ISSN 2499-4553

IJCoL

Italian Journal of Computational Linguistics

Rivista Italiana di Linguistica Computazionale

Volume 10, Number 2
december 2024
Special Issue

Natural Language for Artificial Intelligence in the Era of LLMs

aA Accademia university press

Registrazione presso il Tribunale di Trento n. 14/16 del 6 luglio 2016

Rivista Semestrale dell'Associazione Italiana di Linguistica Computazionale (AILC)

Associazione Italiana di Linguistica Computazionale

direttore responsabile
Michele Arnese

isbn 9791255001430

Accademia University Press
via Carlo Alberto 55
I-10123 Torino
info@aAccademia.it
www.aAccademia.it/IJCoL_10_2

IJCoL

Volume 10, Number 2
december 2024
Special Issue

Natural Language for Artificial Intelligence in the Era of LLMs

editors: *Elisa Bassignana, Dominique Brunato, Marco Polignano, Alan Ramponi*

CONTENTS

Unipa-GPT: Large Language Models for university-oriented QA in Italian

Irene Siragusa*
Università di Palermo

Roberto Pirrone**
Università di Palermo

This paper illustrates the architecture and training of Unipa-GPT, a chatbot relying on a Large Language Model, developed for assisting students in choosing a bachelor/master degree course at the University of Palermo. Unipa-GPT relies on `gpt-3.5-turbo`*, it was presented in the context of the European Researchers' Night (SHARPER night). In our experiments we adopted both the Retrieval Augmented Generation (RAG) approach and fine-tuning to develop the system. The whole architecture of Unipa-GPT is presented, both the RAG and the fine-tuned systems are compared, and a brief discussion on their performance is reported. Further comparison with other Large Language Models and the experimental results during the SHARPER night are illustrated. Corpora and code are available on GitHub*[1].

1. Introduction

Natural Language Processing (NLP), Artificial Intelligence (AI) and Large Language Model (LLM) are highly interesting topics in the scientific community, but they have become also very popular since ChatGPT[2] by OpenAI was released and it was made publicly accessible. Despite the debate about the level of consciousness in ChatGPT and the claims about its standing for a new generation of Artificial General Intelligence (AGI) (Dwivedi et al. 2023; Borji 2023), a plethora of new applications are being developed that are based on either GPTs family (Brown et al. 2020; OpenAI 2024) or others LLMs proposed in the last year, such as the ones belonging to LLama family (Touvron et al. 2023; Llama Team 2024).

In view of the previous considerations, we investigated the use of ChatGPT for building a virtual assistant for the users of a large public institution like the University of Palermo, and in this paper we propose Unipa-GPT, a chatbot that can guide secondary school students to explore the huge information provided by the University institutional website, when they are choosing a bachelor degree course, by answering to questions related to the available courses, application procedure, tax payments, access to scholarships and more. Unipa-GPT was presented to the wide public in the context of the SHARPER Night event[3].

Unipa-GPT has been developed mainly as a Retrieval Augmented Generation (RAG) system (Lewis et al. 2020) based on `gpt-3.5-turbo` that collects the target

* Department of Engineering - Viale delle Scienze, Ed. 6, 90128, Palermo, Italy
E-mail: irene.siragusa02@unipa.it
** Department of Engineering - Viale delle Scienze, Ed. 6, 90128, Palermo, Italy
Email: roberto.pirrone@unipa.it
1 https://github.com/CHILab1/UnipaGPT-23
2 https://openai.com/index/chatgpt/
3 https://www.sharper-night.it/

information from `unipa-corpus`, a document corpus that has been built purposely by scraping the University website. The aim of this work is to explore the behavior and the limitations of LLMs when they are engaged in a Question-Answer (QA) task where precise domain knowledge is required. Obviously, relying on the OpenAI API for inference and fine-tuning of the model, also a fine-tuned version has been built where the corpus has been modified with the aim of saving computational resources i.e. use the as few tokens as possible[4], and a mixed strategy has been adopted where RAG was coupled with fine-tuning to avoid the train step on very detailed information such as the educational objectives of each single class.

Both base and fine-tuned models, have been previously qualitatively tested by very few students and thanks to them, a first comparison of the developed models was done relying on their judgement on two reference chats along with a discussion of the results. Due to the aforementioned token-related costs, we did not conduct an extensive quantitative evaluation on the fine-tuned model. Best model, according to users' judgements, was used by real users in an unsupervised context during the SHARPER (SHAring Researchers' Passion for Enhanced Roadmaps) night, i.e. the European Researchers' Night, where questions, answers and feedback were collected. Starting from the obtained data, a small QA data set was mannally derived, along with the golden document used to generate the golden answer. Objective of this small dataset is to conduct a detailed comparison with open-source LLMs in contrast with ChatGPT evaluating their generation capabilities.

The paper is arranged as follows: a state of the art analysis is reported in Section 2. Section 3 illustrates the different corpora we set up for building both the RAG and the fine-tuned Unipa-GPT. The detailed architecture of both systems is reported in Section 4, while the experimental results are reported and discussed in Section 5. Concluding remarks are drawn in Section 6.

2. Related works

The scientific community had an interest in chatbots since 1960s, when ELIZA (Weizenbaum 1966) was developed. Continuous and increasing studies in deep learning architectures, lead to the development of Transformers and the attention mechanism (Vaswani et al. 2017), that, together with BERT (Devlin et al. 2019), can be considered a turning point in NLP. Another crucial moment is the birth of ChatGPT models that are built from GPT-3 (Brown et al. 2020), and GPT-4 (OpenAI 2024), whose extraordinary generation capabilities overcome pre-existing models, and they can be regareded as a new SOTA level, reaching interest of both researchers and common people.

Since ChatGPT was released, its usage in education has been carefully explored, since a balance is required between potential benefits and drawbacks of this technology. In fact, very much care has to be devoted to the generation of inaccurate or incorrect information, and the possible cheating when a student claims AI-generated text as her/his original production (Gill et al. 2024). Generally speaking, it is considered as a valuable tool that can be used together with traditional methodologies and whose usage cannot be prohibited or completely not considered (Sharma and Yadav 2023). Different use of AI-based chat-bots involves assistants for instructors (Lo 2023), such as in generating questionnaires (Rodriguez-Torrealba, Garcia-Lopez, and Garcia-Cabot 2022), exploring pedagogical abilities of conversational models (Tack and Piech 2022),

4 OpenAI API are available on a token-based fee for inference and fine-tuning

or helping students with interactive learning approaches that can be personalized for their specific needs (Rahman and Watanobe 2023).

Scientific interest in developing domain-based chatbot using a RAG approach, involves not only educational domain, as for Bio-Eng-LMM AI (Forootani, Aliabadi, and Thraen 2024), but also other domains, where chatbots interaction can provide various benefits, such as chatbots providing information about restaurant industry (Bhat et al. 2024), or assistance to victims of sexual harassment (Vakayil et al. 2024) or to patient in healthcare context (Vidivelli, Ramachandran, and Dharunbalaji 2024; Wang et al. 2024). These applications share similar methodologies, consisting in collecting all relevant documents for the domain of interest, pre-process and store them in a vectordatabase, that is sub-sequentially queried for retrieve the closest documents given a question from the user. Retrieved documents and user question, serve as input to a generative LLM that outputs an answer.

According to our knowledge, despite the increasing interest from the Italian community in developing Italian LLMs such as LLaMantino (Basile et al. 2023), Camoscio (Santilli and Rodolà 2023), Fauno (Bacciu et al. 2023), Anita (Polignano, Basile, and Semeraro 2024), and Minerva (Orlando et al. 2024), we did not find a domain specific chatbot application using a RAG approach starting from Italian data (e.g. not translated). To fill this gap, in this work we aim to explore the capabilities of a GPT-based Italian virtual assistant for secondary school students in a chat-bot configuration, which makes use of external specific knowledge injected via a RAG approach in generation phase.

3. Corpora

In this section we outline the differences between the versions of the `unipa-corpus` used for developing the RAG-only and the fine-tuned system.

3.1 `unipa-corpus` for RAG

The corpus used for Unipa-GPT, called `unipa-corpus`, is a collection of documents that were collected directly from the website of the University of Palermo during Summer 2023. A manual selection of the most interesting pages was made with reference to the target audience of secondary-school students and two main sections of the corpus were identified, that correspond to the *Education* and the *Future Students* sub-trees in the institutional website. Since the expected questions are in Italian, the generated corpus is in Italian.

