UniQA: an Italian and English Question-Answering Data Set Based on Educational Documents

Irene Siragusa[1,2,*], Roberto Pirrone[1]

[1]Department of Engineering, University of Palermo, Palermo, 90128, Sicily, Italy
[2]Department of Computer Science, IT University of Copenhagen, København S, 2300, Denmark

Abstract

In this paper we introduce UniQA, a high-quality Question-Answering data set that comprehends more than 1k documents and nearly 14k QA pairs. UniQA has been generated in a semi-automated manner using the data retrieved from the website of the University of Palermo, covering information about the bachelor and master degree courses for the academic year 2024/2025. Data are both in Italian and English, thus making the data set suitable for QA and translation models. To assess the data, we propose a Retrieval Augmented Generation model based on `Llama-3.1-instruct`. UniQA can be found at https://github.com/CHILab1/UniQA.

Keywords

Question Answering, RAG, Large Language Modell

1. Introduction

The even more increasing interest towards both the implementation and the usage of Large Language Models (LLM), involves not only the scientific community, but also users who are already acquainted with models such as Chat-GPT [2] and Gemini [3], that allow for chat-based interaction. Despite a general trust in those systems, it is clear that they are not so precise in answering to domain-specific questions, at least without the usage of external methodologies such as fine-tuning or integrating external knowledge via a Retrieval Augmented Generation (RAG) approach [4]. Moreover, the development and the evaluation of chat-based applications aimed at providing the users with precise answers in a specific domain is a quite difficult task. This is due to the lack of domain-specific annotated and high-quality data sets, such as Question-Answer (QA) pairs.

To overcome this issue, we built **UniQA**[1], a balanced Italian and English QA data set suitable for a domain-specific QA task where external knowledge is required. Our data set comprehends also a corpus of 1000 documents extracted through a scraping procedure over the website of the University of Palermo (UniPA), from which nearly 14k QA pairs have been generated in a semi-automatic manner. In addition, we evaluated UniQA by building a RAG-based QA architecture based on the `Llama-3.1` model [5] as text generator.

The paper is arranged as follows: related works are reported in Section 2, while details on the building process of UniQA are reported in Section 3, and the experimental results obtained using `Llama-3.1`, are reported and discussed in Section 4. Concluding remarks and future works are drawn in Section 6.

2. Related works

QA is a classical task in Natural Language Processing where a model is asked to answer to a question relying on a given context. Unfortunately, annotated QA data sets and specifically the Italian ones are not so common. A valuable example is SQuAD-it [6], derived by the English QA data set SQuAD [7], that collects more than 60k QA pairs obtained a via semi-automatic translation procedure. Generally

NL4AI 2024: Eighth Workshop on Natural Language for Artificial Intelligence, November 26-27th, 2024, Bolzano, Italy [1]
*Corresponding author.
irene.siragusa02@unipa.it (I. Siragusa); roberto.pirrone@unipa.it (R. Pirrone)
0009-0005-8434-8729 (I. Siragusa); 0000-0001-9453-510X (R. Pirrone)

[1]https://github.com/CHILab1/UniQA

speaking, data sets obtained via translations can be useful when large quantity of native data in Italian are not available but in general they do not have the quality of manually annotated ones. On the other side, QUANDHO [8] represents a closed-domain QA data set built from native Italian texts that collects 627 questions manually classified, thus reaching a high level of data quality but its size is moderate. In our work we want to fill this gap by creating a new QA data set with a considerably large number of manually generated prompts for both questions and answers, which rely on structured data in both Italian and English without using any translation procedure.

QA is faced using LLMs by means of RAG to reduce both hallucinations and out-of-topic answers. RAG-based applications [4] mainly present the same architectural structure, where a retrieval component, typically a vector store, is used to save and retrieve documents related with the input, and a LLM-based generator infers the answers according to a suited prompt strategy for the target application. Mostly of the applications involves English, data and they are suitable for developing chat-bots or QA systems. Interesting works in this field that use Italian involve applications whose main focus is building a virtual assistant to help users in diverse tasks such as retrieving information about pregnancy [9] or gaining suggestions about how to write an Italian Funding Application [10], or obtaining real-time data in a industrial context [11].

3. Data set description

To build UniQA, we started from a web scraping procedure using both Selenium[2] and Beautiful Soup[3], over the website of the University of Palermo[4] thus collecting a total of 1048 documents containing information about the bachelor and master degree courses for the academic year 2024/2025. In Table 1 are reported the number of documents collected in both the Italian and English splits along with the total ones (JOINT). Both Italian and English documents are original ones, that were scraped from the corresponding pages of the UniPA website either in the Italian or the English version, thus no translation has been made from Italian to English to create the data set.

Table 1
Overview of the document splits built from scraping the UniPA website.

