HHAI 2025
D. Pedreschi et al. (Eds.)

doi:10.3233/FAIA250637

Towards Building a Trustworthy RAG-Based Chatbot for the Italian Public Administration

Chandana Sree MALA[*] [a,b,1], Christian DI MAIO[*] [a,c], Mattia PROIETTI [d], Gizem GEZICI [b], Fosca GIANNOTTI [b], Stefano MELACCI [c], Alessandro LENCI [d] and Marco GORI [c]

[a] *Department of Computer Science, University of Pisa*
[b] *Department of Informatics, Scuola Normale Superiore*
[c] *Department of Information Engineering, University of Siena*
[d] *Department of Philology, Literature and Linguistics, University of Pisa*

Abstract.
Building a Trustworthy Retrieval-Augmented Generation (RAG) chatbot for Italy's public sector presents challenges that go beyond selecting an appropriate Large Language Model. A major issue is the retrieval phase, where Italian text embedders often underperform compared to English and multilingual counterparts, hindering precise identification and contextualization of critical information. Regulatory constraints further complicate matters by disallowing closed source or cloud based models, forcing reliance on on-premise or fully open source solutions that may not fully address the linguistic complexities of Italian documents. In our study, we evaluate three embedding approaches using a publicly available Italian dataset: a monolingual Italian approach, a translation based method leveraging English only embedders with backward reference mapping, and a multilingual framework applied to both original and translated texts. Our methodology involves chunking documents into coherent segments, embedding them in a high dimensional semantic space, and measuring retrieval accuracy via top-k similarity searches. Our results indicate that the translation based approach significantly improves retrieval performance over Italian specific models, suggesting that bilingual mapping can effectively address both domain specific challenges and regulatory constraints in developing RAG pipelines for public administration.

Keywords. Trustworthiness, Generative AI Chatbot, Public Sector AI, Robustness, Reliability, Retrieval-Augmented Generation (RAG), LLM, Data Privacy.

1. Introduction

The advent of Large Language Models (LLMs) built upon the Transformer architecture [1] has fundamentally reshaped how information is processed across text, vision, and audio modalities [2]. In particular, advancements in Generative AI have paved the

[*] These authors contributed equally to this work.
[1] Corresponding Author: Chandana Sree MALA, chandana.mala@sns.it

way for sophisticated chatbot systems [3], which are now widely implemented in both public and private sectors. However, despite their potential, these models are not without limitations. They remain prone to hallucinations—generating false or incoherent outputs [4]—and often face difficulties in handling domain-specific knowledge, a challenge that is especially critical in the public sector [5].

Retrieval-Augmented Generation (RAG) [6] has emerged as an effective framework for enhancing the performance of LLMs by retrieving external knowledge and incorporating it into the generation process, all without the need for additional model training. However, its success is largely dependent on the quality of the underlying embedding models, which must accurately represent semantic meaning to ensure effective retrieval. Implementing RAG in non-English and domain-specific contexts, such as the Italian public sector, introduces additional challenges. Regulatory restrictions limit access to advanced proprietary models, while Italian-language embeddings generally underperform compared to their English counterparts. Moreover, public sector data often comprises legal, administrative, and technical documents that require highly precise and context-aware retrieval. These challenges highlight the need for optimizing both embedding models and retrieval strategies to develop reliable and effective RAG applications in multilingual and specialized domains.

In this study, we explore three different approaches for embedding Italian texts for retrieval purposes. The first approach uses English-only embedding models, where Italian text is translated into English, and embeddings are generated from the translated text while preserving references to the original Italian. The second approach utilizes models specifically trained on Italian corpora, and the third approach applies multilingual embedding models directly to the Italian text. By systematically comparing these methods, we aim to highlight their respective strengths and limitations, thereby contributing to more effective Italian text retrieval in future AI applications.

