open proceedings

HistoricRAG: Evidence-Centered Newspaper Retrieval for Misinformation-Resilient Question Answering

Stergios Konstantinidis
University of Lausanne
Dept of Information Systems
Lausanne, Switzerland

Min-Yen Kan
National University of Singapore
Dept of Computing
Singapore, Singapore

Michalis Vlachos
University of Lausanne
Dept of Information Systems
Lausanne, Switzerland

Abstract

This demonstration presents HistoricRAG, an iPad-based application that lets users ask questions about historical or other events and immediately inspect the documentary evidence behind each answer. The system couples semantic retrieval over a large collection of OCR-processed Swiss newspapers with a generation component that summarizes the retrieved passages while preserving their provenance. Every response is accompanied by the original scanned articles, which can be explored along an interactive timeline or viewed through a map that reveals the geographic context of reporting. These features encourage users to cross-check claims directly against primary sources rather than relying on unverified narratives. By uniting efficient vector search, structured query interpretation, and evidence-centered presentation within a lightweight mobile interface, HistoricRAG offers a practical path toward misinformation-resilient question answering and demonstrates how generative models can be grounded in curated archival material for research, education, and public history.

Keywords

LLM, RAG, OCR, historical archives, library, information retrieval

1 Introduction

Many widely believed historical claims originate from anecdote rather than grounded evidence, and once embedded in public memory, they can be remarkably persistent. Two illustrative examples are the supposed mass panic triggered by Orson Welles's 1938 War of the Worlds broadcast and the long-standing assertion that the Tuskegee Airmen "never lost a bomber" during World War II. In the first case, it is still reiterated that listeners fled their homes in terror after mistaking a fictional Martian invasion for real news. Archival analysis of contemporaneous police logs, FCC correspondence, listener letters, and local press coverage, however, shows little evidence of widespread hysteria, and later scholarship demonstrates that the panic narrative was largely amplified by newspapers seeking to discredit radio [9, 13]. The second case involves the widely repeated claim (media and public speeches) that the Tuskegee Airmen unit never lost a single bomber they escorted in WWII. Detailed examination of historical records revealed that, while the Airmen performed exceptionally well, they did lose several bombers, and the perfection myth emerged only later through selective retellings [3].

These examples demonstrate how misinformation about real historical events can take hold and propagate for decades when later narratives overshadow primary sources. Motivated by this challenge, our work aims to support evidence-based answers grounded directly in historical documents, enabling users to verify claims against the archival record.

We present a system that answers natural-language questions using digitized historical newspapers. Our prototype leverages the scanned newspaper archives of the Canton of Vaud in Switzerland to support evidence-based exploration of regional history. We formulate the problem as an embedding-based retrieval augmented generation (RAG) task: each scanned article is transcribed through OCR, encoded into a high-dimensional semantic embedding space using a large language model, and indexed for efficient similarity search. User queries are embedded in the same space, and the top-k nearest articles are retrieved as candidate evidence from which the system synthesizes an answer. To promote transparency and verifiability, the original retrieved articles are always displayed alongside the generated response. Users can browse these primary sources either on a temporal axis, visualizing the evolution of events over time, or on a geographical map that highlights spatial relationships relevant to the query. To ensure both high performance and correctness, we employ a metric-tree index structure that guarantees no false negatives in k-nearest-neighbor retrieval while supporting low-latency exploration at scale. By integrating data management techniques with RAG, our system offers an interpretable and provenance-aware alternative to modern information systems, enabling fact checking directly against historical documents rather than relying solely on model-generated text.

2 Background and Motivation

False information, today also known as fake news, is an important research topic in information integrity. The terms of misinformation and disinformation are two terms often used interchangeably but represent different phenomena. Misinformation refers to inaccurate or misleading claims shared without deceptive intent, typically arising from misunderstanding, limited context, or cognitive biases that shape how individuals evaluate evidence [8]. Disinformation, in contrast, is the deliberate creation or strategic dissemination of falsehoods aimed at influencing public opinion or undermining trust in institutions, often amplified through coordinated online behaviors [5, 14]. Prior work catalogs how both processes propagate across digital media ecosystems, how structural features of platforms shape their reach [12], and also how fake news can be detected using either traditional machine learning methods [1, 4, 16] or via LLMs [7].