	# IT-split	**# EN-split**	**# JOINT-split**
Course info	262	262	524
Course outline	262	262	524
# total	524	524	1048

For each available course, two documents have been generated, namely *Course info* and *Course outline* that share an equal header, collecting general information about the course such as the type of degree, the Department of affiliation, and the access rules. In *Course info* are reported also the educational objectives, the professional opportunities, and the final examination rules for the specific course. Despite the University offers a total of 190 bachelor and master degrees, we collected 262 document couples. Provided that a course can have multiple *curricula*, which differ from either some classes or the location where the course is held, it was necessary to build both documents for each of them, causing small overlapping and repetitions as for the *Course info* documents. Documents of the same type follow the same architecture, thus allowing a semi-automated information-extraction for the generation of the QA data set. In addition, we added to each document the following phrases *"For more information visit the course website [link]"*, and *"Per maggiori informazioni consulta il sito del corso [link]"* according to the specific language split and reporting the link to the web-page of the course.

Particularly, ten different QA prompts were generated (five prompts for each language split) that are reported in Table 2 and 3, and refer to the common header shared by each *Course info-Course*

[2] https://www.selenium.dev/
[3] https://www.crummy.com/software/BeautifulSoup/bs4/doc/
[4] https://offertaformativa.unipa.it/offweb/public/corso/ricercaSemplice.seam

outline document couple. Moreover, six prompts (three prompts for each language split) were generated specifically for each *Course info* document, that are reported in Table 4 and 5, and four prompts (two prompts for each language split) for each *Course outline document* (see Table 6 and 7).

Table 2
List of the generated QA pairs, which leverage the common textual header in each document couple in the English split.

Course info / Course outline - English split	
Questions	*Answers*
What are the available curriculum for the `bachelor` degree in `course name`?	There are no available curriculum for the `bachelor` degree in `course name`.
Is the `master` degree in `course name` a open or closed access course?	The `master` degree in `course name` is a free access course.
Where do the lessons of the `bachelor` degree in `course name` take place?	Lessons are held in `location` at the `department name`.
Is it possible to obtain a double degree with a `master` degree in `course name`?	No, it is not possible to obtain a double degree, but it is possible to participate at the Erasmus program.
Provide me some information regarding the `bachelor/master` degree in `course name`.	The `bachelor` degree in `course name` for the academic year 2024/2025 is a `n`-year course held in `location` at the `department name`. It is possible to choose among one of the following curriculum: `curriculum list`. It is a closed access course with `number` seats available. It is possible to obtain a double degree with `affiliate university`.

Table 3
List of the generated QA pairs, which leverage the common textual header in each document couple in the Italian split

Course info / Course outline - Italian split	
Questions	*Answers*
Quali sono i curriculum disponibili per il corso di laurea `triennale` in `nome corso`?	Non sono disponibili curriculum per il corso triennale in `nome corso`.
Il corso di laurea `magistrale` in `nome corso` è a numero chiuso o ad accesso libero?"	Il corso di laurea `magistrale` in `nome corso` è ad accesso libero.
Dove si svolgnono le lezioni del corso di laurea `triennale` in `nome corso`?	Le lezioni si svolgono presso la sede di `luogo` del `nome dipartimento`.
È possibile conseguire il doppio titolo con il corso di laurea `magistrale` in `nome corso`?	"No, non è possibile conseguire il doppio titolo, ma è possibile partecipare al programma Erasmus.
Dammi delle informazioni sul corso di laurea `triennale/magistrale` in `nome corso`.	Il corso di laurea `triennale` in `nome corso` per l'anno accademico 2024/2025 è un corso della durata di `n` anni presso la sede di `luogo` del `nome dipartimento`. È possibile scegliere uno dei seguenti curriculum: `lista dei curriculum`. Il corso è a numero chiuso, sono disponibili `n` posti.

In the last group of prompts, the former asks for generic information about the list of classes held in a target year, while the latter requests for specific information about a target class. As a consequence, the number of generated QA pairs for each document depends on both the number of years of the bachelor or master course and on the number of classes themselves. The following phrases *"For more information visit the course website [link]"*, and *"Per maggiori informazioni vai su [link]"* are further concatenated to each answer of the generated QA pairs in both the English and the Italian split.

We are aware about the limitations and redundancy of the generated data set with a small amount of manually annotated templates for questions and answers. Despite this, our focus was towards generating a data set suitable for fine-tuning a LLM for making it able to generate answers that are not exact frames of the documents they have been generated from, but a re-paraphrased version. At the

Table 4
List of the generated QA pairs for the *Course info document*, English split.

