# Towards a RAG, GraphRAG and MCP Implementation Over the EHRI-KG to Improve LLM Responses on Holocaust Archival Data

Herminio García-González[1,*], Xiao Xu[2] and Mike Bryant[2]

[1]Kazerne Dossin, Goswin de Stassartstraat 153, 2800 Mechelen, Belgium
[2]NIOD Institute for War, Holocaust and Genocide Studies, Herengracht 380, 1016 CJ Amsterdam, The Netherlands

**Abstract**

WWII led to the fragmentation of primary materials about the Holocaust across different nations, thereby complicating historical reconstruction. To help researchers navigate this patchwork, EHRI offers a trans-national and centralised catalogue of archival finding aids, recently made available as a knowledge graph. At the same time, LLMs offer new methods for research amid promises of improving researchers' productivity despite their current limitations and the ethical questioning around their usage, particularly within the humanities. In this work we explore the use of three context-retrieval techniques (RAG, GraphRAG and MCP) over the EHRI Knowledge Graph, which can offer curated and grounded information about the Holocaust to the LLM of choice; we evaluate them as part of an LLM-as-a-judge methodology and deliver a conversational web interface implementing them. Our early results show that RAG performs best for our domain, GraphRAG requires a more complex implementation to reach its full potential, and MCP, while promising, is overshadowed by the non-determinism involving its invocation. We frame this study within a wider pathway toward determining how AI tooling can be integrated into existing solutions and serve the purposes of researchers on the Holocaust in a trustworthy, reliable and ethical manner.

**Keywords**

Holocaust, LLMs, EHRI, RAG, GraphRAG, MCP

## 1. Introduction

The European Holocaust Research Infrastructure (EHRI) developed its Portal website in an attempt to make identifying and locating archival material relevant to the study of the Holocaust easier for researchers in the field. The pursuit of truly trans-national Holocaust research is a complex endeavour, in no small part due to the fragmentation of primary source material that resulted from destruction of evidence, mass migration, and political upheaval during and after WWII [1, 2]. The EHRI Portal [3] attempts to provide a single entry point through which this landscape can begin to be comprehended, linking together information about collections held by institutions large and small on a global basis. More recently, the EHRI Knowledge Graph (EHRI-KG) appeared as a semantic-enabled replica of the EHRI Portal built on top of Semantic Web technologies and leveraging the recently-released Records in Contexts Ontology (RiC-O)[1] [4, 5].

The surge of AI, and more specifically Large Language Models (LLMs) [6], has provided researchers with new methods for retrieving, searching and analysing information, making them invaluable tools when used properly [7, 8, 9, 10]. Yet the longer-term effects on the productivity of researchers, *vis-à-vis*

*RAGE-KG 2026: The Third International Workshop on Retrieval-Augmented Generation Enabled by Knowledge Graphs, co-located with ISWC 2026, October 25–29, 2026, Bari, Italy*

*Corresponding author.
†These authors contributed equally.
herminio.garciagonzalez@kazernedossin.eu (H. García-González); x.xu@niod.knaw.nl (X. Xu); m.bryant@niod.knaw.nl (M. Bryant)
https://herminiogarcia.com/ (H. García-González); https://www.niod.nl/en/staff/xiaoxu (X. Xu); https://www.niod.nl/en/staff/mikebryant (M. Bryant)
0000-0001-5590-4857 (H. García-González); 0000-0003-2325-8306 (X. Xu); 0000-0003-0765-7390 (M. Bryant)



[1]https://www.ica.org/standards/RiC/ontology

archival sources, will likely take some time to understand. In the short term, however, one probable impact on research infrastructures like EHRI, which seek to make information more easily accessible, is that a certain portion of our user base will become increasingly accustomed to using conversational interfaces for information-gathering purposes. Nevertheless, LLMs suffer from different problems and limitations, ranging from the tendency to hallucinate [11] or confabulate [12] to the inevitable cut-off date of the data that is used to train them or the unintentional bias (and lack of curation) introduced in the training set – including a prevalence of sycophancy [13] in models fine-tuned using Reinforcement Learning from Human Feedback (RLHF) [14]. As such, these behaviours can result in reputational risk for the providers of services based on them, particularly when dealing with sensitive topics such as the Holocaust. All this has deepened the humanistic rejection of LLMs – and AI in general – and fuelled a call-out against their use in historical contexts due to the risks of distortion, denialism and bias [15].

This work presents the first steps of our research into making the data from the EHRI-KG – and by extension from the EHRI Portal – accessible via a conversational text interface and discusses three techniques, Retrieval-Augmented Generation (RAG) [16], GraphRAG [17] and the Model Context Protocol (MCP)[2], for grounding the output of LLMs using verified and curated information from EHRI data providers, thus reducing their tendency to output inaccurate or unsourced information. Moreover, we outline this initiative in a wider human-centred approach to the adoption of LLMs in Holocaust-related settings where we seek to understand better how these tools can benefit practitioners in this field in a trustworthy, reliable and ethical manner.

The rest of this paper is structured as follows: Section 2 reviews the background and related work in the field, while Section 3 introduces the EHRI data model, the implemented architecture and how the various components of our chatbot prototype work together. Section 4 discusses our early findings and the methodology used to establish an early evaluation of the output of each retrieval method. Finally, Sections 5 and 6 respectively examine our plans for future work and provide a conclusion.

## 2. Background and Related Work

This section explores the background of this work, related work employing RAG, GraphRAG and MCP solutions in the cultural heritage field, and how LLMs in general are currently used for Holocaust research, highlighting the main associated problems and criticism in the field.

### 2.1. RAG, GraphRAG and MCP in Archives and Cultural Heritage

RAG was introduced by Lewis et al. [16] to combine a parametric language model with a non-parametric external store, retrieving relevant passages at inference time to ground generation and reduce hallucination; subsequent surveys chart its rapid evolution and recurring failure modes such as retrieval errors and chunking artefacts [18, 19]. In cultural heritage and historical fields, RAG has already been used to build question-answering chatbots; e.g., the Artistic Chatbot answered open questions at a live art exhibition from a curated document base [20], and similar pipelines have been built to query historical newspaper collections [21, 22]. However, such systems inherit the limits of vector-only retrieval [19]; answers are assembled from a fixed set of independently embedded vectors, which makes it hard to connect them to stable archival identifiers and tends to conflate sources based on literal semantic proximity instead of curated structures. Even with relevant context retrieved, LLMs could still introduce unsupported or contradictory statements [23], the effectiveness is sensitive to the relevance and arrangement of the retrieved passages [24], and their conflated answers may contribute to a biased historical perspective transformation [25].

