Retrieval-Augmented Generation systems for enhanced access to digital archives

Michele Ciletti[1]
[1] University of Foggia – michele_ciletti.587188@unifg.it

ABSTRACT (ENGLISH)

This study investigates the potential of Retrieval-Augmented Generation (RAG) to enhance access to digital archives related to humanities topics, using a case study on historical newspaper data. Over the past years, Large Language Models (LLMs) have shown remarkable capabilities in generating natural language; however, they still suffer from hallucinations and context window limitations. As a solution, RAG frameworks have emerged, linking a language model to external databases in order to provide grounded, reliable sources. In this work, an experimental RAG pipeline was tested on the "Foggia Occupator Dataset", which includes articles from a 1945–1946 periodical published by the US forces occupying Foggia, Italy, at the end of World War II. To assess performance, domain experts constructed a set of ten questions on topics such as historical figures, examined events, and stylistic aspects of the articles. The corresponding "ground truth" answers were used for both quantitative and qualitative evaluations. Results show that the RAG system performs well in precisely defined scenarios, retrieving accurate information and correctly identifying named entities. However, broader questions occasionally led to incomplete or slightly erroneous answers, pointing to potential areas for algorithmic refinement—particularly in optimizing retrieval for complex or multi-article queries. The study concludes that RAG can significantly improve the searchability and reliability of digital archives, but continued improvements in metadata enrichment, query strategies, and retrieval algorithms are needed.

Keywords: Retrieval Augmented Generation; Artificial Intelligence; archiving; newspapers; Large Language Models

ABSTRACT (ITALIAN)

Retrieval-Augmented Generation per un migliore accesso agli archivi digitali. Questo studio analizza il potenziale della Retrieval-Augmented Generation (RAG) per migliorare l'accesso ad archivi digitali relativi a tematiche umanistiche, utilizzando un caso di studio sui dati di un giornale storico. Negli ultimi anni, i Large Language Models (LLM) hanno dimostrato notevoli capacità di generare linguaggio naturale; tuttavia, soffrono ancora di allucinazioni e limitazioni delle finestre di contesto. Come soluzione, sono emersi i framework RAG, che collegano un modello linguistico a database esterni per fornire fonti fondate e affidabili. In questo studio, una pipeline RAG sperimentale è stata testata sul "Foggia Occupator Dataset", che comprende articoli di un periodico del 1945-1946 pubblicato dalle forze statunitensi che occupavano Foggia, in Italia, alla fine della Seconda Guerra Mondiale. Per valutarne le prestazioni, degli esperti hanno realizzato una serie di dieci domande su argomenti quali personaggi storici, eventi esaminati e aspetti stilistici degli articoli. Le corrispondenti risposte di "ground truth" sono state utilizzate per delle valutazioni quantitative e qualitative. I risultati mostrano che il sistema RAG ottiene ottimi risultati in scenari definiti con precisione, recuperando informazioni accurate e identificando correttamente le entità nominate. Tuttavia, domande più ampie hanno occasionalmente portato a risposte incomplete o leggermente errate, indicando potenziali aree di perfezionamento dell'algoritmo, in particolare nell'ottimizzazione del reperimento di query complesse o multiarticolate. Lo studio conclude che la RAG può migliorare significativamente la ricercabilità e l'affidabilità degli archivi digitali, ma sono necessari continui miglioramenti nell'arricchimento dei metadati, nelle strategie di interrogazione e negli algoritmi di recupero.

Keywords: Retrieval Augmented Generation; Intelligenza Artificiale; archivistica; giornali; Large Language Models

1. INTRODUCTION

In recent years, we are witnessing a growing interest in Large Language Models, both from the scientific community and the general population. Since OpenAI's GPT-3.5 model demonstrated unprecedented natural language generation capabilities in November 2022, considerable efforts in terms of time and funding have been undertaken to develop new, increasingly high-performance models. Today, the latest models obtain excellent results on specific tasks, including programming and solving mathematical problems (Jiang et al., 2024; Li et al., 2024).

