Cambridge Large Two (2020), 1–21

RESEARCH ARTICLE

Topic-RAG for Historical Newspapers: Enhancing Information Retrieval in Humanities Research through Topic-Based Retrieval-Augmented Generation

Keerthana Murugaraj,[1] Salima Lamsiyah,[1] Marten During,[2] and Martin Theobald[1]

[1]University of Luxembourg, Department of Computer Science (DCS), Faculty of Science, Technology and Medicine (FSTM), Esch-sur-Alzette, 4364, Luxembourg
[2]University of Luxembourg, Centre for Contemporary & Digital History (C^2DH),Esch-sur-Alzette, 4364, Luxembourg
Author for correspondence: Keerthana Murugaraj, Email: keerthana.murugaraj@uni.lu.

Abstract

The exploration and retrieval of information from large, unstructured document collections remain challenging. Unsupervised techniques such as clustering and topic modeling provide only a coarse overview of thematic structure, while traditional keyword searches often require extensive manual effort. Recent advances in Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG) introduce new opportunities by enabling focused retrieval of relevant documents or chunks tailored to a user's query. This allows for dynamic, chat-like interactions that streamline exploration and improve access to pertinent information. This paper introduces *Topic-RAG*, a chat engine that integrates topic modeling with RAG to support interactive and exploratory document retrieval. Topic-RAG uses BERTopic to identify the most relevant topics for a given query and restricts retrieval to documents or chunks within those topics. This targeted strategy enhances retrieval relevance by narrowing the search space to thematically aligned content. We utilize the pipeline on 4,711 articles related to nuclear energy from the Impresso historical Swiss newspaper corpus. Our experimental results demonstrate that Topic-RAG outperforms a baseline RAG architecture that does not incorporate topic modeling, as measured by widely recognized metrics such as BERTScore (including Precision, Recall, and F1), ROUGE, and UniEval. Topic-RAG also achieves improvements in computational efficiency for both single and batch query processing. In addition, we performed a qualitative analysis in collaboration with domain experts, who assessed the system's effectiveness in supporting historically grounded research. Although our evaluation is focused on historical newspaper articles, the proposed approach more generally integrates topic information to enhance retrieval performance within a transparent and user-configurable pipeline effectively. It supports the targeted retrieval of contextually rich and semantically relevant content while also allowing users to adjust key parameters such as the number of documents retrieved. This flexibility provides greater control and adaptability to meet diverse research needs in historical inquiry, literary analysis, and cultural studies. Due to copyright restrictions, the raw data cannot be publicly shared. Data access instructions are provided in the repository, and the replication code is available on GitHub: link.

Keywords: Topic Modeling, Large Language Models, Retrieval-Augmented Generation, Information Retrieval, Semantic Chunking.

Plain Language Summary

In the digital age, researchers often face the challenge of searching through vast collections of documents to find relevant information, a process that can be time-consuming. This is particularly difficult when dealing with large volumes of unstructured texts, such as historical newspapers and books. Traditional keyword searches often fall short of quickly providing the most pertinent results. However, recent advancements in Natural Language Processing (NLP) have created new opportunities for more efficient information retrieval.

This research aims to address these challenges by introducing *Topic-RAG*, a novel system designed to help historians and researchers quickly locate relevant information, surpassing the limitations of traditional search methods. Topic-RAG employs topic modeling to analyze large document collections and group them by underlying themes. When a user submits a query, Topic-RAG retrieves relevant documents and generates detailed answers based on the most pertinent content.

The key innovation of Topic-RAG lies in its use of advanced methods for understanding the themes within documents. It focuses only on the most relevant sections to provide quicker and more accurate responses. This system is particularly beneficial for those working with complex and extensive historical or literary collections, helping them find the right information to support their research. With Topic-RAG, researchers can save time and obtain precise, contextually relevant answers, making it a valuable tool for scholars across various fields, particularly in the humanities.

1. Introduction

The increasing availability of digitized and born-digital historical documents presents both significant challenges and opportunities for researchers. While these extensive collections offer immense potential to move beyond the limitations of traditional close-reading-based research practices, their sheer volume necessitates advanced tools for efficient retrieval and comprehension (Sá and Maia 2021). Traditional document retrieval systems often struggle to extract contextually relevant information from such collections, especially in specialized fields like historical research. To overcome these limitations,

https://doi.org/10.1017/chr.2025.10018 Published online by Cambridge University Press

researchers are increasingly exploring hybrid models that integrate state-of-the-art NLP techniques. Among these are Retrieval-Augmented Generation (RAG) frameworks (Lewis et al. 2020), which combine document retrieval and text generation into a unified workflow.

RAG (Lewis et al. 2020) combines retrieval methods with Large Language Models (LLMs) to enhance their generation capabilities by incorporating relevant information from a specific corpus as context. This allows RAG to produce more contextually accurate answers and serves as a cost-effective alternative to fine-tuning LLMs on new corpora. However, existing RAG models primarily rely on vector similarity matching, which limits their ability to capture latent semantic relationships between queries and documents. To address this limitation, we propose *Topic-RAG*, a framework that integrates topic modeling into the RAG pipeline to rerank retrieved results semantically. By leveraging latent topics in both queries and documents, Topic-RAG enhances the precision and semantic relevance of the retrieval process.

Several topic modeling methods exist, including traditional models such as Latent Semantic Analysis (LSA) (Deerwester et al. 1990), Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003), and Non-Negative Matrix Factorization (NMF) (Lee and Seung 1999). These methods have demonstrated effectiveness in various NLP tasks, such as text categorization, summarization, and sentiment analysis (Blei 2012; Wu et al. 2017), making them valuable techniques for processing textual data. However, they are sometimes inadequate at capturing nuanced relationships within texts, particularly when dealing with complex, unstructured datasets like historical archives or large-scale literary collections. This limitation becomes particularly apparent when thematic coherence and topic organization are essential for understanding the content (AlSumait, Barbará, and Domeniconi 2008). Traditional techniques like LDA, while widely adopted, focus on word co-occurrence patterns and assume a fixed number of topics, which can lead to challenges in modeling the rich semantic structures found in modern corpora (Chang et al. 2009). Moreover, these methods cannot fully utilize the semantic depth offered by modern sentence embeddings (Reimers and Gurevych 2019), which excel at encoding contextual relationships across documents. The need for systems that go beyond surface-level similarity is clear. Effective document retrieval should leverage deeper thematic understanding, enabling the selection of contextually relevant and semantically rich documents. Addressing these gaps could lead to significant advancements in information retrieval, particularly for large, complex datasets.

To bridge this gap, we propose a novel RAG framework that integrates BERTopic (Grootendorst 2022), a state-of-the-art topic modeling technique that generates topic representations based on contextual embeddings. BERTopic has been shown to outperform traditional methods like LDA and LSA in capturing nuanced thematic structures. Our approach enhances the retrieval of contextually relevant documents by leveraging advanced sentence embeddings for a more refined and precise retrieval process. By integrating modern embedding techniques with topic modeling, our method dynamically identifies topics and retrieves documents that are closely aligned with the user's query.

Our contributions are summarized as follows.

- We introduce the *Topic-RAG* framework, which combines the strengths of RAG with advanced topic modeling techniques, specifically BERTopic. This framework integrates contextual topic modeling using Jina AI's BERT-based embedding model[a] to generate document embeddings and employs a class-based ranking technique (C-TF-IDF) (Grootendorst 2022) to produce class-specific topic embeddings. This approach enables more precise document retrieval by leveraging both semantic content and thematic relevance.
- We propose *Topic-RAG+*, designed to handle long documents more effectively. Our implementation incorporates a chunking strategy within the indexing system, which first groups documents based on their topics and then stores them after performing semantic chunking.
- We present an interactive, user-driven retrieval system that allows flexible adjustment of the number of documents retrieved for a given query based on Cosine similarity, providing an alternative to the traditional top-k strategy. For response generation, we leverage a quantized version of the Llama 3.1 8B Instruct[b] model, which balances performance and computational efficiency while delivering contextually rich outputs.
- We introduce a tailored prompt for a systematic evaluation of our approaches, asking the underlying Llama 3.1 8B model to generate diverse questions and answers based on a set of input documents. We design the prompt to instruct the model to create questions of varying types (fact-based, inference-based, analytical, causal, and opinion-based), covering key aspects such as main events, actions taken, consequences, technical details, and impact. We manually curate the generated questions, and the final results produced by Topic-RAG are evaluated using a variety of metrics, including BERTScore, ROUGE, and UniEval.
- We conducted a qualitative evaluation of our Topic-RAG+ method using nine challenging questions that require in-depth analysis from a historical perspective, showcasing its potential and identifying key areas for future improvement. In this evaluation, we leveraged the method's capability to retrieve a broader range of relevant documents and generate comprehensive responses to complex queries.

The proposed framework offers significant value for Humanities researchers, particularly historians working with large collections of unstructured text. By integrating topic modeling, document retrieval, and advanced text generation, *Topic-RAG* enables users to retrieve documents that are both content-relevant and thematically aligned with their research. It allows natural language queries, overcoming the limitations of

[a] https://huggingface.co/jinaai/jina-embeddings-v2-base-en
[b] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

https://doi.org/10.1017/chr.2025.10018 Published online by Cambridge University Press

keyword-based searches, while the flexibility of user-driven retrieval ensures access to the most pertinent documents. Additionally, the framework supports local processing, addressing data privacy and legal restrictions by keeping sensitive collections securely stored on local hardware.

This work is part of a collaboration between the Digital History and Computer Science departments at the University of Luxembourg, aimed at fostering interdisciplinary research between historians and computer scientists. Our experiments utilize a subset of 4,711 digitized Swiss newspaper articles from 1971 to 1986, which contain keywords linked to themes surrounding "nuclear power" and "nuclear safety". These were selected from the broader Impresso corpus [c] of historical newspapers (Ehrmann et al. 2020; Düring, Bunout, and Guido 2024). Importantly, the selection was applied to the original French and German texts, rather than to any machine-translated versions.

The rest of the paper is organized as follows. Section 2 reviews the background and related work. Section 3 describes the proposed system. Section 4 presents experimental results, including dataset details, ground-truth Q&A preparation, and quantitative and qualitative evaluations of Topic-RAG against baseline RAG. Section 5 highlights the system's advantages, Section 6 concludes the paper, and Section 7 discusses future research directions.

2. Background & Related Work

In this section, we provide a brief introduction to LLMs and RAG, followed by a discussion of keyword search limitations and NLP applications in historical research.

2.1 Large Language Models (LLMs)

LLMs are a subclass of deep neural networks based on the Transformer architecture (Vaswani et al. 2017), which were trained on large-scale corpora of raw text, often consisting of hundreds of billions to trillions of tokens. These models process input texts as sequences of subword tokens, which are embedded as vectors and passed through multiple layers of multi-head self-attention to capture contextual dependencies across the entire sequence. LLMs are typically developed as *foundation models*, general-purpose pre-trained models that can be adapted to a wide range of downstream tasks. These models are autoregressive, meaning they are trained to predict the next token in a sequence, including their own previously generated tokens.

Recent years have witnessed an unprecedented expansion in the scale and capability of LLMs. Notable models include GPT-3 and GPT-4 (Brown et al. 2020; Achiam et al. 2023), PaLM (Chowdhery et al. 2023), Llama (Touvron et al. 2023), Mistral (Jiang et al. 2023), Claude (Anthropic 2024), and DeepSeek (DeepSeek-AI et al. 2025). These models have demonstrated strong zero-shot and few-shot performance across a wide range of NLP tasks, often surpassing task-specific systems. Instruction-tuned variants, such as InstructGPT (Ouyang et al. 2022), Llama 3 Instruct, and others, extend the capabilities of foundation models by fine-tuning them on curated datasets of prompts and responses. These models are optimized to follow human instructions more effectively across tasks like summarization, reasoning, and dialogues. A key component of this fine-tuning process is known as *Reinforcement Learning from Human Feedback* (RLHF), in which human preferences are used to rank model outputs. A reward model is then trained on this data, guiding the base LLM to produce more aligned and helpful responses through reinforcement learning. This shift has propelled LLMs from general language modeling toward interactive systems capable of following complex instructions, reasoning, and multi-turn dialogues.

Recent research has leveraged LLMs for solving complex domain-specific problems: for example, "BioMedLM" (Bolton et al. 2024) for biomedical question answering, "MathPrompter" (Imani, Du, and Shrivastava 2023) for improving mathematical reasoning, novel methods for historical text summarization, "HistBERTSum-Ext" (Lamsiyah, Murugaraj, and Schommer 2023) and "HistBERTSum-Abs" (Murugaraj, Lamsiyah, and Schommer 2025) were proposed for extractive and abstractive approaches, respectively, and "CodeT5+" (Wang et al. 2023) for advanced code generation and understanding. In education, models like "EduChat" (Dan et al. 2023) enable personalized and compassionate intelligent education, serving teachers, students, and parents, while in legal and policy analysis, LLMs have been applied to extract and reason over complex legislative documents (Yao et al. 2024). HiST-LLM (Hauser et al. 2024) is a benchmark for evaluating large language models on structured historical knowledge. It uses the Seshat Global History Databank[d], which spans thousands of historical data points across various periods and world regions. The Qur'an QA 2023 Shared Task (Malhas, Mansour, and Elsayed 2023) evaluated systems, including LLMs, for their ability to extract semantically accurate answers from religious and historical texts in Arabic, using the Holy Qur'an as a benchmark corpus.

Moreover, open-source ecosystems like Hugging Face's Transformers[e] API and inference-acceleration libraries such as Unsloth[f], Groq[g], and Replicate[h] have democratized access to powerful LLMs, fostering experimentation in low-resource and domain-specific applications. The evolution of LLMs continues to reshape fields like education, healthcare, and digital humanities through models specialized in question answering, summarization, and RAG.