The limits motivate researchers to build retrieval systems that are more explicitly aware of the relationships between records. Two approaches have been developed for mitigation. GraphRAG [26] operates over graph elements (entities, triples, paths or subgraphs) rather than text chunks, which require curated Knowledge Graphs (KGs). Alternatively, MCP standardises how LLMs invoke external

[2] https://modelcontextprotocol.io/specification/

tools; though not a dedicated retrieval method by itself, it can expose KG navigation as callable tools and delegate retrieval decisions to the model. In this paper, we situate RAG, GraphRAG and MCP within a single system over the EHRI-KG, allowing them to be compared as a progression from static retrieval towards dynamic, graph-based, and model-driven access to our archives.

### 2.2. LLMs in Holocaust Research

LLMs are increasingly applied in Holocaust research for both retrieval and user interaction. On the information retrieval and extraction side, LLMs have been applied in domain-specific named-entity recognition and linking, making archives and documents more discoverable and interconnected. EHRI released a multilingual dataset and models for detecting entities in testimonies and archival descriptions and linking them to controlled vocabularies [27], and compared LLM-driven and classical machine learning approaches for automated subject indexing of archival descriptions [28]. Beyond EHRI, LLMs have been used for named entity and relation recognition from Holocaust testimonies [29, 30], and RAG has been applied to the analysis of children's diaries [31]. On the interactive side, systems that let users communicate with survivor records conversationally have been built, e.g., the time-offset interaction with Holocaust survivors [32] and proposals for generative "digital duplicates" in Holocaust testimony [33].

These applications also bring risks and controversies. The interactive applications are highly contested and have drawn sustained ethical critique [34, 35]. The risks on the retrieval side include fabrication, distortion [36, 37, 38], and the staleness of models unaware of newly surfaced or undigitised holdings [39]. Together, these concerns have led to broader scepticism and rejection of LLMs in the historical context [15]. Grounding the retrieval process in domain-specific data also risks becoming an attack surface for poisoned passages [40] – or even more intentional and sophisticated adversarial attacks [41, 42], which call for highly curated sources, additional safeguards, or even a different training and evaluation model for AI [43].

RAG, GraphRAG and MCP are all methods developed to mitigate risks on the retrieval side. However, the benefits of such grounding are bounded as added context can stop helping past a certain point [44]. Since RAG, GraphRAG and MCP leave the model with progressively more discretion over the retrieval process, they also constrain its output progressively less. Therefore, our approach presented in this paper seeks to reduce the retrieval-side risks by grounding responses from a curated KG of EHRI's data, while remaining explicit that the interaction layer and the model-driven effects of the three methods still retain the concerns mentioned above.

## 3. Architecture & User Interface

In this section we introduce the EHRI Portal and EHRI-KG data models, to then describe the architecture used to implement the three retrieval techniques and the rationale for each approach. Afterwards, a web-based interface is presented, which will serve as the basis for future experiments with intended users.

### 3.1. Data Model

The EHRI Portal consists of three main entities: countries or country reports, which serve as a top-level entry point and provide an overview of a specific nationally-centred research situation; institutions within each country as custodians of archival material; and archival descriptions created by institutions' staff as inventories of their holdings. The EHRI Portal also provides: a set of vocabularies representing Holocaust-relevant subject headings, camps and ghettos respectively, acting as thematic links among archival descriptions; a set of corporate bodies' and persons' data, which can be linked as subjects or creators of a certain archival description; and copy/original links which indicate if a certain holding has been copied or is a copy of another one. The EHRI-KG[3], for its part, follows the same structure but

[3] https://lod.ehri-project-test.eu/

is aligned to RiC-O and our own EHRI Ontology[4]. A SPARQL endpoint[5] for the EHRI-KG is publicly available and is the basis for the techniques explained in the following sections.

### 3.2. RAG

The first step before being able to retrieve relevant data based on a user prompt is to index the data in a vector space, so that the application is able to do a semantic search based on a similarity metric (e.g., cosine similarity or Euclidean distance). For this step we employ two different techniques depending on the entity type. For countries, in which the content is less structured and the bulk of the information is contained in lengthier text passages, we leverage a chunking strategy by sentence including an overlap with the preceding and succeeding ones. For institutions and archival descriptions, which contain much more concise information and cannot be treated as a single coherent document, a set of key-value pairs is extracted for each entity. In each case, the data is retrieved from the SPARQL endpoint using a set of pre-defined queries, transformed following the described method and converted to its counterpart embeddings using the all-MiniLM-L6-v2 model under sentence-transformers [45]. All the embeddings are then persisted into a FAISS [46] index file which uses the default method for similarity metrics (i.e., squared Euclidean distance) and a retrieval JSON file is generated to restore the chunks for a specific index.

On retrieval the user prompt is embedded using the same model and the top 6 results are returned based on the equivalences file. All the retrieval happens locally and the timeliness of the data is strictly linked to the date on which the index was generated.

### 3.3. GraphRAG

In contrast to RAG, the retrieval in GraphRAG is complemented by navigating over neighbouring nodes in a KG. In our implementation, we have followed a similar procedure to the RAG one for the embedding phase, though in this case we have limited the extraction to a handful of attributes for each entity pertaining to its name, type, URI and description, which are then stored in a key-value pair set per entity. Moreover, given that our KG does not contain a large number of distinct classes, we created a separate index for each (i.e., countries, institutions and archival descriptions). As in the RAG method, the indexes are persisted in a FAISS vector store and a conversion file is also created for later retrieval.

The first part of the retrieval is identical to the RAG one, but in this case a heuristic method of retrieval is used in which the top 2 countries, the top 6 institutions and the top 10 archival descriptions are recovered. From this point, an exploration of the graph based on a set of pre-defined SPARQL queries is launched per type:

- Country: The attributes for the given country are queried against the KG as well as the list of the first 10 institutions connected to that country (including a name and a link to expand the information about them).
- Institution: The most relevant attributes are recovered from the KG alongside the 10 most prominent archival descriptions (once more pulling only the title and a link to expand information).
- Archival description: The most relevant information is retrieved from the KG and the first 10 archival descriptions, which are noted as copies or original sources, are pulled (only including the title and a link to expand information).