The limitations of Large Language Models, however, persist. Parallel to the systematic expansion and improvement of the amount of training data, work is being done on the design of new frameworks and techniques that can mitigate the inherent flaws of LLMs. Sometimes accused of being little more than "stochastic parrots", Large Language Models produce content on the basis of statistical patterns derived from their training data (Bender et al., 2021). The absence of internal logic in favour of linguistic consistency results in a tendency to hallucinate, i.e. to produce untrue information confidently presented as factual (Perkovic et al., 2024). This leads to considerable risks, as users may fail to notice an unreported error in a large amount of text, and consequently spread misinformation.
Some companies have tackled the problem by expanding their context windows, or by equipping their models with the ability to surf the Internet in order to refine their output. Promising, but still embryonic, appears instead the new tendency to produce specialised models on reasoning, which reflect on their output and make corrections before presenting them to the user (Jaech et al., 2024). Another strategy, particularly appropriate for the application of LLM to third-party contexts (companies, public bodies), is Retrieval-Augmented Generation (Gao et al., 2023). The technique makes it possible to "link" a language model to a database of proprietary data, which may consist of a series of digitised books, a list of products on sale, a series of services available to a customer, or much more: what matters is that this database is converted into embeddings, multidimensional vectors useful for encoding natural language in a computational manner. In practice, when a user queries the model, their query is also converted into embeddings, which are then compared with the database: the most semantically similar documents - usually identified by calculating the cosine similarity between the vectors - are added to the input and passed to the model, which answers the query on the basis of the information contained in the documents themselves. This solution can provide a concrete and reliable database for the language model, without limiting its generative capabilities.
These considerations give rise to the research questions of this paper: can Retrieval-Augmented Generation be implemented in digital archives to facilitate the search and consultation of their materials? What are the guidelines to follow and good practices to adopt? Can sufficiently reliable and informative results be obtained in this way? To provide answers, it was decided to carry out a case study on a recently created database: the Foggia Occupator Dataset, which contains articles published between 1945 and 1946 in the Foggia Occupator, a generalist newspaper printed by the US forces that occupied the city of Foggia in the closing years of World War II (Ciletti et al., 2024). While similar experiments have been tried across various fields (Gupta et al., 2024), this study aims to provide a specific perspective on archives related to the humanities, while also focusing on elaborating insights for practical implementations.

2. METHODOLOGY

The chosen dataset contains 874 articles, totalling 216,015 words. Its content, which is not only historical, but also social and customary, lends itself favourably to thorough investigations and presents elements that require in-depth analysis to be fully interpreted. Its very recent publication also means that the likelihood of an LLM containing fragments of it in its training dataset is almost non-existent. However, possible limitations lie in the narrow scope of the articles, which deal with a short period, with events that took place in a restricted geographical area and are written entirely in a single language (English). The aim, therefore, is to offer an example case that can provide guidelines and practical feedback in terms of performance for those wishing to implement similar solutions in other contexts, rather than to fully evaluate the effectiveness and convenience of RAG across multiple realities.
The articles constituting the dataset, originally collected in a JSON file with simple metadata, were embedded using the text-embedding-3-large model (OpenAI, 2024) and stored in a Chroma database, an AI-native open-source vector storage solution (Chroma, 2025). The database was then linked to the OpenAI API. The RAG system was implemented via a script, which, once a query or request was received as input, converts it into embedding via the same model and identifies the five documents with the closest cosine similarity to the input. The number is indicative and was chosen in consultation with the experts involved in the creation of the dataset, based on an estimate of a reasonable number of relevant documents for each given topic. The request and documents are then sent together with a pre-determined system prompt to OpenAI's GPT-4o model (Hurst et al., 2024), which produces a response. It was decided to choose one of the best performing models at the time of writing, with the aim of testing the real capabilities of an advanced LLM under optimal conditions. This was done in the belief that, in a short time, the general performance of LLMs will be further improved and that equally satisfactory results will be

obtainable at a lower cost and, perhaps, even through open-source models. However, future studies comparing multiple models are certainly desirable.
To test the effectiveness of the answers generated, experts were asked to produce a set of questions and answers concerning certain topics covered in the dataset, accompanied by a variable number of specific relevant articles. In drafting the question-and-answer pairs, the need to investigate multiple aspects was taken into account, namely:

- The ability to identify and provide information on real people, political affairs and social events;
- The ability to reflect on the causes of events and political and social situations in a holistic manner;
- The ability to analyse and provide information on linguistic and stylistic choices.