While previous studies focus on detecting or modeling credibility signals, our approach focuses on historical grounding: rather than inferring truthfulness from patterns of spread or linguistic cues, the system retrieves primary passages from digitized newspapers, synthesizes the answer to a question from them, and provides access to the archival record itself for additional verification. This emphasis on transparent sourcing positions

EDBT '26, Tampere (Finland)

10.48786/edbt.2026.65

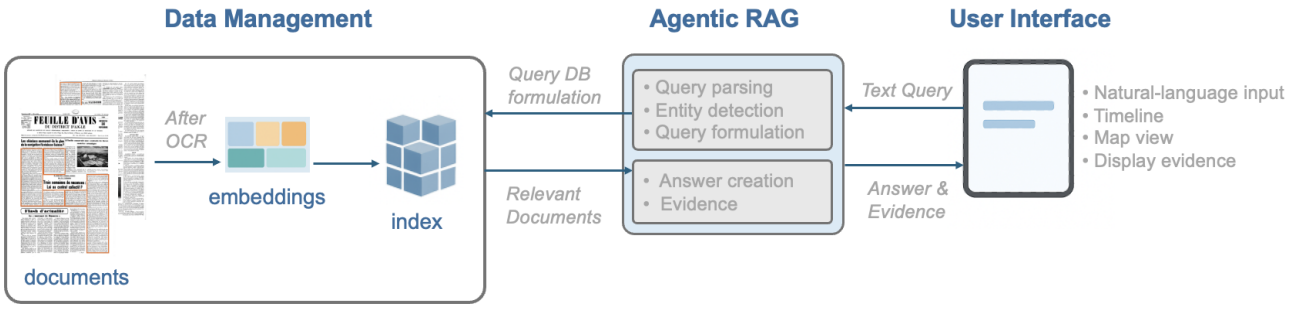

Figure 1: Overview of our system

the prototype as an alternative to traditional credibility interventions, shifting the goal from filtering untrustworthy content to enabling users to explore the historical record from which interpretations should arise.

Recent advances in LLMs introduce new challenges in this landscape. Because LLMs can generate fluent, coherent, and contextually tailored text at scale, they lower the barrier for producing fabricated narratives, persuasive false explanations, and synthetic news articles. Studies show that model-generated content can appear credible even when factually incorrect, contributing to the rapid spread of unverified claims when circulated without grounding in primary sources [11, 15, 17]. These findings highlight the importance of retrieval-augmented and provenance-aware approaches that anchor generated responses in verifiable documents, particularly as generative models become more capable and widely accessible.

3 System Overview and Implementation

Our system consists of:

- A data management component that stores the source data and their vector representations in the form of embeddings. It additionally maintains an index that enables efficient k-NN retrieval.
- A RAG agentic API service that parses the textual query, extracts key entities such as temporal constraints, converts the query into an internal structured form, retrieves the k-nearest neighbors, and generates the textual response together with the supporting evidence.
- A user interface that receives the user query, forwards it to the API, and displays the final answer along with image evidence from archival newspapers. The evidence can be viewed as a timeline ordered by publication date or as geospatial information on a map when the article contains a location.

Data Management: The dataset consists of scanned images of pages from historical newspapers published in the French-speaking canton of Vaud in Switzerland. The cantonal library provides digitized newspapers dating back to the eighteenth century, offering extensive material on both local and global news, always through the lens of the newspapers of that era. After scanning, the pages undergo OCR processing and are segmented into titles, articles, tables, and other components. The text of each segment is converted into a high-dimensional embedding and stored in a metric-tree index [6]. Such an index structure supports fast k-NN search without false negatives, which may be the case when approximate nearest neighbors methods are used [2], ensuring that all true nearest neighbors are found, which is essential for evidence-based retrieval.

RAG Service: When the user submits a query, it is processed by the RAG search API. The query is decomposed into key entities such as "time", "location", "search concept", and "periodical" in order to support flexible filtering. For instance, in a query like "Find what Swiss newspapers say about [topic] after the 1950s", the phrase "after the 1950s" is interpreted as a temporal filter that can be applied prior to embedding search. Likewise, for a query such as "What does *Le Temps* mention in 1982 about [event]", both temporal and periodical constraints must be recognized. A lightweight agentic system identifies and isolates these entities, then formulates the corresponding structured query for the internal data management layer, implemented in MySQL. For the underlying LLM powering the agentic component, we use OpenAI's GPT 4.1 mini because of its speed and cost effectiveness. Once the main "search concept" is isolated, its embedding is computed and then its k-Nearest-Neighbors are retrieved using the data index. Not all k neighbors are relevant, so we discard those that are not. We assign each article d_i a cross encoder relevance score [10]

$$\mathrm{CES}(q, d_i) = f_\theta(q, d_i),$$

where q is the query, d_i is the passage, and f_θ is the cross encoder with parameters θ. We also compute a Relative Relevance Threshold $\mathrm{RRT}(i) = \frac{s_i}{s_1}$, with s_i the embedding similarity for d_i and s_1 the highest similarity among neighbors. Passages are dropped when their cross encoder score is below a minimal value or when $\mathrm{RRT}(i) < \tau$. The RAG service returns both the synthesized response and all scanned pages that pass the threshold test as evidence.

