Unipa-GPT: a framework to assess open-source alternatives to Chat-GPT for Italian chat-bots

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

This paper illustrates the implementation of Open Unipa-GPT, an open-source version of the Unipa-GPT chat-bot that leverages open-source Large Language Models for embeddings and text generation. The system relies on a Retrieval Augmented Generation approach, thus mitigating hallucination errors in the generation phase. A detailed comparison between different models is reported to illustrate their performance as regards embedding generation, retrieval, and text generation. In the last case, models were tested in a simple inference setup after a fine-tuning procedure. Experiments demonstrate that an open-source LLMs can be efficiently used for embedding generation, but none of the models does reach the performances obtained by closed models, such as `gpt-3.5-turbo` in generating answers. Corpora and code are available on GitHub[1]

Keywords

RAG, ChatGPT, LLM, Embedding

1. Introduction

The increasing development of bigger and bigger Large Language Models (LLM), reaching 70B parameters as for Meta LLMs (Llama 2 [1] and `Llama 3` [2]) and more as for OpenAI ones (GPT-3 [3] and GPT-4 [4][1]), requires a significant computational resources for training, fine-tuning or inference. OpenAI models are accessible only upon payment via OpenAI API and cannot be downloaded in any way, while the open-source models by Meta are available also in the 8B and 13B parameters versions, and they can either be fine-tuned via Parameter-Efficient Fine-Tuning techniques (PEFT) [5] such as LoRA [6], or they can make direct inference using a 8-bit quantization [7] keeping the computational resources relatively small.

The availability of open-source small-size LLMs is crucial for developing Natural Language Process (NLP) applications that leverage a fine-tuning procedure over a specific domain or language, as for `Anita` [8], an Italian 8B adaptation of `Llama 3`.

Nevertheless, GPT and Llama models cannot be considered as truly open-source since their training data set is not available and, as for GPT models, and also their actual architecture is not accessible. `Minerva` [9] model, on the other side, is an Italian and English LLM whose architecture, weights, and training data are accessible, but it can be considered as an exception in the LLM landscape.

Starting from this premises, in this paper we propose Open Unipa-GPT, an open-source-based version of Unipa-GPT [10], that is a virtual assistant that uses a Retrieval Augmented Generation (RAG) approach [11] to answer university-related questions issued by secondary school students. Open Unipa-GPT has been developed upon the same architecture of Unipa-GPT, and uses open-source LLMs for embedding generation, retrieval, and text generation. Our models are small, compared to the ones used in our original version, namely `text-embedding-ada-002` and `gpt-3.5-turbo` from OpenAI.

The paper is arranged as follows: related works are reported in Section 2, while the architecture of Open Unipa-GPT is described in Section 3, and an overview of the data set is provided in Section 4. Experiments and related results are reported in Section 5. Finally, concluding remarks are drawn in Section 6.

2. Related works

The increasing interest in developing Language Models (LM) for the Italian language, starts when BERT [12] was first released and adapted models, such as AlBERTo [13] were developed. After ChatGPT was made public [3, 4], an increasing interest in developing and using LLMs, and in generative AI based on decoder-only model, was crucial, also for the Italian NLP community, thus leading to the development of foundational models based on `Llama 2` [1] and `Llama 3` [2]. Among those models, `LLaMantino` (chat version) [14] and `Fauno` [15], are based on `Llama 2` fine-tuned for chat purposes, while `Camoscio` [16] and `Anita` [8] are a fine-tuned Italian version of the instruct version of `Llama 2` and `Llama 3`, respectively.