An attempt was also made to produce questions on specific topics related to the history of the city and the newspaper, avoiding in most cases overly general topics on which there might be information in the model's traning dataset. In any case, a very low temperature was set in the model settings to ensure that the answers were as deterministic and reproducible as possible, as well as based as much as possible on the information contained in the documents. The set of questions and answers (n=10) is outlined in Table 1.

Query ID	Question	Expert answer	Model answer
1	Who was Savino Bufalo?	Savino Bufalo was the photographer of the Foggia Occupator.	Savino Bufalo was a staff photographer who took photos at events such as the Foggia Fair and other local happenings.
2	Who authored the column 'Static from your AES'?	The column 'Static from your AES' was authored by Ken Eill, however, in a single issue it was authored by Bob Dewart.	The column 'Static from your AES' was authored by Cpl. Ken Eill.
3	What were the main events of the 1945 Christmas celebrations in Foggia?	During the 1945 Christmas celebrations, Christmas trees were brought into town, soldiers dressed up as Santa Claus and gave gifts to children staying at the hospital. The EM Red Cross organised most of these celebrations. A party at the gym was also organised. An orchestra led by Frank De Grazia performed musical pieces across those days. The amount of children showing up to the gift distributions created some unrest.	The main events of the 1945 Christmas celebrations in Foggia included parties for children, where gifts were given to the sons of prisoners of war, and these celebrations continued through Christmas Day. The American celebrations featured Franck DeGrazia and his orchestra, who performed at various venues around town, including night spots and serenades on Christmas Eve. There were also Christmas dinners and dances that involved many local women and guests, including important Foggia civil officials and high-ranking Allied Forces officers. Additionally, Frank De Grazia and his group went on a serenade, spreading the Spirit of Christmas with their music throughout the town.
4	How was the Italian educational system structured in 1945-1946?	In 1945-1946, the Italian educational system saw children start elementary school at age 6, possibly after kindergarten. Then, there was the possibility to attend a vocational or classical school. Vocational schools were attended by the poorer children, who were then ready to go to work. Poverty, especially in Southern Italy, meant that many	The main events of the 1945 Christmas celebrations in Foggia included parties for children, where gifts were given to the sons of prisoners of war, and these celebrations continued through Christmas Day. The American celebrations featured Franck DeGrazia and his orchestra, who performed at various venues around town, including night spots and serenades on Christmas Eve. There were also Christmas dinners and dances that involved many local women and guests, including important Foggia civil officials and high-ranking Allied Forces officers. Additionally, Frank De Grazia and his group went on a serenade, spreading the Spirit of Christmas with their music throughout the town.