2. Background: Challenges & Opportunities

Our use case centers on creating a generative AI-powered chatbot specifically designed for employees of a Government Organization (GO), with the goal of offering accessible and user-friendly support. The chatbot leverages a dataset of technical user manuals for GO's internal applications, with the primary objective of establishing a trustworthy generative AI pipeline to serve as an assistant for GO's employees. This pipeline is designed to generate clear, non-technical, and easily understandable responses to employee queries by extracting relevant information from the user manuals. A critical requirement is to ensure that the generated answers are accessible and comprehensible, even to users with no technical background, thereby improving both accessibility and usability. For these reasons, the development choices in such a scenario must be the result of a trade-off between performance, reliability, trustworthiness, and resource feasibility. Several approaches and technologies could be considered to achieve this goal. The most straightforward and immediately available option would be to use a pre-trained LLM, potentially enhancing its performance through advanced prompting techniques like Few-Shot learning or Chain-of-Thought. However, state-of-the-art models suitable for these applications are often proprietary and accessible only via API, which makes them unsuitable for use in the public sector. Another option would be to fine-tune an LLM to specialize

in the target domain and specific downstream tasks. While this approach could address the limitations of off-the-shelf models—especially by utilizing open-source models—it is still largely impractical due to the high computational and financial costs involved. A more feasible alternative is provided by RAG, which we explore in detail below.

Retrieval Augmented Generation. As outlined in Section 1, RAG frameworks have been developed to produce responses that extend beyond the pre-trained knowledge of the LLM, grounding the model's output in factual knowledge stored within a curated knowledge base. More formally, the RAG framework operates by processing a collection of data (which we assume to be unstructured text documents), denoted as $t_i \in T$, for $i = 1, ..., m$ (with m documents). Each t_i is divided into a set of smaller chunks, denoted as $c_j \in C_{t_i}$, for $j = 1, ..., n$ (with n chunks), forming the knowledge base $\mathrm{K} = \bigcup_{i=1}^{m} C_{t_i}$.

Each c_j is transformed into a high-dimensional vector through an embedding module, $\texttt{Enc}$, yielding $\mathbf{e}_j \in \mathbb{R}^{d_{\texttt{Enc}}}$. All the resulting vectors are then stored in a vector database. During inference, a user query q is encoded as $\mathbf{q} = \texttt{Enc}(q) \in \mathbb{R}^{d_{\texttt{Enc}}}$, and its similarity to the stored vectors is assessed using similarity measures such as cosine similarity. The system then retrieves the top-k most relevant chunks, denoted as C_q, where $|C_q| = k$. Finally, the LLM, denoted as M, uses the query q and the retrieved chunks C_q to generate a response y.

This framework offers several possible usage combinations:

- *RAG-only*: The system can be implemented using a standard RAG pipeline, relying on off-the-shelf models for both retrieval and generation.
- *RAG + Fine-Tuning*: Alternatively, the RAG framework can be enhanced by fine-tuning the embedding module ($\texttt{Enc}$), the generator (M), or both components to improve performance for specific tasks or domains.

Considering the trade-offs in our use case, a RAG-only architecture stands out as the most effective solution. It facilitates the integration of external knowledge sources, such as user manuals, to generate accurate and up-to-date responses while minimizing reliance on proprietary models. This approach eliminates the need for large-scale fine-tuning, reducing both computational costs and the effort required for data preparation. Recent studies also indicate that RAG can outperform traditional fine-tuning methods [7], making it a practical and scalable option for real-world deployments.

3. Related Work

Embedding models are essential in retrieval systems as they transform text into dense vector representations that capture semantic meaning [8]. The development of these models has evolved from word-based approaches, such as Word2Vec [9] and GloVe [10], to contextualized models like BERT [11] and Sentence-BERT (SBERT) [12], which enable improved text representations. Multilingual embedding models, including MPNet [13] and Sentence-BERT [14], have demonstrated cross-lingual capabilities [15, 16]. More recent models, such as E5-Multilingual [17] and BGE-Multilingual [18], are trained on multilingual corpora and are specifically designed to generate embeddings that are applicable across different languages. For Italian text retrieval, research has focused on

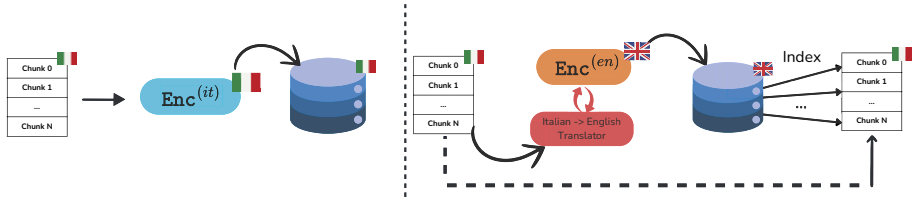

Figure 1. Comparison of two vector store construction strategies for a RAG pipeline in an Italian-language. The standard approach (left) involves chunking and embedding documents directly in Italian before storing them in a vector database. In contrast, the bilingual approach (right) translates chunks into English, computes embeddings using English-language models, and stores them alongside backward references to the original Italian text for later retrieval.