The *Education* section is the main part of the corpus and it is a collection of all the available courses at the University for the academic year 2023/2024. For each course and each curriculum two documents are obtained: *details* is the document that collects all the general details of the course, like name, department of affiliation, typology of course (Bachelor or Master degree), restriction of access and a colloquial description of the course, including its educational objectives and professional opportunities; *course outline* is the second document that collects the course outline divided by year, and the number of credits, the teaching professor, the teaching period and the scientific sector are specified for each class. Three different versions of the *course outline* document were generated, namely `clear`, `full` and `emb`. The `clear` version is the one described above, the `full` version adds a new document for every class in a course and reports its peculiar educational objectives. The `emb` version is a mix of the previous ones where the classes' educational objectives are added directly in the document containing the outline

of the course. Despite this corpus is called *embedded*, it does not contains embeddings: the term *embedded* refers to the educational objectives that are inserted in the document containing the course outline. As a consequence, both the `clear` and the `emb` corpus have the same number of documents but different information, while the `full` contains the same information of the `emb` corpus, but it is arranged in a different number of documents.

The *Future Student* section is the same for the three versions of the corpus, and it is a mix of documents coming from the related section of the University website. The information contained in these documents is addressed to the future students of the University, including the academic calendar, the tax rules and reductions, scholarships, University enrolment procedure, and facilities offered to the students. The statistics of each corpus are reported in Table 1.

Table 1
Number of documents in each version of `unipa-corpus`.

	Education	*Future Students*	*Total documents*	*Total tokens*
`unipa-corpus-clear`	506	104	610	2059638
`unipa-corpus-full`	5794	104	5898	890598
`unipa-corpus-emb`	506	104	610	2063020

3.2 `unipa-corpus` for fine-tuning

The `unipa-corpus` was modified to be in the form required for fine-tuning `gpt-3.5-turbo`. As already mentioned above, our intent in fine-tuning was lowering the computational resources as much as possible that is using the minimum tokens for training the model. Besides the economic aspect, in the case of ChatGPT fine-tuning, this is a crucial topic when dealing with LLMs because also relatively small LLMs like LLama-2-7B (Touvron et al. 2023) require huge computational resources for their full fine-tuning or retraining.

To overcome this problem, the so-called Parameter-Efficient Fine-Tuning (PEFT) techniques as LoRA (Hu et al. 2022) have been developed, that reduce the required computational resources for a satisfactory LLM fine-tuning, but it is not available for closed systems like gpt-3.5-turbo. Since our purpose is primary towards stressing the capabilities of ChatGPT, we decided to test its performance after a fine-tuning phase over the `unipa-corpus-clear`, the smallest created corpus in terms of tokens to be trained thus reaching a trade-off in terms of API costs.

The required format for fine-tuning is a sequence of prompt, question and answer. No additional context (e.g. the closest documents for an effective reply of the model) was added to the prompt for fine-tuning, thus maintaining low the costs of the whole process. The prompt used is a simple instruction of the chatbot behavior:

> *Sei Unipa-GPT, un assistente virtuale che risponde alle domande sull'università di Palermo*
>
> (*You are Unipa-GPT, a virtual assistant that answers to questions related with the University of Palermo*).

Questions and answer were automatically generated with different generation rules for each section of the corpus.

QA pairs were automatically generated from the documents falling in the *Education* section, by asking gpt-3.5-turbo to describe a specific degree course starting from the corresponding *details* document. Moreover, the LLM was asked to provide the topics of a specific degree course starting from the corresponding *course outline* document. In both cases, the corresponding document was given along with the question, and the given answer was considered as an answer for the fine-tuning corpus.

As regards the *Future Students* section, QA pairs were extracted directly from the documents already containing a FAQ section, while the other pairs were manually generated. In the second case, a clear question related to a document's section was formulated whenever it was possible, and the answer was either a precise text or the whole document. Otherwise a generic request was formulated like *parlami di ... (speak about ...)*. Some documents were not considered in their entirety since the information contained was highly specific and it was related to non-relevant topics.

A validation set was also expunged from the training data by changing questions and/or sampling most important questions. A QA pair was randomly picked for each degree course among the *details* and the *course outline* documents in the *Education* section. The statistics of the corpus for fine-tuning are reported in Table 2. Despite this arrangement can be suitable for creating also a test split of the data set, to further evaluate fine-tuning performances of the model, our primary focus in this phase was towards determining the models that better matches users' preferences. Moreover, the principal objective of fine-tuning was to inject domain-specific knowledge in the model thus allowing it to generate more precise answers when queried.

`gpt-3.5-turbo` was the only powerful model available at the time of the experiments to automatically generate QA pairs. As it can be inferred from Table 2, manual generation of the total number of QA pairs was impractical. We are aware of the possible the biases implied in fine-tuning a model with self-generated samples.

Moreover, `gpt-3.5-turbo` generated QA pairs following the simple one-shot strategy described above, in order to fine-tune the model to achieve a terminological knowledge of the domain through the presentation of a small subset of the facts contained in the `unipa-corpus`. We are convinced that this approach does not introduce any relevant bias because the fine-tuned model is inserted in a RAG architecture that is queried using an instruction prompt strategy.

Table 2
Number of documents in the `unipa-corpus-clear` re-arranged for fine-tuning.

	Education	*Future Students*	*Total QA pairs*	*Total tokens*
Training set	506	269	775	259772
Validation set	253	133	386	123118

4. System architecture

Unipa-GPT is developed as a RAG architecture (Lewis et al. 2020) made up of two main components, as shown in Figure 1: the retrieval and the generator module.

Figure 1
The architecture of the RAG version of Unipa-GPT.

The *retrieval module* consists of a vector database provided by the LangChain library[5], which makes use of the Facebook AI Similarity Search (FAISS) library (Johnson, Douze, and Jégou 2019). The vector database is filled with the documents in `unipa-corpus` conveniently divided into chunks of 1000 tokens with an overlap of 50 tokens whose embeddings were extracted using `text-embedding-ada-002` by OpenAI[6].

The *generator module* consists of an instance of `gpt-3.5-turbo` (Brown et al. 2020), a generative LLM based on Transformers (Vaswani et al. 2017). The LLM is queried with a custom prompt in which the behavior of the system is explained, and the question of the user is passed along with the most related documents. The expected answer of the system is a reply to the question of the user according to the prompt that rules the Unipa-GPT overall behavior, and the domain-knowledge given by the retriever. `gpt-3.5-turbo` makes inferences using a temperature hyper-parameter equal to 0 thus its behavior is as much deterministic as possible, and the system is prevented to be creative. Finally no limits *a priori* were put on the maximum tokens available for the answer, in order to prevent broken answers. The chat-bot behavior was implemented via LangChain to keep the chat history and simulate the ChatGPT behavior through the `gpt-3.5-turbo` API.

Both `gpt-3.5-turbo` and `text-embedding-ada-002` were invoked via Azure call to the OpenAI API. Two different Italian prompts were built for `gpt-3.5-turbo`: a *custom prompt* and a *condensed prompt*, as it is shown in Table 3 (English version can be found in Table 1 in Appendix A). The *custom prompt* is the explanation of the behavior of the chat-bot where both the previous conversation and the new question are concatenated to the prompt itself. On the contrary, the *condensed prompt* adds to the *custom prompt* another instruction to condense the previous conversation and re-arrange it as a new single question that will be answered accordingly to the *custom prompt*. In addition to the RAG version illustrated above, a fine-tuned version was implemented with a

5 `https://python.langchain.com/docs/get_started/introduction`
6 `https://openai.com/blog/new-and-improved-embedding-model`

custom fine-tuned version of `gpt-3.5-turbo` where the `unipa-corpus` explicitly rearranged, as described in Section 3.2, was used. The same prompt instances mentioned above, were used on the fine-tuned model, and the also the RAG architecture was used to compensate for the reduced tokens used in this step. The whole system is shown in Figure 2.

Figure 2
The schema illustrates the architecture of the proposed system where `gpt-3.5-turbo-fine-tuned` is used after a fine-tuning process over `unipa-corpus`.

