From RAGs to rich responses

Enhancing LLM reliability through retrieval-augmented generation

Holle Meding and Aurel Daugs

1. Introduction

Retrieval-Augmented Generation (RAG) offers a technical response to a key problem in the use of Large Language Models (LLM) across the humanities: their tendency to hallucinate facts in the absence of verified sources. By supplementing the generative process with targeted retrieval from external resources such as archival corpora, knowledge graphs and document databases, RAG shifts the epistemic burden from the model's internal weights to verifiable information. This article reports on the early-stage implementation and critical evaluation of a RAG pipeline developed at the Chair for Digital History at Humboldt-Universität zu Berlin for historical text analysis using a broad *DER SPIEGEL* magazine corpus containing more than 100,000 articles. While this collection covers a wide variety of themes, we illustrate its use here via a case study on decolonisation in Asia and Africa (1940s–1970s) (Hiltmann et al., 2025).

The pipeline comprises four key stages: (1) *Chunking*, which segments documents into coherent units that preserve semantic meaning within token limits; (2) *Embedding*, using a pre-selected *Sentence Transformer* model to encode each *chunk* into high-dimensional vector representations, which were then stored in *ChromaDB*; (3) *Retrieval*, using cosine-similarity-based matching to identify and return thematically relevant text fragments in response to user queries; and (4) *Generation*, in which the retrieved content was inserted into the prompt of an LLM (HU-LLM-3 and GPT-4o) to produce contextually grounded responses.

The RAG pipeline was developed in a collaborative research project with students aimed at exploring the potential and limitations of LLMs for historical research, while being designed with broader applicability in mind. Its modular design, corpus-driven architecture and transparent evaluation, with all the challenges that entails, allow it to be adapted for a wide range of use cases across the humanities and social sciences, wherever source-critical text generation is required.

The article addresses an interdisciplinary readership, including historians who are beginning to integrate LLMs into their research workflows, as well as scholars with more advanced experience in computational methods. The subject is complex, and we have sought to provide a transparent account of RAG's retrieval and generation workflow, while maintaining accessibility for readers with varying degrees of technical proficiency. Explanatory boxes clarify key terms to support those less familiar with the field without oversimplifying the challenges.

This contribution offers an overview and evaluation of RAG's practical utility in historical scholarship, while also addressing challenges common to its wider use across the humanities and social sciences. We argue that its effectiveness depends on informed prompt engineering based on high-quality retrieval sources and ongoing domain-specific evaluation.

2. Retrieval-augmented generation (RAG)

Retrieval-augmented generation, or RAG for short, refers to an architecture that combines LLMs with external knowledge bases, for example, digitised archives, institutional records or curated corpora, to produce more context-sensitive responses. Rather than relying solely on the implicit knowledge encoded in the LLM's training, a RAG system retrieves relevant texts in real time and uses them to contextually ground its outputs. The core objective of RAG is to reduce hallucinations and enhance the factual reliability of generated content by introducing a retrieval mechanism, without needing to retrain the base LLM from scratch (Lewis et al., 2020; Schuster et al., 2021).

Most commonly, RAG systems employ either sparse or dense retrieval methods. The first is keyword-based and operates strictly at the lexical level, while the second embeds both the user query and segments of the knowledge base into a shared semantic vector space, retrieving relevant passages based on similarity metrics (e.g., cosine similarity). Because both approaches have distinct strengths and limitations, hybrid retrieval, which combines both methods, is often used to balance performance, improving precision on rare or exact terms while maintaining strong recall for semantically similar content (Fan et al., 2024: 6492f.).

RAG has been successfully applied across multiple domains and shows particularly promising results in healthcare. For instance, MMed-RAG (Xia et al., 2024) introduced a versatile multimodal RAG system for medical vision LLMs, achieving an average improvement of 43.8% in factual accuracy compared to baseline generation. In a real-world deployment, Nguyen-Duc et al. (2025) present a multi-agent, retrieval-augmented chatbot system designed for university admissions counseling in Vietnam (MARAUS). Their system demonstrates high accuracy, averaging 92% across more than 6,000 user interactions, while reducing hallucination rates from 15% to just 1.45%.

In the field of history, philosophy and sociology of science (HPSS), however, RAG remains a relatively recent development with only a handful of projects exploring its potential. One such example is the work by Sergeev et al. (2025), who developed a RAG-based chatbot for searching and summarising large, heterogeneous humanities datasets. Through a hybrid strategy combining full-text and semantic retrieval, their findings demonstrate promising results in improving access to archival materials such as the Prozhito diary entries from 1900–1916, though the work lacks user-centered evaluation by historians or archivists. Similarly, Hill (2026) introduces the 'data interview', a hybrid methodological approach that integrates RAG into qualitative workflows. Drawing on a corpus of 2.5 million anonymised Facebook comments, Hill explores how LLMs can be integrated into reflexive, interpretative research, thereby offering a critical template for integrating automation into social inquiry. Her approach both leverages RAG to enhance qualitative analysis and contextualises LLMs within the broader trajectory of knowledge production.