2.2 Retrieval-Augmented Generation (RAG)

Building and fine-tuning LLMs "from scratch" is practically infeasible for most research groups around the world due to the enormous amounts of training data and computational

[c] https://impresso-project.ch

[d] https://seshat-db.com/

[e] https://huggingface.co/transformers/

[f] https://unsloth.ai/

[g] https://groq.com/

[h] https://replicate.com/

https://doi.org/10.1017/chr.2025.10018 Published online by Cambridge University Press

resources they require. RAG (Lewis et al. 2020) therefore presents a fascinating fusion of classic retrieval techniques with LLM-based text generation. In essence, it enables custom text data to be integrated into an LLM by converting an input query (typically a question formulated in natural language) into an appropriate vector representation known as an "embedding" (Mikolov et al. 2013; Devlin et al. 2019). This query vector is then compared to a pre-indexed set of document vectors from a domain-specific corpus or application. The most similar document vectors are then subsequently fed into the LLM's context to generate answers that are tailored to both the background knowledge of the LLM and the retrieved documents. Notable contributions in this area include Yahoo's NGT[i] and Facebook's FAISS[j] (Douze et al. 2024) index structures, which facilitate an efficient form of Approximate Nearest-Neighbor (ANN) search among the query and document vectors. A general RAG workflow is depicted in Figure 1.

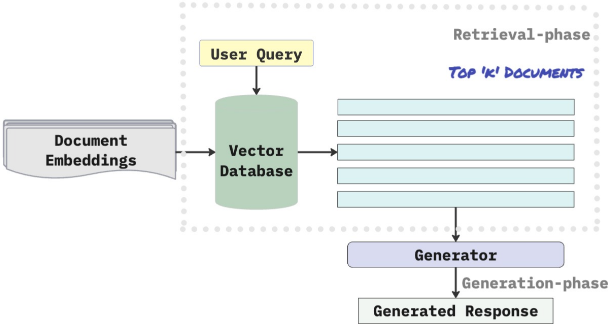


Figure 1. Baseline RAG workflow.

2.3 *Keyword Search Challenges & NLP in Historical Research*

Traditional keyword searches in humanities contexts often require scholars to engage in multiple rounds of query refinement, significantly increasing the time and cognitive effort spent on finding the relevant information. Oberbichler and Pfanzelter 2021 explicitly discusses the problems posed by keyword-based corpus building in humanities research. They note that simple keyword queries often fail to capture nuanced or ambiguous terms, especially in historical texts, leading to both excessive noise and unintentional omissions. This serves as their motivation for introducing text-mining methods to construct a more representative, topic-specific newspaper corpus. Kroll, Mainzer, and Balke 2022 demonstrates that when researchers attempt to retrieve narrative-level information, such as tracking a historical event across different sources, they must decompose their inquiry into a series of discrete keyword searches and then manually reassemble the results, a process that both fragments contexts and introduces a semantic gap. Hoeber et al. 2024 observed that humanities researchers typically reformulate their keyword queries multiple times to filter out false positives and recover relevant variants, underscoring how keyword-based search interfaces fail to support complex exploratory tasks.

Given these challenges, the extensive digitization of historical records has transformed how historians analyze large textual datasets by enabling the use of NLP techniques. NLP facilitates the automated processing and analysis of vast document collections, enabling researchers to address tasks that would be impractical to perform manually. Foundational techniques in NLP, particularly relevant to the analysis of historical texts, include word count analysis or dictionary-based named-entity recognition (NER). These methods provide essential tools for quantifying language use, identifying key actors or concepts, and uncovering thematic trends across large document corpora. For example, word-count analysis has proven instrumental in tracking the rise and fall of specific terms across historical periods.

This approach has significantly contributed to numerous historical studies, including the groundbreaking quantitative analysis of Culturomics (Michel et al. 2011), which utilizes millions of digitized books to uncover patterns in cultural and linguistic trends. As a result, the Google N-Gram Viewer[k] was developed by Google in collaboration with various authors (Michel et al. 2011). This multilingual online tool enables users to explore linguistic and cultural trends by analyzing changes in the frequency of words and phrases over time. The tool utilizes *n*-grams extracted from a vast corpus of digitized books spanning the period from 1500 to 2022.

Greenfield (Greenfield 2013) utilized Google's N-Gram Viewer to explore theories on how cultural and physiological attributes have changed over two centuries between 1800 and 2000. The author found a significant increase in the usage of words associated with individualism (e.g., "choose", "unique") and self-focus (e.g., "self", "get"), indicating a cultural shift toward individualistic values and personal autonomy. Conversely, words tied to community, social responsibility, and traditional values (e.g., "obliged", "give") declined, reflecting a move away from collectivist norms. The authors attribute these changes to societal transformations, including industrialization, urbanization, and advancements in technology and education, which emphasize independence and self-expression.

Another study (Lansdall-Welfare et al. 2017) analyzed a large corpus of historical British newspapers using NLP techniques, including *n*-gram frequency analysis, NER, named-entity disambiguation (NED), and temporal trend analysis to uncover broad trends in history and culture. These trends included patterns in gender bias, geographical focus, technology, and politics, as well as precise identification of dates for key events. The authors presented that simple content analysis enables the detection of specific events, such as wars, epidemics, coronations, and conclaves, with high accuracy. Advanced NLP techniques are neural-network–based methods that go beyond foundational approaches like counting word frequencies or rule-based extractions. Neural-based NER and neural topic modeling were employed to further enhance these works by moving beyond word counting to identify mentions

[i] https://github.com/yahoojapan/NGT
[j] https://github.com/facebookresearch/faiss
[k] https://books.google.com/ngrams/

https://doi.org/10.1017/chr.2025.10018 Published online by Cambridge University Press

of specific names, places, and other entities, thereby providing deeper insights into the corpus. Several studies (Marjanen et al. 2021; Indukaev 2020) have focused on applying classical topic-modeling techniques, particularly LDA (Blei, Ng, and Jordan 2003) and NMF (Lee and Seung 1999), to analyze discourse dynamics and conduct thematic analysis. More recently, the authors (Hills and Miani 2023) have investigated unsupervised techniques for thematic analysis and topic modeling. They highlight three main approaches in Historical NLP: counting words or documents, analyzing the semantic meaning of words (e.g., sentiment analysis), and organizing data through methods like topic modeling, with a particular emphasis on LDA. (Zundert et al. 2022). Thus, to summarize, recent studies in the historical domain continue to strongly rely on LDA and NMF (Oiva 2020; Marjanen et al. 2021; Maltseva et al. 2021; Bodrunova 2021; Uban, Caragea, and Dinu 2021; Grant et al. 2021; Gryaznova and Kirina 2021; Lin and Peng 2022; Karamouzi, Pontiki, and Krasonikolakis 2024; Chappelle, Auelua-Toomey, and Roberts 2024).

Despite their popularity, classical models often fail to capture nuanced semantics in historical texts, and neural topic modeling methods have gained popularity for capturing complex text relationships using deep learning. Neural topic modeling methods—such as Top2Vec (Angelov 2020) and BERTopic (Grootendorst 2022), which uses contextual embeddings and clustering to yield more coherent topics. However, only a handful of studies have explored neural approaches in historical topic modeling. (Arseniev-Koehler et al. 2020; Zundert et al. 2022; Cvejoski, Sánchez, and Ojeda 2023; Ginn and Hulden 2024). Moreover, BERTopic has been shown to outperform classical models in various domains (Egger and Yu 2022a; Orr, Van Kessel, and Parry 2024; Rajwal et al. 2024; Murugaraj et al. 2025) yet their adoption in historical research remains scarce. We believe that this gap presents an opportunity for further exploration, motivating us to employ BERTopic, a cutting-edge topic-modeling method, and integrating it into an RAG framework to further enhance the accuracy and contextual relevance of its response generation. To the best of our knowledge, this is the first work that systematically investigates a combination of topic modeling and RAG to enhance historical newspaper analysis.

3. Topic-RAG Methodology

We propose *Topic-RAG*, a novel RAG framework that integrates topic modeling to improve the retrieval of contextually relevant documents in response to user queries. The inclusion of topic modeling enables a deeper understanding of the underlying themes within a query, ensuring more precise and relevant document retrieval. As illustrated in Figure 2, the proposed framework comprises five main steps: 1) data preparation, 2) topic modeling, 3) vector indexing, 4) retrieval of topics & documents, and 5) response generation. These steps are detailed below.

3.1 Data Preparation

The data preparation step integrates data pre-processing and document embedding, ensuring that the documents are cleaned, transformed, and represented in vector form for subsequent analysis. The input documents—in our use case, a collection of historical newspaper articles—are pre-processed and vectorized to apply a topic modeling algorithm, which generates topic representations. The first step involves *Data Pre-Processing*, including minimal normalization to remove unwanted patterns from the documents. To achieve this, we utilized Python's Regex library to remove noisy patterns and irrelevant elements commonly found in unstructured text.

Special characters (e.g. #, $, and @) were removed, excessive whitespace was normalized, irrelevant alphanumeric codes were discarded, and all text was lower-cased for consistency. By limiting this pre-processing to essential steps, we retain core information while ensuring the documents remain well-structured for embedding and retrieval tasks. We deliberately avoided aggressive pre-processing techniques, such as stopword removal or lemmatization, to preserve semantically relevant content and maintain the original textual structure. These pre-processing steps were specifically tailored to the characteristics of our historical newspaper dataset; other datasets may require different or additional cleaning procedures depending on their structure, quality, and content. Our method is designed to be broadly applicable across various textual data types, not limited to historical sources. The core requirements for data preparation in our system are as follows. First, the input texts should be grammatically coherent and easily interpretable by modern language models to ensure language clarity. Second, the document collection must offer comprehensive topical coverage of the subject areas relevant to the expected queries. Finally, each document must have a unique identifier, with optional metadata such as title, date, or category, to support traceability. These minimal requirements enable robust retrieval and generation without the need for domain-specific pre-processing.

We next employ Jina AI's BERT-based embedding model to create *Document Embeddings* for the entire collection. This model leverages the capabilities of the "Bidirectional Encoder Representations from Transformers" (BERT) framework (Devlin et al. 2019) to create dense, high-dimensional vector representations that capture the semantic content and contextual nuances of each document. These embeddings frequently serve as a foundation for further downstream tasks, such as clustering, indexing, and efficient retrieval, by enabling a machine-readable, context-aware representation of the document corpus.

3.2 Topic Modeling

Using document embeddings generated with the Jina AI model, we leverage BERTopic (Grootendorst 2022) to extract latent topics within the collection. This state-of-the-art topic modeling technique has been shown to outperform traditional models like LDA (Blei, Ng, and Jordan 2003) and NMF (Lee and Seung 1999) across several benchmarks (Eg-

https://doi.org/10.1017/chr.2025.10018 Published online by Cambridge University Press

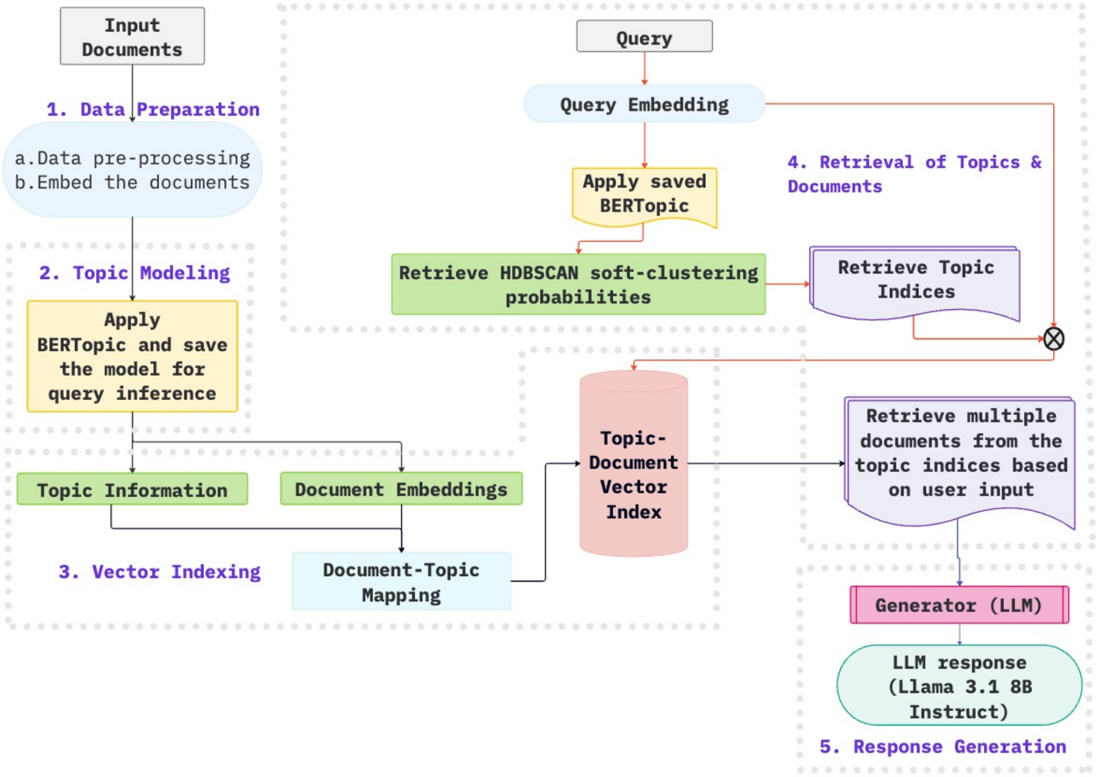


Figure 2. Overview of the proposed Topic-RAG framework.

ger and Yu 2022b; Mendonça and Figueira 2024; Cheddak et al. 2024). BERTopic introduces an innovative approach to topic extraction by employing a so-called "Class-based Term Frequency-Inverse Document Frequency" (C-TF-IDF) scoring technique[l]. Moreover, it also utilizes advanced sentence-based embedding models to represent documents as high-dimensional dense vectors, followed by dimensionality reduction using UMAP (McInnes, Healy, and Melville 2020). Clusters are then identified using HDBSCAN (McInnes, Healy, and Astels 2017), which dynamically determines the optimal number of topics, enabling precise and adaptive topic modeling.