Query ID	Question	Expert answer	Model answer
		children did not attend compulsory school.	
5	What kinds of sports were practised in Foggia in 1945-1946?	The main sports practiced in Foggia in 1945-1946 were baseball, basketball and boxing, all of which had structured tournaments and thorough match reports in the Foggia Occupator. However, softball and handball were practiced too.	The main events of the 1945 Christmas celebrations in Foggia included parties for children, where gifts were given to the sons of prisoners of war, and these celebrations continued through Christmas Day. The American celebrations featured Franck DeGrazia and his orchestra, who performed at various venues around town, including night spots and serenades on Christmas Eve. There were also Christmas dinners and dances that involved many local women and guests, including important Foggia civil officials and high-ranking Allied Forces officers. Additionally, Frank De Grazia and his group went on a serenade, spreading the Spirit of Christmas with their music throughout the town.
6	Did protests, riots or violent events ever happen in Foggia, and if so, why?	Protests and riots happened, mainly because of rifts between locals and soldiers, strikes, food prices, requisitions and grain shortages.	Yes, protests and riots have occurred in Foggia. On March 31, there was an all-day rioting incident where one Italian was killed and eighteen were injured. Thousands of people, led by Foggia building workers, attacked a train and searched the city for black-market grain. This unrest was caused by an extreme shortage of grain in Southern Italy. The building workers were protesting for food and work, and the demonstrators also entered the local ration board and burned the records.
7	What was the political situation like in Italy, and especially in Foggia, in June 1946, during the days of the institutional referendum?	In June 1946, Italy voted for the establishment of the republic in a narrow referendum. The outcome was celebrated in Foggia, even though the monarchy won in the province by a considerable margin. While there was civic unrest in Italy, the situation in Foggia remained peaceful. Representatives from several parties were elected to the National Constituent Assembly. The Christian Democrats had the most, but there were also Socialists, Qualunquists and Communists.	In June 1946, Italy was undergoing a significant political transition. The country was recovering from World War II and the fall of the Fascist regime. A key event during this period was the institutional referendum held on June 2, 1946, where Italians voted to decide between maintaining the monarchy or establishing a republic. This was a crucial moment in Italian history, as it led to the abolition of the monarchy and the establishment of the Italian Republic. In Foggia, the political situation reflected the broader national context. The region participated in the elections for the National Constituent Assembly, which was tasked with drafting a new constitution for Italy. The election results in Foggia Province showed a diverse political landscape, with representation going to Christian Democrats, Communists, Socialists, and Qualunquists. Despite the national vote favoring the monarchy, Foggia Province had a majority voting for the monarchy by more than two to one. However, the transition to a republic proceeded without major incidents in the area, as indicated by the peaceful nature of public events like parades.
8	What does ‘Siponto’ mean, and in what context is it mentioned?	‘Siponto’ is the name of a maritime location near Foggia. It is mentioned in the Foggia Occupator because it was a popular destination for civilians and soldiers looking for leisure, who	"Siponto" refers to a location where there are clubs, specifically the Siponto Officers Club and the Siponto Beach Club. In the context provided, it is mentioned as a place where club managers are preparing for the summer season by constructing outdoor terraces for relaxation and building a new pier at the Siponto Beach Club. Additionally, discussions and activities related to the Italian Peace Treaty and Army vehicles are taking place at the Siponto Beach Club.

Query ID	Question	Expert answer	Model answer
		often went to the Siponto Beach Club.	
9	What are the contents of the column 'Nights' Round Foggia, and who is its author?	The column 'Nights 'Round Foggia' gives information about various entertainment events in Foggia, such as club events, concerts, parties and shows.	The column "Nights 'Round Foggia" is authored by Eddie M. Dunaj. It includes an account of the grand opening of the Castle Bar, a new officers' nightclub in Foggia. The column also features what club managers have to say about the appearance of "Nights 'Round Foggia" from the previous week.
10	How would you describe the writing style of the 'Girl About Town' column?	The column 'Girl About Town' is written in an informal and playful tone, reflecting its contents which concern social life in Foggia and local gossip.	The writing style of the column "Girl About Town" appears to be light-hearted and conversational, with a focus on social tidbits and gossip. The column seems to have a playful tone, as indicated by the repetition of the title and the mention of the writer's eagerness to please the editors or fill up space. It suggests a casual and engaging approach, likely aimed at entertaining readers with social news and anecdotes.

Table 1. The queries and answers set

All queries were subjected to the model. To assess the effectiveness in selecting relevant documents, classic measures such as precision and recall appeared inappropriate due to the incompatibility between the fixed number of documents selected for each query (5) and the variable number of relevant documents for each specific query selected by the experts. To overcome this, the selection of documents for each application was classified according to the following formula: (Number of relevant documents retrieved) / (min(Total number of relevant documents, 5)).
Secondly, the quality of the responses generated was assessed in relation to those provided by the experts. A first assessment was made by calculating the cosine similarity between the embedding of the answers generated by the model and those of the experts. Subsequently, the experts themselves qualitatively rated each question generated by the model on a scale from 1 (not at all satisfactory) to 5 (completely satisfactory), also providing a brief comment on the strengths and weaknesses of each one. Automatic evaluation systems, such as RAGAs (Es et al., 2024), were considered but ultimately discarded because of the manageable size of the questions and answers set. However, such approaches may be beneficial when dealing with larger datasets and future studies are encouraged to implement them.