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy
*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]online rumors refers to 175B and 1T parameter for `gpt-3.5-turbo` and `gpt-3.4` respectively

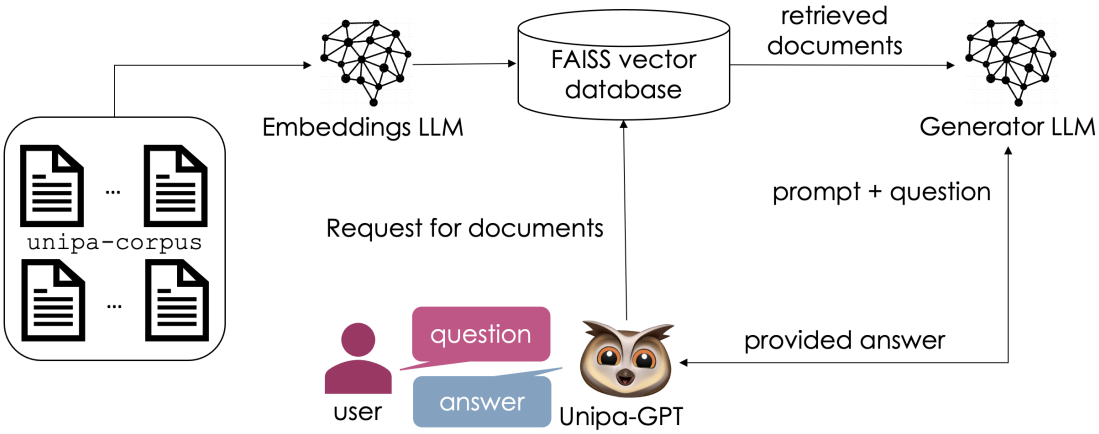

Figure 1: Overview of the Open Unipa-GPT architecture

RAG is used in developing chat-bots which are grounded in various domains where the models need to be deeply guided in generation to avoid hallucination in their answers. Various examples can be found in the educational domain as for AI4LA [17], an assistant to students with Specific Learning Disorders (SLDs) like Dyslexia, Dysorthographia, and Dyscalculia, or as assistant providing information about restaurant industry [18] or as chat-bot for Frequently Asked Questions (FAQ) [19]. Also chat-bots for the Italian language were implemented for real-wold applications, namely as assistant for Italian Funding Application [20], or in the medical domain [21] or in industrial context [22]. The aforementioned works share the same architecture with the one we used to implement our model. In contrast with them, we decided to stress capabilities of open-source LLMs and do not rely on GPT-based models, that are used as baseline reference for text generation (`gpt-3.5-turbo`) and as an external judge to evaluate performances of the other models (`gpt-4.5-turbo`).

3. System architecture

Open Unipa-GPT relies on two main components as it is shown in Figure 1 that is the *Retriever* and the *Generator*. In the following, the two components are detailed.

3.1. Retriever

The Retriever is made up of a vector database built using the LangChain framework[2], which makes use of the Facebook AI Similarity Search (FAISS) library [23]. The vector database is filled with the documents belonging to the `unipa-corpus` (Appendix A), that are divided into 1K token chunks with an overlap of 50 tokens. Split documents are then processed by a LLM (the *Embedding LLM*) to generate the corresponding embedding, and store them in the vector database. Different LLMs were used for embedding generation: we selected the best models according to the Massive Text Embedding Benchmark (MTEB) [24] for Information Retrieval[3]. We selected only models that explicitly state that they were trained and tested also with Italian data. In the end, we selected the following models: `BGE-M3` (BGE) [25], `E5-mistral-7b-instruct` (E5-mistral) [26], `sentence-bert-base-italian-xxl-uncased`[4] (BERT-it) and `Multilingual-E5-large-instruct` (m-E5) [27] .

A vector database was built for each model, and their corresponding embedding spaces were compared to each other and with `text-embedding-ada-002`, the embedding model from OpenAI, to asses their retrieval performances (Section 5).

3.2. Generator

The Generator uses the following Italian isntruction prompt to answer to user questions:

> *Sei 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. 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 associati; Se non*

[2]https://www.langchain.com

[3]as in https://huggingface.co/spaces/mteb/leader-board in June 2024

[4]https://huggingface.co/nickprock/sentence-bert-base-italian-xxl-uncased

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.

Below the English version:

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.