BERTopic's combination of contextual embeddings, dimensionality reduction, and its advanced topic representation method, C-TF-IDF, makes it highly effective in handling complex datasets compared to traditional approaches. C-TF-IDF builds on the traditional (TF-IDF) weighting scheme by incorporating cluster information, enabling more precise topic representations. It is implemented using the `TfidfTransformer`[m] library from Scikit-learn. Specifically, the C-TF-IDF method (Grootendorst 2022) aggregates all documents from a specific class c into a single virtual document. The score for a term t in class c is then calculated as the product of the normalized class-based term frequency $tf_{t,c}$ and the inverse term frequency across all classes tf_t, as follows.

$$C\text{-}TF\text{-}IDF_{t,c} = tf_{t,c} \cdot \log\left(\frac{A}{1 + cf_t}\right) \tag{1}$$

where:

$$tf_{t,c} = \frac{\text{Frequency of term } t \text{ in class } c}{\text{Total number of terms in class } c};$$
$$A = \text{Average number of words per class;}$$
$$cf_t = \text{Frequency of term } t \text{ across all classes.}$$

This process results in a list of topic representations, with each document being assigned to the closest topic based on its embedding. These topic representations consist of the top-k keywords that best capture the theme of each topic. We then save the BERTopic model using the built-in methods provided in the same library. The saved model includes the document embeddings, topics, and all other associated parameters, making it easy for future use.

3.3 Vector Indexing

We next create a vector index to store document vectors using the FAISS library (Douze et al. 2024), which is designed for efficient vector indexing and similarity search. Instead of the flat embedding storage structure commonly used in baseline RAG systems, we propose a more organized indexing framework called the *"Topic-Document Index."* This framework leverages topic information from BERTopic to group document embeddings by their corresponding topics before storing them in a FAISS index. By utilizing BERTopic's topic-document

[l] https://www.maartengrootendorst.com/blog/ctfidf/
[m] https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfTransformer.html

https://doi.org/10.1017/chr.2025.10018 Published online by Cambridge University Press

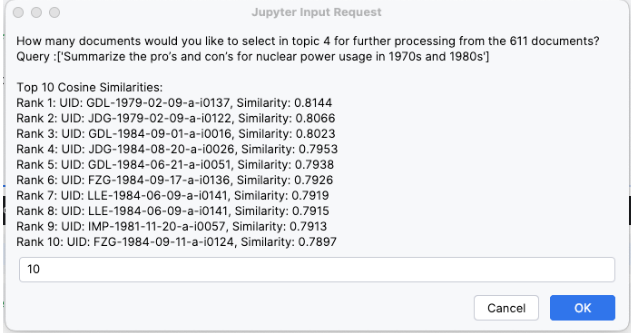
mapping, this structured approach reduces redundancies and improves retrieval efficiency through topic-aware searches. Clustering documents under their respective topics facilitates rapid and precise retrieval of the most relevant embeddings for a given query, significantly enhancing the effectiveness of the retrieval process.

The document index functions similarly to a dictionary, where topic IDs serve as keys, while their associated document embeddings are grouped as values. This structure provides flexibility to refine searches based on the context of the user's query, thereby improving the precision and relevance of the results. By organizing documents according to their topics, the *Topic-Document Index* ensures that the retrieval process considers both the content of the documents and their thematic relationships. This approach delivers more contextually relevant responses to user queries, thereby enhancing the quality of the generated answers.

3.4 *Retrieval of Topics & Documents*

The Topic-RAG workflow is initiated when a user submits a query, as depicted in Figure 2 under the *"Retrieval of Topics & Documents"* We use the same embedding model employed for encoding the input documents and utilize the saved BERTopic model, as outlined in Section 3.1 and Section3.2, to generate query embeddings and topic representations for the user query.

While BERTopic typically assigns a single topic to a query, we enhance query interpretation by leveraging the soft-clustering probabilities computed during HDBSCAN clustering, as provided by BERTopic. These probabilities indicate the degree of association between the query and each topic, enabling the identification of multiple relevant topics while filtering out weaker associations. This approach prioritizes topics that significantly contribute to the query's semantic meaning, ensuring that irrelevant or marginally related topics are excluded. To refine our topic selection, we first identify the dominant topic that is assigned to the query by using BERTopic. We then use the highest topic probability as a reference point to set a threshold, usually 50% of the dominant topic's probability, to include additional relevant topics. In other words, we treat the top-assigned topic as an anchor, and include other topics whose probabilities are at least 50% of the dominant topic's score. This strategy helps us to capture adjacent or semantically similar topics that may also be relevant to the query, without allowing irrelevant or noisy topics to enter the context. The 50% threshold represents a conservative yet inclusive setting that provided a stable balance between capturing secondary relevance and avoiding topic drift. Based on practical observations during development, we found that higher thresholds often excluded meaningful secondary topics, while lower thresholds risked introducing unrelated content. Once the relevant topics are selected, we use their indices to retrieve the corresponding documents from the *Topic-Document* index.



How many documents would you like to select in topic 4 for further processing from the 611 documents?
Query :['Summarize the pro's and con's for nuclear power usage in 1970s and 1980s']

Top 10 Cosine Similarities:
Rank 1: UID: GDL-1979-02-09-a-i0137, Similarity: 0.8144
Rank 2: UID: JDG-1979-02-09-a-i0122, Similarity: 0.8066
Rank 3: UID: GDL-1984-09-01-a-i0016, Similarity: 0.8023
Rank 4: UID: JDG-1984-08-20-a-i0026, Similarity: 0.7953
Rank 5: UID: GDL-1984-06-21-a-i0051, Similarity: 0.7938
Rank 6: UID: FZG-1984-09-17-a-i0136, Similarity: 0.7926
Rank 7: UID: LLE-1984-06-09-a-i0141, Similarity: 0.7919
Rank 8: UID: LLE-1984-06-09-a-i0141, Similarity: 0.7915
Rank 9: UID: IMP-1981-11-20-a-i0057, Similarity: 0.7913
Rank 10: UID: FZG-1984-09-11-a-i0124, Similarity: 0.7897

10

Figure 3. Example of a user's input prompt, allowing the user to adjust the number of documents retrieved for the given query.

Unlike traditional methods that retrieve a fixed number of documents (e.g., the top-*k*), our framework provides flexibility by allowing users to specify the number of documents they wish to retrieve for a given query. This is achieved by further ranking the sets of documents within the filtered topics based on the Cosine similarities of the documents' embedding vectors with the embedding vector of the query, as shown by the user prompt in Figure 3. This user-driven approach enables tailored document retrieval, allowing adjustments based on the specific needs and preferences of the query, rather than relying on a fixed top-*k* retrieval strategy. This flexibility is particularly advantageous when sending retrieved documents to the LLM-based generator model, ensuring a better understanding of and response to the query.

Additionally, we ensure traceability in the Topic-RAG framework by maintaining a link between the retrieved documents and their original sources. All relevant metadata, including user input, document IDs, and topic IDs, are preserved throughout the RAG process. This traceability allows users to cross-check sources and verify information, a feature that is often absent in default RAG systems.

3.5 *Response Generation*

In the final step, the retrieved documents, along with the original query and metadata, are sent to the LLM to generate a concise and coherent response, depicted in Figure 2 as the *"Response Generation"*. For this purpose, we utilize the instruction-finetuned version of Llama 3.1 8B, which is known for its strong benchmark performance across various natural language processing tasks. This model excels in metrics such as Accuracy, Exact Match, and F1 score, making it a robust choice for generating contextually rich and informative responses.

The selection of the 8B variant is a strategic decision, offering an optimal balance between performance and computational efficiency. Its smaller size enables deployment in diverse computing environments without requiring high-end hardware, ensuring accessibility. Additionally, its streamlined architecture facilitates rapid inference, providing a seamless user experience while maintaining high-quality outputs. While this balance makes it a versatile and efficient option, it is worth noting that larger models typically exhibit higher accuracy and provide more nuanced responses in complex scenarios, as evidenced by their superior benchmark scores [n].

[n] https://huggingface.co/meta-llama/Llama-3.1-8B

https://doi.org/10.1017/chr.2025.10018 Published online by Cambridge University Press

4. Experimental Results

In this section, we describe our experimental setup, the dataset used, and the data preparation steps, including the creation of both human-generated and synthetic ground-truth data. We also outline the evaluation metrics employed to assess and compare our approach with the baseline RAG method. Finally, we present the results, emphasizing the performance of our proposed method compared to the baseline.

4.1 Experimental Setup

All experiments in this study were conducted on a personal laptop, specifically a MacBook Pro Max M3 with 36 GB of RAM, highlighting that further high-performance computing (HPC) resources are not necessary for performing our neural-network tasks. For topic modeling, we utilized BERTopic version 0.16.3 to extract and organize topics from document collections. The proposed method was implemented in Python, leveraging its extensive library ecosystem for NLP. For vector indexing, we employed FAISS (using the CPU version), a highly efficient library for vector similarity search that enabled the fast retrieval of relevant documents. Document embeddings were generated using the Hugging Face Jina AI base model. For the generation task within the Topic-RAG system, we used a quantized version of the Llama 3.1 8B Instruct model, which is optimized for generating contextually relevant responses based on the retrieved documents. This setup seamlessly integrated topic modeling, efficient document retrieval, and high-quality response generation, forming the backbone of the proposed approach.

4.2 Dataset

The dataset used in this research comprises 4,711 digitized Swiss newspaper articles and adverts published between 1971 and 1986 on topics surrounding “nuclear power” and “nuclear safety”. The dataset is a subset of a keyword-based collection of ca. 160,000 articles written in French and German, which were extracted from the Impresso[o] corpus of historical media collections (Ehrmann et al. 2020; Düring, Bunout, and Guido 2024). We relied on French and German keywords to detect as many relevant articles with references to nuclear power as possible. An extensive OR-query was compiled of manually selected keywords and keyword suggestions generated via French and German corpus-based word embeddings (as described in (Düring, Bunout, and Guido 2024)). This approach yielded a list of 42 keywords, which included synonyms and spelling variations such as those caused by poor OCR (e.g., "nucliör", "atornique"). We acknowledge that for a keyword-based approach, flaws in OCR and article layout recognition can lead to both false positive and false negative results.

Each of the articles and adverts in the dataset is assigned a unique identifier (UID), facilitating efficient tracking and retrieval. These UIDs serve as key references throughout the research, allowing for easy cross-referencing between different analyses and findings. The articles were originally published in French and German and translated into English using the Google Translate[p] library in Python. This library provides unrestricted access to the Google Translate API, enabling seamless bulk language translation and detection. To ensure the quality of the translations, we employed an LLM (gpt-3.5-turbo [q]) to automatically assess semantic alignment between the original texts and their translated counterparts. The model was prompted to generate structured evaluation reports covering dimensions such as factual consistency, fluency, terminology, and tone. To validate these automated assessments, we manually reviewed approximately 100 randomly selected translation reports across source languages. These manual checks confirmed that the translations were generally correct, fluent, and semantically faithful to the source texts. We did not observe any critical translation errors or omissions that would justify exclusion.

Moreover, the articles contained various formatting issues such as unwanted patterns, characters, tags, and other inconsistencies that could hinder analysis. We therefore initiated the data pre-processing phase discussed in Section 3.1, which involved identifying and removing these extraneous elements. This included eliminating irrelevant patterns, cleaning up unnecessary characters, and removing unwanted tags, ensuring that the dataset was refined and suitable for further analysis and processing.

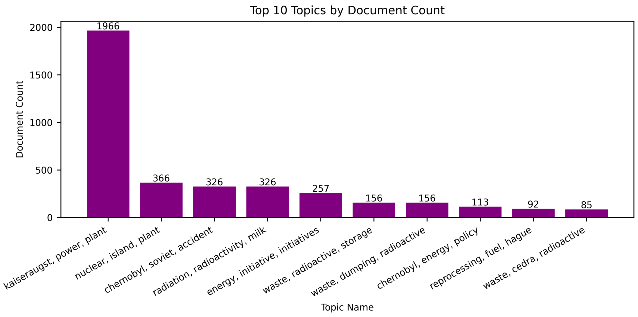


Figure 4. Top-10 topic distribution.

Utilizing BERTopic’s self-optimizing framework, the model automatically inferred that 20 topics most effectively captured the latent semantic structure of the corpus. Rather than being manually predefined, this number emerged from the algorithm’s internal optimization process, which balances topic coherence and cluster compactness to yield a thematically rich, data-driven representation of the content. Among the identified topics, the top-10 ranked by document frequency are presented in Figure 4, while the corresponding top-10 topic terms are shown in Figure 5.