Several recent studies, while demonstrating significant improvements in factual accuracy compared to baseline model outputs, identify the retrieval phase as a critical area for further development. As RAG systems remain vulnerable to noisy or fabricated content in the retrieved documents, they can, in some cases, introduce additional hallucinations based on irrelevant or misleading retrieval results (Fan et al., 2024: 6498; Yang et al., 2024; Asai et al., 2024). Taking a more critical stance, Chatzikyriakidis (2025) examines the presumed generalisability of RAG systems and their claimed performance gains in the domain of semantic historical event extraction. Using Thucydides' *History of the Peloponnesian War* as a controlled test case, he shows that purely inferential generation not only outperforms enhanced RAG configurations in extensively documented domains overall but also delivers markedly higher accuracy. Indeed, RAG can sometimes diminish output quality by distorting the semantic relationships between historical actors and events (8).

In summary, these concerns underscore the need for a robust evaluation step, not only for the specific implementation of our SPIEGEL-RAG pipeline, but more broadly as a response to the epistemic opacity of such systems. Consistent with the principles of the historiography and HPSS, this evaluation should elucidate both the technical components (e.g. retrieval algorithms and similarity metrics) and the interpretative frameworks embedded in the system prompts, thereby keeping methodological transparency central to historical research (Simons et al., 2026; Oberbichler and Petz, 2025).

In its basic form, the framework follows a multi-stage processing pipeline:

1. **Database setup**: Relevant documents are first indexed within a structured database.
2. **Retrieval phase**: Using dense retrieval, embedded document representations are then semantically fetched based on the user's query.
3. **Prompt construction**: An input prompt is crafted that incorporates both the user's original question and the retrieved content from the knowledge base.

This sequential design allows RAG systems to selectively search targeted collections and feed relevant excerpts into the language model, resulting in responses grounded in verifiable context. Additionally, it enables source attribution by linking generated outputs to specific retrieved documents, allowing users to trace the origins of the information and verify its accuracy.

3. Practical applications of RAG in the historical sciences

As part of the LLM project at Humboldt-Universität zu Berlin involving researchers and graduate students, we developed a RAG pipeline to support the analysis of historical texts from the West German news magazine *DER SPIEGEL*. Thematically, the project aimed to trace how the process of decolonisation in Asia and Africa was represented in *SPIEGEL*'s reporting between the late 1940s and the 1970s. Through individual case studies, students explored both the potential and the limitations of integrating LLMs into the historical research process.

One of the key challenges in this historical analysis was the implicit nature of references to decolonisation in the analysed articles. Rather than using the term 'Dekolonisation' ('decolonisation') directly, SPIEGEL authors discussed the topic through contemporary language such as 'Befreiungsbewegung' ('liberation movement'), 'Dritte-Welt-Bewegung' ('Third World movement'), or 'Entwicklungspolitik' ('development policy').

This diversity in terminology poses serious difficulties for traditional, rule-based text-mining approaches which rely on keyword matches. Conversely, this challenge highlights the particular strength of LLMs when used in conjunction with a RAG pipeline: In this setup, historical terms can be embedded into a high-dimensional semantic space, allowing the system to retrieve text passages that are not lexically identical but semantically similar. The LLM can then analyse these retrieved texts through what Simons et al. (2026: 5) term 'interpretive sparring', helping to identify subtle conceptual overlaps and latent discursive patterns, potentially revealing new perspectives on how decolonisation was framed or narrated, even when the term itself was not explicitly mentioned (McGillivray et al., 2024).

In our configuration, retrieval is not primarily a fact-checking or citation mechanism, but a tool for exploratory discovery. This design offers two key advantages. First, semantic retrieval expands the scope of inquiry by surfacing relevant sources that conventional keyword searches would miss, thereby drawing attention to documents that engage with related ideas without using expected terminology. Second, by making the connection between retrieved texts and generative output explicit, the system takes a necessary step towards transparency in alignment with academic integrity. Researchers can verify the sources underlying the model's interpretations, ensuring that computational insights remain accountable to the historical record. In this way, the pipeline supports large-scale interpretative analysis while preserving the core scholarly principle of source-based verification.

4. Understanding the RAG architecture: chunking, embeddings and retrieval

To lay the groundwork for analysing and implementing a RAG pipeline, an external knowledge base is required. In our case, this took the form of a large corpus of 105,257 articles from the *DER SPIEGEL* online archive, extracted via web scraping and converted into a structured CSV database. The corpus comprises approximately 84 million *tokens*, an extensive body of historical material that would be too large to analyse thoroughly without computational methods.

However, this dataset also presents a number of qualitative limitations that must be made explicit. As with most digitisation projects, a degree of information loss is unavoidable (Porter, 2021). Original images have been removed and the physical layout of the articles is no longer preserved, making the visual experience that a contemporary reader would have had impossible to reconstruct. In addition, the text contains occasional OCR errors, especially in the rendering of foreign names with diacritics or non-German characters such as ç, ğ, ł, or ñ, which are often misrecognised or replaced with incorrect symbols. No official documentation has been released by *SPIEGEL* detailing the digitisation process undertaken in 2007.

While this lack of transparency complicates assessments of data provenance and technical procedures, it is not an isolated case. On the contrary, it reflects a broader challenge faced by historians working with digitised source material. The absence of metadata or processing protocols is symptomatic of the infrastructural realities that shape much of today's digital source landscape. That such a large corpus has been made available at all should not be taken for granted. It is the outcome of a major archival effort from which historical research continues to benefit, even as we, as researchers, grapple with its limitations.