5. Experimental results

This section reports the experimental results obtained in three different scenarios: sub-section 5.1 illustrates the preliminary development phase, that took place during the summer 2023. In 5.2 we analyze the results obtained during the SHARPER Night 2023, when the best model resulting from the previous phase was used in free interaction with the public. Finally, sub-section 5.3 reports the comparison of the original Unipa-

Table 3
The prompts used in the system.

prompt type	*prompt text*
custom prompt	Sei unipa-gpt, il chatbot e assistente virtuale dell'Università degli Studi di Palermo. Rispondi cordialmente e in forma colloquiale alle domande che ti vengono poste. Se ricevi un saluto, rispondi salutando e presentandoti. Se ricevi una domanda riguardante l'università degli studi di Palermo, rispondi in base ai documenti che ti vengono dati insieme alla domanda. Se non sai rispondere, scusati e suggerisci di consultare il sito web, non inventare risposte. Question: {question} Documenti: {context}
condensed prompt	Data la seguente conversazione e la domanda successiva, riformula la domanda successiva in modo tale sia una domanda singola. Conversazione: {chat_history} Domanda successiva: {question} Domanda singola:

GPT that uses `gpt-3.5-turbo` as generative LLM, with other versions of the system, that were built till June 2024, and use more recent generative LLM architectures. In the last experimental phase, a suitable test set has been adopted, that consists of a subset of the questions posed directly by the users during the SHARPER night. In the first two scenarios, we used a qualitative metric for evaluation, since the principal intent was to investigate the preferences of the users as regards the feedback provided by the chat-bot during the interaction. Quantitative evaluation was performed in the third scenario, as we manually distilled the *golden labels* for our test set from the true questions of users. Of course, the University domain did not required a domain expert for this task. `gpt-4-turbo` was used as external oracle for computing metrics with the RAGAS[7] framework (Es et al. 2024).

5.1 Setting up the system

This section reports the results of the very early evaluation runs performed using both versions of Unipa-GPT, namely the RAG-only and the fine-tuned one, before deploying the system to the public during the SHARPER Night 2023. These first experiments were aimed at gaining at least qualitative information on the structure of the prompt to be used for querying the system. Moreover, we wanted to assess the best performing Unipa-GPT version, provided that the size of the data set was very small when compared to the training data used for `gpt-3.5-turbo`.

Since the chat-bot was designed to answer questions posed by secondary school students, we devised two lists of Italian questions that had not been showed in advance to the system, and were issued by two secondary school students, during an Open Day activity at the University of Palermo. The students were asked to chat with the demo version of Unipa-GPT, and to express a qualitative evaluation of the received answers. The argument of the questions was the same for both students to collect their qualitative judgement on the Unipa-GPT performance.

The first list, that we called Chat1, consists of 14 questions about subscription procedure, courses at the Department of Engineering, suggestions related to the Bachelor course to apply for in order to access a specific Master course, information about University admission, test and questions about classes in the first year. Conversely, the second one, called Chat2, contains 10 questions with typos and grammatical errors. Questions in Chat2 are expressed more like a Google search or a list of keywords. The involved topics in Chat2 are related to taxes, Erasmus programs, scholarships and Student Desk.

Both Chat1 and Chat2 were issued using different Unipa-GPT configurations as regards the corpus used in the retriever (`clear`, `full`, `emb`) the prompt (*custom prompt*, *condensed prompt*) and the LLM training procedure (`gpt-3.5-turbo`, `gpt-3.5-turbo-fine-tuned`) for a total of 12 runs. Each student was asked to simply classify each answer as "excellent", "good" and "bad". Answers with "good" label are not optimal and precise ones, but they contain a partially good answer, and may add non requested pieces of information that are less related to the initial question. Answers with "bad" are either answers that provide wrong information or out of topic replies along with any unexpected/undesired behavior of the system. Also, replies to questions that the system cannot answer are considered bad. Both Chat1 and Chat2 contain questions that are strictly related with the previous answers from the system:this is

7 `https://docs.ragas.io/en/latest/index.html\#`

Table 4
Evaluation results of the different runs for both Chat1 and Chat2. Here and in the next tables FT stands for "fine-tuned".

Custom prompt				***Custom prompt FT***			
	excellent	good	bad		excellent	good	bad
Chat1-`emb`	6	1	7	Chat1-`emb`	4	6	4
Chat1-`full`	6	4	4	Chat1-`full`	4	4	6
Chat1-`clear`	8	4	2	Chat1-`clear`	6	4	4
Chat2-`emb`	4	2	4	Chat2-`emb`	5	1	4
Chat2-`full`	3	3	4	Chat2-`full`	6	3	1
Chat2-`clear`	8	1	1	Chat2-`clear`	6	1	3
Condensed prompt				***Condensed prompt FT***			
	excellent	good	bad		excellent	good	bad
Chat1-`emb`	5	1	8	Chat1-`emb`	5	2	7
Chat1-`full`	10	2	2	Chat1-`full`	4	1	10
Chat1-`clear`	10	1	3	Chat1-`clear`	6	0	8
Chat2-`emb`	6	2	2	Chat2-`emb`	5	1	4
Chat2-`full`	8	2	0	Chat2-`full`	6	1	3
Chat2-`clear`	9	0	1	Chat2-`clear`	8	0	2

the case when the user asks for further information. As an example, if the user asks for the admission test to a particular Bachelor course, and the system replies that the course is a free admission one, the (expected) question regarding test rules is not applicable thus it is not posed to the system, and the missed reply is labeled as a "bad" answer. Table 4 shows the evaluation results.

The two best performing models use both the *condensed prompt* and `gpt-3.5-turbo`, while they differ for the corpus because they use `unipa-corpus-full` and `unipa-corpus-clear` respectively for the retrieval. We selected the one which uses the `-clear` corpus because this is the one with the highest number of "excellent" labels independently of the prompt (18 labels in total). The lack of additional information regarding classes' educational objectives in the retrieval corpus, makes it more precise. On the other hand, the use of the *condensed prompt* produces a less degrading conversation compared with the use of the *custom prompt* where the conversation is passed as it is, without refinements. None of the models performs bad for the entire conversation.

No improvements in performances were found in the fine-tuned models, revealing that this strategy is not effective for the target task, and a performance decrease is observed in the *condensed prompt* configurations. Generally speaking, answers of the fine-tuned models are very long and they tend to be repetitive (Holtzman et al. 2020) and to generate loops and hallucinations (Ji et al. 2023). Fine-tuned models enrich their answers also with extra pieces of information that sometimes are pertinent, e.g. the question regarding taxes suggests the possibility for payments reduction as reported in Table 5 (English version in Table 2 in Appendix). In other cases, answers are completely out of topic or the chat-bot contradicts itself, as shown in Table 6 (English version in Table 3 in Appendix).

With reference to the prompts, the *condensed prompt* performs best, along with both the `unipa-corpus-full` and the `unipa-corpus-clear` corpus for the retrieval.

Table 5
An extract of Chat2 is reported were the same question is made to systems that use *condensed prompt* and the `unipa-corpus-emb` for the retrieval. Note that the output of the second model is not entirely reported due its length: in the missing text the model continued generating other cases of taxes reduction.

condensed-emb	*condensed-FT-emb*
cos'è il contributo omnicomprensivo	
Il contributo omnicomprensivo per l'immatricolazione all'Università degli Studi di Palermo per l'anno accademico 2023/24 è calcolato in base alla classe di reddito ISEE dello studente e alle prestazioni erogate nell'ambito del diritto allo studio universitario. Il pagamento delle tasse e del contributo onnicomprensivo è suddiviso in tre rate, la prima delle quali deve essere pagata entro il 30 settembre 2023 per l'immatricolazione ai corsi di studio ad accesso libero e per le iscrizioni ad anni successivi al primo. Per conoscere i termini e gli importi del pagamento delle tasse e del contributo omnicomprensivo, si consiglia di consultare il Calendario didattico e il Regolamento sulla contribuzione studentesca vigenti. In caso di dubbi o domande specifiche, si suggerisce di contattare direttamente la Segreteria Studenti dell'Università degli Studi di Palermo.	Il contributo omnicomprensivo è il pagamento delle tasse universitarie che gli studenti devono effettuare all'atto dell'immatricolazione o iscrizione. Il pagamento è suddiviso in tre rate: la prima rata, da versare all'atto dell'immatricolazione o iscrizione, nella misura del 40% dell'importo dovuto; la seconda rata, da versare entro il 29 dicembre 2023, nella misura del 30% dell'importo dovuto; la terza rata, da versare entro il 29 febbraio 2024, nella misura del 30% dell'importo dovuto. Gli studenti che si iscrivono per la prima volta alla Laurea triennale o Laurea Magistrale a Ciclo Unico, diplomatisi con voto compreso tra 95 e 100, usufruiscono, per il primo anno, di una riduzione del 25% fino ad un massimo di 350,00 euro. Coloro che si diplomano con voto pari a 100 e lode usufruiscono, per il primo anno, di una riduzione del 50% fino un massimo di 500,00 euro. [...]