Italian-specific embedding models, such as those fine-tuned on Italian corpora [19–21]. The Italian language features distinct linguistic characteristics, including complex morphology and syntax, which may not be fully captured by multilingual models. As embedding models continue to evolve, ongoing research in both multilingual and language-specific approaches is essential for refining text representations across diverse linguistic contexts [22]. When applied to the public sector, the challenge of developing reliable and trustworthy AI systems is further amplified. The public administration domain is characterized by specialized and technical language, demanding high precision in information retrieval and generation. Legal and regulatory documents, in particular, pose additional challenges for LLMs [23, 24], highlighting that fine-tuned models tailored for specific domains typically outperform general-purpose models, albeit at the expense of large volumes of training data and substantial computational resources.

4. Methodology

The success of RAG pipelines depends on two critical factors: the accuracy of text embedders in mapping chunks into a high-dimensional semantic space that effectively captures their meaning, and the ability of the LLM to interpret contextual information and integrate it to generate accurate responses. To address this, following the approach proposed by Iscan et al. [25], we explore whether translating Italian documents into English and leveraging English-language embedding models can serve as a viable alternative to using a fine-tuned model for Italian. This approach (i) eliminates the need for language-specific fine-tuning, and (ii) allows the use of widely available English embedding models, which are the majority of publicly released resources.

Bilingual Mapping Approach. Let $T^{(\mathrm{it})}$ be a set of m Italian documents to be stored. As illustrated in Figure 1, our goal is to construct a knowledge base optimized for English-based retrieval, denoted as $\mathrm{K}^{(\mathrm{en})}$, while preserving a reliable mapping to the original-language content for response generation. To this end, each document $t_i^{(\mathrm{it})}$ is translated into English using a deterministic offline translation function, resulting in $t_i^{(\mathrm{en})} = \mathrm{Tr}_{\mathrm{it}\to\mathrm{en}}(t_i^{(\mathrm{it})})$. This offline approach ensures translation consistency and simplifies downstream processing by standardizing the pipeline in a single language from the outset. Following the standard RAG framework, each translated document $t_i^{(\mathrm{en})}$ is segmented into a set of chunks $c_k^{(\mathrm{en})} \in C_{t_i}^{(\mathrm{en})}, \quad k = 1, \ldots, n$, where n is the number of chunks per document.

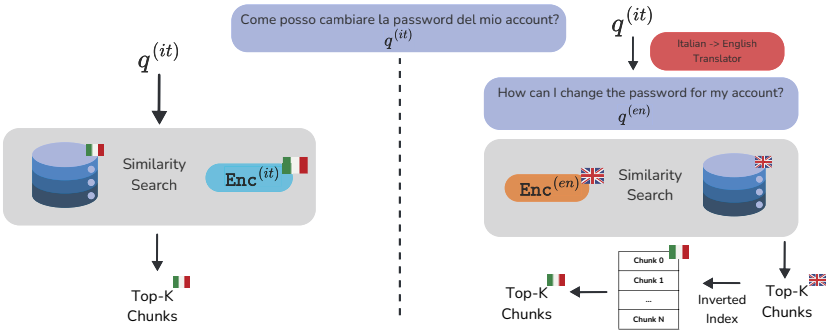

Figure 2. Comparison of two retrieval strategies in a RAG pipeline for an Italian-language public administration chatbot. The standard approach (left) embeds the Italian query and retrieves relevant chunks from an Italian embedding space. The bilingual strategy (right) translates the query into English, embeds it using an English encoder, and performs retrieval in an English vector store, mapping results back to Italian via an inverted index.