3. RESULTS AND DISCUSSION

The performance of the RAG system is displayed in Figure 1. Overall, the cosine similarity between the model and expert responses is consistently high, with a mean of 0.92668 and a standard deviation of 0.031383. Retrieval performance is perfect in half of the cases, while it has a value of 0.8 in two others and 0.6 in three others. Three answers were rated as completely satisfactory by the experts (5 out of 5), four others scored 4 out of 5, two scored 3 out of 5 and one scored 2 out of 5. No response was rated as not at all satisfactory (1 out of 5).
Analysing the answers given by the model in relation to ground truth and the experts' comments, clear strengths and weaknesses appear. The model excels in the more specific answers, with clearly delineated areas and themes: query 4, which refers to a long article on the Italian education system, is particularly precise and detailed. On the contrary, broader queries, which ask for an overall analysis on phenomena spanning several articles and periods, sometimes lack precision or completeness. This is the case with query 3, which received the lowest score from the experts because, in the course of a long and otherwise valid answer, it placed a social event of February 1946 during the Christmas celebrations of 1945. This is the only case of completely wrong information that was found. Query 6, on the other hand, produced an accurate answer, but limited in its analysis, as it analysed a single protest event, when there were many others in the database. Problems of this kind are certainly attributable to the retrieval algorithm, which in this experiment is rather basic and for demonstration purposes: a weighting system of certain keyword-related embeddings could certainly mitigate the situation. The word 'Foggia', for example, should be weighted negatively, since almost all articles refer to the city context, but the algorithm, when the name of

the city is specified in the query, is inclined to prioritise documents that explicitly mention it. Not being able to rely so much on the users' ability to construct optimal prompts (which is why there was no intervention on the queries written by experts), algorithmic performance is certainly a point to invest in. Likewise, it was observed that the addition of a system prompt to guide the model significantly improves understanding of the general context in which the model operates.
On the other hand, an excellent performance was observed in interpreting stylistic elements of a given group of articles, especially with regard to query 10. Furthermore, a better performance was noted in queries that referred to specifically named headings, which leads one to think that the results can be further improved with the implementation of accurate and complete metadata, a practice that is already widely encouraged in the construction of digital archives.