Despite these challenges, the corpus provided a workable and sufficiently rich basis for building and testing the RAG pipeline. Before the texts could be processed by the language model, however, they needed to be adapted to the technical requirements of semantic retrieval and generation. We first segmented the texts into smaller, yet semantically coherent units – commonly referred to as *chunks* – in order to remain within the context window limits of the models. This step was essential not only for technical compatibility but also for structuring the data in a way that preserves interpretative coherence. To that end, we employed a *context-overlap* strategy, that is, the final portion of each *chunk* was repeated at the beginning of the subsequent one. This approach preserves continuity between sections, ensuring that key semantic relationships were not lost during processing.

> A *token* represents the smallest unit of processing in Natural Language Processing (NLP). Depending on the *tokenisation* method used, it may correspond to an entire word, a word fragment, or a punctuation mark.

For the RAG pipeline, we applied *recursive chunking* with a *chunk size* of 500 *tokens* and an *overlap* of 50 tokens.

> *Chunking* is a method of progressively breaking down text, where a text is divided into smaller segments according to a fixed number of characters or by using delimiters (e.g., paragraphs, sentences). In recursive chunking, the text is not simply divided once into fixed sections (e.g., paragraphs), but if the initial division does not yield the desired segment size or structure, smaller units based on delimiter criteria (e.g., sentences) are used step by step. The aim is to produce semantically coherent and roughly equally sized *chunks* without imposing rigid boundaries, thereby preserving the context within the segments as effectively as possible.

Subsequently, we converted each of these *chunks* into *sentence embeddings* using a *Sentence Transformer model* (*all-MiniLM-L6-v2*) and stored them in the open-source vector database

ChromaDB. Instead of representing words merely as sequences of characters, the model captures relationships between words and sentences and expresses them in a numerical format.

> *Embeddings* are high-dimensional numerical vector representations of text that capture semantic meaning relative to context. Because language models do not 'understand' text as humans do, they instead rely on *embeddings* to translate language into a mathematical form suitable for computation. *Embeddings* are essential for semantic search, enabling systems to retrieve conceptually similar text even without shared keywords.

The retrieval module identified the most semantically relevant *chunks* in response to each user query. We converted the query into *embeddings* and compared them with the stored vectors using *cosine similarity*, allowing the module to fetch not only exact matches but also conceptually related passages. These retrieved *chunks* ultimately served as the contextual basis for generating the model's response.

> *Cosine similarity* is a metric used to measure the semantic proximity between two vector representations by calculating the cosine of the angle between them. The result always falls between –1 (vectors pointing in exactly opposite directions) and +1 (identical direction), with 0 indicating orthogonality. Higher scores (closer to +1) correspond to stronger semantic similarity, while values towards –1 indicate increasing dissimilarity.

5. Towards reliable LLM output: system prompts for uncertainty and source traceability

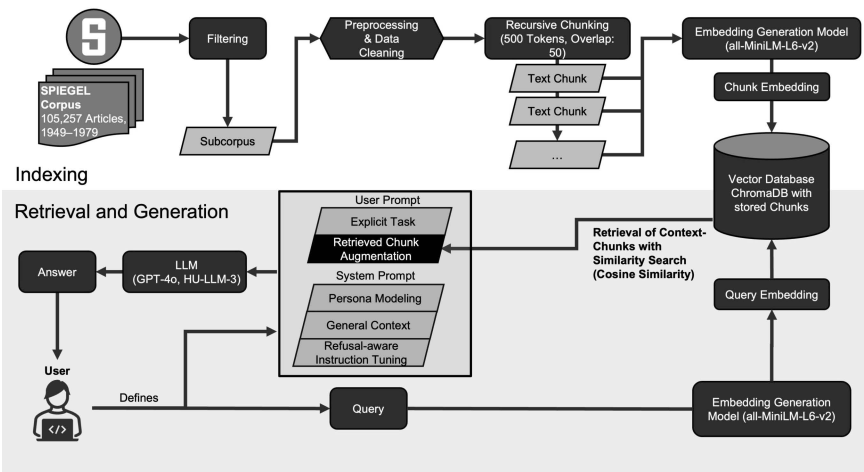
Ensuring the reliability of LLM-generated responses requires careful prompt design (Chen et al., 2025). We implemented a *system prompt* that grounds the model's answers strictly in the documents retrieved during the search phase (Box 1). To reduce hallucinations, the prompt explicitly instructs the model to acknowledge uncertainty. When no relevant information is found, it responds with 'I cannot answer that', rather than speculating. In an ideal scenario, this approach maintains the integrity of the output and reinforces source traceability within the workflow. In addition, we employed role-based prompting, in which the model was instructed to act as a historian with specialised expertise in the thematic case study under investigation and to contextualise its responses with reference to the relevant time frame (1940s–1970s). This framing is intended to guide the model towards historically grounded, context-sensitive responses.

> A *system prompt* is a high-level instruction that defines fundamental rules for an LLM such as its behaviour, tone, and scope of responses. It is included with every API call alongside the user prompt and any prior messages, ensuring that its influence persists throughout the interaction.
>
> In contrast, a *user prompt* is formulated individually for each query. While the system prompt governs the model's overall behaviour, the user prompt determines the specific content of a given interaction.