The retrieval-ready knowledge base is then defined as $\mathrm{K}^{(\mathrm{en})} = \bigcup_{i=1}^{m} C_{t_i}^{(\mathrm{en})}$. Meanwhile, the corresponding original-language chunks are retained as $c_j^{(\mathrm{it})} \in C_{t_i}^{(\mathrm{it})}, \quad j = 1, \ldots, n$. By design, each English chunk $c_k^{(\mathrm{en})}$ has a one-to-one alignment with its Italian counterpart $c_j^{(\mathrm{it})}$ in the same position, such that: $c_k^{(\mathrm{en})} = \mathrm{Tr}_{\mathrm{it}\to\mathrm{en}}(c_j^{(\mathrm{it})})$. Each English chunk is indexed and associated with a backward reference $\mathrm{Ref}(k) \mapsto c_j^{(\mathrm{it})}$, preserving traceability to the source content. All chunks in $\mathrm{K}^{(\mathrm{en})}$ are embedded into a d-dimensional vector space using an English encoder $\texttt{Enc}^{(\mathrm{en})}$, yielding $\mathbf{e}_k^{(\mathrm{en})} = \texttt{Enc}^{(\mathrm{en})}(c_k^{(\mathrm{en})}) \in \mathbb{R}^d$. These pairs $\left(c_k^{(\mathrm{en})}, \mathbf{e}_k^{(\mathrm{en})}\right)$ are stored in a vector database for efficient retrieval. At inference time (Figure 2), an Italian query $q^{(\mathrm{it})}$ is translated into English $q^{(\mathrm{en})} = \mathrm{Tr}_{\mathrm{it}\to\mathrm{en}}(q^{(\mathrm{it})})$, and embedded as $\mathbf{q}^{(\mathrm{en})} = \texttt{Enc}^{(\mathrm{en})}(q^{(\mathrm{en})})$. Similarity is computed between $\mathbf{q}^{(\mathrm{en})}$ and each chunk embedding $\mathbf{e}_k^{(\mathrm{en})}$ using cosine similarity or ℓ_2 distance. The top-k most relevant chunks, denoted $C_{q^{(\mathrm{en})}}$, are retrieved and mapped back to their original Italian form via the stored references, allowing responses grounded in the source language.

5. Experiments

We conducted our evaluation using a synthetic dataset provided by ReDix Informatica [26], built from Wikipedia passages with questions automatically generated by a proprietary model. The dataset comprises approximately 105k entries, each consisting of a *question*, *context*, and *answer*. To obtain a balanced and computationally feasible subset, we embedded all contexts using an oracle embedding model and applied a clustering algorithm with $k = 10$, from which we sampled 50 complete triplets per cluster, yielding a final evaluation set of 1k entries. The goal of the experiment was to assess the retrieval performance of models not explicitly trained on Italian, particularly when working with translated content. To isolate this aspect, we restricted the evaluation to *question-context* pairs, excluding the answer field. As shown in Table 1, the selected embedding models were trained on Italian, English, or multilingual corpora, allowing for a direct comparison across language-specific and cross-lingual settings. Retrieval was evaluated using the

Table 1. Overview of embedding models

Model	Language	Variant	Embedding Dims	Max Tokens	#Parameters	Reference
Bertino	Italian	base	768	512	66M	[20]
Gattina	Italian	base	768	512	109M	[19]
Mmarco	Italian	base	768	512	109M	[21]
GTE-EN	English	base	768	512	109M	[27]
BGE	English	base	768	512	109M	[18]
mGTE	Multilingual	base	768	8.192	305M	[28]
KaLM-E	Multilingual	–	896	131.072	494M	[29]

Retrieval Accuracy metric, which assigns a positive score when the ground-truth context is retrieved as top-1 for a given question. Accordingly, each model was tested on the version of the data that matched its training—original Italian for Italian models, translated English for English models, and both versions for multilingual ones.

Table 2. Comparison of Retrieval Accuracy

Model	Original (IT)	Translated (EN)
BERTino	0.81	–
Gattina	0.70	–
Mmarco	0.66	–
GTE-en	–	0.88
BGE	–	0.79
mGTE	0.88	**0.89**
KaLM-E	0.86	0.72

The results, summarized in Table 2, indicate that the best overall performance is achieved by the multilingual GTE model when applied to the translated version of the documents. Interestingly, the same model also performs best on the original Italian texts, demonstrating strong cross-lingual capabilities. Close behind is its monolingual variant, GTE-en, which ranks second overall despite not being trained on Italian data. In contrast, models specifically trained on Italian consistently underperform compared to both English and multilingual alternatives. Although this outcome may appear counterintuitive, it reinforces the relative strength of English and multilingual embedding models in retrieval tasks, even when applied to non-English corpora. In this context, an Italian-specific model ranks only as the fourth-best option for encoding the target document collection.