The articles cover a wide range of topics related to nuclear power, including discussions on the construction of nuclear power plants, energy policies, and the authorization of new plants. They also address the aftermath of major nuclear ac-

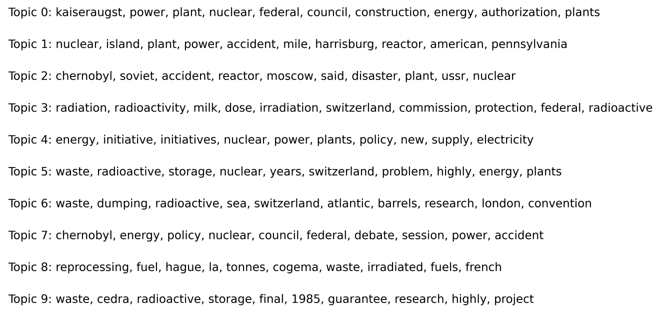
[o] https://impresso-project.ch/app/

[p] https://pypi.org/project/googletrans/3.1.0a0/

[q] https://platform.openai.com/docs/models/gpt-3.5-turbo

https://doi.org/10.1017/chr.2025.10018 Published online by Cambridge University Press

cidents, such as the Three Mile Island incident and the Chernobyl disaster, highlighting the global impact and concerns about nuclear safety. Additional topics include radiation protection, public health concerns related to radioactive exposure, and debates on energy initiatives aimed at securing a reliable power supply. Furthermore, there were articles about the challenges of managing radioactive waste, including storage and disposal issues, along with international research and agreements concerning the reprocessing of nuclear fuel and waste management. These topics reflect the complexity and diversity of perspectives on nuclear power as represented in the media.



Topic 0: kaiseraugst, power, plant, nuclear, federal, council, construction, energy, authorization, plants

Topic 1: nuclear, island, plant, power, accident, mile, harrisburg, reactor, american, pennsylvania

Topic 2: chernobyl, soviet, accident, reactor, moscow, said, disaster, plant, ussr, nuclear

Topic 3: radiation, radioactivity, milk, dose, irradiation, switzerland, commission, protection, federal, radioactive

Topic 4: energy, initiative, initiatives, nuclear, power, plants, policy, new, supply, electricity

Topic 5: waste, radioactive, storage, nuclear, years, switzerland, problem, highly, energy, plants

Topic 6: waste, dumping, radioactive, sea, switzerland, atlantic, barrels, research, london, convention

Topic 7: chernobyl, energy, policy, nuclear, council, federal, debate, session, power, accident

Topic 8: reprocessing, fuel, hague, la, tonnes, cogema, waste, irradiated, fuels, french

Topic 9: waste, cedra, radioactive, storage, final, 1985, guarantee, research, highly, project

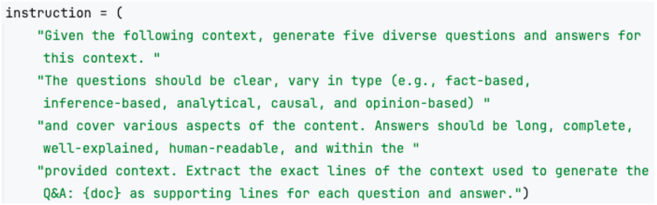
Figure 5. Top-10 topics and corresponding terms.

4.3 *Synthetic Q&A Generation*

Due to the lack of annotations in our original dataset and to systematically evaluate our Topic-RAG system, we generated synthetic Q&A pairs automatically by using an LLM. This approach allowed us to conduct a controlled and scalable evaluation, while acknowledging that it does not replicate the complexity of questions typically crafted by domain experts or historians. As an initial step, our focus was on generating single-document-answerable questions to validate the core functionality of our pipeline before extending it to more complex cross-document scenarios. To this end, we instructed the LLM to generate five diverse Q&A pairs for the first 100 documents, resulting in over 500 Q&A pairs in total. These questions were designed to cover a wide range of types including *fact-based*, *inference-based*, *analytical*, *causal*, and *opinion-based* questions, addressing various aspects of the documents, such as key events, consequences, actions taken, involved parties, technical details, and their impact on the population or environment.

To support the evaluation, we manually selected a representative subset of 50 Q&A pairs as ground truth for a detailed analysis. The 50 Q&A pairs were selected to reflect a representative range of question types and clarity levels rather than simply choosing the most easily answerable cases. While care was taken to ensure diversity, we acknowledge the potential for unintentional selection bias, which we plan to mitigate in future iterations. To ensure transparency and traceability, we extracted the exact lines from the original article that the LLM used as the basis for each answer, labeling these as supporting lines. These were verified manually to minimize hallucinations and to ensure alignment between the generated Q&A pairs and the source text.

We employed prompting techniques to create these relevant Q&A pairs, leveraging the recent Unsloth[r] library, which facilitates faster training and inference for LLMs. For this task, we utilized the Llama 3.1 8B Instruct[s] model to generate the synthetic Q&A pairs. A sample of the prompt design is shown in Figure 6. Since each Q&A pair was generated from a single document, no question required synthesis across multiple articles, and this constraint defines our current scope of evaluation.



```
instruction = (
    "Given the following context, generate five diverse questions and answers for
     this context. "
    "The questions should be clear, vary in type (e.g., fact-based,
     inference-based, analytical, causal, and opinion-based) "
    "and cover various aspects of the content. Answers should be long, complete,
     well-explained, human-readable, and within the "
    "provided context. Extract the exact lines of the context used to generate the
     Q&A: {doc} as supporting lines for each question and answer.")
```

Figure 6. Sample prompt for generating synthetic Q&A pairs.

4.3.1 *Human Validation of Synthetic Q&A Pairs*

To guarantee the quality and reliability of the generated Q&A pairs, we conducted a manual evaluation of a randomly sampled subset of approximately 100 pairs from the full set. We reviewed each pair for relevance, factual correctness, and alignment with the supporting lines extracted from the original documents. In addition, the supporting lines themselves were verified to eliminate potential hallucinations. This assessment verified that the answers were grounded in the source text and that the questions were clear, coherent, and contextually appropriate. Not all Q&A pairs were fully accurate; in such cases, they were flagged and excluded from further analysis. This human validation process served as an essential quality control step, both to assess the overall reliability of the synthetic Q&A generation pipeline and to support the construction of a reliable ground-truth dataset, which provides a quantitative benchmark for evaluating the performance of our Topic-RAG system.

4.3.2 *Ground-Truth Data Preparation*

Once the human validation process is finished, the ground-truth Q&A pairs are constructed through a refinement and filtering process. This involves incorporating human feedback to improve or enhance the phrasing, structure, and accuracy of the LLM-generated Q&A pairs. An illustrative example of this process is provided below.

- **Generated Question**: *"What are the implications of the IAEA's offer of advice and assistance to the Soviet Union in dealing with the aftermath of the Chernobyl accident?"*
- **Generated Answer**: *"The IAEA's offer of advice and assistance demonstrates a willingness to collaborate and provide expertise in addressing the crisis. This could potentially lead to a more effective and efficient response to the accident, and*

[r] https://github.com/unslothai/unsloth
[s] https://huggingface.co/unsloth/Meta-Llama-3.1-8B-Instruct

may also help to rebuild trust between the Soviet Union and the International community."

- **Supporting Lines**: *"...the IAEA, which brings together 112 members, including the USSR, offered its advice and any necessary assistance to the soviet union last Wednesday to deal with the aftermath of the accident..."*
- **Human Validation**: *"The answer could be expanded to include additional context about the IAEA, such as its membership of 112 countries, to emphasize the significance of its global expertise and influence."*
- **Modified Answers**: *"The IAEA, comprising 112 member countries, including the USSR, actively offers advice and assistance, showcasing a commitment to collaboration and expertise in managing the crisis. This approach could facilitate a more effective and efficient response to the incident while also contributing to rebuilding trust between the Soviet Union and the International community."*

The questions and answers were reformulated with the help of the supporting lines extracted during the synthetic Q&A generation as mentioned in Section 4.3. We manually selected 50 Q&A pairs as the "ground truth" based on these adjustments to ensure they are accurate and relevant. These ground-truth pairs act as the standard references for evaluating our Topic-RAG method with the baseline RAG method. By comparing the RAG responses to the ground truth, we can measure key aspects like accuracy, relevance, and coherence. This approach helps us thoroughly evaluate how well our RAG method retrieves and generates information, improving the overall user experience in practical applications.

4.4 Evaluation Metrics

In this subsection, we outline the various evaluation metrics we employed to assess the performance of our proposed method. We discuss both traditional metrics, such as BERTScore and ROUGE, and more recent advanced metrics like UniEval, each providing valuable insights into specific performance characteristics.

4.4.1 BERTScore

BERTScore (Zhang et al. 2020) is a widely adopted automatic evaluation metric in various NLP tasks, and we use it for assessing the quality of text generated by LLMs. It uses contextual embeddings to represent the reference and candidate tokens and then computes token-level cosine similarity between them. BERTScore also calculates Precision (P_{BERT}), Recall (R_{BERT}), and F1 (F_{BERT}) using the Cosine similarity values and by the greedy matching technique. Precision measures how well the tokens in the candidate text ($\hat{x}$) are represented in the reference text (x). Recall measures how well the tokens in the reference text (x) are captured in the candidate text ($\hat{x}$). F1 is the harmonic mean between the Precision and Recall. Scores range from 0 to 1, with higher values indicating better alignment[t].

Specifically, given the reference sentence $x = \langle x_1, x_2, \ldots, x_n \rangle$ and a corresponding candidate sentence $\hat{x} = \langle \hat{x}_1, \hat{x}_2, \ldots, \hat{x}_m \rangle$, it uses contextual embeddings to represent the tokens using the pre-trained BERT based embedding model. The reference sentence is first tokenized, and the embedding model converts it into a sequence of vectors $\langle \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \rangle$. Similarly, the candidate sentence is tokenized and converted into embedding vectors $\langle \hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \ldots, \hat{\mathbf{x}}_m \rangle$. Then the similarity between each reference token x_i and candidate token $\hat{x}_j$ is computed using their respective vectors $(\mathbf{x}_i, \hat{\mathbf{x}}_j)$ as follows: (Zhang et al. 2020).

$$Cosine(\mathbf{x}_i, \hat{\mathbf{x}}_j) = \frac{\mathbf{x}_i \cdot \hat{\mathbf{x}}_j}{\|\mathbf{x}_i\| \cdot \|\hat{\mathbf{x}}_j\|} \tag{2}$$

This reduces to $\mathbf{x}_i \cdot \hat{\mathbf{x}}_j$ as both $\mathbf{x}_i$, $\hat{\mathbf{x}}_j$ are pre-normalized vectors. Then, for each token x_i in x, it finds the token $\hat{x}_j$ in $\hat{x}$ that has the maximum Cosine similarity to calculate P_{BERT}. Similarly, for each token $\hat{x}_j$ in $\hat{x}$, it finds the token x_i in x that has the maximum Cosine similarity to compute R_{BERT}. Finally, $F1_{\text{BERT}}$ combines P_{BERT} and R_{BERT} using harmonic mean as follows.

$$P_{\text{BERT}} = \frac{1}{|\hat{x}|} \sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} \mathbf{x}_i \cdot \hat{\mathbf{x}}_j \tag{3}$$

$$R_{\text{BERT}} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} \mathbf{x}_i \cdot \hat{\mathbf{x}}_j \tag{4}$$

$$F_{\text{BERT}} = 2 \cdot \frac{P_{\text{BERT}} \cdot R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}} \tag{5}$$

Extensive experiments by the authors of Zhang et al. (2020) have demonstrated that BERTScore, which measures token-level semantic similarity between a generated candidate text and a reference text, correlates strongly with human judgment. In this work, we use BERTScore to evaluate the semantic fidelity of generated answers against ground-truth references. By computing BERTScore Precision, Recall, and F1, we aim to provide a nuanced evaluation of the effectiveness of our proposed method.

4.4.2 Recall-Oriented Understudy for Gisting (ROUGE) Evaluation

ROUGE (Lin 2004) is a set of metrics popularly used for evaluating the quality of the generated text, particularly in NLP tasks like text generation, summarization, and machine translation. It measures the overlap between n-grams, word sequences, and pairs of words between the gold standard text and the generated text. ROUGE has several variants like ROUGE-N, ROUGE-L, ROUGE-W, and ROUGE-S. In this work, we use ROUGE-N, ROUGE-L, and ROUGE-LSum[u], which are widely used for evaluating text generation tasks.

[t] https://huggingface.co/spaces/evaluate-metric/bertscore

[u] https://huggingface.co/spaces/evaluate-metric/rouge/blob/main/README.md

https://doi.org/10.1017/chr.2025.10018 Published online by Cambridge University Press

- **ROUGE-N: N-Gram Based Overlap** quantifies the overlap of *n*-grams (e.g., unigrams, bigrams) between the generated text and the reference text. In this study, we specifically compute ROUGE-1 and ROUGE-2, corresponding to unigram (single word) and bigram (two-word sequence) overlaps, respectively. The formula for ROUGE-N is given as:

$$\text{R-N} = \frac{\text{Number of overlapping } n\text{-grams}}{\text{Total number of } n\text{-grams in reference}} \tag{6}$$

  where $N = 1, 2, \ldots$. With its multiple variants, we can evaluate the generated text regarding content overlap, fluency, and coherence.
- **ROUGE-L: Longest Common Subsequence** measures the Longest Common Subsequence (LCS) between the reference and generated text. The advantage of using ROUGE-L is that it does not need consecutive matches between the words instead it looks for in-sequence matches. Unlike the *n*-gram-based measure, it automatically includes the longest in-sequence common *n*-grams without explicitly mentioning the pre-defined *n*-gram value. ROUGE-L captures sentence-level structure similarity by identifying the longest subsequence of words that appear in both texts, and it measures Precision, Recall, and F1 as follows.

$$\text{R-L(Precision)} = \frac{\text{Length of LCS}}{\text{Length of generated text}} \tag{7}$$

$$\text{R-L(Recall)} = \frac{\text{Lenght of LCS}}{\text{Length of reference text}} \tag{8}$$

$$\text{R-L(F1)} = \frac{(1+\beta^2)\cdot \text{Precision} \cdot \text{Recall}}{\beta^2 \cdot \text{Precision} + \text{Recall}} \tag{9}$$

  Here, β is the weighting factor to balance Precision and Recall and is normally set to 1.
- **ROUGE-Lsum** is the variant of ROUGE-L developed to capture the summary-level similarity. ROUGE-L considers the entire input text as a single unit, ignoring sentence boundaries like newlines. ROUGE-Lsum builds upon ROUGE-L splits the text into sentences based on newlines and then computes ROUGE-L for each sentence pair in the reference and generated text. Finally, it averages these scores across all sentence pairs:

$$\text{R-Lsum(Precision)} = \frac{\sum \text{LCS across all sentences}}{\text{Total length of generated text}} \tag{10}$$

$$\text{R-Lsum(Recall)} = \frac{\sum \text{LCS across all sentences}}{\text{Total length of reference text}} \tag{11}$$

$$\text{R-Lsum(F1)} = \frac{(1+\beta^2)\cdot \text{Precision} \cdot \text{Recall}}{\beta^2 \cdot \text{Precision} + \text{Recall}} \tag{12}$$

  where β is the weighting factor to balance Precision and Recall, and is normally set to 1.