In contrast to the RAG approach, while the set of persisted entities refers to that of the date in which the FAISS index was created, the retrieval is able to offer more up-to-date information about the relevant entities and provides a richer contextual window at the cost, however, of a higher consumption of tokens.

[4] http://lod.ehri-project-test.eu/ontology
[5] https://lod.ehri-project-test.eu/query/

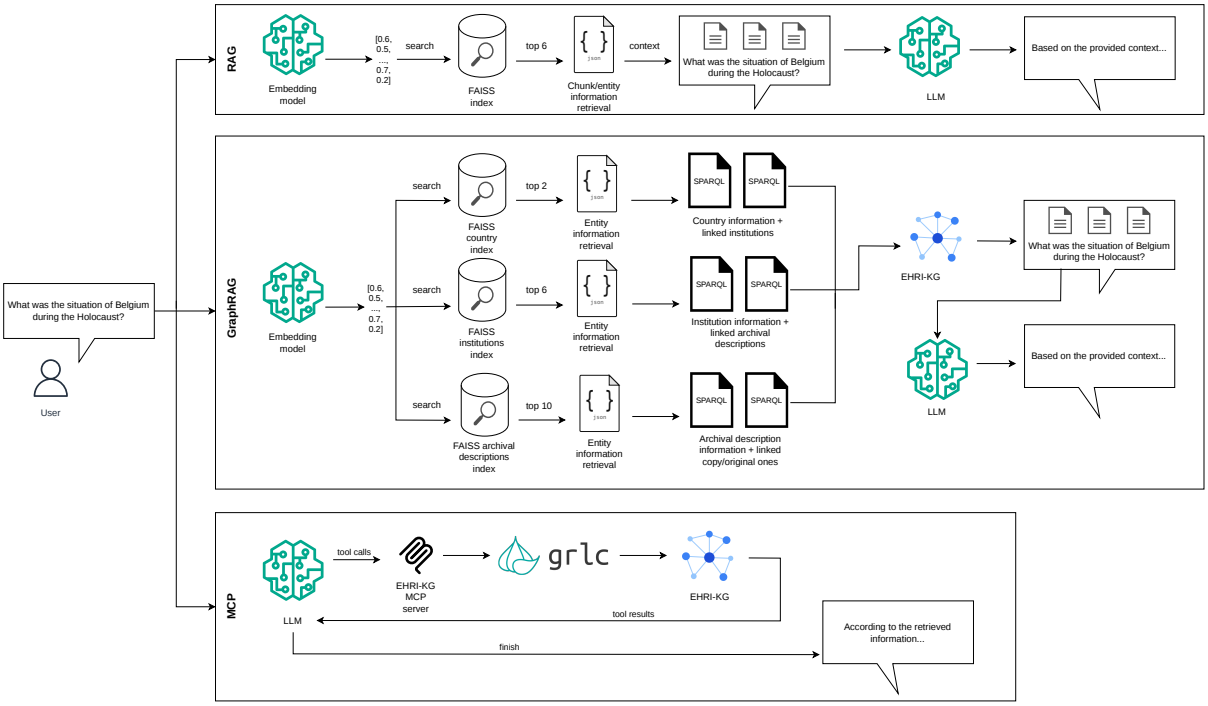



**Figure 1:** Diagram representing how each of the retrieval methods works upon a user request.

## 3.4. MCP

MCP was introduced as an open standard method to allow LLMs to call external tools. As mentioned before, it has a more general purpose than RAG but can be leveraged for similar purposes [47, 48].

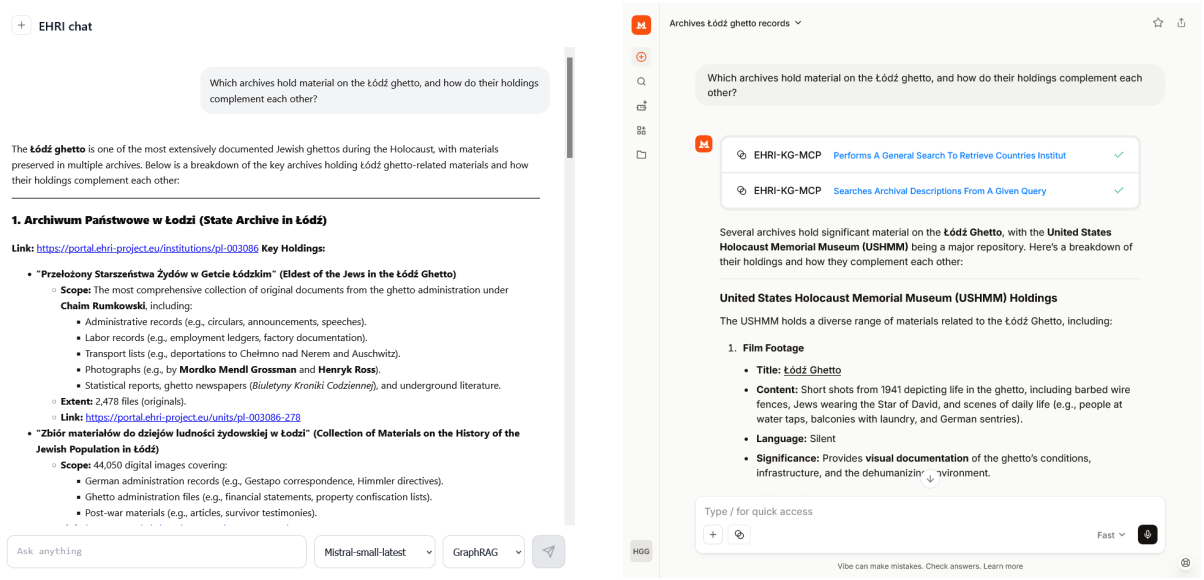
In order to build an MCP server (using the Streamable HTTP transport layer) we designed a set of SPARQL queries[6] that cover the same entities and navigational pathways as those in the GraphRAG implementation. We then made use of grlc[7] [49] to build an OpenAPI-compliant REST API which was later converted into an MCP server using FastMCP[8]. One caveat is that queries must be designed to take into account that LLMs are not good at traversing different methods nor using identifiers for obtaining the results, and they expect endpoints to follow a more natural-language-based retrieval approach with fewer operations.[9] In practice, we created two sets of calls, one with identifiers (as a REST API) and the other using free text as inputs (and tagged as "mcp"). When deploying with FastMCP only those tagged as "mcp" are incorporated in the MCP server[10]. Once set up, the MCP server can be registered in any compatible AI provider tools or used in combination with any compatible tool-supporting model, for which in many cases the development of an agentic loop is required.