Both the question and the related relevant context are passed as input to the model, along with the prompt. As regards the Generator LLM, we used Transformer-based models [28]. We choose not to use LLMs based on `Llama 2` and deeply focused our work towards the most recent models, covering both Llama- and Mistral-based architectures. In particular, `Llama-3-8B-instruct` [2] was used along with its adapted version for Italian, `Anita-8B` [8], and `Minerva-3B` [9], which is a Mistral-based architecture [29]. All the generation LLMs were evaluated both in their base version and in the instruction-tuned one. The last ones were obtained via a three-epochs fine-tuning procedure with the Alpaca-LoRA [6] strategy testing the Alpaca-LoRA hyper-parameters[5] for both 20 and 50 epochs. In the generation phase, models were asked to output at most 256 tokens. We manually generated a small set of Question-Answer (QA) pairs for evaluation starting from the real questions issued by the public during the 2023 SHARPER European Researchers' Night where Unipa-GPT was demonstrated. The procedure for building these QA pairs is reported in Section 4. We developed the entire system on a server with 2 Intel(R) Xeon(R) 6248R CPUs, 384 GB RAM, and two 48 GB NVIDIA RTX 6000 Ada Generation GPUs.

4. The data set

The Italian documents data set built for Unipa-GPT is called `unipa-corpus` [10], and it has been generated from scraping either HTML pages or PDF documents that are publicly available on the website of the University of Palermo, and it includes information about all the available Bachelor/Master degree courses in the academic year 2023/2024 along with practical information for future students, e.g. how to pay taxes, the enrollment procedure, and the related deadlines. Starting from this data set, a QA data set was created with a semi-supervised procedure to allow instruction-tuning over general-purpose LLMs. Further information about the `unipa-corpus` is reported in Appendix A.

As already mentioned The original Unipa-GPT was available for public unsupervised QA during the European Researchers' Night in 2023, where a total of 165 questions was collected, along with feedback of users. On average, an interaction with the chat-bot was two questions long, and we collected qualitative evaluation of the user experience through a suitable questionnaire people were requested to fill on line just after having chatted with Unipa-GPT. Questionnaires were further analyzed, and resulted in a general positive evaluation of the system's performances by the majority of the users, which were mostly University students.

To generate the golden QA pairs used to assess the different performances of each generator LLM, we devised six typologies by the direct inspection of collected questions. Particularly we groupte questions in Generic Information, Courses' Information, Other University-related, Services and Structures, Taxes and Scholarships, University Environment, and Off-topic. Next, we picked one question per typology, discarding the Off-topic ones, and a golden answer was manually built for each of them by leveraging the actual relevant documents contained in the corpus, thus marking them as golden documents. Note that if an answer can be elicited by multiple documents, all of them have been marked as golden. The detailed list of the Italian QA pairs is reported in Appendix B in Table 4, while the English version is reported in Table 5. Note that the English version is reported here for full readability purposes, while only Italian data were used for evaluation.

5. Experimental results

The proposed model is intended to work in an open QA context, where correct answers are not known, thus, after a previous phase of qualitative evaluation [10] as in [17, 20, 21, 22], we opted for a quantitative analysis, relying on the small QA data set described in Section 4 to evaluate the performances against a set of golden labels in terms of both retrieval and answering capabilities [30, 19, 18].

For each QA test pair, we retrieved the four most relevant documents from each vector database related to one of the open Embedding LLMs under investiga-

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

Table 1
Context Relevancy scores over different Embedding LLMs. Bold values refer to the most relevant documents selected by RAGAS among the first four documents retrieved using the RAG. Underlined values refer to the golden documents.