Table 6
An extract of Chat1 is reported where the same question is made to systems that use the *custom promt*. Note that the fine-tuned model contradicts itself.

custom-full	*custom-FT-clear*
la triennale in ingegneria meccanica è a numero chiuso o a numero aperto?	
La triennale in ingegneria meccanica presso l'Università degli Studi di Palermo è a numero aperto, con un numero di posti in programmazione locale pari a 200. Questo è indicato nel documento relativo all'anno accademico 2023/2024 per il curriculum meccanico. Se hai bisogno di ulteriori informazioni, ti suggerisco di consultare il sito web dell'università	La triennale in ingegneria meccanica è a numero aperto.
devo fare test per entrare alla triennale in ingegneria meccanica?	
Sì, per accedere alla triennale in ingegneria meccanica all'Università degli Studi di Palermo è necessario sostenere un test di accesso. Il Centro Orientamento e Tutorato (COT) dell'Università organizza diverse iniziative per affrontare al meglio i test di accesso e superarli con successo, come i workshop "Come affrontare il test di accesso" e la simulazione dei test di accesso. Inoltre, sono disponibili corsi di preparazione alle prove di accesso che inizieranno il 17 luglio 2023 e si svolgeranno esclusivamente online sulla piattaforma Microsoft Teams. Per maggiori informazioni, ti consiglio di consultare i documenti forniti insieme alla domanda.	Sì, sono previsti test per entrare alla triennale in ingegneria Meccanica.

Both corpora provide very pertinent and similar answers. On the contrary the models with the *custom prompt* tend to be less precise, as reported in Table 7 (English version in Table 4 in Appendix).

Table 7
An extract of Chat1 is reported where the same question is made to system that use `unipa-corpus-clear` as the retrieval corpus. Both the answers are correct but the one reported by the RAG-only model is more precise.

condensed-clear	*custom-FT-clear*
quale corso triennale mi consigli di frequentare per poi passare alla magistrale di aerospaziale?	
Per passare alla laurea magistrale in Ingegneria Aerospaziale all'Università degli Studi di Palermo, il corso triennale consigliato è la Laurea in Ingegneria Meccanica con curriculum Aeronautico. Ti consiglio di consultare il piano di studi del corso di laurea triennale e di contattare il Dipartimento di Ingegneria per ulteriori informazioni.	Ti consiglio di frequentare il corso di Laurea triennale in Ingegneria Meccanica per poi passare alla magistrale in Ingegneria Aerospaziale. Il corso triennale ti fornirà le basi necessarie per affrontare al meglio il percorso magistrale.

All models share the same negative aspects: all the provided answers are well written in Italian, and an uninformed user can easily consider a wrong answer as truthfully by looking at both the lexicon and the grammar used by the system. Sometimes, the

prompt structure makes the system to answer with an out of context presentation of Unipa-GPT, and a repetition of the question is necessary to obtain the expected reply. In this situation, prompt tuning plays a key role in achieving the best performance (Brown et al. 2020; Zhao et al. 2021). The reported results are the optimal ones after several tests with different prompts. A test with an English-written prompt was performed but it did not show significant improvements.

Unipa-GPT performs in free QA domain where the user can be anyone with different background, thus questions submitted to the system are deeply variable; we found that the developed models result more sensible and perform better or worse depending on the specific question. In Chat2, questions about scholarships and classes were made as a request and not as a question. This led the system either to answer badly or not to provide an answer at all by making its excuses, as reported in Table 8 (English version in Table 5 in Appendix). In Table 8 the behavior of the fine-tuned version of `gpt-3.5-turbo` is also reported with the same prompt used in training phase and without the access to `unipa-corpus`, to test the effectiveness of fine-tuning: the fine-tuned model provides an answer to both the proposed questions (even to the no-question) but both are wrong, despite they are credible.

Table 8
An extract of Chat2 is reported where the same questions are posed to systems that use the general `gpt-3.5-turbo` and *condensed prompt*, system on the left retrieves documents from the `unipa-corpus-emb`, while the one at the centre uses `unipa-corpus-full`. At the right, the same questions were proposed to the fine-tuned version of `gpt-3.5-turbo`: the prompt given is the same used in training phase and no access was provided to `unipa-corpus`.

condensed-emb	*condensed-full*	*pure fine-tuned model*
lista esami primo anno di medicina		
Gli esami del primo anno del corso di medicina all'Università degli Studi di Palermo per l'anno accademico 2023/24 non sono specificati nei documenti forniti. Ti suggerisco di consultare i bandi di concorso per maggiori informazioni. Inoltre, per quanto riguarda l'immatricolazione, i corsi di studio ad accesso libero possono essere immatricolati senza mora dall'1 agosto [...]	Mi dispiace, ma non ho trovato informazioni riguardanti gli esami del primo anno del corso di medicina all'Università degli Studi di Palermo per l'anno accademico 2023/24. Ti consiglio di consultare i bandi di concorso per maggiori informazioni.	La lista degli esami del primo anno del corso di laurea in Medicina e Chirurgia presso l'Università di Palermo prevede i seguenti esami: Anatomia Umana con Esercitazioni, Biologia Applicata, Lingua Inglese, Chimica Biologica con Elementi di Biologia Molecolare, Fisica, Matematica, Statistica e Informatica, Istologia Umana con Elementi di Biologia Cellulare, Fisiologia Umana con Esercitazioni, e Patologia Generale.
quali sono le materie del primo anno di medicina		
Le materie del primo anno del corso di medicina all'Università degli Studi di Palermo per l'anno accademico 2023/24 sono: Fisica I, Matematica I e Chimica Generale ed Organica C.I. Invece, per il corso di laurea in Scienze Biologiche, le materie del primo anno sono: Matematica con esercitazioni, Citologia e Istologia con esercitazioni, Chimica Generale con esercitazioni e Botanica Generale e istematica con esercitazioni.	Le materie del primo anno del corso di medicina all'Università degli Studi di Palermo per l'anno accademico 2023/24 sono: Fisica I, Matematica I, Chimica Generale ed Organica C.I., Chimica Organica, Biologia, Embriologia ed Istologia C.I., Genetica e Microbiologia C.I., Informatica, Bioinformatica e Inglese Tecnico Scientifico C.I. Ti consiglio di consultare i documenti forniti per maggiori informazioni.	Le materie del primo anno di medicina sono: Anatomia Umana, Istologia, Embriologia, Biochimica, Biologia Cellulare, Genetica, Fisica, Chimica, Lingua Inglese, Metodologia del Lavoro Intellettuale, e Cultura e Valori in Sanità.

5.2 The SHARPER night

Unipa-GPT was presented during the European Researchers' Night (SHARPER night) when visitors were allowed to interact with the chat-bot making questions about the University of Palermo. The created interface, shown in Figure 3, was designed to be as user-friendly as possible with an avatar that mimics the answer of the system by moving itself and reproducing an audio version of the answer, thus making it more accessible and interactive. As shown in 3 the chat-bot always starts a conversation displaying a welcome message in Italian, here reported in its English version:

Figure 3
The GUI of Unipa-GPT presented at the SHARPER night.

> Hi! I am Unipa-GPT, a virtual assistant at the University of Palermo. I am not a person, but an artificial intelligence developed to answer your questions about the university. If you need any information, just ask me!

Some users were confident during the interaction, while others needed an external human-guidance to interact with the chat-bot, thus demonstrating that skeptical visitors about AI, in a real-context, are less prone to give a chance to this type of systems. The model that was actually used during the SHARPER Night is the one that uses *condensed prompt* with `gpt-3.5-turbo` and `unipa-corpus-clear` as the retrieval corpus: due to further experiments, few changes were made in the prompt, as shown in Table 9 (English version in Table 6 in Appendix).