In contrast to the other two methods, the decision about what methods to call and how to retrieve the contents is entirely delegated by the developer to the LLM. This ensures, however, that the data returned is completely up-to-date (due to the avoidance of the static embedding phase) and its inclusion in more general AI tooling becomes more straightforward.

## 3.5. User Interface

In order to facilitate the use of the different retrieval methods and set the stage for a future user study, we have developed a web interface that allows the user to choose between different models (at the time of writing Mistral-small-latest and Gemini-Flash-2.5) and retrieval methods (i.e., vanilla, to use the model

[6] https://github.com/EHRI/ehri-kg-sparql-queries-grlc
[7] https://grlc.io/
[8] https://gofastmcp.com/integrations/openapi
[9] https://www.philschmid.de/mcp-best-practices
[10] https://lod.ehri-project-test.eu/mcp




(a) EHRI chat interface using the GraphRAG method (b) Mistral Vibe using the EHRI-KG MCP server

**Figure 2:** Screenshots showing: (a) the EHRI chat interface answering a user prompt using the GraphRAG retrieval method, and (b) the MCP server being called from other AI tools (in this example, Mistral Vibe) to retrieve data from the EHRI-KG.

without any retrieval; RAG; GraphRAG and MCP). The web interface is based on nanoChat[11], adapted to our specific use case and needs, and as such allows using the full chat history for the generation (although only the latest prompt will be used for RAG and GraphRAG retrieval). It also allows the modification of some of the generation parameters (e.g., temperature and top-k). Figure 2a shows the chat interface while Figure 2b shows how the MCP server can be leveraged from AI providers' existing tools.

## 4. Early Findings & Discussion

With the aim of testing the overall architecture, the different context-retrieval approaches, and their ability to answer users' queries based on the EHRI-KG data, we have conducted an evaluation following an LLM-as-a-judge methodology. Due to the absence of a ground truth at this stage, which we seek to establish in the future with the aid of domain experts through a user study, this methodology allows us to establish an initial assessment of the benefits that this solution may have and, at the same time, establishes a framework that can be used for continuous validation throughout further refinement and development of this tool. It also allows us to increase the number of questions evaluated in parallel with expanding the number of domain entities provided for. Recent literature suggests that despite inherent biases that LLMs may possess, scores attributed by them tend to align to a high degree with those of human annotators/reviewers [50], making this technique a more scalable and less resource-intensive alternative.

Thus, for our experimental set-up we have followed the three metrics proposed in the RAG Triad[12] and adopted by evaluation frameworks like RAGAs [50], ARES [51] and Ragchecker [52]: Faithfulness or Groundedness, Answer Relevance, and Context Relevance. Then, we asked a more powerful LLM (in our case, Mistral-large-latest) than those used for generation (i.e., Mistral-small-latest and Gemini-2.5-Flash) to rate the answers alongside the corresponding context for each of the three retrieval methods, using a scale going from 0 to 3 (0 denoting no relevance, not grounded, or irrelevant, and 3 high relevance,

[11] https://github.com/karpathy/nanochat
[12] https://www.trulens.org/getting_started/core_concepts/rag_triad/

| Entity type: Countries | |
|---|---|
| Question | Complexity analysis |
| *What was the situation of Belgium during the Holocaust?* | The information is contained in Belgium's entry, but additional archival descriptions or institutional information may add relevant data. |
| *What was the fate of Roma in France during the Holocaust?* | The question asks about the Roma ethnic group, which can create confusion for the LLM. France's country report contains a specific paragraph about this, notwithstanding further information on archival descriptions. |
| *Can you establish a relation between France and Italy in terms of collaboration with the Nazis?* | Both country reports need to be retrieved and compared, forcing the LLM to analyse the data in contrast to just serving the pulled context. |
| Entity type: Institutions | |
| Question | Complexity analysis |
| *List the most relevant institutions for the Holocaust in Belgium.* | The correct institutions need to be recovered and the relevancy assessed according to their descriptions. |
| *Can you locate museums for the Holocaust in the Netherlands?* | This query forces the method to filter the retrieved institutions to those that also operate or are mainly a museum. This may be subtle depending on the descriptions provided by the partner institutions. |
| *Locate camp sites that hold an archive within their premises in Germany these days.* | In comparison with the previous query, this question adds a double filter as candidate institutions need to be former camp sites and hold an archive. |
| Entity type: Archival descriptions | |
| Question | Complexity analysis |
| *I need to retrieve information about deportations of Jews in Antwerp.* | This query forces the tool to filter the descriptions based on a specific location and group. The latter can prove complex if the archival descriptions assume this group and do not specifically mention it. |
| *Can you locate archival material relating to raids in Amsterdam.* | Similar to the previous query, but for a different city and a more generally included term. Given the broad term, the LLM may be forced to select among numerous likely valid results. |
| *Do you have information on what countries issued visas for Jewish refugees during the War?* | The retrieval method should return any archival description touching (even slightly) the topic, while the LLM has to elaborate a list of countries based on the provided context and avoid inferring from training information. |
| Entity type: General information | |
| Question | Complexity analysis |
| *Who was Hermann Göring?* | Information about this person cannot be retrieved from the EHRI-KG, as the retrieval mechanisms for this entity are not implemented yet. However, the LLM should be able to answer this based on prior knowledge (given the general availability of information about him on the Internet) and a number of archival descriptions may contribute to shaping the response. |

**Table 1**
Set of questions used for the LLM-as-a-judge-based evaluation, grouped by the type of entity that should be retrieved for answering them and including a complexity analysis as a subjective indication of the effort required to retrieve the requested information.

high groundedness, or fully relevant).[13] Despite the possible misalignment of the metrics with the vanilla mode (i.e., that without any context retrieval), we decided to include it in our evaluation in order to have a sort of baseline, allowing us to confirm the improvement of the answers when additional context is provided over those based solely on LLMs' prior knowledge. We have also captured the usage of input and output tokens for each method and calculated an estimated running cost (based on the prices in force at the time of writing). We have included three questions per entity type (countries, institutions, and archival descriptions) and one about a person, whose entity type is not implemented in the retrieval process yet, but for whom relevant information should already be known by the LLMs or easily retrieved from related archival descriptions (see Table 1 for the full set of questions). For the generation process, we used, when available[14], the same parameters for both models: temperature set to 0.1, top-k to 20 and maximum tokens to generate to 4096. Table 2 shows the results obtained from this evaluation averaged by entity type and by the full set of questions.