At the same time, the results show how a bilingual-mapping approach as the one described in Sec. 4 is a promising route to cope with the shortcomings posed by the Italian-specific models landscape. In fact, enabling the use of English-only text embedders through the translate-and-map procedure enables access to a broader selection of open-source models, beyond those specifically trained on Italian or in a multilingual setting.

6. Conclusion and Future directions

In this work, we have addressed key challenges in developing RAG-powered systems for the Italian language, particularly focusing on the retrieval module. We reviewed three possible approaches for embedding and retrieving Italian texts: (i) monolingual Italian models, (ii) monolingual English models via translation and bilingual mapping, and (iii) multilingual models applied to both scenarios. Our results indicate that monolingual Italian models underperform compared to both English and multilingual alternatives. This highlights the limitations of currently available Italian-specific models but also suggests bilingual mapping as a promising strategy. By leveraging bilingual representations, we can expand the pool of efficient open-source models, enhancing retrieval performance without resorting to costly fine-tuning on the target language.

Beyond embedding quality, the effectiveness of retrieval also depends on how text is chunked. In Public Administration, retrieval must respect logical or sequential structures (e.g., ordered tutorial steps: $c_1 \prec c_2 \prec \cdots \prec c_N$). A purely vector-based approach may fail to maintain these relationships, leading to two key issues: (i) retrieving a relevant chunk c_j may exclude its logical successor c_{j+1}, and (ii) overlapping or branching structures may be lost. To address this, future work will focus on integrating structured retrieval policies into RAG-based agents to preserve order and logical dependencies. Additionally, we will explore efficient offline translation methods to further improve retrieval robustness in multilingual settings.

7. Limitations

Since the private dataset lacks ground-truth passages, we do not have corresponding results for it and rely solely on the WikipediaQA dataset. However, this approach may lead to discrepancies, as the private dataset contains domain-specific knowledge with specialized terminology that may not match the general knowledge found in Wikipedia.

Acknowledgements

The authors gratefully acknowledge the Regione Toscana Team—especially Luca Ciprani, Marco Caldini, Silvio De Magistris, Davide Bruno, and Gianluca Vannuccini—for their valuable expertise and contributions to this work in the field of public administration.

This work has been supported by the European Union under ERC-2018-ADG GA 834756 (XAI), by the Partnership Extended PE00000013 - "FAIR - Future Artificial Intelligence Research" - Spoke 1 "Human-centered AI".

The acknowledgment is extended to Rete SAIHUB (Siena, Italy), which, jointly with the Italian Ministry of University and Research (DM 117/2023, PNRR, Missione 4, Componente 2, Investimento 3.3), funded the scholarship of Christian Di Maio.

References

[1] Vaswani A. Attention is all you need. Advances in Neural Information Processing Systems. 2017.

[2] Cahyawijaya S, Lovenia H, Fung P. LLMs Are Few-Shot In-Context Low-Resource Language Learners. arXiv preprint arXiv:240316512. 2024.

[3] Al-Amin M, Ali MS, Salam A, Khan A, Ali A, Ullah A, et al. History of generative Artificial Intelligence (AI) chatbots: past, present, and future development. arXiv preprint arXiv:240205122. 2024.

[4] Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems. 2024.

[5] Yu H, Gan A, Zhang K, Tong S, Liu Q, Liu Z. Evaluation of retrieval-augmented generation: A survey. In: CCF Conference on Big Data. Springer; 2024. p. 102-20.

[6] Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems. 2020;33:9459-74.

[7] Lakatos R, Pollner P, Hajdu A, Joó T. Investigating the Performance of Retrieval-Augmented Generation and Domain-Specific Fine-Tuning for the Development of AI-Driven Knowledge-Based Systems. Machine Learning and Knowledge Extraction. 2025. Available from: https://api.semanticscholar.org/CorpusID:276266216.

[8] Karpukhin V, Oguz B, Min S, Lewis PS, Wu L, Edunov S, et al. Dense Passage Retrieval for Open-Domain Question Answering. In: EMNLP (1); 2020. p. 6769-81.

[9] Mikolov T, Chen K, Corrado GS, Dean J. Efficient Estimation of Word Representations in Vector Space. In: International Conference on Learning Representations; 2013. .

[10] Pennington J, Socher R, Manning CD. Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP); 2014. p. 1532-43.