	Q1				Q2				Q3			
model	D1	D2	D3	D4	D1	D2	D3	D4	D1	D2	D3	D4
open-ai-ada	**0,1**	**0,1**	0,0833	0,0714	0,0909	0,125	**0,625**	0,333	0,1	**0,111**	**0,111**	**0,111**
e5-mistral	0,0217	**0,0345**	**0,0345**	0,0233	**0,5**	0,0526	0,0333	0,025	0,0345	0,0154	0,0185	**0,0435**
bge	0,0345	0,0217	**0,0385**	0,0233	0,0526	**0,312**	0,0333	0,0909	0,0345	**0,0667**	0,0435	0,0154
bert-it	**0,125**	**0,125**	0,0345	0,0345	0,25	**0,333**	0,125	0,143	**0,172**	0,125	0,143	0,0192
m-e5	**0,125**	0,0217	0,1	0,0833	0,25	0,333	**0,5**	**0,5**	**0,333**	0,0185	0,037	0,0345
	Q4				Q5				Q6			
model	D1	D2	D3	D4	D1	D2	D3	D4	D1	D2	D3	D4
open-ai-ada	0,167	0,0588	0,1	**0,333**	0,429	**0,5**	0,143	0,111	**1**	0,1	0,167	0,5
e5-mistral	0,0417	0,05	0,05	**0,276**	0,154	0,04	**0,5**	0,0667	**0,333**	0,111	**0,333**	0,111
bge	0,241	0,0417	**0,333**	0,1	0,154	0,04	**0,333**	0,05	0,182	0,111	**0,444**	0,333
bert-it	**0,152**	0,0303	**0,152**	0,0303	**0,5**	0,04	0,0385	0,0769	0,111	**0,333**	0,25	0,25
m-e5	0,333	**0,5**	**0,5**	**0,5**	0,154	0,333	0,167	**0,5**	0,333	**0,5**	0,111	0,333

tion. Then we scored the retrieved documents in terms of their context relevancy with respect to the provided question using the RAGAS framework [31] that exploits `gpt-4-turbo` for the evaluation. Results are reported in Table 1, and they include also the performances of the original vector database using OpenAI embeddings (`text-embedding-ada-002`, referred as `open-ai-ada`). The overall scores are not so high, and also the highest relevancy do not always correspond to the golden document used for generating the corresponding answer. In Table 1 the underlined values are the ones associated with golden documents, while the bold ones are the highest RAGAS values. A model is considered to perform correctly if the highest context relevancy score is assigned to one of the golden documents. This evaluation procedure led to select E5-mistral as the best performing Embeddings LLM among the ones we investigated.

Superior performances of E5-mistral are also confirmed by a deep analysis on the embeddings space by means of two different clustering procedures. We clustered the embeddings generated by each LLM starting from the documents belonging to both sections *Educational Offer* and *Future Students* of the UniPA website. The firs group of documents is the list of all the available courses at the University, while the second group contains useful information for future students who want to enroll in a degree course. We clustered the embedding spaces according to the either the course degree typology (bachelor/master degree) or the Department where a degree course is affiliated to. Quantitative measures of the clustering goodness are reported in Table 2, where the Silhouette Coefficients [32] have been computed for each model, and again E5-mistral is the best performing one. In Appendix C, we report the scatter plots of the embedding spaces for each Embeddings LLM (Figure 3 and Figure 4). Plots have been obtained through a 2D dimensionality reduction using t-SNE [33].

We used the six QA test pairs to obtain also a quantitative evaluation of the correctness of the answers provided by all the Generation LLMs under investigation. Comparison was carried out against both the golden answers and the ones generated via `gpt-3.5-turbo` (GPT) in the original Unipa-GPT set up. The proposed evaluation task, can be regarded as an open QA one where, despite a golden answer is provided for a given question, diverse correct answers can be proposed with different linguistic nuances, according to Italian diaphasic variation [34]. To evaluate both *strict* and *light* correctness of the generated answers, we employed traditional QA metrics such as BLEU [35] (Figure 2.a) and ROGUE-L score [36] (Figure 2.b) and novel metrics leveraging the RAGAS framework [31] to evaluate Faithfulness (Figure 2.c) and Correctness (Figure 2.d) of the generated output. Such measures request an external LLM acting as a "judge", and we we used `gpt-4-turbo` in this respect. More specifically, Faithfulness measures the factual consistency of the generated answer against the given context, while 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.

Both BLEU and ROUGE scores are generally low, but we assume that this is mainly related to the fact that an exact match cannot be reached between the golden answer and the generated one, and a more semantically comparison should be taken into account. Overall, answers generated by `gpt-3.5-turbo` can be considered as the best ones as they attain highest values. By contrast, fine-tuning did not provided a desired improvement in the open-source models: all BLEU scores are almost zero, except for `Anita-8B` . ROUGE scores are higher than the corresponding BLEU ones, and again the base ver-

Table 2
Silhouette Coefficients for each Embedding LLM with reference to the two proposed clustering schemes, that is the degree courses typology and their affiliation to a particular Department.