During the event, we collected a total of 165 questions that can be divided in seven categories, namely Generic Information, Courses' Information, Other University-related, Off-topic, Services and Structures, Taxes and Scholarships and University Environment (Figure 4.a). On average, the interaction produced a two questions long chat, and we asked each visitor to give a feedback via a Google Form thus collecting 31 replies. The majority of the collected feedback came from either University students or professors, i.e. people that are familiar with the academic world (Figure 4.b) and the general rate of the chat-bot is very positive, despite two bad rankings (Figure 4.c). One bad evaluation came from a conversation where a student asked about a specific parameter in the scholarship rank: in this particular case the model provided a false answer, saying that the asked parameter is not considered for the rank, despite the truth. We suppose that this is a case in which the model made use of an incomplete

Table 9
The prompts used in the system.

prompt type	*prompt text*
custom prompt	Sono Unipa-GPT, chatbot e assistente virtuale dell'Università degli Studi di Palermo che risponde cordialmente e in forma colloquiale. Ai saluti, rispondi salutando e presentandoti; Rispondi alla domanda con la dicitura "Risposta: " Ricordati che il rettore dell'Università è il professore Massimo Midiri. Se la domanda riguarda l'università degli studi di Palermo, rispondi in base alle informazioni e riporta i link ad esse associate; Se non sai rispondere alla domanda, rispondi dicendo che sei un'intelligenza artificiale che ha ancora molto da imparare e suggerisci di andare su https://www.unipa.it/, non inventare risposte. Domanda: {question} Informazioni: {context}
condensed prompt	Data la seguente conversazione e la domanda successiva, riformula la domanda successiva in modo tale sia una domanda singola. Conversazione: {chat_history} Domanda succesiva: {question} Domanda singola:

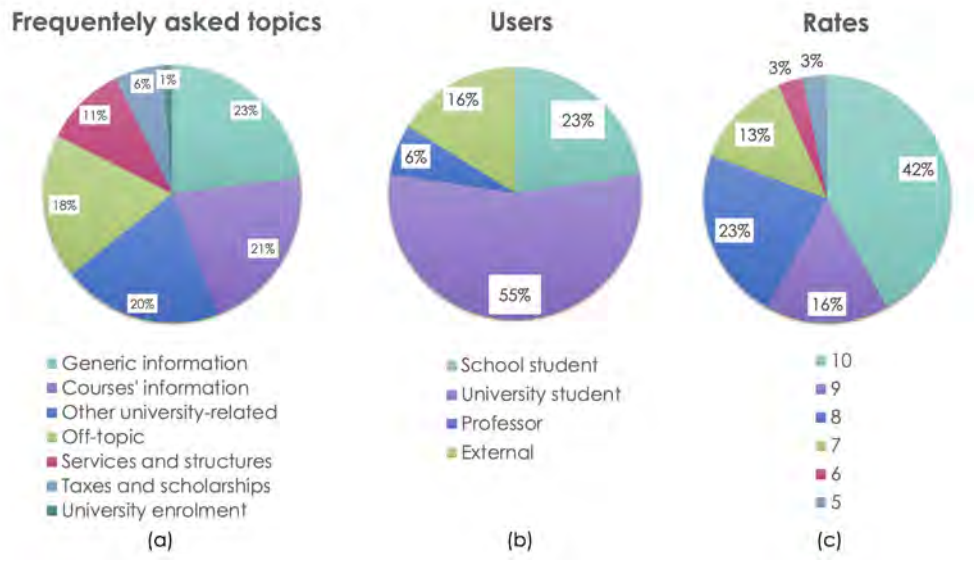

Figure 4
Charts illustrate the frequently asked topics to the model during the SHARPER Night (a); among the 31 feedback collected, an overview of the users is reported (b) and the rates to the application (c).

document. The second bad evaluation came from a conversation where information about the final test for a bachelor course were asked to the model. Here the problem is the bachelor course name: the user did not used the complete official course name but the one to which it is usually referred to. Moreover, bachelor and master courses share the same name, and this may cause an ambiguity in the model that can be fixed by adding external information. In this case, the final test is peculiar of the bachelor

course while thesis refers to the master course: we assume that if this kind of general knowledge is provided to the model, in addition to the document, the correct answer may be generated.

5.3 gpt-3.5-turbo vs other open LLMs

During the first semester of 2024 we carried out more experiments to evaluate the quality of the answers provided by Unipa-GPT, and we compared the performances of `gpt-3.5-turbo`, considered as a baseline model, with other LLMs averaging from 3B to 8B parameters, considering both foundation and fine-tuned models. The choice of the models sizes relies only on computational constraints as we did not have access to any cloud with HPC facility at that time, and we used local resources.

We selected one question for each typology reported in Figure 4.a) picking them from the real ones posed by the users to Unipa-GPT during the SHARPER Night, and excluding the off-topic requests. We built manually a golden answer to each question as it is shown in Table 10 (Table 7 in Appendix for the English version).

The following foundation models were selected: `Llama-2` (Touvron et al. 2023), `LLaMAntino-2` (Basile et al. 2023), `Llama-3` (Llama Team 2024), `Minerva` (Orlando et al. 2024).

We considered both chat-oriented and purely fine-tuned models: `Llama-2-chat` (Touvron et al. 2023), `LLaMAntino-2-chat` (Basile et al. 2023), `Fauno` (Bacciu et al. 2023), and instruction-tuned models: `Camoscio` (Santilli and Rodolà 2023), `Llama-3-instruct` (Llama Team 2024), and `LLaMAntino-3-ANITA` (Polignano, Basile, and Semeraro 2024) henceforth referred to as `Anita-3`.

All models, except for `Minerva`, that is a `Mistral`-based model (Jiang et al. 2023), are `Llama`-based ones. In particular, `LLaMAntino-2`, `Fauno` and `Camoscio` are fine-tuned version of `Llama-2`, while `LLaMAntino-2-chat` is a build upon `Llama-2-chat`, and `Anita-3` is a fine-tuned version of `Llama-3-instruct`.

Evaluation was performed through instruction prompting using the *custom prompt*, and models were forced to generate a maximum of 256 new tokens. Except for `gpt-3.5-turbo`, all the evaluation runs were performed on a 48 GB NVIDIA RTX 6000 Ada Generation without any quantization. Even if the same prompt was used for all the models, it was provided differently to each of them following its authors' specific prompt strategies for either fine-tuning or inference. As an example, `### Istruzione:` was added before the prompt for `Camoscio`, `Minerva-3B`, `LLaMAntino-2`, `Llama-3` and `Llama-2`, while the `[/INST]` tag was used for `Llama-2-chat` . Moreover, exactly the same retriever was used in all the runs in order to provide each model with the same documents of the othr ones, thus measuring its capability to paraphrase, and re-elaborate the provided context for generating a helpful answer.

Correctness of the answers with respect to the golden labels was measured both with the BLEU (Papineni et al. 2002) and the ROGUE-L score (Lin 2004). It is worth noticing that BLEU and ROUGE are not specifically tailored for evaluating a text generation task, but they can be considered as a good starting point in automatic evaluation, given a golden answer.

The results are reported in Table 11 and Table 12, and they show clearly that `gpt-3.5-turbo` performs significantly better than both foundation and chat-oriented fine-tuned models that attain 0. Only instruction-tuned models obtain a non-zero BLEU score, but such values remain far low compared with the results for `gpt-3.5-turbo`.

Table 10
List of the six QA pairs used for evaluation purposes

ID	Question	Answer
Q1	*Chi è il professore di Intelligenza Artificiale 1 per il corso di Laurea Magistrale in Ingegneria Informatica?*	Il professore di Intelligenza Artificiale 1 del corso di Laurea Magistrale Magistrale in Ingegneria Informatica è il professore Gaglio e verrà erogata durante il primo semestre. Per maggiori informazioni vai su http://www.unipa.it/struttura.html?id=721
Q2	*Quali sono le scadenze di iscrizione ad un corso di laurea?*	La presentazione delle domande di iscrizione per l'Anno Accademico 2023/2024 varia in base alla tipologia di corso. Per i corsi di studio triennali e magistrali a ciclo unico, la domanda può essere presentata dall'1 agosto al 30 settembre 2023, mentre per i corsi magistrali, dall'1 agosto al 30 novembre 2023, in entrambi i casi, è richiesto il pagamento della prima rata delle tasse universitarie. Per i corsi di studio ad accesso programmato locale, con prenotazione online o ad accesso programmato nazionale, le immatricolazioni vanno effettuate entro i termini previsti dal relativi bandi di concorso.
Q3	*Come funziona la magistrale di chimica?*	Il corso di laurea magistrale in Chimica presso l'Università degli Studi di Palermo, ha sede a Palermo e ha una durata di 2 anni. L'obiettivo del corso è la formazione di laureati con una solida preparazione di base, che li ponga in grado di affrontare con competenza ed autonomia sia attività professionali sia quelle di ricerca accademica ed industriale. Per questo motivo il corso di studio è progettato con lo scopo di completare la cultura scientifica nelle discipline chimiche fondamentali e di introdurre conoscenze più avanzate. Quindi lo studente può completare il proprio piano di studi, attenzionando gli insegnamenti che meglio si adattano ai propri interessi specifici, come la chimica dei materiali, dell'ambiente e dei beni culturali, la chimica supramolecolare e la didattica della chimica. Per maggiori informazioni vai su http://www.unipa.it/struttura.html?id=766
Q4	*ciao! sono un ragazzo appena uscito dal liceo che è interessato al settore legale,in particolare alle leggi sulle aziende. dove potrei iscrivermi?*	Ciao! Il corso di Laurea in Consulente Giuridico d'Impresa sembra proprio fare al caso tuo! Il corso di laurea ha sede a Trapani e ha una durata di 3 anni e ti fornirà le conoscenze in ambito giuridico ed economico-aziendalistico. Una volta terminato il percorso di studio potrai svolgere attività interdisciplinari che richiedono competenze giuridiche, aziendalistiche e organizzativo-gestionali. Per maggiori informazioni vai su http://www.unipa.it/struttura.html?id=1557
Q5	*come posso prenotare un appuntamento in segreteria?*	È possibile recarsi in segreteria il lunedì, mercoledì e venerdì dalle 10.00 alle 12.00, martedì e giovedì dalle 15.00 alle 17.00. Puoi prenotare il tuo turno attraverso la App SolariQ. Per maggiori informazioni vai su https://www.unipa.it/servizi/segreterie/
Q6	*Come si pagano le tasse?*	Il pagamento delle tasse deve essere effettuato esclusivamente mediante sistema PAgoPA (Pagamenti della Pubblica Amministrazione) al quale si accede tramite portale di ateneo. Dopo aver compilato la pratica online, è possibile pagare direttamente online con il sistema PAgoPA o stampare il bollettino e pagare presso una ricevitoria abilitata PAgoPA. Ulteriori informazioni sul pagamento via PAgoPA sono reperibili qui https://immaweb.unipa.it/immaweb/public/pagamenti.seam, mentre è disponibile il Regolamento in materia di contribuzione studentesca https://www.unipa.it/servizi/segreterie/.content/documenti/regolamenti_calendari/2023/5105144-def_regolamento-contribuzione–studentesca-2023—24-2.pdf