Box 1 **Sample system prompt used in the RAG pipeline, own translation (see Appendix A, Box 1 for the original system prompt).**

System prompt (English translation): You are a historian with specialised expertise in the Algerian War and the role of Charles de Gaulle. Your task is to analyse articles from the West German news magazine *DER SPIEGEL*, published between the 1940s and the 1970s, and to answer the following question as accurately and factually as possible. Your response must rely exclusively on the content of the retrieved *SPIEGEL* articles. Do not draw on external sources or make inferences beyond what is stated in the texts. If the retrieved material does not provide sufficient or clear information, make this explicit and respond with: 'I cannot answer that.'

Fig. 1: Meding and Daugs: Early-stage RAG Pipeline for the Analysis of Historical Texts from the SPIEGEL Online-Archive (1940s–1970s).



To generate the final output, the *system prompt*, the *user query* with a specific research question, and the relevant context retrieved by the RAG system were passed to each model via the API. Since the sources were in German, all inputs and outputs were also formulated and generated in German to ensure linguistic alignment. However, for future applications, it would be worthwhile to test prompt strategies in English and evaluate whether English, German or a multilingual approach yields better performance. Moreover, although the chosen prompts were tested for effectiveness, more extensive evaluations, including longer prompts, would help to determine which elements contribute most to plausible and relevant responses.

In summary, this pipeline created a controlled environment for generation. It ensured the retrieval of relevant text passages and ensured the traceability of the sources used by the model for its response. In the next phase, each retrieved *chunk*, the model's output and its source-boundedness could be individually evaluated.

6. Challenges in evaluating RAG for historical research

Language models generate output through probabilistic next-token prediction, and therefore their use inevitably introduces a degree of uncertainty, posing challenges for their reliable integration into research workflows (Meding and Daugs, 2026). The step of evaluation is therefore vital. Every prompt needs to be clearly defined, and model responses must be checked against expected results.

The same principle applies to RAG systems. While the citations in the generated responses correspond to actual sources retrieved from the database, these sources can still be irrelevant or only loosely connected to the query, which affects the reliability of the output. Thus, it is not sufficient to judge the quality of the output alone. The relevance and appropriateness of the material retrieved for the model to generate its response are equally important.

In historical research, this evaluation is especially critical. The selection and retrieval of source material, as well as the relevance and plausibility of the generated analysis, require close examination and assessment. Neglecting either risks accepting outputs that may appear credible but ultimately misrepresent historical evidence.

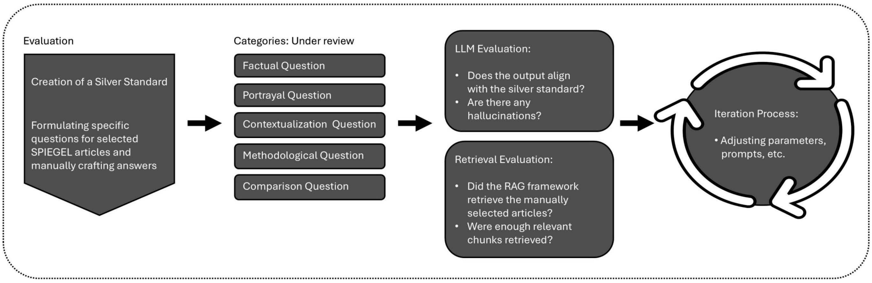
6.1 Establishing a silver standard in the absence of objective ground truth

Evaluating the retrieval component of a RAG system requires a clear reference framework to judge the quality of its results. Ideally, this would be a *ground truth*, that is, a fully validated and objective standard against which outcomes can be measured. Yet in historical research, such a definitive standard rarely exists. History is always shaped by perspective and is inescapably framed by the vantage point from which it is written. Every narrative bears the imprint of its author's cultural assumptions, interpretative choices and the intellectual climate of the era. Historical truth or objectivity, therefore, is best understood as a theoretical ideal rather than a fixed reality (Jordan, 2010; Hiltmann, 2024: 205f.), complicating efforts to evaluate in this field.

Given this difficulty, we developed a *silver standard* as a practical alternative. This standard consists of a curated set of research questions and reference answers aligned with selected articles from *DER SPIEGEL*. We began by defining a clear thematic scope and formulating specific questions connected to previously identified source material. For each question, we manually composed plausible answers grounded in the information provided by the sources. This *silver standard* thus defines the minimum set of relevant articles, key statements and contextual details necessary to adequately address each question.It provides a structured baseline against which to measure the effectiveness of the retrieval component, while recognising the interpretative nature of historical scholarship.

Although we do not claim that these reference answers represent absolute historical truths, the framework supports a meaningful evaluation of document retrieval relevance and content quality. Building on this, we developed question categories to assess the RAG system's performance across different types of historical inquiry.

Fig. 2: Meding and Daugs: Evaluation of the RAG framework for analysing historical texts from the SPIEGEL archive (1940s–1970s).



6.2 Evaluating the RAG pipeline in practice: the portrayal of Charles de Gaulle and the Algerian war in *DER SPIEGEL*

To illustrate the evaluation process and its specific challenges, we present a practical case study focused on Charles de Gaulle's rise to power in 1958 and its portrayal in *DER SPIEGEL*, particularly in relation to the Algerian War and the representation of the French diaspora and the Muslim population living in Algeria at the time.