Retriever	Silhouette score typology	Silhouette score Departments
openAI-ada	-0.0915	-0.0627
E5-mistral	**-0.0194**	**-0.0048**
BGE	-0.0422	-0.0708
BERT-it	-0.0221	-0.0367
m-E5	-0.0982	-0.0503

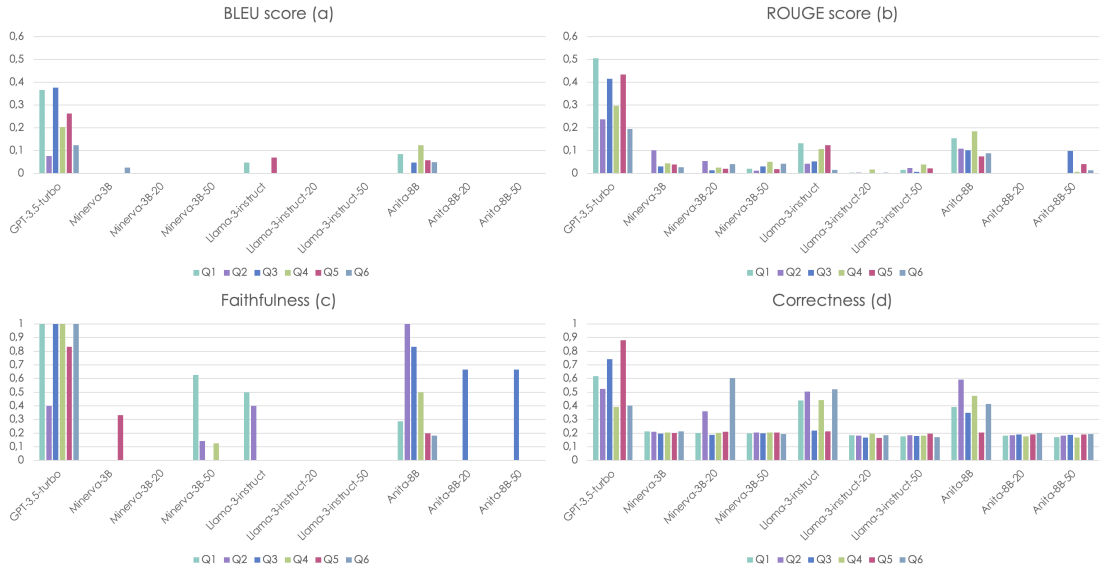

Figure 2: Inference results over the generated answers according the following scores: (a) BLEU, (b) ROUGE, (c) Faithfulness and (d) Correctness. Due to displaying reasons, (a,b) are represented in a [0, 0.6] range, while (c,d) in a [0,1] range.

sion of each LLM performes better than the fine-tuned ones. Generally speaking, `Anita-8B` and `Llama-3-8B-instruct` outperform `Minerva`, since both reach comparable scores, but we assume that the tailored Italian fine-tuning over Llama-3 to obtain `Anita-8B` was crucial to make it the best performing open-source model during this first automatic evaluation phase.

`gpt-3.5-turbo` exhibits the best Faithfulness scores despite being surpassed by `Anita-8B` in question Q2, and also these results confirm the previous considerations about BLEU scores. Something changes in evaluating models in terms of their Correctness: in this case `gpt-3.5-turbo` is the best model in three answers out of six, followed by `Anita-8B` (two best results) and `Minerva-3B-20` (one best result). We are aware that gpt-based evaluation may lead to a preference over GPT models themselves, but `gpt-4-turbo` was the only high quality generative model we had access to at the time of making the experiments.

Overall results confirm that a (moderate) fine-tuning is not significantly beneficial in terms of performance increase for any model and, even if it does not reach the same performances, `Anita-8B` seems to be the most valuable alternative to GPT.