Course info - English split	
Questions	*Answers*
What are the educational objectives of the `master` degree in `course name`?	`Educational objectives from the Course info document.`
What are the professional opportunities of the `bachelor` degree in `course name`?	`Professional opportunities from the Course info document.`
What are the features of the final examination of the `master` degree in `course name`?	`Features of the final examination from the Course info document.`

Table 5
List of the generated QA pairs for the *Course info document*, Italian split.

Course info - Italian split	
Questions	*Answers*
Quali sono gli obiettivi formativi del corso di laurea `magistrale` in `nome corso`?	`Obiettivi formativi dal documento course info.`
Quali sono gli sbocchi occupazionali che il corso di laurea `triennale` in `nome corso`?	`Sbocchi occupazionali dal documento course info.`
Quali sono le caratteristiche della prova finale del corso di laurea `magistrale` in `nome corso`?	`Caratteristiche della prova finale dal documento course info.`

Table 6
List of the generated QA pairs for the *Course outline document*, English split.

Course outline - English split	
Questions	*Answers*
What are the subjects of the `target year` of the `bachelor` degree in `course name curriculum name`?	Subjects of the `target year` of the `bachelor` degree in `course name curriculum name` are: `subjects list`. A thesis is also expected to be conducted. It is possible to choose among the following teachings as for optional subjects `optional subject list`.
Provide me some details regarding the teaching of `subject` from the `master` degree in `course name`.	`subject` is a n-ECTS subject of the `teaching year` of the `master` degree in `course name`. Teaching is held by professor `surname`. Lessons will take place during the `target` semester.

end of the generation process, we collected a total of 13742 QA pairs, equally split in 6871 Italian pairs and 6871 English pairs, as in Table 8.

4. Experiments

4.1. Data split

To stress the scientific interest of the developed data set, we provided also a list train-test split of the data set that are interleaved with the language split. The resulting available splits are reported in Table 9. Starting from the *Course info* documents, we first selected all the unique bachelor and master degrees, so that courses with multiple curricula were counted once, thus a set of 190 courses was obtained. Then

Table 7
List of the generated QA pairs for the *Course outline document*, Italian split.

Course outline - Italian split	
Questions	*Answers*
Quali sono le materie del `anno target` del corso di laurea `triennale` in `nome corso` `nome curriculum`?	Le materie del `anno target` del corso di laurea `triennale` in `nome corso` `nome curriculum` sono: `lista delle materie`. È inoltre previsto lo svolgimento della tesi. All'interno del Gruppo di attività formative opzionali è possibile scegliere tra le seguenti materie: `lista materie opzionali`.
Dammi informazioni sulla materia `nome materia` del corso di laurea `magistrale` in `nome corso`.	`nome materia` è una materia di n CFU del `anno insegnamento` del corso di laurea magistrale in `nome coros`. L'insegnamento è tenuto dal professore `cognome`. Le lezioni si terranno nel `target` semestre.

Table 8
Distribution of the generated QA pairs in the two language splits.

	# IT-split	**# EN-split**	**# JOINT-split**
Course info QA	1520	1520	3040
Course outline QA	5351	5351	10702
# total	6871	6871	13742

the courses were sub-grouped with respect to the Department they belong to, and for each sub-group, a random 80-20 split was done to generate train and test groups. This procedure was implemented to ensure that:

- bachelor and master courses with multiple *curricula* are considered as a unique block, and then put together in either a train or test split;
- courses belonging to the same Department are equally divided to prevent any bias on the trained models.

Due to computational constraints on the training procedure of the generation LLM in our RAG architecture, we created also a *reduced* split of the data set whose global input is less than 3000 tokens, and it is 16% smaller than the original one, thus providing a not so significant reduction in performance.

Table 9
List of the UniQA train-test splits grouped in Italian, English and Joint split

	#Train	**#Test**	**#All**
Documents-IT	398	126	524
Documents-EN	398	126	524
Documents-JOINT	796	252	1048
QAs-IT	5738	1709	7447
QAs-EN	5738	1709	7447
QAs-JOINT	11476	3418	14894
QAs-IT-reduced	4293	1249	5542
QAs-EN-reduced	5303	1556	6859
QAs-JOINT-reduced	9596	2805	12401

In all the splits we included the QA pairs as well as the original documents, thus allowing the data set to be suitable for a large variety of NLP tasks, such as translation and QA with support of external knowledge (QA-EK). In this paper we report the performance on a QA-EK task of a RAG-based architecture based on `Llama 3.1` both in the Foundational and Instruct version [5].