Results in Table 12 for the ROUGE-L score confirm the aforementioned comments for BLEU: `gpt-3.5-turbo` performs significantly better compared to other models, and instruction-tuned models reach the highest results in comparison with the other `Llama`- and `Mistral`-based ones.

Table 11
Results with BLEU score. Bold values are the higher ones.

	Q1	Q2	Q3	Q4	Q5	Q6
Baseline models						
gpt-3.5-turbo	**0.334**	**0.151**	**0.309**	**0.213**	**0.207**	**0.197**
Foundation models						
Llama-2	0.0	0.0	0.0	0.0	0.0	0.0
LLaMantino-2	0.0	0.0	0.0	0.0	0.0	0.0
Llama-3	0.0	0.0	0.0	0.0	0.0	0.0
Minerva-3B	0.0	0.0	0.0	0.0	0.0	0.0
Chat-oriented fine-tuned models						
Llama-2-chat	0.0	0.0	0.0	0.0	0.0	0.0
LLaMantino-2-chat	0.0	0.0	0.0	0.0	0.0	0.0
Fauno	0.0	0.0	0.0	0.0	0.0	0.0
Instruction-tuned models						
Camoscio	0.0	0.0	0.0	0.0	0.019	0.0
Llama-3-instruct	0.047	0.0	0.114	0.0	0.091	0.066
Anita-3	0.048	0.0	0.066	0.063	0.0	0.068

Table 12
Results with ROUGE-L score. Bold values are the higher ones.

	Q1	Q2	Q3	Q4	Q5	Q6
Baseline models						
gpt-3.5-turbo	**0.472**	**0.295**	**0.423**	**0.301**	**0.187**	**0.284**
Foundation models						
Llama-2	0.054	0.049	0.007	0.010	0.020	0.008
LLaMantino-2	0.035	0.042	0.050	0.014	0.045	0.044
Llama-3	0.052	0.012	0.020	0.051	0.020	0.009
Minerva-3B	0.023	0.030	0.0	0.0	0.054	0.043
Chat-oriented fine-tuned models						
Llama-2-chat	0.010	0.009	0.013	0.009	0.008	0.019
LLaMantino-2-chat	0.0	0.05	0.039	0.0	0.0	0.0
Fauno	0.013	0.046	0.013	0.011	0.033	0.047
Instruction-tuned models						
Camoscio	0.010	0.016	0.019	0.019	0.031	0.017
Llama-3-instruct	0.109	0.052	0.131	0.091	0.092	0.074
Anita-3	0.133	0.124	0.118	0.078	0.096	0.089

We also assessed the answers provided by each model through manual inspection. A tendency comes up to generate correct English answers (`Llama-2-chat`), out of topic answers bot in Italian (`LLaMantino-2` models and `Minerva`) and other languages (`Llama-2` and `Camoscio`) or including code (`Fauno` and `Llama-3`).

`Llama-3-instruct` and `Anita-3` generate the most satisfactory answers, since they provide topic-related answers in Italian apart from few English ones.

Further evaluations were carried out using RAGAS to compute *context relevancy*, *faithfulness*, and *answer correctness*. The RAGAS metrics need a "judge", and we opted for `gpt-4-turbo` as is the most performative model we had access to at the time of the experiments.

Context relevancy is measured based on both the question itself and the retrieved documents that form the context. The values fall within the range of (0, 1) with higher values indicating better relevancy. Results are reported in Table 13, where the scores in each row refer to the first four documents retrieved by the RAG to answer the corresponding question. We starred the values obtained for the "golden documents", that is the ones used for generating each of the golden answers, in order to highlight divergences. The results are extremely low, and a correct retrieval (i.e. the document with highest context relevancy coincides with a golden document) is verified in just three out of the six considered questions. We argue that this finding could be caused by the huge documents' size compared with the length of the question, and the actual relevant piece of information.

Table 13
Evaluation results with RAGAS context relevancy metric. Starred results refer to the document that was used for manual generation of the respective golden answer, while bold values refer to the highest scores.

	D1	D2	D3	D4
Q1	***0.1***	**0.1**	0.083	0.071
Q2	*0.091*	0.125	***0.625***	0.333
Q3	*0.1*	**0.111**	**0.111**	**0.111**
Q4	***0.167***	*0.059*	0.1	0.333
Q5	*0.429*	**0.5**	0.143	0.111
Q6	**1.0**	*0.1*	0.167	0.5

The faithfulness measures the factual consistency of the generated answer against the given context, while answer correctness involves gauging the accuracy of the generated answer when compared to the ground truth. Both metrics range from 0 to 1 and better performances are associated with higher scores. The results are reported in Table 14 and 15 for the faithfulness and the answer correctness, respectively.

The highest faithfulness results are reached by `gpt-3.5-turbo`, followed by `Llama-3-instruct`, and `Anita-3`, confirming the previous results for both BLEU and ROUGE. The two models performed far better than `gpt-3.5-turbo` in Q5. Similar behavior can be found in answer correctness, where `Anita-3` outperformed `gpt-3.5-turbo` in Q4, while `Llama-3-instruct` received the same score as `gpt-3.5-turbo` in Q2. After manual inspection we found that the answers generated by `Anita-3` convey the correct meaning, but they are not well spelled, and are grammatically wrong, thus being penalized both with traditional metrics and with RAGAS ones.

The presented results show clearly that `gpt-3.5-turbo` was a SOTA model for the time Unipa-GPT was released, and in general the GPT family is an ideal one for the kind of architecture we presented in this work. We can not exclude that 13B to

Table 14
Evaluation results with RAGAS faithfulness metric. Bold values are the highest ones.

	Q1	Q2	Q3	Q4	Q5	Q6
Baseline models						
gpt-3.5-turbo	**1.0**	**1.0**	**1.0**	**1.0**	0.333	**0.875**
Foundation models						
Llama-2	0.083	0.0	0.0	0.0	0.0	0.0
LLaMantino-2	0.0	0.0	0.0	0.0	0.0	0.0
Llama-3	0.2	0.5	0.0	0.0	0.0	0.0
Minerva-3B	0.0	0.0	0.0	0.0	0.0	0.0
Chat-oriented fine-tuned models						
Llama-2-chat	0.0	0.2	0.0	0.0	0.0	0.0
LLaMantino-2-chat	0.0	0.0	0.0	0.0	0.0	0.0
Fauno	0.0	0.0	0.0	0.0	0.25	0.286
Instruction-tuned models						
Camoscio	0.0	0.0	0.0	0.0	0.0	0.0
Llama-3-instruct	0.143	**1.0**	0.25	0.0	0.667	0.0
Anita-3	0.429	0.6	0.667	0.444	**1.0**	0.333

Table 15
Evaluation results with RAGAS answer correctness metric. Bold values are the highest ones.