4.4.3 UniEval

UniEval (Zhong et al. 2022) is the automatic multi-dimensional evaluation framework for Natural Language Generation (NLG) tasks. It aims to overcome the limitations of traditional evaluation metrics (e.g., BLUE (Papineni et al. 2002)), which often fail to assess all the necessary aspects of generated text. The evaluation dimensions in UniEval[v] include Coherence, Consistency, Fluency, and Relevance. ***Coherence*** evaluates whether the sentences in the generated text are logically connected and form a clear, understandable narrative or argument. It ensures that the text flows smoothly and makes sense as a whole. ***Consistency*** evaluates the factual alignment between the generated text and the reference text. ***Fluency*** measures the quality of the sentences in terms of readability and grammar. Finally, ***Relevance*** determines how well the generated text captures the key information from the original document. In this study, we employed UniEval-sum, a pre-trained evaluator specifically designed for text summarization tasks. Further details regarding the training process of this evaluator can be found in the original paper (Zhong et al. 2022). The authors highlighted that, based on extensive experiments conducted across three NLG tasks, UniEval demonstrates a better correlation with human judgments compared to existing evaluation metrics.

4.5 Quantitative Evaluation with Baseline Method

In this subsection, we present a comparison of the performance of our proposed method, Topic-RAG, with a well-established baseline RAG method. We structure our evaluation into two main components: **retrieval quality** is assessed through lexical and semantic similarity between the query and retrieved documents, reflecting how well the system locates relevant material; **generation quality** is evaluated using BERTScore, ROUGE, and UniEval as discussed in Subsection 4.4. By assessing both methods based on the same evaluation criteria, we aim to highlight the areas where our method excels and demonstrates superior results or introduces innovative benefits. The purpose of this comparison is to objectively evaluate the performance of our proposed approach relative to existing methods, highlighting both its strengths and areas for improvement to provide a clear and comprehensive understanding of its overall effectiveness.

We implemented a modified version of the original Retrieval-Augmented Generation (RAG) architecture as proposed by Lewis et al. (2020), and we refer to it as *Base-RAG*. For the retrieval component, we analogously utilized the FAISS[w] library, which supports efficient vector indexing and several methods for similarity search. For the generation component, we employed the quantized version of Llama 3.1 8B Instruct to generate coherent and contextually relevant text based on the retrieved documents. This setup leverages the strengths of FAISS for fast and scalable retrieval and the Llama model's robust capabilities for generating high-quality text.

[v] https://huggingface.co/MingZhong/unieval-sum
[w] https://github.com/facebookresearch/faiss

https://doi.org/10.1017/chr.2025.10018 Published online by Cambridge University Press

4.5.1 *Retrieval Quality*

To evaluate retrieval quality, we assessed both *lexical* and *semantic* alignment between the retrieved documents and the ground truth documents across a set of 50 queries. The ground truth documents refer to the original source passages used to generate synthetic Q&A pairs via an LLM (as described in subsection 4.3). For each query, we examine whether the retrieval system is able to return the same document that was originally used to generate the Q&A pair, thereby serving as a reference for evaluating retrieval accuracy. A boxplot summarizes a distribution by displaying its median, quartiles, and range, making it easy to compare performance and variability across methods.

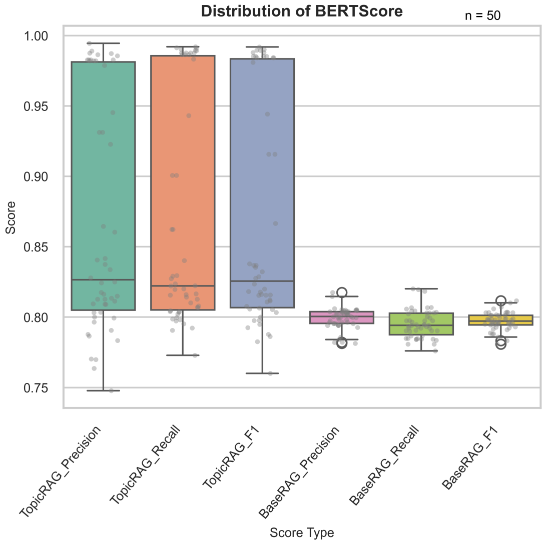


Figure 7. Evaluation of retrieval quality using BERTScore for BaseRAG and TopicRAG across 50 queries.

In comparing retrieval quality between TopicRAG and BaseRAG, both semantic (BERTScore) and lexical (ROUGE) metrics consistently favor TopicRAG as the superior system. As shown in Figure 7, TopicRAG achieves markedly higher BERTScore values across all three submetrics, including Precision, Recall, and F1. While BaseRAG's scores cluster around the 0.78–0.81 range with wider interquartile ranges and longer lower whiskers, TopicRAG exhibits tighter distributions with many individual scores near 1.0. Notably, despite an unexpected dip in the median of TopicRAG's Precision (approximately 0.83), its Recall and F1 scores remain substantially higher than those of BaseRAG, indicating that it retrieves more semantically relevant content overall. The dense concentration of high-scoring points and narrower variance further affirm TopicRAG's reliability and semantic robustness. The difference is even more pronounced when evaluating lexical alignment using ROUGE metrics, as seen in Figure 8. Although TopicRAG consistently scores above 0.9 in ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-Lsum, their medians typically vary, indicating moderate token-level overlap. In contrast, BaseRAG performs substantially worse: its ROUGE-1 median hovers near 0.3, while ROUGE-2 and ROUGE-Lsum are close to 0.1 or lower, with wide variability and multiple low outliers. This dramatic gap in ROUGE performance illustrates that BaseRAG frequently retrieves documents with little to no lexical overlap with the ground truth, undermining its effectiveness in scenarios that depend on token-level alignment.

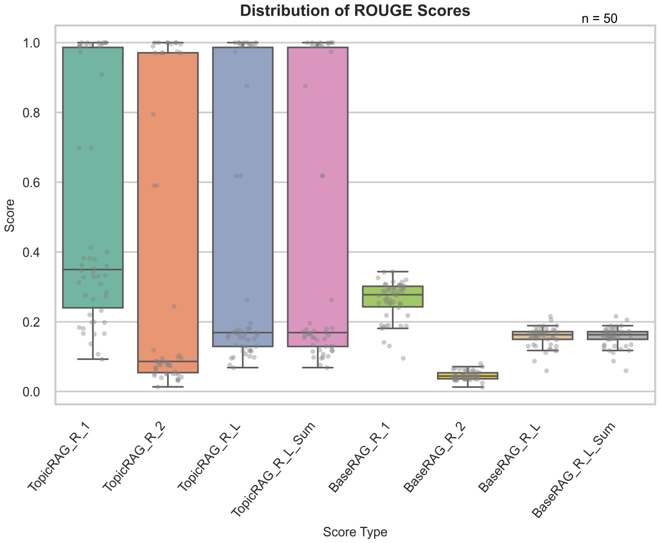


Figure 8. Evaluation of retrieval quality using ROUGE for BaseRAG and TopicRAG across 50 queries.

Although ROUGE scores remain relatively low across the retrieved outputs, BERTScore results offer a compelling indication of TopicRAG's retrieval quality. ROUGE evaluates exact token overlap between the retrieved document and the ground truth, and its low values suggest limited lexical matching. However, this does not necessarily imply poor retrieval, particularly in settings where documents may express similar meaning using different phrasings. In contrast, BERTScore, by leveraging contextual embeddings, provides a more robust measure of semantic similarity. The consistently high BERTScore values across Precision, Recall, and F1 suggest that TopicRAG retrieves documents that are semantically close to the reference passages, even if their surface forms differ. This pattern validates TopicRAG's effectiveness in retrieving content that captures the intended meaning of the source, supporting its utility in applications where semantic alignment is more critical than exact wording.

4.5.2 *Generation Quality*

We evaluated our proposed method by comparing the generated answers with a set of 50 ground truth Q&A pairs. These pairs were prepared using synthetic data-generation techniques and have been validated by human experts. We evaluated by comparing the performance of our method with the baseline using various metrics, including BERTScore (Precision, Recall, and F1), ROUGE (R-1, R-2, R-L, R-Lsum) in terms of F1 score, and UniEval. We visualized our results using box plots. Each box represents the inter-quartile range (IQR), which contains the middle 50% of the values. The box

https://doi.org/10.1017/chr.2025.10018 Published online by Cambridge University Press

spans from the first quartile (Q1, 25th percentile) to the third quartile (Q3, 75th percentile), with a line at the median (50th percentile). The whiskers extending from the box indicate the range of the scores, excluding outliers. Circles outside the whiskers represent outliers, which are values that are significantly higher or lower than the rest of the data. The central line within each box indicates the median, which represents the middle value of the distribution.

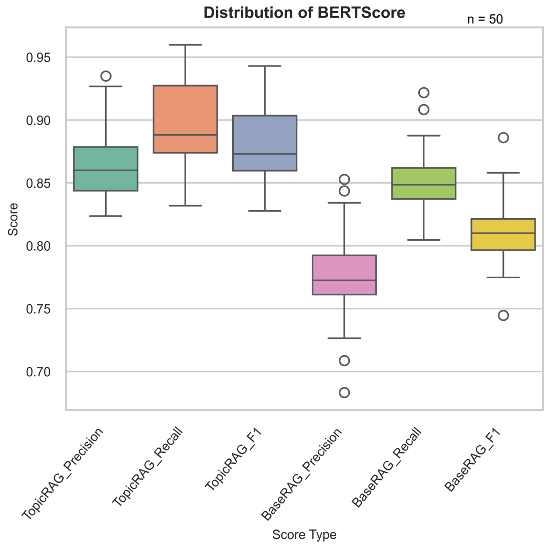


Figure 9. Evaluation of generation quality using BERTScore for BaseRAG and TopicRAG across 50 queries.

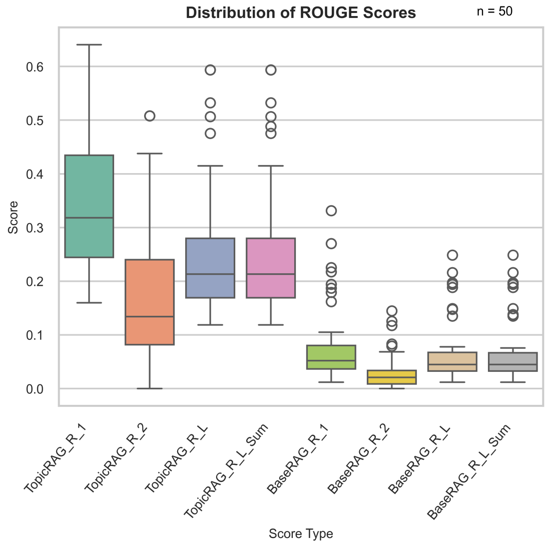


Figure 10. Evaluation of generation quality using ROUGE for BaseRAG and TopicRAG across 50 queries.

The three box plots in Figure 9, Figure 10, and Figure 11 visualize the performance of Topic-RAG versus Base-RAG using three different evaluation metrics: BERTScore, ROUGE scores, and UniEval scores.Topic-RAG shows a high BERTScore precision, recall, and F1 score compared to Base-RAG. Topic-RAG's Recall reaches up to 0.95, while Precision and F1 remain above 0.85 for most of the samples. Base-RAG shows lower precision, recall, and F1 scores, while the outliers (circles) occasionally indicate better performance, but the median is lower than for Topic-RAG. Moreover, Topic-RAG outperforms Base-RAG across all ROUGE metrics, particularly in ROUGE-1 and ROUGE-2, where the median scores for Topic-RAG are higher than those for Base-RAG, indicating that it generates responses with better word overlap at the unigram and bigram levels. For ROUGE-L and ROUGE-Lsum, Topic-RAG's median is around 0.2–0.3, suggesting that it performs better in matching longer, more meaningful text sequences than Base-RAG, whose median is much lower. Topic-RAG also performs better than Base-RAG in terms of the UniEval metric. Topic-RAG has tighter box plots with fewer outliers, indicating a very stable performance across all queries. Overall, our approach outperformed the Base-RAG model, demonstrating the advantages of incorporating topic modeling in the retrieval phase.