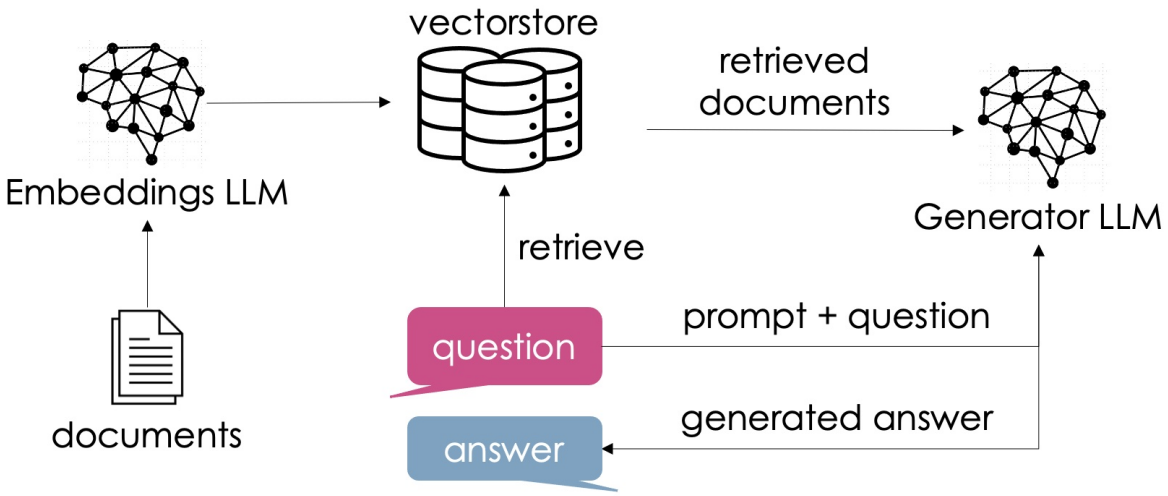

Figure 1: Schema of the implemented architecture for QA.

4.2. Experimental setup

We implemented a RAG-based architecture to perform QA-EK tasks on the UniQA data set in order to evaluate the quality of our data with respect to the correctness of the provided answers that is also related to the retrieval accuracy of the related documents. Such evaluations can be easily performed since golden answers are known. Finally, this type of architecture, can be easily queried with domain-related questions that are not in UniQA data set: in this case, answers can be generated from the retrieved documents, but evaluation can be trivial due to the lack of the corresponding golden answer.

The implemented RAG-based architecture is illustrated in Figure 1, where two main components can be distinguished: the *retriever module* and the *generator LLM*.

Retriever module The retriever module is composed by a vector store and an *Embeddings LLM*. To build it, we implemented a FAISS-based vector store [12] where the generated documents, both from the train and test split, were injected after being splitted in 1000 token chunks with 100 overlapping tokens, using tiktoken[5] as the tokenizer. The token chunks are then processed by a LLM tailored for embedding generation (Embeddings LLM), with retrieval capabilities, that supports both English and Italian. Accordingly, we selected the best models that meet our constraints on computational resources from the Massive Text Embedding Benchmark (MTEB) [13][6] namely `BGE-M3` (BGE) [14], `gte-Qwen2-7B-instruct` (GTE) [15] and `Multilingual-E5-large-instruct` (m-E5) [16]. All of them where trained on multilingual data, including English and Italian: actually, it is not so simple to select models that explicitly state that were trained also on Italian data. As the internal architecture, all of them are build upon Transformers encoder [17], and both BGE and m-E5 are small 5M models, while GTE is a 7B one. One vector database for each model was built using the LangChain framework[7], and their retrieval performances are reported in Section 5.

Generator LLM We decided to stress the capabilities of `Llama-3.1` 8B models [5], the last decoder-only generative model of the Llama models family, that has a native support for Italian and English as well, and it is freely available. We tested both Foundational and Instruct models providing two different English prompts: *Prompt 1* is designed as standard instruction-prompt, and it is suitable for Foundational models, while *Prompt 2*, follows the instruction prompt suggested for both Instruction tuning and inference by the authors of the `Llama 3.1` models [5]. Prompts are reported below.

Prompt 1

[5]https://github.com/openai/tiktoken
[6]https://huggingface.co/spaces/mteb/leaderboard
[7]https://www.langchain.com/langchain

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction: *You are Unipa-GPT, the chatbot and virtual assistant of the University of Palermo. Provide a cordially and colloquially answers to the questions provided. If you receive a greeting, answer by greeting and introducing yourself. If you receive a question concerning the University of Palermo, answer relying on the documents given to you with the question. If you do not know how to answer, apologize and suggest that you consult the university website [https://www.unipa.it/], do not invent answers. If the question is in English, answer in English. If the question is in Italian, answer in Italian.*

Input:

Question: question

Documents: context

Response:

Prompt 2

<|start_header_id|>system<|end_header_id|>

You are Unipa-GPT, the chatbot and virtual assistant of the University of Palermo. Provide a cordially and colloquially answers to the questions provided. If you receive a greeting, answer by greeting and introducing yourself. If you receive a question concerning the University of Palermo, answer relying on the documents given to you with the question. If you do not know how to answer, apologize and suggest that you consult the university website [https://www.unipa.it/], do not invent answers. If the question is in English, answer in English. If the question is in Italian, answer in Italian.