[13]The implementation of this evaluation can be consulted on: https://github.com/herminiogg/ehri-chat/tree/main/evaluations
[14]top-k is not available for Mistral-small-latest

| Mode: Vanilla (baseline) | Mistral-small-latest | | | | | | Gemini-2.5-Flash | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Query group | CR | AR | G | IT | OT | Price | CR | AR | G | IT | OT | Price |
| Countries (n=3) | 1 | 3 | 1 | 320 | 987.7 | $0.00034 | 1.3 | 2.3 | 3 | 314 | 1492 | $0.00382 |
| Institutions (n=3) | 2.7 | 3 | 2.7 | 319.7 | 523 | $0.00019 | 2.3 | 2.3 | 3 | 313.7 | 318.7 | $0.00089 |
| Archival descriptions (n=3) | 1 | 2.3 | 1 | 320.7 | 409.3 | $0.00015 | 0.3 | 1.3 | 1 | 314.3 | 69 | $0.00027 |
| General information (n=1) | 3 | 3 | 3 | 313 | 643 | $0.00022 | 3 | 3 | 3 | 307 | 95 | $0.00033 |
| All (n=10) | 1.7 | 2.8 | 1.7 | 319.4 | 640.3 | $0.00022 | 1.5 | 2.1 | 2.4 | 313.3 | 573.4 | $0.00153 |

(a) Results of the evaluation for the vanilla mode against Mistral-small-latest and Gemini-2.5-Flash.

| Model: Mistral-small-latest | RAG | | | | | | GraphRAG | | | | | | MCP | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Query group | CR | AR | G | IT | OT | Price | CR | AR | G | IT | OT | Price | CR | AR | G | IT | OT | Price |
| Countries (n=3) | 2 | 3 | 3 | 1517 | 428.3 | $0.00028 | 1.7 | 2 | 2.7 | 10005 | 569 | $0.00117 | 1.7 | 3 | 2 | 11525 | 1403.7 | $0.00157 |
| Institutions (n=3) | 2.3 | 3 | 3 | 1092 | 552.3 | $0.00027 | 2 | 2.3 | 2.3 | 8602 | 697.7 | $0.00107 | 1.7 | 2.3 | 2.7 | 40722.3 | 1303.3 | $0.00446 |
| Archival descriptions (n=3) | 2 | 3 | 3 | 1131 | 516.7 | $0.00027 | 2 | 2.3 | 2.7 | 7350 | 845.7 | $0.00099 | 2.3 | 2.7 | 3 | 9739.7 | 483.7 | $0.00112 |
| General information (n=1) | 3 | 3 | 3 | 2323 | 494 | $0.00038 | 3 | 3 | 3 | 4577 | 719 | $0.00067 | 3 | 3 | 3 | 10402 | 1590 | $0.00152 |
| All (n=10) | 2.2 | 3 | 3 | 1354.3 | 498.6 | $0.00029 | 2 | 2.3 | 2.6 | 8244.8 | 705.6 | $0.00104 | 2 | 2.7 | 2.6 | 19636.3 | 1116.2 | $0.00230 |

(b) Results of the evaluation for the three context-retrieval techniques against Mistral-small-latest.

| Model: Gemini-2.5-Flash | RAG | | | | | | GraphRAG | | | | | | MCP | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Query group | CR | AR | G | IT | OT | Price | CR | AR | G | IT | OT | Price | CR | AR | G | IT | OT | Price |
| Countries (n=3) | 2 | 2.7 | 3 | 1519 | 196.3 | $0.00095 | 1.3 | 1.3 | 2 | 9899.7 | 465.7 | $0.00413 | 1.3 | 2 | 3 | 17557.3 | 259 | $0.00591 |
| Institutions (n=3) | 2.3 | 2.3 | 3 | 1074 | 122.7 | $0.00063 | 2 | 2.7 | 3 | 8508.3 | 417.7 | $0.00360 | 1.3 | 1.3 | 3 | 21321.7 | 160.7 | $0.00680 |
| Archival descriptions (n=3) | 2.3 | 2.3 | 3 | 1137.3 | 138.3 | $0.00069 | 2 | 2 | 3 | 7364.7 | 336.3 | $0.00305 | 2 | 1.7 | 2.7 | 17443.3 | 252.3 | $0.00586 |
| General information (n=1) | 3 | 3 | 3 | 2331 | 732 | $0.00253 | 3 | 3 | 3 | 4517 | 1354 | $0.00474 | 3 | 2 | 3 | 17750 | 628 | $0.00690 |
| All (n=10) | 2.3 | 2.5 | 3 | 1352.2 | 210.4 | $0.00093 | 1.9 | 2.1 | 2.7 | 8183.5 | 501.3 | $0.00371 | 1.7 | 1.7 | 2.9 | 18671.7 | 264.4 | $0.00626 |

(c) Results of the evaluation for the three context-retrieval techniques against Gemini-2.5-Flash.

**Table 2**
Results for the evaluation using an LLM as a judge alongside the consumption of tokens and associated cost calculations. (a) shows the results for the vanilla mode against both tested models, while (b) and (c) collect the results for the evaluation of the context-retrieval methods against Mistral-small-latest and Gemini-2.5-Flash respectively. Both models' outputs, context retrieval for all modes, as well as the input queries were scored by Mistral-large-latest using the same prompts. CR stands for Context Relevance, AR for Answer Relevance, G for Groundedness, IT for Input Tokens, and OT for Output Tokens. Metrics in red represent the best value among the evaluations of the retrieval techniques – i.e., results under (b) and (c), while those in blue denote the best value within a model group for the same techniques. The vanilla mode is provided as a baseline and as an orientation of the degree of improvement achieved by each context-retrieval method.

