titoli e contenuti degli articoli, riassumili e integrali in un unico testo. Input: Title: Vertici Ue, l'euro un successo ma ora per noi nuove sfide 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. Ma il nostro lavoro non è finito. Perché ci troviamo di fronte a nuove sfide che i Paesi Ue non possono affrontare da soli e i cittadini guardano all'Europa per trovare risposte". Lo scrivono i vertici delle istituzioni Ue nel loro messaggio congiunto agli europei. "Siamo di fronte a tensioni geopolitiche crescenti, non ultima la guerra della Russia contro l'Ucraina, che richiede decisioni collettive coraggiose. Ci troviamo di fronte a una crescente crisi climatica che possiamo risolvere veramente solo insieme: le emissioni di carbonio non si fermano alle frontiere. E dobbiamo affrontare sfide senza precedenti per la nostra competitività a causa delle politiche energetiche e industriali di altre parti del mondo", scrivono ancora nella loro lettera comune Charles Michel (Presidente del Consiglio europeo), Paschal Donohoe (Presidente dell'Eurogruppo), Christine Lagarde (Presidente della Banca centrale europea), Roberta Metsola (Presidente del Parlamento europeo) e Ursula von der Leyen (Presidente della Commissione europea). "Ciò significa che questioni come la difesa, così come la transizione verde e quella digitale, sono diventate urgenti e di interesse comune. Stesso vale per il modo in cui affrontiamo il finanziamento dei massicci investimenti necessari per decarbonizzare le nostre economie, rendere più sicure le nostre catene di approvvigionamento e aggiornare le nostre tecnologie: nell'Ue la sola transizione verde richiederà investimenti per 620 miliardi di euro all'anno fino al 2030", si legge ancora. "Le soluzioni devono abbracciare la scala resa possibile dalla collaborazione in Europa. Costruire una autentica unione dei mercati dei capitali che abbracci il continente per mobilitare i finanziamenti privati. Utilizzare gli strumenti e le politiche europee per rafforzare la nostra competitività e la nostra sicurezza, ad esempio rafforzando le strutture esistenti attraverso regole fiscali rinnovate e una unione bancaria più solida. E portare la stessa moneta unica nell'area digitale, preparando le basi per un potenziale euro digitale che possa integrare il denaro contante", sottolineano i vertici dell'Ue. "I cittadini europei sanno che il mondo sta cambiando. E sanno che l'unione fa la forza. Circa due terzi degli europei sono convinti che l'Ue sia un baluardo di stabilità. Dimostriamo loro che l'Europa può dare forma a questo cambiamento e soddisfare le loro aspettative", concludono: Vertici Ue, l'euro un successo ma ora per noi nuove sfide 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. Ma il nostro lavoro non è finito. Perché ci troviamo di fronte a nuove sfide che i Paesi Ue non possono affrontare da soli e i cittadini guardano all'Europa per trovare risposte". Lo scrivono i vertici delle istituzioni Ue nel loro messaggio congiunto agli europei. "Siamo di fronte a tensioni geopolitiche crescenti, non ultima la guerra della Russia contro l'Ucraina, che richiede decisioni collettive coraggiose. Ci troviamo di fronte a un'accelerazione della crisi climatica che possiamo risolvere veramente solo insieme: le emissioni di carbonio non si fermano alle frontiere. E dobbiamo affrontare sfide senza precedenti per la nostra competitività a causa delle politiche energetiche e industriali di altre parti del mondo", scrivono ancora nella loro lettera comune Charles Michel (Presidente del Consiglio europeo), Paschal Donohoe (Presidente dell'Eurogruppo), Christine Lagarde (Presidente della Banca centrale europea), Roberta Metsola (Presidente del Parlamento europeo) e Ursula von der Leyen (Presidente della Commissione europea). "Ciò significa che questioni come la difesa, così come la transizione verde e quella digitale, sono diventate urgenti e di interesse comune. Stesso vale per il modo in cui affrontiamo il finanziamento dei massicci investimenti necessari per decarbonizzare le nostre economie, rendere più sicure le nostre catene di approvvigionamento e aggiornare le nostre tecnologie: nell'Ue la sola transizione verde richiederà investimenti per 620 miliardi di euro all'anno fino al 2030", si legge ancora. "Le soluzioni devono abbracciare la scala resa possibile dalla collaborazione in Europa. Costruire un'autentica unione dei mercati dei capitali che abbracci il continente per mobilitare i finanziamenti privati. Utilizzare gli strumenti e le politiche europee per rafforzare la nostra competitività e la nostra sicurezza, ad esempio rafforzando le strutture esistenti attraverso regole fiscali rinnovate e un'unione bancaria più solida. E portare la stessa moneta unica nell'era digitale, preparando le basi per un potenziale euro digitale che possa integrare il denaro contante", sottolineano i vertici dell'Ue. "I cittadini europei sanno che il mondo sta cambiando. E sanno che l'unione fa la forza. Circa due terzi degli europei sono convinti che l'Ue sia un baluardo di stabilità. Dimostriamo loro che l'Europa può dare forma a questo cambiamento e soddisfare le loro aspettative", concludono.