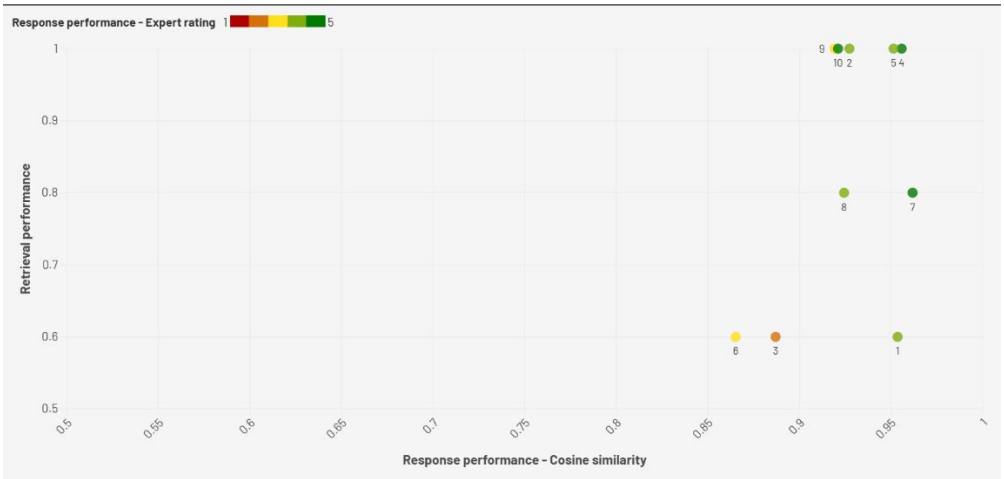

Figure 1. Performance analysis of the RAG system

4. CONCLUSIONS

Through the application of a RAG system to the Foggia Occupator dataset, the strengths and weaknesses of such approaches in making digital archives more accessible and easily searchable were reflected upon. Clear opportunities arise in making the contents of archives more immediate and effective, but doubts remain about accuracy and reliability. Points to work on to mitigate these problems are highlighted, mainly the enrichment of metadata, algorithmic refinement and the judicious choice of contexts to which to apply the systems. Further research testing this method against larger, more diverse text corpora are also desirable.
From the point of view of practical implementation, substantial obstacles remain in the costs of using LLM and the technical requirements needed to build, and then host, RAG-based systems. However, it has been demonstrated that moderate initial deployments are feasible even with relatively modest resources, and the hope is that costs and computational requirements will continue to decrease. Likewise, the proliferation of educational content on machine learning in general, which is often freely accessible, can only foster the development of digital skills necessary for the application of LLM-based systems, even by people who are not necessarily formally trained in computer science, such as staff in the GLAM sector. Effective implementation still requires competency across multiple domains including search engines, embeddings, document preprocessing, and LLM configuration—which is why collaboration between domain experts and technical scientist remains fundamental to bridge knowledge gaps. This collaborative approach becomes more feasible as technical literacy becomes increasingly attainable through accessible learning resources. It is hoped that future research will continue to analyse techniques and strategies to foster the implementation of machine learning technologies in humanistic contexts, with a focus on practical aspects and possible methodological improvements.

ACKNOWLEDGEMENTS

The authors have no competing interests to declare that are relevant to the content of this article.

REFERENCES

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? 🦜. In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency (pp. 610-623).

Ciletti, M., di Furia, M., Guarini, P., & Toto, G. A. (2024). Foggia Occupator: a case study on the creation of an Open Educational Resource through the digitization of a historical newspaper. In HELMeTO 2024 Book of Abstracts (p. 235).

Chroma. (2025). Chroma: The AI-native open-source embedding database [Computer software]. GitHub. https://github.com/chroma-core/chroma

Es, S., James, J., Anke, L. E., & Schockaert, S. (2024). Ragas: Automated evaluation of retrieval augmented generation. In Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations (pp. 150-158).

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Guo, Q., Wang, M., & Wang, H. (2023). Retrieval-Augmented Generation for Large Language Models: A Survey. ArXiv, abs/2312.10997. https://doi.org/10.48550/arXiv.2312.10997.

Gupta, A., Shirgaonkar, A., Balaguer, A. D. L., Silva, B., Holstein, D., Li, D., ... & Benara, V. (2024). RAG vs Fine-tuning: Pipelines, Tradeoffs, and a Case Study on Agriculture. arXiv preprint arXiv:2401.08406.

Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., ... & Kivlichan, I. (2024). Gpt-4o system card. arXiv preprint arXiv:2410.21276.

Jaech, A., Kalai, A., Lerer, A., Richardson, A., El-Kishky, A., Low, A., ... & Kaiser, L. (2024). OpenAI o1 System Card. arXiv preprint arXiv:2412.16720.

Jiang, J., Wang, F., Shen, J., Kim, S., & Kim, S. (2024). A Survey on Large Language Models for Code Generation. ArXiv, abs/2406.00515. https://doi.org/10.48550/arXiv.2406.00515.

Li, Q., Cui, L., Zhao, X., Kong, L., & Bi, W. (2024). GSM-Plus: A Comprehensive Benchmark for Evaluating the Robustness of LLMs as Mathematical Problem Solvers., 2961-2984. https://doi.org/10.48550/arXiv.2402.19255.

OpenAI. (2024). GPT-4o API. Retrieved from https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo.

OpenAI. (2024). text-embedding-3-large. Retrieved from https://platform.openai.com/docs/guides/embeddings.

Perkovic, G., Drobnjak, A., & Boticki, I. (2024). Hallucinations in LLMs: Understanding and Addressing Challenges. 2024 47th MIPRO ICT and Electronics Convention (MIPRO), 2084-2088. https://doi.org/10.1109/MIPRO60963.2024.10569238.