A manual inspection of the generated answers, outlines a common issue related to the tokenization of the generated output: despite of its semantic correctness, the generated text is outputted as a unique word without any spaces, as

> *Glielezionidelcorsosaracondottoattraversounaprocesso*

in `Llama-3-8B-instruct`, or it is over-splitted as

> *e-domandre d'i-s-c-r-i-z-ion-e-per-l-A.-A.–2023–/—-cor-so-n-d'-l-a-u-re-'-(M-agg-is-t-ra-le)–a-dd-ac-ce-o-lib-ro*

in `Anita-8B`. These errors make the models not suitable for human interaction, since it is not possible read the generated answers. We argue that a deeper analysis on

the tokenizer that has been used and, a hyper-parametrs tuning in the generator, may lead to an increase of performances. Models tend also to answer in other languages as

> ** La durada édié depresso àdue años, * Accesso libre! * Dipartment of Physics & Chemistry "Emilo Segré" Codice course : 21915*

in `Llama-3-8B-instruct`. We argue that this trouble can be related to the memory of multi-lingual models that uses texts also in French and Spanish despite the Italian fine-tuning. It is worth noticing that those languages are linguistically close to Italian and together belong to the Romance Languages [37]. Thus, even if the output has to be considered wrong, a linguistic connection can be highlighted.

The most unsatisfactory results are reported for `Minerva-3B`: the model does not generate any answer related to the given question, and it seems that answers where generated with samples from model's training set. As stated before, a tuning of the generator hyperparameters may help in this case.

Despite the promising results, in some cases answers by both `Anita-8B` and `Llama-3-8B-instruct` are not good from a grammatical point of view, since they are full of mistakes, thus making them not yet ready to be used in real-world applications compared to OpenAI's ones.

6. Conclusions and future works

In this paper we presented Open Unipa-GPT, a virtual assistant, which is based solely on open-source LLMs, and uses a RAG approach to answer Italian university-related questions from secondary school students. The main intent of the presented research was setting up a sort of framework to test open-source small size LLMs, with either moderate or no fine-tuning at all, to be used for generating the embeddings and/or as text generation front-end in a RAG set up.

Our study led us to devise `E5-mistral-7b-instruct` as a valuable open-source alternative to OpenAI's embeddings, while none of the considered models attain a generation performance comparable to `gpt-3.5-turbo`, even after a fine-tuning procedure. The most promising Generation LLM, when plunged in our architecture, appears to be `Anita-8B`, but it still shows some issues related to both the tokenization and the grammatical correctness of the output. We are currently working to deep exploration of different fine-tuning approaches along with the use of huge size open-source LLMs for text generation.

References

[1] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al., Llama 2: Open foundation and fine-tuned chat models, arXiv preprint arXiv:2307.09288 (2023).

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

[3] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.

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

[5] L. Xu, H. Xie, S.-Z. J. Qin, X. Tao, F. L. Wang, Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment, 2023. URL: https://arxiv.org/abs/2312.12148. arXiv:2312.12148.

[6] 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).

[7] T. Dettmers, M. Lewis, Y. Belkada, L. Zettlemoyer, Llm.int8(): 8-bit matrix multiplication for transformers at scale, 2022. URL: https://arxiv.org/abs/2208.07339. arXiv:2208.07339.

[8] M. Polignano, P. Basile, G. Semeraro, Advanced natural-based interaction for the italian language: Llamantino-3-anita, 2024. arXiv:2405.07101.

[9] R. Orlando, L. Moroni, P.-L. Huguet Cabot, S. Conia, E. Barba, R. Navigli, Minerva technical report, 2024. URL: https://nlp.uniroma1.it/minerva/.

[10] I. Siragusa, R. Pirrone, Unipa-gpt: Large language models for university-oriented qa in italian, 2024. URL: https://arxiv.org/abs/2407.14246. arXiv:2407.14246.

[11] 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.

[12] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL:

https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[13] M. Polignano, P. Basile, M. Degemmis, G. Semeraro, V. Basile, Alberto: Italian bert language understanding model for nlp challenging tasks based on tweets, in: Italian Conference on Computational Linguistics, 2019. URL: https://api.semanticscholar.org/CorpusID:204914950.