User Interface: The user interface provides a natural-language search field through which the user submits queries. When the RAG service produces a response, the interface displays the synthesized answer, a timeline view of the retrieved evidence, and a map representation of locations extracted from the OCR text. The timeline is ordered by the publication date of each periodical, and the map is populated using any location entities mentioned in the corresponding articles. If no explicit location appears in an article, the system defaults to the location of the newspaper or periodical in which the article was published.

4 Demo Description

In the demo, we will showcase the user interface and provide several examples of the system responses. The interface is an iPad application written in Swift and SwiftUI.

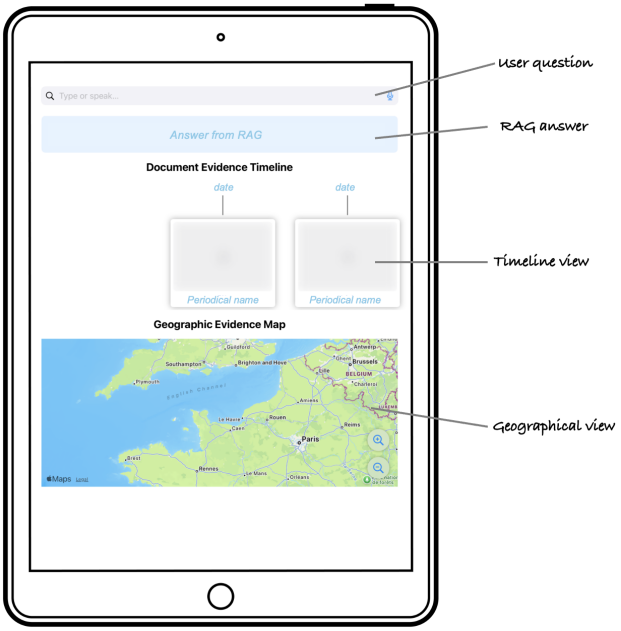

Figure 2: The iPad user interface

Below we show the response of the system to three queries. A video of these use-cases which demonstrates the fluidity of the interface is available at https://youtu.be/H_T2Yn4I7aM.

5 Use Cases

Query 1: Price of a Mazda 626 in Switzerland in year 2000. This is a numerical query for which the RAG retrieves a relevant advertisement and provides a precise answer of 28,950 Swiss francs (Figure 3). The user can click on any retrieved page to see it in higher resolution. For comparison, we also asked the same question to ChatGPT 5.1, which could not provide an official list price, but was able to produce an estimated range of 18,000 to 25,000 francs based on web search results.

Figure 3: Asking the price of a Mazda model in year 2000.

Query 2: What do different places report about the *Fête des Vignerons* in the 1900s? The Fête des Vignerons is a multi day celebration honoring Swiss winemakers, held approximately every 25 years. The results for this query demonstrate the usefulness of the timeline view; it highlights publications from 1905, 1927, 1955, 1977, and 1999, the years in which the celebration occurred.