	Q1	Q2	Q3	Q4	Q5	Q6
Baseline models						
gpt-3.5-turbo	**0.616**	**0.727**	**0.563**	0.401	**0.531**	**0.666**
Foundation models						
Llama-2	0.21	0.183	0.175	0.173	0.18	0.29
LLaMantino-2	0.191	0.186	0.18	0.188	0.186	0.187
Llama-3	0.385	0.429	0.199	0.199	0.191	0.184
Minerva-3B	0.183	0.186	0.176	0.189	0.195	0.203
Chat-oriented fine-tuned models						
Llama-2-chat	0.175	0.314	0.179	0.196	0.19	0.216
LLaMantino-2-chat	0.187	0.184	0.182	0.188	0.188	0.191
Fauno	0.175	0.213	0.152	0.162	0.182	0.219
Instruction-tuned models						
Camoscio	0.287	0.175	0.173	0.171	0.18	0.181
Llama-3-instruct	0.396	0.22	0.322	0.217	0.224	0.605
Anita-3	0.395	0.362	0.541	**0.444**	0.515	0.224

70B models built on both the `Llama` and the `Mistral` foundation ones using either intstuction tuning or explicit fine-tuning could reach comparable performances.

6. Conclusions and future works

In this paper we presented Unipa-GPT, a virtual assistant capable of answering to secondary school students who need information for accessing and studying at the University of Palermo. The developed system relies on a RAG architecture that uses documents from a corpus purposely scraped from the University institutional website and comes in two versions that make use of either `gpt-3.5-turbo` or a fine-tuned model `gpt-3.5-turbo-fine-tuned` where the corpus has been reduced to keep low the computational resources needed for fine-tuning. Moreover, different versions of the system have been tested as regards both the prompt and the strutcure of the documents used in the RAG corpus. Significant improvements were not found in the fine-tuned model, and the best performing system was the one that uses the so called *condensed prompt* where the previous conversation and the current question are reformulated into a unique prompt. Such a prompt induces `gpt-3.5-turbo` to summarize the conversation at each question, and then it behaves as it was instructed instructed using our base *custom prompt* which is tailored for the application purposes. Moreover, this system uses the `unipa-corpus-clear` for the retrieval where the specifico educational objectives for each class are not reported. We argue that this "light" version of the corpus provides more compact and precise information to the LLM thus generating the best answers. The presented model has been compared against several open LLMs based on both the `Llama` and the `Mistral` foundation ones using different metrics, and the results proved that the `gpt-` family outperforms the others in this task. At the time of writing the manuscript, we moved to the integration of `gpt-4.o-` models in the Unipa-GPT framework. Future work will be devoted to build an open framework where all kinds of models can be integrated and tested. Finally, it is well known that integrating structured knowledge in RAG can mitigate hallucination (Casheekar et al. 2024). Following this assumption, we are developing a hybrid RAG for the Unipa-GPT framework, that relies on both a vector store, and either a knowledge graph obtained by document indexing or a formal OWL ontology if available to describe the domain under investigation.

Acknowledgments

We thank all the CHILab team, in particular Dr. Salvatore Contino and Dr. Paolo Sortino, and PhD students Luca Cruciata and Gaetano Pottino, that contributed in generating the corpus with the scraping and implemented the demonstration interface.

References

Bacciu, Andrea, Giovanni Trappolini, Andrea Santilli, Emanuele Rodolà, and Fabrizio Silvestri. 2023. Fauno: The italian large language model that will leave you senza parole! *arXiv preprint arXiv:2306.14457*.

Basile, Pierpaolo, Elio Musacchio, Marco Polignano, Lucia Siciliani, Giuseppe Fiameni, and Giovanni Semeraro. 2023. Llamantino: Llama 2 models for effective text generation in italian language. *arXiv preprint arXiv:2312.09993*.

Bhat, Vani, Sree D. Cheerla, Jinu R. Mathew, Nupur Pathak, Guannan Liu, and Jerry Gao. 2024. Retrieval augmented generation (rag) based restaurant chatbot with ai testability. In *2024 IEEE 10th International Conference on Big Data Computing Service and Machine Learning Applications (BigDataService)*, pages 1–10, Shanghai, China, July.

Borji, Ali. 2023. A categorical archive of chatgpt failures. *arXiv preprint arXiv:2302.03494*.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Casheekar, Avyay, Archit Lahiri, Kanishk Rath, Kaushik S. Prabhakar, and Kathiravan Srinivasan. 2024. A contemporary review on chatbots, ai-powered virtual conversational

agents, chatgpt: Applications, open challenges and future research directions. *Computer Science Review*, 52:100632.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Dwivedi, Yogesh K., Nir Kshetri, Laurie Hughes, Emma L. Slade, Anand Jeyaraj, Arpan K. Kar, Abdullah M. Baabdullah, Alex Koohang, Vishnupriya Raghavan, Manju Ahuja, et al. 2023. "so what if chatgpt wrote it?" multidisciplinary perspectives on opportunities, challenges and implications of generative conversational ai for research, practice and policy. *International Journal of Information Management*, 71:102642.

Es, Shahul, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. RAGAs: Automated evaluation of retrieval augmented generation. In Nikolaos Aletras and Orphee De Clercq, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158, St. Julians, Malta, March. Association for Computational Linguistics.

Forootani, Ali, Danial E. Aliabadi, and Daniela Thraen. 2024. Bio-eng-lmm ai assist chatbot: A comprehensive tool for research and education. *arXiv preprint arXiv:2409.07110*.

Gill, Sukhpal S., Minxian Xu, Panos Patros, Huaming Wu, Rupinder Kaur, Kamalpreet Kaur, Stephanie Fuller, Manmeet Singh, Priyansh Arora, Ajith K. Parlikad, et al. 2024. Transformative effects of chatgpt on modern education: Emerging era of ai chatbots. *Internet of Things and Cyber-Physical Systems*, 4:19–23.

Holtzman, Ari, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*, Addis Ababa, Ethiopia, April.

Hu, Edward J., Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, Online, April.

Ji, Ziwei, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Survey*, 55(12), mar.

Jiang, Albert Q., Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra S. Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Johnson, Jeff, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Lewis, Patrick, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Lin, Chin-Yew. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July. Association for Computational Linguistics.

Llama Team, AI @ Meta. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Lo, Chung K. 2023. What is the impact of chatgpt on education? a rapid review of the literature. *Education Sciences*, 13(4).

OpenAI. 2024. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Orlando, Riccardo, Luca Moroni, Pere-Lluís H. Cabot, Edoardo Barba, Simone Conia, Sergio Orlandini, Giuseppe Fiameni, and Roberto Navigli. 2024. Minerva llms: The first family of large language models trained from scratch on italian data. *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), Pisa, Italy*, December.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July.

Polignano, Marco, Pierpaolo Basile, and Giovanni Semeraro. 2024. Advanced natural-based interaction for the italian language: Llamantino-3-anita. *arXiv preprint arXiv:2405.07101*.

Rahman, Md. Mostafizer and Yutaka Watanobe. 2023. Chatgpt for education and research: Opportunities, threats, and strategies. *Applied Sciences*, 13(9).
Rodriguez-Torrealba, Ricardo, Eva Garcia-Lopez, and Antonio Garcia-Cabot. 2022. End-to-end generation of multiple-choice questions using text-to-text transfer transformer models. *Expert Systems with Applications*, 208:118258.
Santilli, Andrea and Emanuele Rodolà. 2023. Camoscio: an italian instruction-tuned llama. *arXiv preprint arXiv:2307.16456*.
Sharma, Sudhansh and Ramesh Yadav. 2023. Chat gpt – a technological remedy or challenge for education system. *Global Journal of Enterprise Information System*, 14(4):46–51, May.
Tack, Anaïs and Chris Piech. 2022. The AI teacher test: Measuring the pedagogical ability of blender and GPT-3 in educational dialogues. In Antonija Mitrovic and Nigel Bosch, editors, *Proceedings of the 15th International Conference on Educational Data Mining*, pages 522–529, Durham, United Kingdom, July. International Educational Data Mining Society.
Touvron, Hugo, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
Vakayil, Sonia, D. Sujitha Juliet, Anitha J., and Sunil Vakayil. 2024. Rag-based llm chatbot using llama-2. In *2024 7th International Conference on Devices, Circuits and Systems (ICDCS)*, pages 1–5, Coimbatore, India, April. IEEE.
Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
Vidivelli, S., Manikandan Ramachandran, and A. Dharunbalaji. 2024. Efficiency-driven custom chatbot development: Unleashing langchain, rag, and performance-optimized llm fusion. *Computers, Materials & Continua*, 80(2).
Wang, Calvin, Joshua Ong, Chara Wang, Hannah Ong, Rebekah Cheng, and Dennis Ong. 2024. Potential for gpt technology to optimize future clinical decision-making using retrieval-augmented generation. *Annals of Biomedical Engineering*, 52(5):1115–1118.
Weizenbaum, Joseph. 1966. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45, January.
Zhao, Zihao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706, Online, July. PMLR.