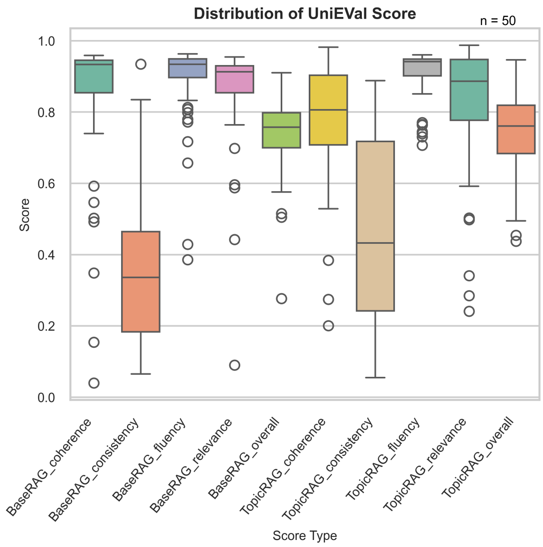


Figure 11. Evaluation of generation quality using UniEval for BaseRAG and TopicRAG across 50 queries.

Additionally, the retrieval time of Topic-RAG was significantly reduced compared to the baseline method. This improvement is due to our method's ability to avoid processing the entire document corpus. Unlike basic RAG approaches that perform exhaustive searches over every document to retrieve relevant information, our method employs a more efficient retrieval strategy. By utilizing topic indices and embedding-based retrieval, it focuses only on the subset of documents most relevant to the query's topic representation. This focused approach reduces the amount of data that needs to be processed, resulting in faster retrieval and generation times without compromising the quality of the output response, as shown in Figure 12 and Figure 13. We measured processing time for both single and multiple queries and also tracked the time based on the number of retrieved documents. Topic-RAG's computation time remains low when retrieving a small number

https://doi.org/10.1017/chr.2025.10018 Published online by Cambridge University Press

of documents, but increases as more documents are retrieved. In contrast, Base-RAG's computation time remains relatively constant, as it performs exhaustive searches regardless of the number of retrieved documents. Our results demonstrate that Topic-RAG offers a substantial improvement in computational efficiency compared to Base-RAG, particularly in scenarios involving large document corpora. By focusing retrieval efforts on a smaller, topic-relevant subset of documents, Topic-RAG reduces processing time while maintaining high-quality responses. This makes it a more scalable and efficient solution for real-time retrieval-augmented generation tasks.

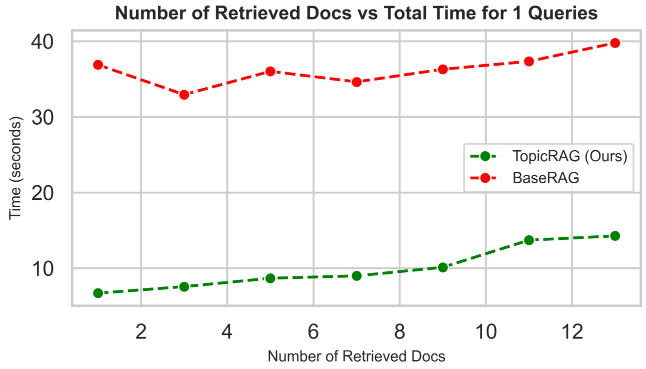


Figure 12. Processing time for a single query.

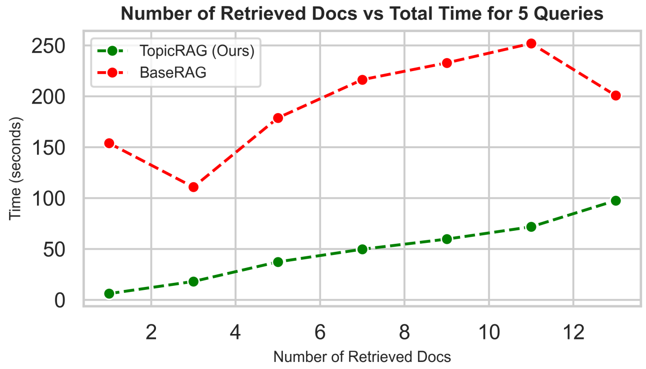


Figure 13. Processing time for 5 queries.

4.6 Topic RAG+ to Handle Long Documents

We also implemented an enhanced version of our Topic-RAG framework by introducing a chunking strategy to handle long documents more efficiently and improve retrieval performance. We refer to this improved version as *Topic-RAG+*. The key difference between this Topic-RAG+ and Topic-RAG lies in the way document embeddings are stored and managed.

In the traditional Topic-RAG, documents are processed and stored based on their overall topic, with each document treated as a single, coherent unit. The embeddings of entire documents are computed, and these embeddings are then indexed based on their topic association. This approach works well for smaller documents, where the entire content can be effectively represented in a single embedding. However, this method encounters limitations when dealing with long documents, as they may contain multiple subtopics, which can result in inefficiencies during the retrieval process.

To address the challenge of efficiently handling long documents and improving retrieval, we incorporated a semantic chunking strategy within the indexing process. Long documents are split into smaller, semantically meaningful chunks using the `SemanticChunker` of LangChain[x], an advanced tool that segments documents based on the semantic coherence of the subsequent sentences. `SemanticChunker` divides each document into semantically coherent segments by analyzing sentence-level embedding similarity. This allows it to detect subtle topic shifts, even within documents that discuss multiple related subjects, thereby improving alignment between chunks and the underlying thematic structure. This chunking process ensures that each chunk represents a specific part of the document, capturing a meaningful portion of the content.

First, we apply BERTopic to our corpus to perform topic modeling, ensuring that the global context and primary themes of the documents are preserved. Once the documents are assigned to topics, we map them accordingly, creating a document-topic mapping. In each topic, we apply the semantic chunking strategy to break down the documents into smaller, topic-relevant chunks. Each chunk is then embedded using the same embedding model as in Topic-RAG, and the resulting embeddings are stored again in the FAISS vector index for efficient retrieval. Additionally, we store essential metadata separately, including the document IDs (UIDs), the chunk IDs (CIDs), the chunk contents, and the chunk embeddings. This process enables the retrieval system to focus on semantically meaningful chunks rather than entire documents, significantly improving retrieval efficiency and accuracy.

The chunking strategy improves the granularity of the retrieval process. Instead of retrieving entire documents that may contain irrelevant information, the system can focus on retrieving the most relevant chunks to the query, ensuring the responses generated are more focused and relevant. During retrieval, each chunk is indexed with this metadata, which also includes the chunk position, ensuring the system can maintain accurate tracking of the chunk's location within the original document.

The chunking strategy also improves the scalability of the Topic-RAG+ framework. Since large documents are broken down into smaller chunks, the retrieval process becomes more efficient even when dealing with extensive document collections. This is particularly valuable for applications that require processing vast amounts of data, such as scientific research papers, legal documents, or news archives. By working with chunks rather than entire documents, the system can scale more effectively to handle larger datasets without sacrificing retrieval performance or accuracy. The combination of semantic chunking and metadata-based storage ensures that Topic-RAG+ remains scalable, adaptable, and capable of han-

[x] https://python.langchain.com/docs/how_to/semantic-chunker/

https://doi.org/10.1017/chr.2025.10018 Published online by Cambridge University Press

dling large datasets, making it a robust solution for complex, long-document retrieval tasks. By preserving both the global document context through topic modeling and the semantic integrity of each chunk, this approach ensures that long documents are handled effectively, making it ideal for large corpora like historical newspaper archives.

4.7 Case Study: Qualitative Evaluation from a Historian's Perspective

In addition to the quantitative analysis of Topic-RAG provided in the previous subsection, we also employed *Topic-RAG+* for a qualitative evaluation over nine complex queries (depicted in Table 1. These queries were formulated by our historian domain experts specifically based on this corpus and research objectives, reflecting the typical lines of inquiry historians follow when analyzing archival newspapers, rather than being derived from a particular published source.

Here, documents were dynamically retrieved based on user input (100 documents), ensuring that the majority of relevant documents were considered. The chunking strategy proved highly effective in efficiently handling such queries. The remainder of this section presents an additional case study evaluating these complex queries, conducted independently by expert historians. This qualitative assessment by domain experts complements our quantitative analysis and reinforces the strength of our method by examining the system's ability to support historically grounded research inquiries.

For this purpose, we identified nine intentionally challenging queries that require a combination of analytical, opinion-based, factual, and inference-based reasoning. From a historiographical perspective, it is essential to recognize newspapers not merely as carriers of factual information but both as reflections of their contemporary world and as active agents that promote interpretations of this world alongside the political and/or religious orientations of their creators. Especially for this case study, newspapers are less relevant for the retrieval of historical facts but rather as carriers of a past "Zeitgeist". Traditionally, historians have sought to approach this goal by close reading and synthesizing large volumes of text. Scholars quickly reach their limits here: the abundance of archived materials has forced those working with traditional methods to either invest considerable time or make rigorous choices regarding the number of newspaper titles or the temporal coverage. LLMs present an exciting opportunity to augment and scale these research practices by delegating tasks. In the context of this case study, we have experimented with the following:

- the classification of themes and viewpoints according to theoretical frameworks and analysis of the distributions across newspaper titles and over time,
- the detection and mapping of opinion-based arguments expressed by different stakeholders such as journalists, activists, or experts,
- the distinction between editorial text striving for neutral coverage and the explicit statement of opinions,
- the detection of subtle expressions of opinion within seemingly neutral coverage, e.g., expressed through omissions and stylistic choices.

These tasks were translated into specific queries to support a research project on the representation of the debate over nuclear power technologies in Switzerland in the 1970s and 80s. Note that these queries would be hard to answer for a human since they require in-depth analysis of a large number of texts and abstract reasoning, and pattern recognition. Results, as presented in Table 1, are overall promising, especially in light of queries for elusive concepts such as "scientific experts", "fear", or "environmentalism", which keyword-based approaches can hardly capture. A closer look at the results serves to better illustrate the strengths and some shortcomings of the system.

Seemingly precise instructions as those in Query 2 (*"Which articles take an explicit position in favor or against the Volksbegehren?"*), yielded summaries of the viewpoints of stakeholders first in full text and then accompanied by a list representation of the same information.

Query 3, with the task of retrieving a list of scientific experts, yielded some good results but also clear mistakes. A positive example is the following, quoted from the output: *"**Mr. Hans-Rudolf Lutz**, Director of Kaiseraugst SA, engaged in discussions with the municipality about the safety of the nuclear power plant."* Other output was flawed: poor article segmentation linked Dr. Josef Amstutz, superior general of the Bethlehem Mission Society, to concerns over nuclear waste. A closer inspection showed that poor article segmentation in the input newspaper data caused the system to incorrectly merge two distinct articles.

The response for Query 5 about the media creating a sense of fear among their readers yielded encouraging results which demonstrates the potential of Topic-RAG+ to generate relevant content that is too elusive for standard keyword queries: *"The Neue Zürcher Zeitung wrote that this kind of anxiety makes minds sensitive to all negative information which is welcomed and promptly disseminated by the mass media, often inflated to the point of grotesqueness or distorted, which further increases the psychological effect."*

Similarly encouraging is the output for Query 8, which asks for experts on nuclear waste. The system correctly identified a number of experts and presented them together with a summary of their representation in the newspapers: *"1. Otto Luscher, an engineer from Winterthur, Switzerland, who suggests that radioactive residues can be used for future applications, such as sterilizing medical and surgical accessories."*

There are, however, also clear limitations with the current setup, which require future work. In its current configuration, the system displays a clear preference and strong performance for summarization. The research-motivated queries, on the other hand, require higher levels of abstraction (ratios, evolution over time, etc.) as well as precise contextualization. To a researcher, it matters who exactly said what, where, and when. In this regard, output in the form of lists and tables with rich contextual information originating from metadata is preferable to mere summaries. Our corpus is dominated

https://doi.org/10.1017/chr.2025.10018 Published online by Cambridge University Press

by Topic 0, containing 1,966 articles, while the other topics contain between 23 and 326 articles. This seemingly resulted in a bias towards Topic 0 when linking queries to topics. More focused queries and more explicit instructions regarding the expected output may help to address the problems we have observed.

Overall, and despite these limitations, the system has shown its potential to comprehend complex queries, to map them to corresponding topics, and to perform complex reasoning. These first experiments demonstrate, however, that the system is in principle capable of delivering such output and may perform significantly better with adjusted prompting.

5. Strengths of the Proposed System

In this section, we highlight the key advantages of our Topic-RAG and Topic-RAG+ systems over baseline RAG. The features outlined below collectively contribute to more accurate, context-aware, and reliable responses, making our approach highly effective for a range of applications, from historical research to domain-specific tasks.

1. **Better Query Understanding:** Our method leverages topic modeling to provide a broader interpretation of queries by using probabilities to identify multiple relevant topics, instead of relying on BERTopic's single-topic assignment. In cases where a query is narrowly focused, the retrieval method naturally prioritizes documents dominated by a single topic. For example, in the query: "Give an overview of the year in which environmental topics were first discussed in the different newspapers. As output, produce a table with the following columns: newspaper title, date of publication, article title, 10-word article summary." Relevant topic IDs are 5, 6, 8, 9, 10, 12, 13, 14, 16, 18, and 19. This diversity highlights the need to capture multiple themes to retrieve a representative and meaningful set of documents. However, in some cases, such as the more focused query: "How did the newspapers of Eastern European countries (Prague, Bucharest, East Berlin, Budapest, and Sofia) report on the Three Mile Island nuclear accident?" Only a single topic (Topic ID: 1) is dominant. In such cases, our multi-topic pipeline naturally prioritizes documents with high relevance to that one topic, effectively behaving like a single-topic retrieval system. This approach captures topic overlaps in real-world queries, resulting in a more accurate and nuanced understanding of query intent.
2. **Flexible Document Retrieval:** Users can retrieve a flexible number of documents based on Cosine similarity scores, providing more control compared to traditional top-k retrieval. This flexibility allows adjusting the number of retrieved documents according to the query's complexity or user preferences.
3. **Contextual Document Retrieval and Noise Reduction:**Our method ensures better coverage and diversity in retrieved documents by capturing the full scope of the query, overcoming the limitations of the single-topic approaches, and filtering out documents that are semantically related but contextually irrelevant. This reduces noise in the retrieval process and enhances the quality of responses.
4. **Traceability:**A key advantage of our Topic-RAG system is its ability to maintain a comprehensive record of the original documents, UID, chunk IDs, and associated metadata throughout the retrieval and generation process. This structured tracking ensures transparency and traceability, enabling users to understand the source, context, and reliability of the retrieved information. Unlike a baseline RAG system, which typically focuses on entire document embeddings and retrieval, our approach delivers richer, context-aware outputs by utilizing this metadata. This feature is especially valuable for applications that require high accountability, such as historical research, legal analysis, or academic work, where understanding the provenance of information is crucial.
5. **Adaptability to Diverse Domains:** We evaluated our proposed method using the Impresso dataset. Although the current experiments are focused on this specific dataset and question type, the proposed Topic-RAG framework is designed to generalize across domains due to its modular and unsupervised architecture. Specifically, BERTopic allows adaptive topic discovery from unstructured corpora, and RAG dynamically retrieves semantically relevant context using dense embeddings, enabling the system to operate effectively across diverse content without requiring retraining of the underlying LLM. Because our method is fully unsupervised and does not rely on annotated data or domain-specific fine-tuning, it is more broadly applicable to new domains and corpora. The method's ability to identify and retrieve thematically coherent document clusters helps reduce noise and improve the efficiency of RAG. Future work will examine this potential by applying the approach to a wider range of datasets and question formats.