[14] P. Basile, E. Musacchio, M. Polignano, L. Siciliani, G. Fiameni, G. Semeraro, Llamantino: Llama 2 models for effective text generation in italian language, 2023. arXiv:2312.09993.

[15] A. Bacciu, G. Trappolini, A. Santilli, E. Rodolà, F. Silvestri, Fauno: The italian large language model that will leave you senza parole!, arXiv preprint arXiv:2306.14457 (2023).

[16] A. Santilli, E. Rodolà, Camoscio: an italian instruction-tuned llama, 2023. arXiv:2307.16456.

[17] S. D'Urso, F. Sciarrone, Ai4la: An intelligent chatbot for supporting students with dyslexia, based on generative ai, in: A. Sifaleras, F. Lin (Eds.), Generative Intelligence and Intelligent Tutoring Systems, Springer Nature Switzerland, Cham, 2024, pp. 369–377.

[18] V. Bhat, D. Sree, J. Cheerla, N. Mathew, G. LIu, J. Gao, Retrieval augmented generation (rag) based restaurant chatbot with ai testability, 2024.

[19] M. Kulkarni, P. Tangarajan, K. Kim, A. Trivedi, Reinforcement learning for optimizing rag for domain chatbots, 2024. URL: https://arxiv.org/abs/2401.06800. arXiv:2401.06800.

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

[21] 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.

[22] 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.

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

[24] 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.

[25] 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. URL: https://arxiv.org/abs/2402.03216. arXiv:2402.03216.

[26] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, F. Wei, Improving text embeddings with large language models, arXiv preprint arXiv:2401.00368 (2023).

[27] 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).

[28] 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).

[29] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, W. E. Sayed, Mistral 7b, 2023. URL: https://arxiv.org/abs/2310.06825. arXiv:2310.06825.

[30] S. Vidivelli, M. Ramachandran, A. Dharunbalaji, Efficiency-driven custom chatbot development: Unleashing langchain, rag, and performance-optimized llm fusion., Computers, Materials & Continua 80 (2024).

[31] S. Es, J. James, L. Espinosa-Anke, S. Schockaert, Ragas: Automated evaluation of retrieval augmented generation, arXiv preprint arXiv:2309.15217 (2023).

[32] 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.

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

[34] G. Berruto, Variazione diafasica, 2011. URL: https://www.treccani.it/enciclopedia/variazione-diafasica_(Enciclopedia-dell'Italiano)/.

[35] 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.

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

[37] T. Alkire, C. Rosen, Romance languages: A historical introduction, Cambridge University Press, 2010.

A. unipa-corpus details

unipa-corpus [10] is a collection of Italian documents that were retrieved directly from the website of the University of Palermo in Semptember 2023. The corpus is divided in two main sections, namely *Education*, that groups the available bachelor and master degree courses, and *Future Students* where important information about taxes payment and enrollment procedure are reported. For fine-tuning purposes, a semi-automatic procedure, involving gpt-3.5-turbo [3], was implemented to build a QA dataset. In Table 3 are reported the statistics of unipa-corpus.

Table 3
Number of documents and QA pairs in unipa-corpus.

	Education	*Future Students*
Documents	506	104
Tokens	1072214	987424
QA pairs train	506	269
Tokens train	191612	68160
QA pairs val	253	133
Tokens val	93443	29675

B. Inference QA pairs

Table 4
Overview of 6 QA pairs manually generated used for evaluation purposes

IDs	Questions	Answers
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 in Ingegneria Informatica è il professore Gaglio e l'insegnamento verrà erogato 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.htmlid=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). Dopo aver compilato la pratica online, è possibile pagare direttamente online con il sistema PAgoPA o stampare il bollettino e pagare presso tabaccai convenzionati o ricevitorie abilitate 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

Table 5
English version of Table 4.