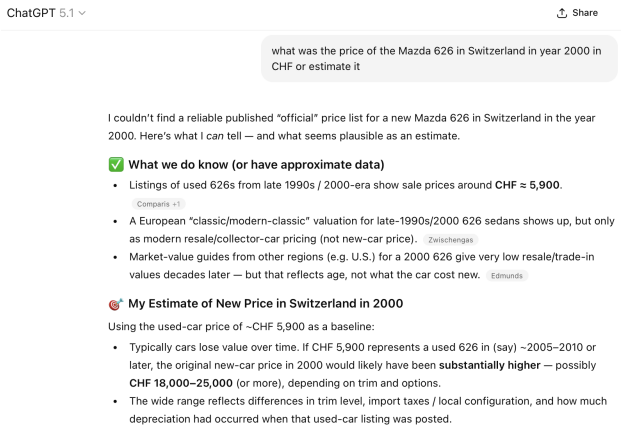

Figure 4: Answer of ChatGPT 5.1 for the same question

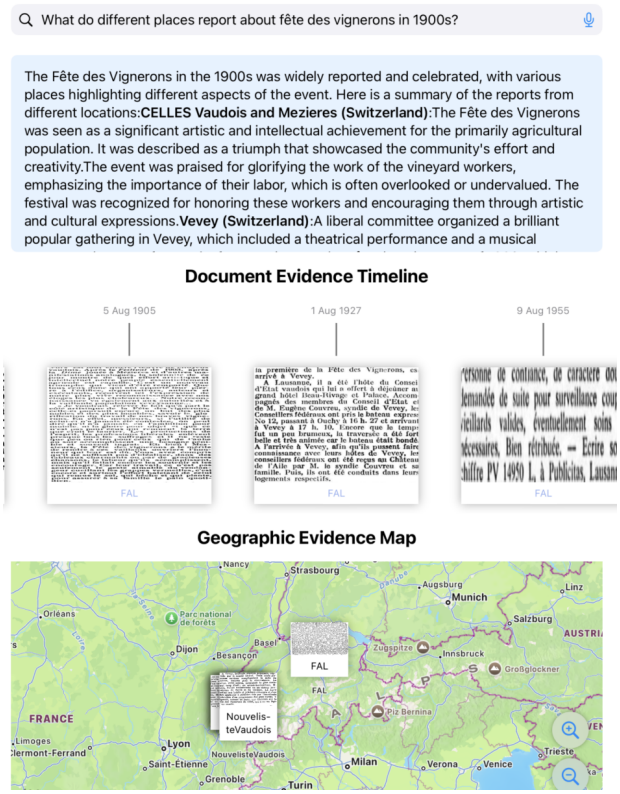

Figure 5: The query is about the "Fête des Vignerons" a generational celebration of wine makers in Switzerland.

Query 3: What did journalists write about the Garibaldi case? The question refers to Giuseppe Garibaldi's 1860 military campaign in Sicily, a pivotal moment in the unification of Italy. Newspapers from the 1860s report the story in contrasting tones that reflected the political and ideological divisions of the time. The interface here highlights these documents through a timeline that displays evidence from May and June of 1860, showing how perceptions evolved as Garibaldi advanced through southern Italy. The geographic map view complements this by illustrating

the Italian regions mentioned in the articles, allowing users to see how journalistic coverage was tied to specific locations involved in the campaign.

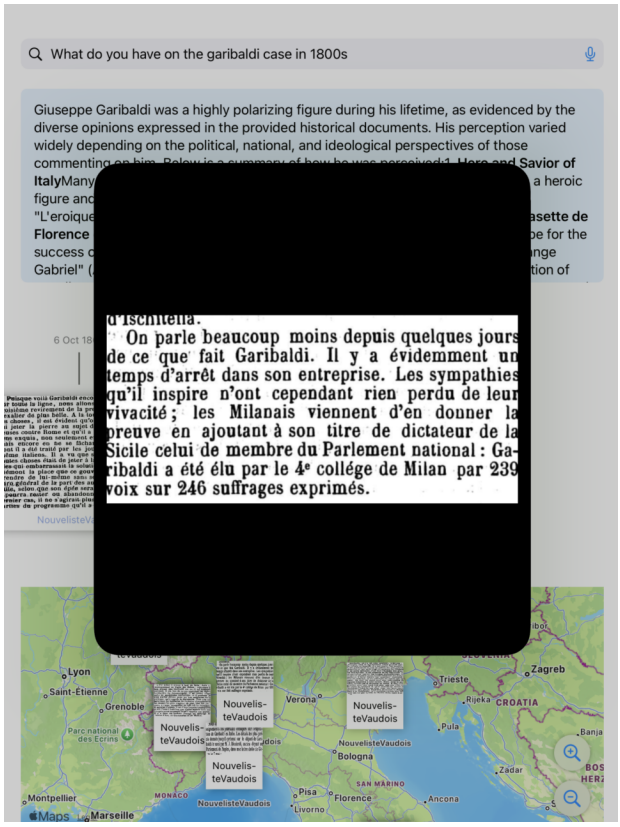

Figure 6: The Garibaldi case in 1860, showing different locations in Italy.