Based on the results obtained, we can see a clear tendency towards better scores when using RAG, which is also the least costly retrieval method for each of the models tested. When it comes to GraphRAG, the current algorithm provides extensive contextual information, but the usefulness of it seems to vary considerably depending on the specific question, an issue perhaps caused by the observed tendency of LLMs to overly rely on information located at the beginning and end of the input context [53] and/or the inclusion of additional noise by this specific method. This suggests that our initial implementation of the GraphRAG retrieval algorithm, while not totally incorrect, needs further fine-tuning to better select the relevant entities and relations, and to reduce the size of the retrieved context. MCP, for its part, surprisingly shows a very degraded performance when retrieving the appropriate context. This is something that we detected in our initial tests as many LLMs show a "lazy" behaviour with regard to the invocation of MCP calls – and tools in general. In order to mitigate this, we added the sentence "Use the EHRI-KG MCP server." at the end of the user prompt when evaluating the MCP method, yet in some cases the LLM disregarded this instruction completely, as well as the system prompt. Interestingly, while the MCP protocol seems like the most native manner to allow LLMs to retrieve additional context – and interoperate with other systems, their non-deterministic nature poses some additional challenges when used as part of a RAG architecture. Nevertheless, in combination with more advanced models –

or even so-called frontier ones – with their enhanced reasoning capabilities, these flaws may become less acute due to these models' ability to iterate over different methods, endpoints and results; and the answers as a consequence may be improved. In spite of this, MCP is the method that shows the highest consumption of tokens and, therefore, the highest price per request.

When comparing between models, we observe that Mistral-small-latest tends to be much more verbose and tries to elaborate an answer more eagerly than Gemini-2.5-Flash, at the cost, however, of losing a bit of groundedness for some of the methods and questions. In contrast, Gemini-2.5-Flash takes a more succinct approach, making its answers more grounded but less relevant. As discussed above, the configuration options for both models were the same, showing that a combined evaluation (between retrieval methods, models, and data) is necessary to determine the best configuration for a specific domain. Apart from this, the biggest differences between models arise when using the MCP server, where the less avid behaviour of Gemini-2.5-Flash prevents it from calling the MCP functions more often, and hence from potentially retrieving answers to the questions posed by the user.

When exposed to a query for which LLMs' internal knowledge may be leveraged, the additional context has a stronger influence on the three evaluation metrics in comparison to the other questions under the same context-retrieval method, suggesting that LLMs may be able to leverage this additional information better when they have prior knowledge about the subject entity. This is in line with studies showing that RAG systems used in combination with LLMs fine-tuned over domain-specific data may be able to generate better responses than just employing a base model [54, 55, 56].

Finally, comparing the context-retrieval methods' scores to those obtained by the same model in the vanilla mode (i.e., without any context injected in the prompt nor retrieved by any other means), we detect a mixed behaviour depending on the entity type. Overall, improvements are generally seen in the context relevance and groundedness of the answer. However, any increase in the latter comes at the cost of the answer relevance. When dealing with queries about countries, Mistral-small-latest improves its scores with the three retrieval methods, while Gemini-2.5-Flash only achieves a noticeable improvement with the RAG one. Notably, the improvement of the RAG metrics is also closely related to a decrease in the price per request, mainly due to a reduction in the normally more-costly output tokens. While we cannot assure the reasons for that, we hypothesise that this is caused by the increase in groundedness and context relevance, which probably prevents the LLM from making up an answer and eagerly completing it based on prior knowledge. Answers for the institutions' category do not show a clear improvement when additional context is injected and only RAG maintains a similar performance to the vanilla mode on the three metrics. This may be due to the fact that these institutions are quite well known to LLMs as many of the organisations' websites are probably crawled during their training phase, and the injected context might introduce additional noise without providing much new information. Moreover, we noticed that the LLM judge tended to score empty contexts as highly relevant whenever the answer was deemed fully grounded, making this metric less reliable in the absence of a context. On the other hand, answers for queries about archival descriptions, for which the LLM may have more limited prior knowledge, benefit from all the context-retrieval methods and increase their metrics with respect to those of the vanilla mode. As expected, the query about an entity already known to LLMs was already attributed with the highest scores for the three metrics in the vanilla mode. However, the three context-retrieval methods maintained these scores for both models (except for the answer relevance for Gemini-2.5-Flash over MCP), while augmenting the length of the generated answer (i.e., the number of output tokens), which suggests that the additional information was also deemed as highly relevant.

## 5. Future Work

As discussed at the beginning of this paper, we see the work described here as a first step towards conducting a more thorough evaluation that includes users and domain experts, allowing us to better define the scope of the tool, the users' expectations and potential pitfalls in use, and whether a similar solution can help them to be more productive. Most importantly, we intend to obtain a set of real-world questions and answers which can be incorporated into the evaluation process described in this work.

Additionally, we aim to complement the LLM-as-a-judge methodology with scores given by domain experts so as to: 1) determine whether the LLM-based evaluation is accurate enough with regard to the judgement of experts in the field, and 2) understand factors that the LLM needs to take into account for the evaluation of each query in order to also assess historical rigour.

Regarding the retrieval techniques, we envision their expansion to cover the full set of entities defined in our data model, that is, implementing additional retrieval options for vocabularies, persons, and corporate bodies. The GraphRAG method is the one that, according to our evaluation, has the biggest scope for improvement and, at the same time, provides a lot of flexibility in its implementation. Hence, we will seek to test different implementations and retrieval techniques to determine which one is most suited to our data. Hybrid techniques will also be studied, like those combining RAG and GraphRAG (also called HybridRAG) or even a slimmed (Graph)RAG phase with a subsequent MCP calling, limiting the number of options in the latter.

Similarly, and due to the promising results when retrieving content from an *a priori* known entity, we will investigate the possibility of fine-tuning a set of open-weight models with a curated selection of the EHRI-KG data, using techniques like LoRA [57] and QLoRA [58]. The aim is to investigate how the use of a fine-tuned LLM over historical data, in combination with the described context-retrieval techniques, may increase the accuracy and relevance of the responses to users' queries.

Finally, we will follow the advancements of alternative techniques closely, such as those developed under the Text2SPARQL umbrella, as they would represent a more direct approach to transforming user queries into actionable context for an LLM.

## 6. Conclusions

The use of LLMs in historical contexts – and in the Holocaust specifically – is faced with reluctance by the community of experts, amid fears of distortion, bias, or even new forms of historical revisionism. In this work, we have presented a context-retrieval implementation over the EHRI-KG, involving RAG, GraphRAG and MCP, which can ground users' queries with context derived from Holocaust structured data and reduce the tendency of LLMs to confabulate, editorialise, or otherwise stray from the task domain. The implementation is offered through a dedicated web interface in which the different modes can be tested and, in the case of the MCP server, integrated into general AI tooling.