6. Conclusion

We implemented a Topic-RAG system that integrates topic modeling within the Retrieval-Augmented Generation (RAG) architecture to enhance both retrieval accuracy and retrieval time. Our Topic-RAG system leverages BERTopic for topic modeling, which allows the system to efficiently focus on document subsets relevant to the query's topic, rather than performing exhaustive searches over the entire corpus. This results in faster retrieval times and more accurate responses. Additionally, we developed a Topic-RAG+ system by incorporating a chunking strategy to handle long documents more effectively. The chunking strategy, powered by a semantic chunker from LangChain, splits lengthy documents into semantically meaningful chunks, ensuring that the system can manage large and complex documents without losing important context.

Topic-RAG outperformed the baseline RAG in terms of evaluation metrics such as BERTScore, ROUGE, and UniEval scores. Additionally, Topic-RAG+ showed significant potential in handling complex queries within a real-time historical research case study. These improvements signify the efficacy of incorporating topic modeling and chunking strategies to

https://doi.org/10.1017/chr.2025.10018 Published online by Cambridge University Press

refine document retrieval and generation processes.

The importance of these advancements is particularly notable in the context of Humanities research, where vast archives of historical, literary, and cultural documents need to be analyzed and interpreted efficiently. Humanities scholars often work with large, unstructured corpora, such as historical newspapers, literary works, or cultural texts, which present unique challenges for retrieval systems. The Topic-RAG and chunking strategies enable more accurate, contextually relevant document retrieval, making it easier to uncover insights from large datasets. This system enhances researchers' ability to quickly access and synthesize information, providing a powerful tool for exploring historical narratives, conducting literary analysis, and gaining deeper insights into cultural trends. Furthermore, by improving retrieval accuracy and reducing computational time, our approach significantly aids scholars in overcoming the limitations posed by large-scale document collections in the humanities.

7. Future Work

While our current study focused on a mid-sized corpus (4.7K documents, future work will aim to scale Topic-RAG to much larger corpora (e.g., 100K–1M documents) to simulate real-world archival research environments. We plan to benchmark retrieval and generation performance across different corpus sizes and document granularities, including cross-document reasoning and multi-hop question answering. Also, we plan to focus on generating more complex and nuanced questions to better align with the requirements for complex reasoning and rich contextualization required in historical research. We plan to test the Topic-RAG system on a broader set of diverse questions to evaluate its generalizability and robustness across different domains and query types. Specifically, we have identified the following areas to address shortcomings we observed during the first evaluations: the quality of relevant topic identification by the system, the interpretation of queries regarding expected output formats (list, summary, table, etc.), the preservation of relevant provenance information, and the integration of recent advances in explainable AI research for more transparency of system out.

Additionally, we intend to experiment with different and forthcoming versions of LLMs, both for retrieval and generation tasks, to assess how model size influences the system's performance in terms of accuracy, retrieval time, and response quality. Moreover, we will explore the integration of fine-tuning on domain-specific corpora to improve the contextual understanding of the model. These future directions will help us enhance the scalability, adaptability, and overall efficacy of our Topic-RAG system, making it a useful tool for various research fields, particularly in the Humanities.

Authorship Contribution Statement(CRediT) K.M.: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Software, Writing – original draft. S.L.: Validation, Investigation, Writing – reviewing & editing. M.D.: Resources, Writing – reviewing & editing. M.T.: Supervision, Project administration, Writing – reviewing & editing.

Competing Interests - We declare no competing interests to disclose.

Funding Disclosure Statement This work is part of the Doctoral Training Unit "Deep Data Science for History"[y], a collaboration between the Luxembourg Centre for Contemporary and Digital History (C^{2}DH) and the Department of Computer Science (DCS) at the University of Luxembourg, promoting interdisciplinary research between historians and computer scientists.

Data Availability Statement — The replication code and detailed information on how to obtain access to the dataset are available in our GitHub repository: link.

References

Achiam, Josh, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774.*

AlSumait, Loulwah, Daniel Barbará, and Carlotta Domeniconi. 2008. On-line LDA: Adaptive Topic Models for Mining Text Streams with Applications to Topic Detection and Tracking. In *Proceedings of the 8th IEEE international conference on data mining (ICDM 2008),* 3–12. IEEE Computer Society. https://doi.org/10.1109/ICDM.2008.140.

Angelov, Dimitar. 2020. Top2vec: distributed representations of topics. *ArXiv* abs/2008.09470. https://api.semanticscholar.org/CorpusID:221246303.

Anthropic. 2024. The claude 3 model family: opus, sonnet, haiku. https://api.semanticscholar.org/CorpusID:268232499.

Arseniev-Koehler, Alina, Susan D. Cochran, Vickie M. Mays, Kai Wei Chang, and Jacob Gates Foster. 2020. Integrating topic modeling and word embedding to characterize violent deaths. *Proceedings of the National Academy of Sciences of the United States of America* 119. https://api.semanticscholar.org/CorpusID:235658060.

Blei, David M. 2012. Probabilistic topic models. *Communications of the ACM* 55 (4): 77–84. https://doi.org/10.1145/2133806.2133826.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* 3:993–1022. https://jmlr.org/papers/v3/blei03a.html.

Bodrunova, Svetlana S. 2021. Topic modeling in russia: current approaches and issues in methodology. In *The palgrave handbook of digital russia studies,* 409–426. Cham: Springer International Publishing. ISBN: 978-3-030-42855-6. https://doi.org/10.1007/978-3-030-42855-6_23. https://doi.org/10.1007/978-3-030-42855-6_23.

Bolton, Elliot, Abhinav Venigalla, Michihiro Yasunaga, David Hall, Betty Xiong, Tony Lee, Roxana Daneshjou, et al. 2024. Biomedlm: a 2.7b parameter language model trained on biomedical text. *ArXiv* abs/2403.18421. https://api.semanticscholar.org/CorpusID:268723860.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33:1877–1901.

[y] https://d4h.uni.lu/

Chang, Jonathan D., Jordan L. Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei. 2009. Reading Tea Leaves: How Humans Interpret Topic Models. In *Advances in neural information processing systems 22: 23rd annual conference on neural information processing systems 2009.* 288–296. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2009/hash/f92586a25bb3145facd64ab20fd554ff-Abstract.html.

Chappelle, Marc, Sakaria Laisene Auelua-Toomey, and Steven O. Roberts. 2024. Sankofa: using topic models to review the history of the journal of black psychology. *Journal of Black Psychology* 50 (1): 9–29. https://doi.org/10.1177/00957984231221028.

Cheddak, Asma, Tarek Ait Baha, Youssef Es-saady, and Mohamed El Hajji. 2024. BERTopic for Enhanced Idea Management and Topic Generation in Brainstorming Sessions. *Information* 15:365. https://doi.org/10.3390/info15060365.

Chowdhery, Aakanksha, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: scaling language modeling with pathways. *Journal of Machine Learning Research* 24 (240): 1–113.

Cvejoski, Kostadin, Ramsés J. Sánchez, and C. Ojeda. 2023. Neural dynamic focused topic model. In *Aaai conference on artificial intelligence.* https://api.semanticscholar.org/CorpusID:252564193.

Dan, Yuhao, Zhikai Lei, Yiyang Gu, Yong Li, Jianghao Yin, Jiaju Lin, Linhao Ye, Zhiyan Tie, Yougen Zhou, Yilei Wang, et al. 2023. Educhat: a large-scale language model-based chatbot system for intelligent education. *arXiv preprint arXiv:2308.02773.*

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Jun-Mei Song, Ruoyu Zhang, et al. 2025. Deepseek-r1: incentivizing reasoning capability in llms via reinforcement learning. *ArXiv* abs/2501.12948. https://api.semanticscholar.org/CorpusID:275789950.

Deerwester, Scott, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science* 41 (6): 391–407. https://doi.org/10.1002/(SICI)1097-4571(199009)41:6\<391::AID-ASI1\>3.0.CO;2-9.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers),* 4171–4186. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423.

Douze, Matthijs, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. *The Faiss library.* arXiv: 2401.08281 [cs.LG]. https://arxiv.org/abs/2401.08281.

Düring, Marten, Estelle Bunout, and Daniele Guido. 2024. Transparent generosity. Introducing the impresso interface for the exploration of semantically enriched historical newspapers. Publisher: Routledge, *Historical Methods: A Journal of Quantitative and Interdisciplinary History* (June): 35–55. ISSN: 0161-5440, accessed June 14, 2024. https://doi.org/10.1080/01615440.2024.2344004. https://www.tandfonline.com/doi/full/10.1080/01615440.2024.2344004.

Egger, Roman, and Joanne Yu. 2022a. A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts. *Frontiers in Sociology* 7. https://api.semanticscholar.org/CorpusID:248530058.

———. 2022b. A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. *Frontiers in Sociology* 7. https://doi.org/10.3389/fsoc.2022.886498.

Ehrmann, Maud, Estelle Bunout, S. Clematide, Marten Düring, Andreas Fickers, Roman Kalyakin, F. Kaplan, et al. 2020. Historical Newspaper Content Mining: Revisiting the Impresso Project's Challenges in Text and Image Processing, Design and Historical Scholarship. In *Digital Humanities Conference.* https://api.semanticscholar.org/CorpusID:220667683.

Ginn, Michael, and Mans Hulden. 2024. Historia magistra vitae: dynamic topic modeling of roman literature using neural embeddings. *arXiv preprint arXiv:2406.18907.*

Grant, Philip, Ratan Sebastian, Marc Allassonnière-Tang, and Sara Cosemans. 2021. Topic modelling on archive documents from the 1970s: global policies on refugees. *Digital Scholarship in the Humanities* 36, no. 4 (March): 886–904. ISSN: 2055-7671. https://doi.org/10.1093/llc/fqab018. eprint: https://academic.oup.com/dsh/article-pdf/36/4/886/41027215/fqab018.pdf. https://doi.org/10.1093/llc/fqab018.

Greenfield, Patricia M. 2013. The changing psychology of culture from 1800 through 2000. *Psychological Science* 24:1722–1731. https://api.semanticscholar.org/CorpusID:6123553.

Grootendorst, Maarten. 2022. *BERTopic: Neural topic modeling with a class-based TF-IDF procedure.* arXiv: 2203.05794 [cs.CL]. https://arxiv.org/abs/2203.05794.

Gryaznova, Ekaterina, and Margarita Kirina. 2021. Defining kinds of violence in russian short stories of 1900-1930: a case of topic modelling with lda and pca. In *Ims,* 281–290.

Hauser, Jakob, Daniel Kondor, Jenny Reddish, Majid Benam, Enrico Cioni, Federica Villa, James S. Bennett, et al. 2024. Large language models'expert-level global history knowledge benchmark (hist-llm). In *Advances in neural information processing systems,* edited by A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, 37:32336–32369. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2024/file/38cc5cba8e513547b96bc326e25610dc-Paper-Datasets_and_Benchmarks_Track.pdf.

Hills, Thomas, and Alessandro Miani. 2023. A short primer on historical natural language processing. https://api.semanticscholar.org/CorpusID:273174599.

Hoeber, Orland, Morgan Harvey, Milad Momeni, Abbas Pirmoradi, and David Gleeson. 2024. Exploratory search in digital humanities: a study of visual keyword/result linking. *Proceedings of the Association for Information Science and Technology* (USA) 61, no. 1 (October): 161–171. https://doi.org/10.1002/pra2.1017. https://doi.org/10.1002/pra2.1017.

Imani, Shima, Liang Du, and Harsh Shrivastava. 2023. Mathprompter: mathematical reasoning using large language models. *arXiv preprint arXiv:2303.05398.*

Indukaev, Andrey. 2020. Studying Ideational Change in Russian Politics with Topic Models and Word Embeddings. *The Palgrave Handbook of Digital Russia Studies,* https://api.semanticscholar.org/CorpusID:234122155.

Jiang, Albert Qiaochu, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, et al. 2023. Mistral 7b. *ArXiv* abs/2310.06825. https://api.semanticscholar.org/CorpusID:263830494.

Karamouzi, Eirini, Maria Pontiki, and Yannis Krasonikolakis. 2024. Historical portrayal of Greek tourism through topic modeling on international newspapers. In *Proceedings of the 8th joint sighum workshop on computational linguistics for cultural heritage, social sciences, humanities and literature (latech-clfl 2024),* edited by Yuri Bizzoni, Stefania Degaetano-Ortlieb, Anna Kazantseva, and Stan Szpakowicz, 121–132. St. Julians, Malta: Association for Computational Linguistics, March. https://aclanthology.org/2024.latechclfl-1.13.