The example is based on the *SPIEGEL* article 'De Gaulle, the God-Sent' (N.A. 1958, 'De Gaulle, der Gottgesandte', *DER SPIEGEL* 36), featuring an interview with François Mauriac, Nobel Prize-winning author and later biographer of de Gaulle. This article served as the basis for a question-and-answer pair created by students as part of the *silver standard*.

The question, classified as a 'representation question', examines how Mauriac portrays Algerians during de Gaulle's rise to power and identifies the rhetorical strategies and linguistic devices used in the article. The minimum plausible answer notes a dehumanising portrayal of Muslim Algerians, depicted through racist and patriarchal-colonial language as a suffering mass to be cared for by the French state. By contrast, French Algerians are framed by Mauriac as the true threat in the conflict, and to French democracy as a whole. Key excerpts from the article support this interpretation, describing Muslims as 'a shifting, formless mass of undernourished and ignorant unfortunates' and emphasising the 'unyielding Muslim liberation front' confronted by the powerful French Algerians.

The retrieval evaluation focused on how well the system identified relevant source material. Using the query 'algerische Muselmanen' (archaic term for Algerian Muslims), the system extracted ten text *chunks* with the highest similarity scores. While relevant sections were retrieved, three significant shortcomings were observed. First, most retrieved *chunks* repeated the same content, with only two providing relevant distinctions (redundancy). Second, despite *recursive chunking* and *overlap*, many segments lacked complete sentences, disrupting semantic coherence (fragmentation). Third, the system sometimes included passages from unrelated articles, misattributing context (offtopic recall and misattribution). These challenges demonstrate that retrieval quality in RAG systems is still the primary bottleneck in RAG workflows. Addressing repetition, incomplete texts, misattribution and semantic incoherence is therefore essential to improving RAG's reliability for historical research. Our pilot evaluation confirms that both the *chunking process* and the similarity-based retrieval require substantial refinement, reaffirming the critical importance of our chosen approach.

Box 2 **Retrieval output of the identified text passages (chunks), own translation (see Appendix B, Box 2 for the original output).**

Relevance: 0.5066
Title: DE GAULLE, DER GOTTGESANDTE
Date: 1958–09-02
URL: https://www.spiegel.de/politik/de-gaulle-der-gottgesandte-a-f1b76d80-0002-0001-0000-000041759098?context=issue
Text: The strength of the one million French in Algeria is far, far greater than that of the eight million Muslims. SPIEGEL: But can the one million French defeat the Muslims? Is it not, above all, the support of the army that gives the Algerian French their power? What would these one million French be without the soldiers whom the forty million French in France send to Algeria? MAURIAC: The strength of the French in Algeria is immense. The Muslims are little more than a formless mass, surging back and forth.
Relevance: 0.5258
Title: DE GAULLE, DER GOTTGESANDTE
Date: 1958–09-02
URL: https://www.spiegel.de/politik/de-gaulle-der-gottgesandte-a-f1b76d80-0002-0001-0000-000041759098?context=issue
Text: MAURIAC: The danger of overexertion is immense, I am the first to concede that to you. The possible consequences of such an exacting policy go beyond anything one can imagine. As far as Algeria and the Muslims are concerned, we would, I think, quickly find ourselves in agreement on all essential points. But the problem is evidently insoluble, because the French of Algeria exist. That is the heart of the matter. Were it not for the enormous pressure exerted by these French in Algeria, the problem would quickly and relatively easily find an acceptable solution.

Nevertheless, the RAG system successfully identified relevant *chunks* that were integrated into the prompt context for model evaluation. Using these retrieved contexts, we compared responses from the HU-LLM-3 and OpenAI's GPT-4o models. Each model received the same *system prompt*, *question prompt* and *context chunks* via their respective APIs.

The HU-LLM-3 model generated a response considered satisfactory against the *silver standard*. GPT-4o's answer was also found to be adequate. However, given its strict adherence to data privacy and transparency measures, particular attention was paid to evaluating the performance of the HU-LLM-3 model.

Box 3 Generated response by the HU-LLM-3 model, own translation (see Appendix C, Box 3 for the original output).

'In the selected excerpts, Algerians – particularly the Muslim population – are portrayed by François Mauriac and *Der Spiegel* magazine in a highly stereotypical and derogatory manner. The following are some of the linguistic devices and semantic terms used to describe the Algerian people:

- **"Moslem' or 'Muselmannen"**: The use of these terms, which refer to the Islamic faith, reduces Algerians to their religious identity and suggests a sense of otherness and foreignness.
- **"The French of Algeria"**: Mauriac and *Der Spiegel* refer to the French settlers in Algeria as a distinct group, which implies a separation from the indigenous Algerian population and positions them as holding a dominant role.
- **"The power of the French in Algeria"**: Here, the strength and authority of the French in Algeria are emphasised, while the Muslim population is described as a "shifting, formless mass", conveying a negative and dehumanising representation.
- **"Pressure on France"**: Algerians are depicted as a threat to France, creating an insoluble crisis through their 'enormous pressure'.
- **"Insoluble problem"**: The very existence of the French in Algeria is portrayed as a core issue that appears insurmountable.
- **"Overexertion"**: Mauriac stresses the potential consequences of a policy that considers the interests of Algerians, suggesting that such a course of action would represent an excessive burden for France.

Taken together, these linguistic choices reveal a clear hierarchy and a colonial mindset in which Algerians are seen as inferior, while the French settlers are framed as the dominant force. The language frequently conveys a tone that is dismissive and dehumanising, obscuring the complex realities of the Algerian population and their struggles for independence and self-determination.