IDs	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

C. Embedding spaces

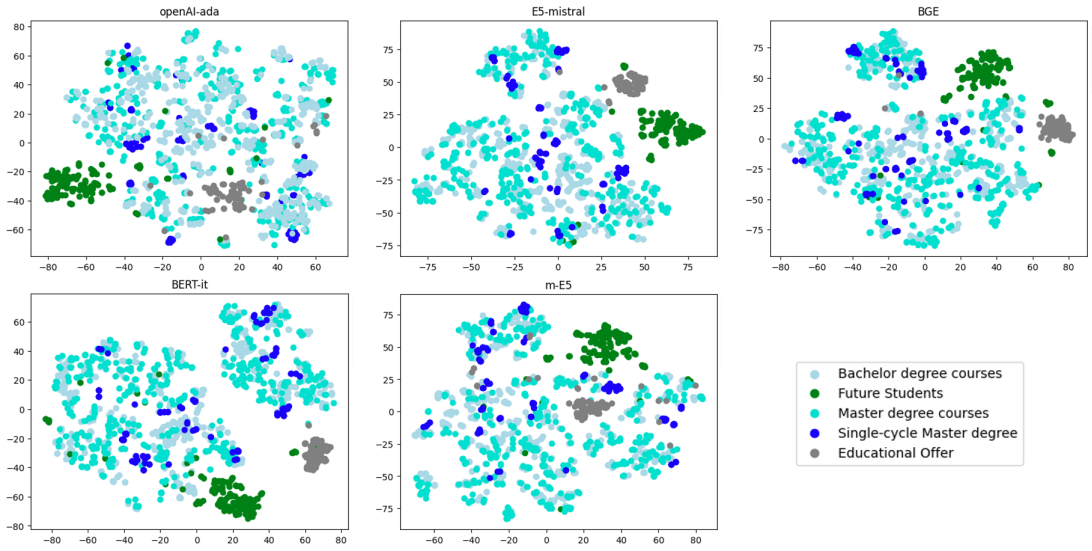

Figure 3: Scatter plots of embedding spaces labeled as for typology

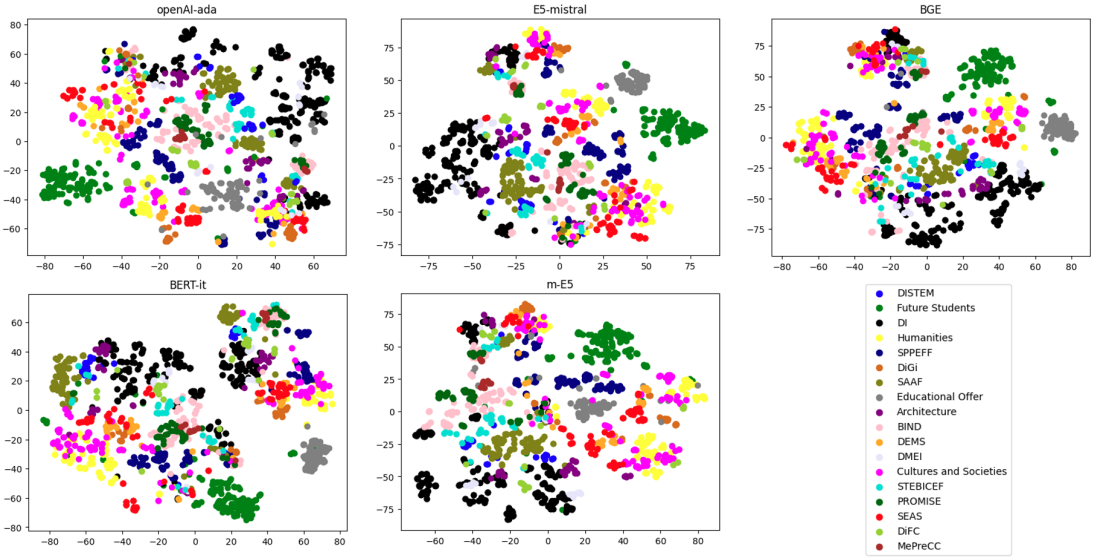

Figure 4: Scatter plots of embedding spaces labeled as for department