<|eot_id|><|start_header_id|>user<|end_header_id|>

Question: question

Documents: context

<|eot_id|><|start_header_id|>assistant<|end_header_id|>

Both the prompts were used for querying Foundational models, while Prompt 2 was used only for Instruct models. Despite the multilingual task, we opted for an English prompt since it can be more flexible in a real-wold application where the language of both the questions and the documents provided as inputs is not known a priori. After evaluating models in their base versions, we proceeded with a 3 epochs fine-tuning procedure over the best performing one, that is *Llama-3.1-Instruct*, using Prompt 2. Fine-tuning was performed with LoRA [18] following the Alpaca-LoRA hyper-parameters[8] and we will refer to this model as *UniQA-3-ft*. We trained the model using the prompt associated with the best model, and providing the golden documents as context. We developed the entire system on a server with two Intel(R) Xeon(R) Gold 6442Y CPUs, 384 GB RAM, and two 48 GB NVIDIA RTX 6000 Ada Generation.

5. Results

To evaluate the retrieval performance, we performed at first a cluster analysis in the embedding space relying on the "native" clustering of the documents, that can be divided by language (Italian or English) or by department. The scatter plots of the embedding spaces for each Embeddings LLM are reported in Figure 2 and 3 are reported. Such plots were obtained after a dimensionality reduction performed using

[8] https://github.com/tloen/alpaca-lora

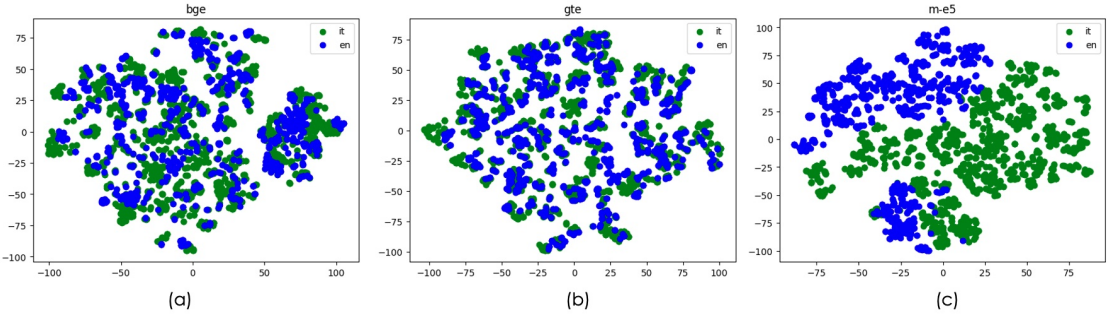

Figure 2: Scatter plots of 2D reduced embedding spaces, labeled for each language.

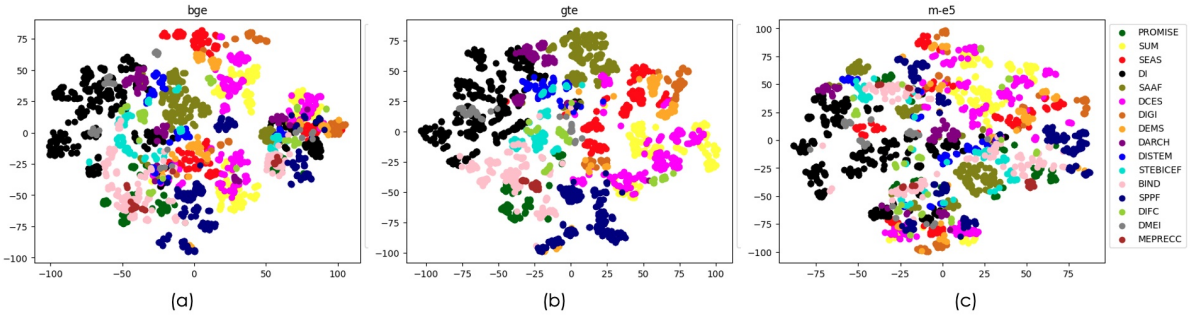

Figure 3: Scatter plots of 2D reduced embedding spaces, labeled for each Department.

t-SNE [19]. Along with the graphical visualization, we provide an analytical analysis, calculating the Silhouette Coefficient [20], as in Table 10.

Both the graphical and the analytical representation highlight that m-E5 has the best clustering capabilities in language separation (Figure 2.c) while the other models tend to overlap the embeddings. Conversely, they perform better in grouping documents in a semantic way that is by Department. Particularly, GTE outperforms the other models (Figure 3.b). Indeed, documents referring to Computer and Mechanical Engineering degree courses, which are taught in the same Department have much more in common than the ones concerning Nursing or Law. Moreover, the Italian description of a degree course contains many English terms, and this can make it harder to cluster documents based on their native language.