It is important to note that this representation is heavily shaped by the perspective of the French colonial power and often silences the voices and lived experiences of Algerians themselves.'

Although HU-LLM-3 did not clearly differentiate between the portrayals of French and Muslim Algerians, or flag the term 'Muselmannen' as archaic, it nonetheless placed relevant terms and phrases successfully within a broader colonial and historical framework. The model also flagged the potentially biased perspective of the article's French viewpoint and recommended consulting Algerian sources for a more comprehensive analysis.

Such a step may seem self-evident to a trained historian, but this simulation of source criticism within the generated output indicates a potential heuristic use of LLMs. The model generated a plausible critique by extrapolating from patterns in its training data, identifying a colonial bias and recommending Algerian perspectives and sources. Treated as a heuristic tool, the output can alert historians to potential interpretative leads and asymmetries in text corpora that warrant closer inspection and suggest complementary sources. In other words, the model can help scaffold an initial analytic pass, provided scholars subject every suggestion to human verification, recognising that its 'criticism' is probabilistic pattern-matching, not a historical method.

Yet this functionality cannot become an excuse to bypass hermeneutic close reading or to engage solely with LLM-highlighted passages while neglecting the original source. Historians remain responsible for returning to the primary text, weighing context, language and authorship against any machine-generated cue. Responsible deployment therefore demands transparent prompts, verifiable retrieval chains and an iterative workflow.

7. Conclusion and discussion

In this article, we have shown that a RAG-enhanced LLM pipeline not only meets a core requirement of historical research, managing extensive datasets like the *SPIEGEL* archive with full provenance, but also transforms how scholars interact with source material. Instead of producing decontextualised responses that are fully reliant on the internal and often opaque parametric knowledge representation of the model, every generated answer is explicitly tied to the retrieved document segments, giving historians the ability to verify and cross-check the model's claims against the original texts.

Beyond this traceability, another strength of the RAG architecture lies in its modularity and adaptability. The system is not constrained by topical boundaries; provided the input corpus is properly indexed and semantically embedded, it can support a wide array of queries within the *SPIEGEL* archive or be extended to other structured collections. This extensibility makes the pipeline highly scalable and domain-transferable, facilitating integration with historical databases or digital corpora. By enabling semantic-level retrieval, RAG supports historians in locating conceptually related documents that would remain inaccessible through manual or analogue methods. From a distant reading standpoint, this opens the door to novel perspectives and encourages scholars to formulate questions that previously might have gone unasked.

Despite these strengths, the limitations of the approach remain evident. Adding external, retrieved material to the prompt narrows the model's focus, but does not eliminate the core problem of hallucinations or inaccuracies in LLM-generated content. If irrelevant or incorrect sources are retrieved, the model may still produce fluent but historically implausible answers. In some cases, RAG may reinforce false confidence by pairing fabricated claims with credible citations. This highlights the ongoing need for analogue source evaluation.

Moreover, the retrieval quality depends heavily on the user's command of the corpus's language and conventions. Crafting effective prompts and queries requires an understanding of historical terminology and period-specific vocabulary. For example, recognising that post-war *SPIEGEL* articles employ the archaic term 'Muselmanen' rather than 'Muslime' is crucial. Without such domain expertise, even the most advanced retrieval engines will struggle to deliver relevant and precise results.

Turning to our evaluation framework, the *silver standard* we created offers a practical method to measure system performance, but faces inherent challenges. Its main limitation stems from the interpretative nature of historiography, where objectivity is never absolute and every answer reflects a particular viewpoint. The manual formulation of plausible responses introduces unavoidable subjectivity and presents one reading among many possible interpretations. This complicates efforts to establish firm benchmarks for model assessment. Moreover, the complexity of many historical questions resists simplification into neat question-and-answer formats.

In light of these considerations, it is clear that the *silver standard* functions as a pragmatic yet limited evaluation benchmark. Expanding its scope and sensitivity to ambiguity would improve its value, but some hurdles, such as historiographical subjectivity and shifting scholarly understandings, remain firmly in place. RAG pipelines enhance LLM transparency by exposing provenance chains and injecting semantically aligned passages. However, whether RAG truly leads to 'rich responses' ultimately depends on historians' expertise in domain-specific examination of both the generated content and the sources they retrieve. Looking ahead, the most promising trajectory lies in hybrid workflows that fuse neural semantics with rule-based temporal filters and corpus-level validation, ensuring that every machine-generated output is re-anchored in the full evidential record before it enters scholarly interpretation.[1]

References

Asai A, Wu Z, Wang Y, Sil A and Hajishirzi H (2024) Self-rag: Learning to retrieve, generate, and critique through self-reflection. In: *The Twelfth International Conference on Learning Representations (ICLR 2024)*, Vienna, Austria, 7 – 11 May 2024: 1–30.

Chatzikyriakidis S (2025) Reasoning with RAGged events: RAG-Enhanced Event Knowledge Base Construction and reasoning with proof-assistants. *arXiv preprint: 1–16.* https://doi.org/10.48550/arXiv.2506.07042.

Chen B, Zhang Z, Langrené N and Zhu S (2025) Unleashing the Potential of Prompt Engineering in Large Language Models: A Comprehensive Review. *Patterns* 6 (6): Article 101260. https://doi.org/10.1016/j.patter.2025.101260.