Using the LLM-as-a-judge methodology we have sought to gauge which of the three retrieval methods provides the most grounded and relevant outputs to a selection of possible user queries. In this evaluation, the RAG method performs the best compared to the other two approaches and tends to outperform the vanilla mode (i.e., the generation without any additional context injected). GraphRAG, however, has substantial scope for improvement given the flexibility of the technique, and MCP usage varies across models, making the developer lose control over its behaviour and consequently the results become less deterministic.

As mentioned in Section 5, we will seek to follow this research up by consulting with prospective users, among them domain experts, to better understand the strengths and weaknesses of conversational approaches to curated dataset exploration in the field of Holocaust research, and ensure that ethical, trust and reliability concerns are fully taken into consideration.

On balance, this work represents the very first step toward implementing an LLM-based system over Holocaust archival data using different context-retrieval techniques for improved groundedness and accuracy. Despite its early state, the tool shows some promising results and paves the way for future initiatives at using LLMs – and other sorts of agentic workflows – in inherently sensitive fields like the Holocaust, where accuracy and explainability come before any other considerations and a failed adoption strategy will condemn even the most technologically-advanced tool to rejection.

## Supplemental Material Availability

The source code for the described architecture, implementation and evaluation is openly available on GitHub: https://github.com/herminiogg/ehri-chat and under the following permanent DOI: https://doi.org/10.5281/zenodo.21280663.


## Funding

This work has been carried out in the context of the OSCARS project, which has received funding from the European Commission's Horizon Europe Research and Innovation programme under grant agreement No. 101129751.


## Declaration on Generative AI

During the preparation of this work, the authors used Claude's Sonnet and Google's Gemini in order to: Grammar and spelling check, Paraphrase and reword, and Improve writing style. Further, the authors used Claude's Sonnet for Figure 1: Generate images (some of the icons). After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

## References


[1] R. Speck, T. Blanke, C. Kristel, M. Frankl, K. Rodriguez, V. V. Daelen, The past and the future of holocaust research: From disparate sources to an integrated european holocaust research infrastructure, in: Evolution der Informationsinfrastruktur: Forschung und Entwicklung als Kooperation von Bibliothek und Fachwissenschaft, University of Goettingen Press, 2014, pp. 157–177. doi:10.48550/arXiv.1405.2407.

[2] P. Links, R. Speck, V. Vanden Daelen, Who holds the key to holocaust-related sources? authorship as subjectivity in finding aids, Holocaust Studies 22 (2016) 21–43. doi:10.1080/17504902.2015.1118193.

[3] T. Blanke, M. Bryant, M. Frankl, C. Kristel, R. Speck, V. V. Daelen, R. V. Horik, The european holocaust research infrastructure portal, Journal on Computing and Cultural Heritage (JOCCH) 10 (2017) 1–18. doi:10.1145/3004457.

[4] H. García-González, M. Bryant, The holocaust archival material knowledge graph, in: T. R. Payne, V. Presutti, G. Qi, M. Poveda-Villalón, G. Stoilos, L. Hollink, Z. Kaoudi, G. Cheng, J. Li (Eds.), The Semantic Web - ISWC 2023 - 22nd International Semantic Web Conference, Athens, Greece, November 6-10, 2023, Proceedings, Part II, volume 14266 of *Lecture Notes in Computer Science*, Springer, 2023, pp. 362–379. URL: https://doi.org/10.1007/978-3-031-47243-5_20. doi:10.1007/978-3-031-47243-5_20.

[5] H. García-González, M. Bryant, Implementing the EHRI-KG: From Declarative Mapping Rules to a Three-Layered KG Construction Architecture, in: Seventh International Workshop on Knowledge Graph Construction@ ESWC2026, To appear, 2026. URL: https://openreview.net/attachment?id=sgxEw3aZmh&name=pdf.

[6] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, 2020. URL: https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

[7] M. Brachman, A. H. El-Ashry, C. Dugan, W. Geyer, How knowledge workers use and want to use llms in an enterprise context, in: F. F. Mueller, P. Kyburz, J. R. Williamson, C. Sas (Eds.), Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, CHI EA 2024, Honolulu, HI, USA, May 11-16, 2024, ACM, 2024, pp. 189:1–189:8. URL: https://doi.org/10.1145/3613905.3650841. doi:10.1145/3613905.3650841.

[8] E. Koutsiana, J. Walker, M. Nwachukwu, B. Zhang, A. Meroño-Peñuela, E. Simperl, Knowledge prompting: How knowledge engineers use generative AI, J. Web Semant. 88 (2026) 100873. doi:10.1016/J.WEBSEM.2025.100873.

[9] A. Simons, M. Zichert, A. Wüthrich, Large language models for history, philosophy, and sociology of science: Interpretive uses, methodological challenges, and critical perspectives, Studies in History and Philosophy of Science 117 (2026) 102151. URL: https://www.sciencedirect.com/science/article/pii/S0039368126000373. doi:10.1016/j.shpsa.2026.102151.

[10] M. Binz, S. Alaniz, A. Roskies, B. Aczel, C. T. Bergstrom, C. Allen, D. Schad, D. Wulff, J. D. West, Q. Zhang, et al., How should the advancement of large language models affect the practice of science?, Proceedings of the National Academy of Sciences 122 (2025) e2401227121. doi:10.1073/pnas.2401227121.

[11] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, T. Liu, A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions, ACM Trans. Inf. Syst. 43 (2025) 42:1–42:55. URL: https://doi.org/10.1145/3703155. doi:10.1145/3703155.

[12] A. L. Smith, F. Greaves, T. Panch, Hallucination or confabulation? neuroanatomy as metaphor in large language models, PLOS Digital Health 2 (2023) 1–3. doi:10.1371/journal.pdig.0000388.

[13] M. Sharma, M. Tong, T. Korbak, D. Duvenaud, A. Askell, S. R. Bowman, E. Durmus, Z. Hatfield-Dodds, S. R. Johnston, S. Kravec, T. Maxwell, S. McCandlish, K. Ndousse, O. Rausch, N. Schiefer, D. Yan, M. Zhang, E. Perez, Towards understanding sycophancy in language models, in: The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024, 2024. URL: https://proceedings.iclr.cc/paper_files/paper/2024/file/0105f7972202c1d4fb817da9f21a9663-Paper-Conference.pdf.