The retrieval performances of the models were evaluated by querying their vector stores with question samples belonging to the test set[9]: if at least one of the retrieved document matches the golden one associated with the question, it was considered a correct retrieval. Thus, an accuracy measure was computed as it is reported in Table 10: here the superiority of GTE is confirmed with an accuracy of almost 86%, while BGE reaches an accuracy just over 81%, and m-E5 attains just 77%.

Table 10
Clustering and retrieval performances of the three embedding LLM that we tested. S-Score is used for clustering, while retrieval performances are expressed in terms of accuracy. Best results are in bold.

Retriever	S-Score language	S-Score departments	Exact matches	Total matches	Accuracy (Total/Exact)
BGE	0.0057	-0.0883	2283	2805	81.3904%
GTE	-0.0048	**0.0470**	2408	2805	**85.8467%**
m-E5	**0.2237**	-0.1007	2159	2805	76.9697%

QA evaluation, was performed by querying the models using the reduced joint version of the test set (Table 11) the English only reduced split (Table 12), and the Italian only one (Table 13). Performance were measured in terms of BLEU [21] and Rouge score [22]. Inference ability in *Llama 3.1 Foundational*, *Llama*

[9]we used the reduced split in this experiment just like in the ones devoted to QA evaluation

3.1 Instruct and *UniQA-3-ft* was assessed without any quantization strategy. *Llama 3.1 Foundational* was queried using both Prompt 1 and Prompt 2, while *Llama 3.1 Instruct* and *UniQA-3-ft* were queried using only Prompt 2, since they are Instruction fine-tuned models. The evaluation consisted in two runs. In the first run, the golden context was provided in a one-shot scenario to the LLM without RAG, while the second one made use of GTE-based retriever module. The former run was aimed at evaluating the model inherent capabilities at generating a correct answer that adheres to the golden one, that is a re-paraphrase of the context. The latter run was aimed at evaluating end-to-end performances of the whole RAG architecture. We will refer to models that make use of RAG using the suffix *retrieved*.

Prompt 3

<|start_header_id|>system<|end_header_id|>

You are Unipa-GPT, the chatbot and virtual assistant of the University of Palermo. Provide a cordially and colloquially answers to the questions provided. If you receive a greeting, answer by greeting and introducing yourself. If you do not know how to answer, apologize and suggest that you consult the university website [https://www.unipa.it/], do not invent answers. If the question is in English, answer in English. If the question is in Italian, answer in Italian.

<|eot_id|><|start_header_id|>user<|end_header_id|>

Question: question

<|eot_id|><|start_header_id|>assistant<|end_header_id|>

To provide a more comprehensive overview of our contribution, we tested our *UniQA-3-ft* fine-tuned model to asses its generation capabilities in a zero-shot scenario without RAG: we will refer to this evaluation configuration as *UniQA-3-ft no-RAG*. A suitable version of the Prompt 2 was devised for this purpose that we called *Prompt 3*, and does not contain any mention to rely on external documents for answer generation. In Tables 11, 12 and 13, best results for each run are in bold, while italicised score values have been used in the third run when *UniQA-3-ft no-RAG* performed better than *UniQA-3-ft retrieved*.

Table 11
BLEU and ROUGE score for the different LLMs used for evaluation of the QAs-JOINT-reduced split.

LLM	Prompt	BLEU	Rouge-1	Rouge-2	Rouge-L	Rouge-Lsum
Llama 3.1	1	0.0043	0.0913	0.0167	0.0557	0.0648
Llama 3.1	2	0.0035	0.0710	0.0132	0.0455	0.0544
Llama 3.1 inst	2	0.0328	0.1636	0.0370	0.1017	0.1291
UniQA-3-ft	2	**0.2730**	**0.5390**	**0.3780**	**0.5217**	**0.5223**
Llama 3.1 retrieved	1	0.0043	0.0914	0.0167	0.0556	0.0647
Llama 3.1 retrieved	2	0.0030	0.0332	0.0036	0.0265	0.0298
Llama 3.1 inst retrieved	2	0.0070	0.1204	0.0250	0.0800	0.0992
UniQA-3-ft retrieved	2	**0.1646**	**0.3548**	**0.2075**	**0.3312**	**0.3332**
UniQA-3-ft no-RAG	3	0.1322	*0.3777*	0.2007	0.3288	*0.3333*

As it was expected, *UniQA-3-ft* and *UniQA-3-ft retrieved* outperform the other models in their respective runs, while their difference in performances is not so significant, and it mainly depends on the quality of the retrieved documents. *Llama 3.1 Foundational* performs a bit better using Prompt 2 with respect to Prompt 1, and *Llama 3.1 Instruct* shows clearly its ability to follow instructions in both settings.