Fan W, Ding Y, Ning L, Wang S, Li H, Yin D et al. (2024) A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models. In: *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (ed Baeza-Yates R and Bonchi F), Barcelona, Spain, 25 – 29 August 2024: 6491–6501.

Hill M (2026) The data interview. Reflexive integration of large language models in qualitative content analysis. In: Simons A, Wüthrich A, Zichert M, et al. (eds) *Understanding Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science*. Bielefeld: transcript, part-5.

1 This chapter was written with support from large language models (LLMs). All model-generated text was reviewed and, where necessary, rewritten by the authors, who remain fully responsible for the final version. For details on the use of LLMs in this volume, see the statement in the volume's introduction.

Hiltmann T (2024) Hermeneutik in Zeiten der KI. Large Language Models als hermeneutische Instrumente in den Geschichtswissenschaften. In: Schreiber G and Ohly L (eds) *KI:Text. Diskurse über KI-Textgeneratoren*. Berlin / Boston: De Gruyter, 201–232.

Jordan S (2010) *Vetorecht der Quellen*. In: *Docupedia-Zeitgeschichte*, February 11. https://dx.doi.org/10.14765/zzf.dok.2.570.v1 (accessed 02 March 2025).

Koszyk K (1999) Presse unter alliierter Besatzung. In: Jürgen W (eds) *Mediengeschichte der Bundesrepublik Deutschland*. Köln / Wien: Böhlau Verlag, 31–58.

Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V and Goyal N (2020) Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In: *Advances in Neural Information Processing Systems* 33 (ed Larochelle H, Ranzato M, Hadsell R et al.), Vancouver BC, Canada, 6 – 12 December 2020: 9459–74.

Mauermann J and Oberbichler S (2025) LLM Biases: Expected and Unexpected Model Design Effects in Historical Newspaper Article Extraction on the Messina Earthquake. In: *DH Lab*, 31. January. https://doi.org/10.58079/137qr.

McGillivray B, Nanni F and Beelen K (2024). Why Does Digital History Need Diachronic Semantic Search? In: Tilton L, Mimno D and Marie J (eds) *Computational Humanities*. Minneapolis / London: University of Minnesota Press, 177–194.

Meding H and Daugs A (2026) From RAGs to rich responses. Enhancing LLM reliability through retrieval-augmented generation. In: Simons A, Wüthrich A, Zichert M, et al. (eds) *Understanding Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science*. Bielefeld: transcript, part-5.

N.A. (1958) DE GAULLE, DER GOTTGESANDTE. *DER SPIEGEL 36*, September 02. https://www.spiegel.de/politik/de-gaulle-der-gottgesandte-a-f1b76d80-0002-0001-0000-000041759098?sara_ref=re-xx-cp-sh (accessed 07 December 2024).

Nguyen-Duc ACVM, Tran BA, Ngo VP, Chi LL and Nguyen AQ (2025) An Empirical Study of Multi-Agent RAG for Real-World University Admissions Counselinh. *arXiv preprint*: 1–12. https://doi.org/10.48550/arXiv.2507.11272.

Oberbichler S and Petz C (2025) Working Paper: Implementing Generative AI in the Historical Studies. *Zenodo*, February 25. https://doi.org/10.5281/zenodo.14924737.

Porter D (2021) *Manuscript Loss in Digital Contexts*. Talk, originally presented at the 14th Annual Lawrence J. Schoenberg Symposium on Manuscript Studies in the Digital Age, November 16. https://www.dotporterdigital.org/manuscript-loss-in-digital-contexts/ (accessed 11 July 2025).

Sergeev A, Goloviznina V, Melnichenko M and Kotelnikov E (2025) Talking to Data: Designing Smart Assistants for Humanities Databases. *arXiv preprint*: 1–15. https://doi.org/10.48550/arXiv.2506.00986.

Shuster K, Poff S, Chen M, Kiela D and Weston J (2021) Retrieval augmentation reduces hallucination in conversation. In: *Findings of the Association for Computational Linguistics (EMNLP 2021)*, Punta Cana, Dominican Republic, November 2021: 3784–3803.

Simons A, Zichert M and Wüthrich A (2026) Large language models for history, philosophy, and sociology of science: Interpretive uses, methodological challenges, and critical perspectives. *Studies in History and Philosophy of Science* 117: 102151. https://doi.org/10.1016/j.shpsa.2026.102151.

Von Lucke A (2024) Seismograph Augstein. DER SPIEGEL als Produkt und Produzent der Liberalisierung – und der neue Ruck nach rechts. In: Lilienthal V (ed) *„Sagen, was* ist“: *Journalismus für eine offene Gesellschaft – Rudolf Augstein zum 100. Geburtstag*. Köln: Herbert von Halem Verlag, 115–131.

Xia P, Zhu K, Li H, Wang T, Shi W, Wang S et al. (2025) MMed-RAG: Versatile Multimodal RAG System for Medical Vision Language Models. *arXiv preprint*: 1–30. https://doi.org/10.48550/arXiv.2410.13085.

Yang X, Sun K, Xin H, Sun Y, Bhalla N, Chen X et al. (2024) CRAG – comprehensive RAG benchmark. *arXiv preprint*: 1–20. https://doi.org/10.48550/arXiv.2406.04744.