#### Output Target Summary in italiano (Mistral 8x7b):

"I vertici delle istituzioni Ue hanno rilasciato un messaggio congiunto agli europei, sottolineando il successo dell'euro ma anche le nuove sfide che i Paesi Ue non possono affrontare da soli. Le tensioni geopolitiche crescenti, come la guerra della Russia contro l'Ucraina, richiedono decisioni collettive coraggiose. Si stanno affrontando anche un'accelerazione della crisi climatica e sfide senza precedenti per la competitività a causa delle politiche energetiche e industriali di altre parti del mondo.

Le questioni come la difesa, la transizione verde e digitale sono diventate urgenti e di interesse comune. Lo stesso vale per il finanziamento dei massicci investimenti necessari per decarbonizzare le economie, rendere più sicure le catene di approvvigionamento e aggiornare le tecnologie. Nell'Ue, la sola transizione verde richiederà investimenti per 620 miliardi di euro all'anno entro il 2030.

Le soluzioni devono abbracciare la scala resa possibile dalla collaborazione in Europa. Costruire un'autentica unione dei mercati dei capitali che abbracci il continente per mobilitare i finanziamenti privati. Utilizzare gli strumenti e le politiche europee per rafforzare la competitività e la sicurezza, ad esempio rafforzando le strutture esistenti attraverso regole fiscali rinnovate e un'unione bancaria più solida. E portare la stessa moneta unica nell'era digitale, preparando le basi per un possibile euro digitale che possa integrare il denaro contante.

I cittadini europei sanno che il mondo sta cambiando e sanno che l'unione fa la forza. Circa due terzi degli europei sono convinti che l'Ue sia un baluardo di stabilità. Dimostriamo loro che l'Europa può plasmare questo cambiamento e soddisfare le loro aspettative."

#### Output Summary finetuned (Gemma 2b):

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 richiedono 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 affrontare queste sfide, i vertici hanno proposto soluzioni che abbracciano la scala resa possibile dalla collaborazione in Europa. Questi includono la costruzione di un'autentica unione dei mercati dei capitali che abbracci il continente per mobilitare i finanziamenti privati, l'utilizzo di strumenti e politiche europee per rafforzare la nostra competitività e la nostra sicurezza, e portare la stessa moneta unica nell'era digitale, preparando le basi per un potenziale euro digitale che possa integrare il denaro contante.

I cittadini europei sanno che il mondo sta cambiando e sanno che l'unione fa la forza. Circa due terzi degli europei sono convinti che l'Ue sia un baluardo di stabilità. I vertici dell'Ue hanno dimostrato loro che l'Europa può dare forma a questo cambiamento e soddisfare le loro aspettative.

## 6.2 Possible improvements

1. **Refinement of the fine-tuning dataset for news articles:** Incorporating a ranking system similar to the one employed in the multilingual dataset could enhance the quality and relevance of the data. This approach could involve assigning accuracy scores or ratings to the news articles, allowing for better filtering and prioritization of high-quality content.
2. **Experimenting with Different Prompts and LLM Models:** To further optimize the evaluation and quality assessment processes, future efforts could explore experimenting with different prompt templates and leveraging a broader range of large language models (LLMs). This would involve evaluating the performance and suitability of various prompt designs and LLM architectures for rating and filtering the fine-tuning datasets. Such an approach could lead to the identification of more effective prompts and models, ultimately enhancing the accuracy and reliability of the quality assessment methodology.

### 6.3 Future Developments

**Expanding Data Collection:** To further enrich the news dataset and improve its comprehensiveness, future developments could focus on expanding the data collection process. This can be achieved by incorporating additional reputable news sources and increasing the frequency of scraping, potentially implementing a daily cadence. By capturing a broader range of news articles from diverse sources on a regular basis, the dataset would better reflect the dynamic nature of current events and provide a more comprehensive representation of news coverage.