Kroll, Hermann, Niklas Mainzer, and Wolf-Tilo Balke. 2022. On dimensions of ;plausibility for ;narrative information access to ;digital libraries. In *Linking theory and practice of digital libraries: 26th international conference on theory and practice of digital libraries, tpdl 2022, padua, italy, september 20–23, 2022, proceedings,* 433–441. Padua, Italy: Springer-Verlag. ISBN: 978-3-031-16801-7. https://doi.org/10.1007/978-3-031-16802-4_43. https://doi.org/10.1007/978-3-031-16802-4_43.

Lamsiyah, Salima, Keerthana Murugaraj, and Christoph Schommer. 2023. Historical-domain pre-trained language model for historical extractive text summarization. August. https://doi.org/10.11159/cist23.152.

Lansdall-Welfare, Thomas, Saatviga Sudhahar, James Thompson, Justin Lewis, FindMyPast Newspaper Team, Nello Cristianini, Amy Gregor, et al. 2017. Content analysis of 150 years of British periodicals. *Proceedings of the National Academy of Sciences* 114 (4): E457–E465. https://doi.org/10.1073/pnas.1606380114.

Lee, Daniel D., and H. Sebastian Seung. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature* 401 (6755): 788–791. https://doi.org/10.1038/44565.

Lewis, Patrick, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, et al. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems.* NIPS '20. Curran Associates Inc. ISBN: 9781713829546.

Lin, Chin-Yew. 2004. ROUGE: a package for automatic evaluation of summaries. In *Text Summarization Branches Out,* 74–81. Association for Computational Linguistics. https://aclanthology.org/W04-1013/.

Lin, King Ip, and Sabrina Peng. 2022. Enhancing digital history – event discovery via topic modeling and change detection. In *Proceedings of the 2nd international workshop on natural language processing for digital humanities,* edited by Mika Hämäläinen, Khalid Alnajjar, Niko Partanen, and Jack Rueter, 69–78. Taipei, Taiwan: Association for Computational Linguistics, November. https://aclanthology.org/2022.nlp4dh-1.10.

Malhas, Rana, Watheq Mansour, and Tamer Elsayed. 2023. Qur'an QA 2023 shared task: overview of passage retrieval and reading comprehension tasks over the holy qur'an. In *Proceedings of arabicnlp 2023,* edited by Hassan Sawaf, Samhaa El-Beltagy, Wajdi Zaghouani, Walid Magdy, Ahmed Abdelali, Nadi Tomeh, Ibrahim Abu Farha, et al., 690–701. Singapore (Hybrid): Association for Computational Linguistics, December. https://doi.org/10.18653/v1/2023.arabicnlp-1.76. https://aclanthology.org/2023.arabicnlp-1.76/.

Maltseva, Anna, Natalia Shilkina, Evgeniy Evseev, Mikhail Matveev, and Olesia Makhnytkina. 2021. Topic modeling of russian-language texts using the parts-of-speech composition of topics (on the example of volunteer movement semantics in social media). In *2021 29th conference of open innovations association (fruct),* 247–253. https://doi.org/10.23919/FRUCT52173.2021.9435475.

Marjanen, Jani, Elaine Zosa, Simon Hengchen, Lidia Pivovarova, and Mikko Tolonen. 2021. Topic Modelling Discourse Dynamics in Historical Newspapers. *Digital Humanities in the Nordic and Baltic Countries Publications* 3:63–77. https://doi.org/10.5617/dhnbpub.11235.

McInnes, Leland, John Healy, and S. Astels. 2017. hdbscan: Hierarchical density-based clustering. *Journal of Open Source Software.* 2:205. https://api.semanticscholar.org/CorpusID:53231359.

McInnes, Leland, John Healy, and James Melville. 2020. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.* arXiv: 1802.03426 [stat.ML]. https://arxiv.org/abs/1802.03426.

Mendonça, Margarida, and Álvaro Figueira. 2024. Topic Extraction: BERTopic's Insight into the 117th Congress's Twitterverse. *Informatics* 11 (1). https://doi.org/10.3390/informatics11010008.

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, et al. 2011. Quantitative analysis of culture using millions of digitized books. *Science* 331 (6014): 176–182. https://doi.org/10.1126/science.1199644.

Mikolov, Tomas, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *International Conference on Learning Representations.* https://api.semanticscholar.org/CorpusID:5959482.

Murugaraj, Keerthana, Salima Lamsiyah, Marten During, and Martin Theobald. 2025. Mining the past: a comparative study of classical and neural topic models on historical newspaper archives. In *Proceedings of the 5th international conference on natural language processing for digital humanities,* edited by Mika Hämäläinen, Emily Öhman, Yuri Bizzoni, So Miyagawa, and Khalid Alnajjar, 452–463. Albuquerque, USA: Association for Computational Linguistics, May. ISBN: 979-8-89176-234-3. https://doi.org/10.18653/v1/2025.nlp4dh-1.39. https://aclanthology.org/2025.nlp4dh-1.39/.

Murugaraj, Keerthana, Salima Lamsiyah, and Christoph Schommer. 2025. Abstractive summarization of historical documents: a new dataset and novel method using a domain-specific pretrained model. *IEEE Access* 13:10918–10932. https://doi.org/10.1109/ACCESS.2025.3528733.

Oberbichler, Sarah, and Eva Pfanzelter. 2021. Topic-specific corpus building: a step towards a representative newspaper corpus on the topic of return migration using text mining methods. *Journal of Digital History* 1 (1).

Oiva, Mila. 2020. Topic modeling russian history. *The Palgrave Handbook of Digital Russia Studies,* https://api.semanticscholar.org/CorpusID:234266287.

Orr, Martin, Kirsten Van Kessel, and David Parry. 2024. Ethical thematic and topic modelling analysis of sleep concerns in a social media derived suicidality dataset. In *Proceedings of the 9th workshop on computational linguistics and clinical psychology (clpsych 2024),* edited by Andrew Yates, Bart Desmet, Emily Prud'hommeaux, Ayah Zirikly, Steven Bedrick, Sean MacAvaney, Kfir Bar, Molly Ireland, and Yaakov Ophir, 74–91. St. Julians, Malta: Association for Computational Linguistics, March. https://aclanthology.org/2024.clpsych-1.6/.

Ouyang, Long, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, et al. 2022. Training language models to follow instructions with human feedback. In *Proceedings of the 36th international conference on neural information processing systems.* NIPS '22. New Orleans, LA, USA: Curran Associates Inc. ISBN: 9781713871088.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics,* 311–318. Association for Computational Linguistics. https://doi.org/10.3115/1073083.1073135.

Rajwal, Swati, Avinash Kumar Pandey, Zhishuo Han, and Abeed Sarker. 2024. Unveiling voices: identification of concerns in a social media breast cancer cohort via natural language processing. In *Proceedings of the first workshop on patient-oriented language processing (cl4health) @ lrec-coling 2024,* edited by Dina Demner-Fushman, Sophia Ananiadou, Paul Thompson, and Brian Ondov, 264–270. Torino, Italia: ELRA / ICCL, May. https://aclanthology.org/2024.cl4health-1.32/.

Reimers, Nils, and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP),* 3982–3992. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1410.

Sá, Gildácio, and José Maia. 2021. Retrieving and processing images from the pages of a historical newspaper and modeling the text topics. *Journal of Digital Information Management* 19:41. https://doi.org/10.6025/jdim/2021/19/2/41-46.

Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: open and efficient foundation language models. *arXiv preprint arXiv:2302.13971.*

Uban, Ana Sabina, Cornelia Caragea, and Liviu P. Dinu. 2021. Studying the evolution of scientific topics and their relationships. In *Findings of the association for computational linguistics: acl-ijcnlp 2021,* edited by Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, 1908–1922. Online: Association for Computational Linguistics, August. https://doi.org/10.18653/v1/2021.findings-acl.167. https://aclanthology.org/2021.findings-acl.167/.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017.* 5998–6008. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Wang, Yue, Hung Le, Akhilesh Deepak Gotmare, Nghi DQ Bui, Junnan Li, and Steven CH Hoi. 2023. Codet5+: open code large language models for code understanding and generation. *arXiv preprint arXiv:2305.07922.*

Wu, Zongda, Ling Lei, Guiling Li, Hui Huang, Chengren Zheng, Enhong Chen, and Guandong Xu. 2017. A topic modeling based approach to novel document automatic summarization. *Expert Systems with Applications* 84:12–23. https://api.semanticscholar.org/CorpusID:13752634.

Yao, Shunyu, Qingqing Ke, Qiwei Wang, Kangtong Li, and Jie Hu. 2024. Lawyer gpt: a legal large language model with enhanced domain knowledge and reasoning capabilities. In *Proceedings of the 2024 3rd international symposium on robotics, artificial intelligence and information engineering,* 108–112. RAIIE '24. Singapore, Singapore: Association for Computing Machinery. ISBN: 9798400718311. https://doi.org/10.1145/3689299.3689319. https://doi.org/10.1145/3689299.3689319.

Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. *Bertscore: evaluating text generation with bert.* arXiv: 1904.09675 [cs.CL]. https://arxiv.org/abs/1904.09675.

Zhong, Ming, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing,* 2023–2038. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.131.

Zundert, Joris J. van, Marijn Koolen, Julia Neugarten, Peter Boot, Willem Robert van Hage, and Ole Mussmann. 2022. What do we talk about when we talk about topic? In *Workshop on computational humanities research.* https://api.semanticscholar.org/CorpusID:254045843.

https://doi.org/10.1017/chr.2025.10018 Published online by Cambridge University Press

Table 1. Qualitative assessment of Topic-RAG's responses to complex historical queries. "Task" outlines the historian's informational intent, "Query Type" categorizes the nature of the question, and "Comments" provide expert feedback on the relevance and quality of the generated responses.

No	Query	Task	Query Type	Comment
1	Which arguments did proponents and opponents of the 1979 Volksbegehren put forward?	Analyze different viewpoints and capture opinion-based arguments from various stakeholders	Analytical, Opinion-based	The query focuses on the national public vote on 18th February 1979, which would have effectively made it very hard to build nuclear power plants in Switzerland[a]. The system, however, chose topic 0 on the Kaiseraugst nuclear plant project, which was historically important in the context of the question. Other topics (11, 15, 17) surrounding public policy, votes, and mobilization would have likely been more appropriate here. Nevertheless, the answer does adequately summarize positions for and against the Kaiseraugst nuclear plant.
2	Which articles take an explicit position in favor of or against the Volksbegehren?	Identify explicit statements.	Fact-based, Opinion-based	Again, the system chose topic 0 to answer a question on the national vote. The query asks for specific articles to be identified, instead the output gives an adequate summary of the positions of the main stakeholders.
3	Give me a name list of all scientific experts mentioned in the articles in the context of the debate about the Kaiseraugst nuclear plant, and in how many articles they are mentioned. Then, for each expert quoted, give a summary of the statement and classify the theme the person spoke about.	Combine fact collection with analysis of mention frequency, summarization, and classification.	Fact-based, Analytical	This complex query asks about the representation of scientific experts in the debate surrounding Kaiseraugst. This time, the chosen topic 0 is appropriate. The system successfully identified five scientific experts and described their roles adequately. The output continues with a general classification of their views and a verbose reflection on their statements, which is, however, too generic to be of use.
4	Give an overview of the year in which environmental topics were first discussed in the different newspapers. As output, produce a table with the following columns: newspaper title, date of publication, article title, and 10-word article summary.	Summarize, track topic emergence over time	Fact-based, Analytical, Causal	The system correctly selected all topics surrounding environmental issues, especially those concerning nuclear waste. The output includes—as expected—article titles with date and a short summary. Results are grouped by topic, not by newspaper title, as expected from the query. With 9,000 words, the output is surprisingly long and hard to process.
5	How did articles create a sense of fear in their readers? Use quotes from the articles to exemplify your answer.	Interpret emotional impact and rhetorical techniques	Inference-based, Analytical, Opinion-based	The system successfully identified two relevant quotes in the articles, but then goes on to point to very generic statements about nuclear power and fear. Again, topic 0 has been identified, but articles about the Gösgen nuclear power plant are referenced, which suggests a broader scope after all.
6	Which articles strongly advocate nuclear power? What are the arguments they put forward, and which articles strongly oppose nuclear power? What are the arguments they put forward? Do arguments change over time	Analyze positions and their evolution over time	Analytical, Opinion-based, Causal	The query asks for individual articles, but the system outputs a summarization of the positions of individual actors in the debate.
7	Give me 10 examples of articles that either subtly argue in favor of or against nuclear power but appear to be neutral at first sight.	Identify subtle bias and retrieve relevant articles.	Inference-based, Analytical, Opinion-based	The query explicitly asks for 10 articles, but the output is again a generic summarization of positions, limited to topic 0.
8	Which experts are linked to the nuclear waste topic?	Identify entities with a specific profile	Analytical, Fact-based	The output is as expected in the form of a list of experts together with a brief summary of their profession and positions in the debate. It is not clear how comprehensive and relevant the list of seven experts is. The query specifies the topic of nuclear waste, but only topic 5 was selected, while topics 6, 8, 9, 10, 12, 13, 14, 16, 18, and 19 should also have been relevant in this context.
9	Describe the ratio of articles concerned with Swiss national politics compared to international politics.	Analyze and compare content distribution	Analytical	The topic selection is wrong, but the output contains an adequate summary of the articles linked to the topic.

[a] https://www.bk.admin.ch/ch/d/pore/va/19790218/det296.html