Appendix

Appendix A. Text-based Box 1: sample system prompt used in the RAG-pipeline

“Du bist ein Historiker mit tiefgehender Expertise zum Algerienkrieg und zu Charles de Gaulle. Deine Aufgabe ist es, die Artikel des westdeutschen SPIEGEL-Nachrichtenmagazins aus den Jahren 1949 bis 1979 zu analysieren und die folgende Frage so präzise und faktenbasiert wie möglich zu beantworten. Dabei stützt du dich ausschließlich auf die in den abgerufenen SPIEGEL-Artikel enthaltenen Informationen. Falls eine Information nicht in den Artikeln zu finden ist, ergänzt du sie nicht aus anderen externen Quellen oder durch Spekulation. Wenn die Artikel keine eindeutigen Informationen zu einer Frage enthalten, stellst du dies klar heraus und antwortest explizit mit: ‘Das kann ich nicht beantworten.’”

Appendix B. Text-based Box 2: generated response by the HU-LLM-3 model

“Relevanz: 0.5258
Titel: DE GAULLE, DER GOTTGESANDTE
Datum: 1958–09-02
URL: https://www.spiegel.de/politik/de-gaulle-der-gottgesandte-a-f1b76d80-0002-0001-0000-000041759098?context=issue
Text: MAURIAC: Die Gefahr der Überanstrengung ist gewaltig, ich bin der erste, Ihnen das zuzugestehen. Die möglichen Konsequenzen einer so anspruchsvollen Politik gehen über das Vorstellbare hinaus. Was Algerien und die Moslem angeht, so würden wir wohl schnell in allen wesentlichen Punkten übereinstimmen. Aber das Problem ist offenbar unlösbar, weil es die Franzosen Algeriens gibt. Das ist der Kern der Sache. Übten diese Franzosen Algeriens nicht den enormen Druck auf Frankreich aus, dann fände das Problem schnell und relativ einfach eine akzeptable Lösung.
Relevanz: 0.5066
Titel: DE GAULLE, DER GOTTGESANDTE
Datum: 1958–09-09
URL: https://www.spiegel.de/politik/de-gaulle-der-gottgesandte-a-f1b76d80-0002-0001-0000-000041759098?context=issue
Text: die Kraft der einen Million Franzosen in Algerien ist viel, viel größer als die der acht Millionen Muselmanen. SPIEGEL: Kann aber die eine Million Franzosen die Moslem besiegen? Ist es nicht vor allein die Hilfe der Armee, die den algerischen Franzosen Macht verleiht? Was wäre diese Million Franzosen ohne die Soldaten, die

die 40 Millionen Franzosen in Frankreich nach Algerien schicken. MAURIAC: Die Kraft der Franzosen Algeriens ist gewaltig. Die Moslem sind nicht viel mehr als eine hin- und herflutende, gestaltlose Masse [...].”

Appendix C. Text-based Box 3: generated response by the HU-LLM-3 model

“Antwort: In den Textauszügen werden die Algerier*innen, insbesondere die muslimische Bevölkerung, von Francois Mauriac und dem SPIEGEL-Magazin in einer sehr stereotypisierenden und abwertenden Weise dargestellt. Hier sind einige sprachliche Mittel und semantische Begriffe, die zur Beschreibung der Algerier*innen verwendet werden:

- ‘Moslem’ oder ‘Muselmanen’: Die Verwendung dieser Begriffe, die sich auf die muslimische Religion beziehen, reduziert die Algerier*innen auf ihre religiöse Identität und suggeriert eine gewisse Fremdheit und Andersartigkeit.
- ‘Die Franzosen Algeriens’: Mauriac und der SPIEGEL bezeichnen die algerischen Franzosen als eine separate Gruppe, was impliziert, dass sie sich von den einheimischen Algerier*innen unterscheiden und eine dominante Rolle einnehmen.
- ‘Kraft der Franzosen Algeriens’: Hier wird die Macht und Stärke der algerischen Franzosen betont, während die muslimische Bevölkerung als ‘hin- und herflutende, gestaltlose Masse’ beschrieben wird, was eine negative und entmenschlichende Darstellung darstellt.
- ‘Druck auf Frankreich’: Die Algerier*innen werden als eine Bedrohung für Frankreich dargestellt, die durch ihren “enormen Druck” eine unlösbare Situation schaffen.
- ‘Unlösbares Problem’: Die Existenz der algerischen Franzosen wird als Kernproblem dargestellt, das unüberwindbar erscheint.
- ‘Überanstrengung’: Mauriac betont die möglichen Konsequenzen einer Politik, die die Interessen der Algerier*innen berücksichtigt, und suggeriert, dass dies eine übermäßige Belastung für Frankreich darstellen würde.

Insgesamt zeigen diese sprachlichen Mittel eine klare Hierarchie und eine koloniale Denkweise, die die Algerier*innen als unterlegen und die algerischen Franzosen als die dominierende Kraft darstellt. Die Sprache ist oft abwertend und entmenschlichend, was die komplexe Realität der algerischen Bevölkerung und ihre Kämpfe für Unabhängigkeit und Selbstbestimmung ignoriert.

Es ist wichtig zu beachten, dass diese Darstellung stark von der Perspektive der französischen Kolonialmacht geprägt ist und die Stimmen und Erfahrungen der Algerier*innen selbst oft ausblendet.”