Appendix A: Translated tables

Table 1
English version of Table 3.

prompt type	*prompt text*
custom prompt	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, apologise and suggest that you consult the website, do not invent answers. Question: {question} Documents: {context}
condensed prompt	Given the following conversation and question, rephrase the next question in such a way that it is a single question. Conversation: {chat_history} Next question: {question} Single question:

Table 2
English version of Table 5.

condensed-emb	*condensed-FT-emb*
what the all-inclusive fee is	
The all-inclusive fee for enrolment at the University of of Palermo for the academic year 2023/24 are calculated according to the student's ISEE income class and the benefits provided as part of the right to study university. The payment of fees and the all-inclusive fee is divided into in three instalments, the first of which must be paid by 30 September 2023 for enrolment in open access study courses and for enrolment to years subsequent to the first. To find out the terms and amounts of payment of fees and the all-inclusive fee, we advise you to consult the Teaching Calendar and the Calendar and the Student Contribution Regulations in force. In case of any doubts or specific questions, contact the Student secretariat directly of the University of Palermo.	The all-inclusive fee is the payment of the tuition fees that students have to pay at the time of university enrolment or registration. Payment is divided into three instalments: the first instalment, payable during the registration or enrolment, is the 40% of the amount due; the second instalment, to be paid by 29 December 2023, is the 30% of the amount due; the third instalment, to be paid by 29 February 2024, is the 30% of the amount due. Students enrolling for the first time to a three-year degree or a single-cycle master's degree, who graduated with a grade between 95 and 100, benefit, for the first year, from a reduction of 20% up to a maximum of €350.00. Those who graduate with a grade of 100 cum laude receive a 50% reduction for the first year, up to a maximum of a maximum of 500.00 euro. [...]

Table 3
English version of Table 6.

custom-full	*custom-FT-clear*
is the bachelor degree course in mechanical engineering open or restricted access?	
The bachelor degree course in mechanical engineering at the University of Palermo is open-access, with a number of places in local programming equal to 200. This is indicated in the document for the academic year 2023/2024 for the mechanics curriculum. If you need further information, I suggest you to consult the university website	The bachelor degree course in mechanical engineering is open-access
do i have to take a test to enter the bachelor degree course in mechanical engineering?	
Yes, to enter the bachelor degree course in mechanical engineering at the University of Palermo, you must take an admission test. The University's Guidance and Tutoring Center (COT) organises various initiatives to help you cope with entrance tests and pass them successfully, such as the 'How to cope with the entrance test' workshops and mock entrance tests. In addition, preparation courses are available for the admission tests, which will start on 17 July 2023 and will take place exclusively online on the Microsoft Teams platform. For more information, I advise you to consult the documents provided with your application.	Yes, there are tests for entry to the bachelor degree course in Mechanical Engineering.

Table 4
English version of Table 7

condensed-clear	*custom-FT-clear*
which three-year course do you recommend I attend to then go on to the master's degree in aerospace engineering?	
To move on to the master's degree in Aerospace Engineering at the University of Palermo, the recommended three-year course is the Bachelor's degree in Mechanical Engineering with Aeronautical curriculum. I advise you to consult the study plan for the three-year degree course and contact the Department of Engineering for further information.	I recommend that you attend the three-year degree course in Mechanical Engineering and then move on to the master's degree course in Aerospace Engineering. The three-year course will provide you with the necessary foundations for the master's degree.

Table 5
English version of Table 8.

condensed-emb	*condensed-full*	*pure fine-tuned model*
first-year medicine course exam list		
The exams for the first year of the medicine course at the University of Palermo for the academic year 2023/24 are not specified in the documents provided. I suggest you consult the call for applications for more information. Furthermore, as far as enrolment is concerned, open access courses can be enrolled without late fees from 1 August [...]	I am sorry, but I have not found any information regarding the first-year exams of the medicine course at the University of Palermo for the academic year 2023/24. I advise you to consult the call for applications for more information.	The list of exams for the first year of the degree course in Medicine at the University of Palermo includes the following exams: Human Anatomy with Exercises, Applied Biology, English, Biological Chemistry with Elements of Molecular Biology, Physics, Mathematics, Statistics and Computer Science, Human Histology with Elements of Cell Biology, Human Physiology with Exercises, and General Pathology.
what are the subjects of the first year of medicine		
The first-year subjects of the medical course at the University of Palermo for the 2023/24 academic year are: Physics I, Mathematics I and General and Organic Chemistry C.I. Instead, for the degree course in Biological Sciences, the first-year subjects are: Mathematics with Exercises, Cytology and Histology with Exercises, General Chemistry with Exercises and General Botany and Histology with Exercises.	The first-year subjects of the medical course at the University of Palermo for the academic year 2023/24 are: Physics I, Mathematics I, General and Organic Chemistry C.I., Organic Chemistry, Biology, Embryology and Histology C.I., Genetics and Microbiology C.I., Computer Science, Bioinformatics and Technical Scientific English C.I. I advise you to consult the documents provided for more information.	The first-year medical subjects are: Human Anatomy, Histology, Embryology, Biochemistry, Cell Biology, Genetics, Physics, Chemistry, English, Methodology of Intellectual Work, and Culture and Values in Healthcare.

Table 6
English version of Table 9.

prompt type	*prompt text*
custom prompt	I am Unipa-GPT, a chatbot and virtual assistant of the University of Palermo, who responds cordially and in a colloquial manner. To greetings, answer by greeting and introducing yourself; Answer the question with the words "Answer: " Remember that the rector of the university is Professor Massimo Midiri. If the question concerns the University of Palermo, answer on the basis of the information and provide the links associated with it; If you do not know how to answer the question, answer by saying that you are an artificial intelligence that still has a lot to learn and suggest that you go to https://www.unipa.it/, do not invent answers. Question: {question} Documents: {context}
condensed prompt	Given the following conversation and question, rephrase the next question in such a way that it is a single question. Conversation: {chat_history} Next question: {question} Single question:

Table 7
English version of Table 10.

ID	Questions	Answers
Q1	*Who is the Artificial Intelligence 1 professor for Computer Engineering Master degree course?*	The Artificial Intelligence 1 professor for the Computer Engineering Master degree course is Professor Gaglio and it will be delivered during the first semester. For more information go to http://www.unipa.it/struttura.html?id=721
Q2	*What are the deadlines for enrolling in a degree programme?*	The submission of applications for the Academic Year 2023/2024 varies according to the type of course. For three-year and single-cycle master's degree courses, applications can be submitted from 1 August to 30 September 2023, while for master's degree courses, from 1 August to 30 November 2023; in both cases, payment of the first instalment of tuition fees is required. For courses with local programmed access, with online booking or national programmed access, enrolment must be carried out by the deadlines set out in the corresponding calls for application.
Q3	*How does the master's degree in chemistry work?*	The Master's degree course in Chemistry at the University of Palermo is based in Palermo and lasts 2 years. The aim of the course is to train graduates with a good background, enabling them to deal competently and independently with both professional activities and academic and industrial research. For this reason, the course is designed with the aim of completing the scientific culture in the fundamental chemical disciplines and introducing more advanced knowledge. Therefore, students can complete their study plan by focusing on the subjects that best suit their specific interests, such as the chemistry of materials, the environment and cultural heritage, supramolecular chemistry and the didactics of chemistry. For more information go to http://www.unipa.it/struttura.html?id=766
Q4	*hello! I'm a guy just out of high school who is interested in law, especially corporate law. where should i apply?*	Hi! The Bachelor of Business Law Consultant programme sounds like it could be just the thing for you! The degree course is based in Trapani and lasts 3 years and will provide you with knowledge in the fields of law and business economics. Once you have completed the course you will be able to carry out interdisciplinary activities requiring legal, business and organisational-managerial skills. For more information go to http://www.unipa.it/struttura.html?id=1557
Q5	*how can i book an appointment at the secretariat?*	You can go to the secretariat on Mondays, Wednesdays and Fridays from 10 a.m. to 12 noon, Tuesdays and Thursdays from 3 p.m. to 5 p.m. . You can book your appointment through the SolariQ App. For more information go to https://www.unipa.it/servizi/segreterie/
Q6	*How do I pay fees?*	Fees must be paid exclusively through the PAgoPA (Public Administration Payments) system, which is accessed through the university portal. After completing the paperwork online, you can either pay directly online via the PAgoPA system or print out the payment slip and pay at a PAgoPA-enabled tax office. Further information on paying via PAgoPA can be found here https://immaweb.unipa.it/immaweb/public/pagamenti.seam, while the Student Contribution Regulations is available here https://www.unipa.it/servizi/segreterie/.content/documents/regulations_calendars/2023/5105144-def_regulation-student-contribution-2023-24-2.pdf