UniQA-3-ft no-RAG reaches comparable performances to *UniQA-3-ft retrieved*, and in some it scores higher than the RAG version. This finding indicates clearly that UniQA is a high quality robust data set that can be used to test both fine-tuned models and RAG architectures. It is worth noticing that the

Table 12
BLEU and ROUGE score for the different LLMs used for evaluation of the QAs-EN-reduced split.

LLM	Prompt	BLEU	Rouge-1	Rouge-2	Rouge-L	Rouge-Lsum
Llama 3.1	1	0.0016	0.1103	0.0230	0.0635	0.0744
Llama 3.1	2	0.0017	0.0913	0.0168	0.0540	0.0664
Llama 3.1 inst	2	0.0313	0.1445	0.0401	0.0945	0.1184
UniQA-3-ft	2	**0.2900**	**0.5813**	**0.4385**	**0.5641**	**0.5648**
Llama 3.1 retrieved	1	0.0016	0.1103	0.0230	0.0635	0.0744
Llama 3.1 retrieved	2	0.0021	0.0442	0.0040	0.0346	0.0395
Llama 3.1 inst retrieved	2	0.0064	0.1047	0.0252	0.0728	0.0888
UniQA-3-ft retrieved	2	**0.1555**	**0.3542**	**0.2169**	**0.3271**	**0.3304**
UniQA-3-ft no-RAG	3	0.1526	*0.4191*	*0.2478*	*0.3687*	*0.3751*

Table 13
BLEU and ROUGE score for the different LLMs used for evaluation of the QAs-IT-reduced split.

LLM	Prompt	BLEU	Rouge-1	Rouge-2	Rouge-L	Rouge-Lsum
Llama 3.1	1	0.0082	0.0677	0.0089	0.0459	0.0527
Llama 3.1	2	0.0055	0.0455	0.0088	0.0347	0.0396
Llama 3.1 inst	2	0.0349	0.1876	0.0331	0.1106	0.1425
UniQA-3-ft	2	**0.2433**	**0.4850**	**0.3029**	**0.4680**	**0.4687**
Llama 3.1 retrieved	1	0.0082	0.0677	0.0090	0.0459	0.0527
Llama 3.1 retrieved	2	0.0036	0.0195	0.0032	0.0164	0.0178
Llama 3.1 inst retrieved	2	0.0079	0.1402	0.0246	0.0890	0.1121
UniQA-3-ft retrieved	2	**0.1701**	**0.3543**	**0.1963**	**0.3357**	**0.3373**
UniQA-3-ft no-RAG	3	0.1023	0.3260	0.1421	0.2789	0.2817

structure of the train-test split guarantees that the answers provided by *UniQA-3-ft no-RAG* leverage only the knowledge acquired during the fine-tuning phase. In fact, when answering to a question belonging to the test set, the model is completely unaware on the degree courses that are not in its training set (Section 4.1). In this configuration, the inference capabilities of the model can be truly tested since it is relying on the acquired knowledge from the QA pairs of similar degree courses in the same Department.

Via a manual inspection of some of the generated answers, we found that both the non fine-tuned models and the fine-tuned ones, tend to output misspelled words, while both *UniQA-3-ft no-RAG* and the non fine-tuned models provide incorrect answers since they have no access to a complete knowledge of the UnPA domain, thus they reply leveraging their native and incmplete knowledge. The non fine-tuned models tend to output verbose answers, and not to provide important information, thus wandering off with a hypothetical course degree outline which is not required, and may be imprecise. Generally speaking, the non fine-tuned models may output some correct information, but in a different format as the one provided in the golden answer, thus making it more difficult to evaluate the overall correctness of the generated replies.

In both the golden answers and the retrieved documents a suggestion is reported for the final user to visit the website of the degree course to get more information: models try to generate links following the structure of the ones provided in the retrieved documents and in the prompt. The non fine-tuned models fail since either no link is generated at all or the generated link does not refer to the UniPA website. Fine-tuned models perform better, but not all the generated links are correct since misspellings are quite common.

6. Conclusions and future works

In this paper we presented UniQA, a high-quality QA data set in Italian and English suitable for translation and question-answering tasks where external knowledge is required. UniQA is a balanced data set among the two languages, and it didn't require any translation because it was scraped from original Italian and English web pages of related to the degree courses issued at UniPA. UniQA counts 1048 documents and 13742 QA pairs generated in a semi-automated manner.