[11] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: North American Chapter of the Association for Computational Linguistics; 2019. Available from: https://api.semanticscholar.org/CorpusID:52967399.

[12] Reimers N, Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In: Inui K, Jiang J, Ng V, Wan X, editors. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics; 2019. p. 3982-92. Available from: https://aclanthology.org/D19-1410/.

[13] Song K, Tan X, Qin T, Lu J, Liu TY. Mpnet: Masked and permuted pre-training for language understanding. Advances in neural information processing systems. 2020;33:16857-67.

[14] Khattab O, Zaharia M. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In: Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval; 2020. p. 39-48.

[15] Pires T, Schlinger E, Garrette D. How Multilingual is Multilingual BERT? In: Korhonen A, Traum D, Màrquez L, editors. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics; 2019. Available from: https://aclanthology.org/P19-1493/.

[16] Feng F, Yang Y, Cer D, Arivazhagan N, Wang W. Language-agnostic BERT Sentence Embedding. In: Muresan S, Nakov P, Villavicencio A, editors. ACL (1). Association for Computational Linguistics; 2022. p. 878-91. Available from: http://dblp.uni-trier.de/db/conf/acl/acl2022-1.html#FengYCA022.

[17] Wang L, Yang N, Huang X, Jiao B, Yang L, Jiang D, et al. Text Embeddings by Weakly-Supervised Contrastive Pre-training. arXiv. 2024. Available from: https://arxiv.org/abs/2212.03533.

[18] Xiao S, Liu Z, Zhang P, Muennighoff N. C-Pack: Packaged Resources To Advance General Chinese Embedding. CoRR. 2023;abs/2309.07597. Available from: http://dblp.uni-trier.de/db/journals/corr/corr2309.html#abs-2309-07597.

[19] @mrinaldi. Flash-IT-HA-Classifier-Cossim; 2024. Available from: https://huggingface.co/mrinaldi/gattina-ha-classifier-cossim.

[20] @efederici. Sentence-BERTino; 2022. Available from: https://huggingface.co/efederici/sentence-BERTino.

[21] @nickprock. MMARCO-bert-base-italian-uncased; 2023. Available from: https://huggingface.co/nickprock/mmarco-bert-base-italian-uncased.

[22] Chirkova N, Rau D, Déjean H, Formal T, Clinchant S, Nikoulina V. Retrieval-augmented generation in multilingual settings. arXiv preprint arXiv:240701463. 2024.
[23] Jeong C. Fine-tuning and Utilization Methods of Domain-specific LLMs. ArXiv. 2024;abs/2401.02981. Available from: https://api.semanticscholar.org/CorpusID:266844751.
[24] Chen ZZ, Ma J, Zhang X, Hao N, Yan A, Nourbakhsh A, et al. A Survey on Large Language Models for Critical Societal Domains: Finance, Healthcare, and Law. arXiv preprint arXiv:240501769. 2024.
[25] Iscan C, Ozara MF, Akbulut A. Enhancing RAG Pipeline Performance with Translation-Based Embedding Strategies for Non-English Documents. In: 2024 Innovations in Intelligent Systems and Applications Conference (ASYU); 2024. p. 1-6.
[26] Informatica RLR. wikipediaQA-ita: An Open Dataset of italian QA from wikipedia documents. ReDiX Labs; 2024. https://https://huggingface.co/datasets/ReDiX/wikipediaQA-ita.
[27] Li Z, Zhang X, Zhang Y, Long D, Xie P, Zhang M. Towards general text embeddings with multi-stage contrastive learning. arXiv preprint arXiv:230803281. 2023.
[28] Zhang X, Zhang Y, Long D, Xie W, Dai Z, Tang J, et al. mGTE: Generalized Long-Context Text Representation and Reranking Models for Multilingual Text Retrieval. In: Dernoncourt F, Preotiuc-Pietro D, Shimorina A, editors. EMNLP (Industry Track). Association for Computational Linguistics; 2024. p. 1393-412. Available from: http://dblp.uni-trier.de/db/conf/emnlp/emnlp2024i.html#ZhangZLXDTLYXHZ24.
[29] Hu X, Shan Z, Zhao X, Sun Z, Liu Z, Li D, et al. KaLM-Embedding: Superior Training Data Brings A Stronger Embedding Model. arXiv. 2025. Available from: https://arxiv.org/abs/2501.01028.