6 Conclusion

We have presented HistoricRAG, a system novel in its integration of RAG with data from a real-world archival library in Switzerland, providing a concrete demonstration of the potential of this combination. The system shows that retrieval-augmented generation grounded in digitized historical newspapers can produce misinformation-resilient answers by linking natural language queries directly to verifiable primary sources. By combining data management, semantic retrieval, and transparent evidence presentation, the prototype enables users to examine historical claims with both accuracy and interpretability. The integration of timeline and geospatial visualizations further supports contextual exploration, making archival materials accessible beyond traditional keyword search. As large language models continue to reshape information use, our approach offers a scalable, provenance-aware framework for fact checking, education, and historical inquiry, underscoring the value of pairing generative models with curated documentary evidence.

Acknowledgments

This work was supported by the Swiss National Science Foundation (SNSF), Switzerland, under Grant No. 237991.

References

[1] Esma Aïmeur, Sabrine Amri, and Gilles Brassard. 2023. Fake news, disinformation and misinformation in social media: a review. *Social Network Analysis and Mining* 13, 1 (2023), 30.
[2] Martin Aumüller, Erik Bernhardsson, and Alexander Faithfull. 2020. ANN-Benchmarks: A benchmarking tool for approximate nearest neighbor algorithms. *Information Systems* 87 (2020), 101374.
[3] Daniel L Haulman. 2017. False claims about the Tuskegee Airmen. *Air Power History* 64, 1 (2017), 47–55.
[4] Md Rafiqul Islam, Shaowu Liu, Xianzhi Wang, and Guandong Xu. 2020. Deep learning for misinformation detection on online social networks: a survey and new perspectives. *Social Network Analysis and Mining* 10, 1 (2020), 82.
[5] David M. J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. 2018. The science of fake news. *Science* 359, 6380 (2018), 1094–1096. doi:10.1126/science.aao2998
[6] David Novak, Michal Batko, and Pavel Zezula. 2011. Metric index: An efficient and scalable solution for precise and approximate similarity search. *Information Systems* 36, 4 (2011), 721–733.
[7] Eleftheria Papageorgiou, Christos Chronis, Iraklis Varlamis, and Yassine Himeur. 2024. A survey on the use of large language models (LLMs) in fake news. *Future Internet* 16, 8 (2024), 298.
[8] Gordon Pennycook and David G. Rand. 2021. The psychology of fake news. *Trends in Cognitive Sciences* 25, 5 (2021), 388–402. doi:10.1016/j.tics.2021.02.007
[9] Jefferson Pooley and Michael J. Socolow. 2013. The Myth of the War of the Worlds Panic. *Slate* (2013). https://slate.com/culture/2013/10/orson-welles-war-of-the-worlds-panic-myth-the-infamous-radio-broadcast-did-not-cause-a-nationwide-hysteria.html Online article.
[10] Ronak Pradeep, Yuqi Liu, Xinyu Zhang, Yilin Li, Andrew Yates, and Jimmy Lin. 2022. Squeezing water from a stone: a bag of tricks for further improving cross-encoder effectiveness for reranking. In *European Conference on Information Retrieval*. Springer, 655–670.
[11] Hannah Rashkin, David Reitter, Gaurav Singh Tomar, and Dipanjan Das. 2021. Increasing faithfulness in knowledge-grounded dialogue with controllable features. *arXiv preprint arXiv:2107.06963* (2021).
[12] Dietram A. Scheufele and Nicole M. Krause. 2019. Science audiences, misinformation, and fake news. *Proceedings of the National Academy of Sciences* 116, 16 (2019), 7662–7669. doi:10.1073/pnas.1805871115
[13] A. Brad Schwartz. 2015. *Broadcast Hysteria: Orson Welles's War of the Worlds and the Art of Fake News.* Hill and Wang.
[14] Karishma Sharma, Feng Qian, He Jiang, Natali Ruchansky, Ming Zhang, and Yan Liu. 2019. Combating fake news: A survey on identification and mitigation techniques. *ACM Trans. on Intelligent Systems and Technology* 10, 3 (2019), 1–42.
[15] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, et al. 2022. Taxonomy of risks posed by language models. In *Proceedings ACM conference on fairness, accountability, and transparency*. 214–229.
[16] Weizhi Xu, Junfei Wu, Qiang Liu, Shu Wu, and Liang Wang. 2022. Evidence-aware fake news detection with graph neural networks. In *Proceedings of the ACM web conference 2022*. 2501–2510.
[17] Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending Against Neural Fake News. In *Advances in Neural Information Processing Systems*.