We also tested a RAG-based architecture for QA with external knowledge tasks whose generation LLMs were both *Llama 3.1 Foundational* and *Llama 3.1 Instruct*. *Llama 3.1* was selected as a proof of concept because it is recognized as a SOTA multilingual LLM, while both the fine-tuning and the inference-only runs required a considerable amount of time on our local computational facilities. At the time of submitting the manuscript, extensive tests are being run using also both Foundation and Instruct LLMs that are based on different architectures than Llama as well as on the most known Italian adaptations of such models.

Future developments of this work are towards both extensive fine-tuning of the models under investigation and on end-to-end training of the whole RAG architecture including the retriever. Finally, a hybrid RAG architecture using both vector and graph databases is under development to encode both (vector) semantic similarity between documents and their closeness with respect to a domain ontology implemented as a graph of semantic relations between the documents in the corpus.

References

[1] G. Bonetta, C. D. Hromei, L. Siciliani, M. A. Stranisci, Preface to the Eighth Workshop on Natural Language for Artificial Intelligence (NL4AI), in: Proceedings of the Eighth Workshop on Natural Language for Artificial Intelligence (NL4AI 2024) co-located with 23th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2024), 2024.

[2] OpenAI, Gpt-4 technical report, 2024. URL: https://arxiv.org/abs/2303.08774. arXiv:2303.08774.

[3] G. Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, et al., Gemini: a family of highly capable multimodal models, arXiv preprint arXiv:2312.11805 (2023).

[4] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al., Retrieval-augmented generation for knowledge-intensive nlp tasks, Advances in Neural Information Processing Systems 33 (2020) 9459–9474.

[5] A. . M. Llama Team, The llama 3 herd of models, 2024. URL: https://arxiv.org/abs/2407.21783. arXiv:2407.21783.

[6] D. Croce, A. Zelenanska, R. Basili, Neural learning for question answering in italian, in: C. Ghidini, B. Magnini, A. Passerini, P. Traverso (Eds.), AI*IA 2018 – Advances in Artificial Intelligence, Springer International Publishing, Cham, 2018, pp. 389–402.

[7] P. Rajpurkar, J. Zhang, K. Lopyrev, P. Liang, Squad: 100,000+ questions for machine comprehension of text, 2016. URL: https://arxiv.org/abs/1606.05250. arXiv:1606.05250.

[8] S. Menini, R. Sprugnoli, A. Uva, "who was pietro badoglio?" towards a QA system for Italian history, in: N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), European Language Resources Association (ELRA), Portorož, Slovenia, 2016, pp. 430–435. URL: https://aclanthology.org/L16-1069.

[9] S. Ghanbari Haez, M. Segala, P. Bellan, S. Magnolini, L. Sanna, M. Consolandi, M. Dragoni, A retrieval-augmented generation strategy to enhance medical chatbot reliability, in: J. Finkelstein, R. Moskovitch, E. Parimbelli (Eds.), Artificial Intelligence in Medicine, Springer Nature Switzerland, Cham, 2024, pp. 213–223.

[10] T. Boccato, M. Ferrante, N. Toschi, Two-phase rag-based chatbot for italian funding application assistance, 2024.

[11] R. Figliè, T. Turchi, G. Baldi, D. Mazzei, Towards an llm-based intelligent assistant for industry 5.0, in: Proceedings of the 1st International Workshop on Designing and Building Hybrid Human–AI Systems (SYNERGY 2024), volume 3701, 2024. URL: https://ceur-ws.org/Vol-3701/paper7.pdf.

[12] J. Johnson, M. Douze, H. Jégou, Billion-scale similarity search with GPUs, IEEE Transactions on Big Data 7 (2019) 535–547.

[13] N. Muennighoff, N. Tazi, L. Magne, N. Reimers, Mteb: Massive text embedding benchmark, arXiv preprint arXiv:2210.07316 (2022). URL: https://arxiv.org/abs/2210.07316. doi:10.48550/ARXIV.2210.07316.

[14] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, Z. Liu, Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, 2024. arXiv:2402.03216.

[15] Z. Li, X. Zhang, Y. Zhang, D. Long, P. Xie, M. Zhang, Towards general text embeddings with multi-stage contrastive learning, arXiv preprint arXiv:2308.03281 (2023).

[16] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, F. Wei, Multilingual e5 text embeddings: A technical report, arXiv preprint arXiv:2402.05672 (2024).

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[18] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Lora: Low-rank adaptation of large language models, arXiv preprint arXiv:2106.09685 (2021).

[19] L. Van der Maaten, G. Hinton, Visualizing data using t-sne., Journal of machine learning research 9 (2008).

[20] P. J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, Journal of computational and applied mathematics 20 (1987) 53–65.

[21] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.

[22] C.-Y. Lin, Rouge: A package for automatic evaluation of summaries, in: Text summarization branches out, 2004, pp. 74–81.