[14] P. F. Christiano, J. Leike, T. B. Brown, M. Martic, S. Legg, D. Amodei, Deep reinforcement learning from human preferences, in: I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, 2017, pp. 4299–4307. URL: https://proceedings.neurips.cc/paper/2017/hash/d5e2c0adad503c91f91df240d0cd4e49-Abstract.html.

[15] A. Dean, B. C. G. Lee, R. M. Ehrenreich, T. Presner, M. H. Goldman, J. Keck, V. G. Richardson-Walden, A. Potter, A. Lerner, A. Bonazzi, et al., Ai and the future of holocaust research & memory, 2026. URL: https://hdl.handle.net/1773/55320.

[16] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive NLP tasks, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, 2020. URL: https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

[17] D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, J. Larson, From local to global: A graph RAG approach to query-focused summarization, CoRR abs/2404.16130 (2024). doi:10.48550/ARXIV.2404.16130.

[18] D. Diefenbach, V. Lopez, K. Singh, P. Maret, Core techniques of question answering systems over knowledge bases: a survey, Knowledge and Information systems 55 (2018) 529–569. doi:10.1007/s10115-017-1100-y.

[19] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, H. Wang, Retrieval-augmented generation for large language models: A survey, CoRR (2023). doi:10.48550/arXiv.2312.

10997.

[20] F. J. Kucia, A. Wróblewska, B. Grabek, S. D. Trochimiak, How to make museums more interactive? case study of artistic chatbot, in: Proceedings of the 34th ACM International Conference on Information and Knowledge Management, 2025, pp. 6654–6658. doi:10.1145/3746252.3761485.

[21] T. T. Tran, C.-E. González-Gallardo, A. Doucet, Retrieval augmented generation for historical newspapers, in: Proceedings of the 24th ACM/IEEE Joint Conference on Digital Libraries, 2024, pp. 1–5. doi:10.1145/3677389.3702542.

[22] K. Murugaraj, S. Lamsiyah, M. During, M. Theobald, Topic-rag for historical newspapers: Enhancing information retrieval in humanities research through topic-based retrieval-augmented generation, Computational Humanities Research (2025) 1–21. doi:10.1017/chr.2025.10018.

[23] C. Niu, Y. Wu, J. Zhu, S. Xu, K. Shum, R. Zhong, J. Song, T. Zhang, Ragtruth: A hallucination corpus for developing trustworthy retrieval-augmented language models, in: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2024, pp. 10862–10878. doi:10.18653/V1/2024.ACL-LONG.585.

[24] F. Cuconasu, G. Trappolini, F. Siciliano, S. Filice, C. Campagnano, Y. Maarek, N. Tonellotto, F. Silvestri, The power of noise: Redefining retrieval for rag systems, in: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2024, pp. 719–729. doi:10.1145/3626772.3657834.

[25] R. Mahadeshwar, A. v. Cranenburgh, T. Caselli, M. Nissim, Evaluating the impact of source diversity for rag in historical research, in: S. Piperidis, N. Bel, H. van den Heuvel, N. Ide, S. Krek, A. Toral (Eds.), Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), European Language Resources Association (ELRA), Palma, Mallorca, Spain, 2026, pp. 716–734. doi:10.63317/4yz9z3uvzasd.

[26] B. Peng, Y. Zhu, Y. Liu, X. Bo, H. Shi, C. Hong, Y. Zhang, S. Tang, Graph retrieval-augmented generation: A survey, ACM Transactions on Information Systems 44 (2025) 1–52. doi:10.1145/3777378.

[27] M. Dermentzi, H. Scheithauer, Repurposing holocaust-related digital scholarly editions to develop multilingual domain-specific named entity recognition tools, in: Proceedings of the First Workshop on Holocaust Testimonies as Language Resources (HTRes)@ LREC-COLING 2024, 2024, pp. 18–28. URL: https://aclanthology.org/2024.htres-1.3/.

[28] M. Dermentzi, M. Bryant, F. Rovigo, H. García-González, Multilingual automated subject indexing: a comparative study of llms vs alternative approaches in the context of the ehri project, DH Benelux Journal Volume 7: Breaking Silos, Connecting Data: Advancing Integration and Collaboration in Digital Humanities (2025) 33–55. doi:10.5281/zenodo.17648380.

[29] I. Anuradha Nanomi Arachchige, L. Ha, R. Mitkov, J.-D. Steinert, Enhancing named entity recognition for holocaust testimonies through pseudo labelling and transformer-based models, in: Proceedings of the 7th International Workshop on Historical Document Imaging and Processing, 2023, pp. 85–90. doi:10.1145/3604951.3605514.

[30] I. Anuradha, R. Mitkov, V. Nahar, et al., Evaluating of large language models in relationship extraction from unstructured data: Empirical study from holocaust testimonies, in: Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing, 2023, pp. 117–123. URL: https://aclanthology.org/2023.ranlp-1.13.

[31] A. Parfenova, Unveiling hidden stories: Automated narrative extraction from holocaust diaries with ensemble llms, in: Proceedings of Text2Story—Eighth Workshop on Narrative Extraction From Texts, volume 3964 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2025, pp. 81–92. URL: https://ceur-ws.org/Vol-3964/paper7.pdf.

[32] R. Artstein, D. Traum, O. Alexander, A. Leuski, A. Jones, K. Georgila, P. Debevec, W. Swartout, H. Maio, S. Smith, Time-offset interaction with a holocaust survivor, in: Proceedings of the 19th international conference on Intelligent User Interfaces, 2014, pp. 163–168. doi:10.1145/2557500.2557540.

[33] A. Kozlovski, M. Makhortykh, Digital dybbuks and virtual golems: the ethics of digital duplicates in holocaust testimony, Memory, Mind & Media 4 (2025) e10. doi:10.1017/mem.2025.10006.

[34] C. K. N. Schultz, Creating the 'virtual' witness: the limits of empathy, Museum Management and Curatorship 38 (2023) 2–17. doi:10.1080/09647775.2021.1954980.
[35] T. Presner, Ethics of the algorithm: Digital humanities and Holocaust memory, Princeton University Press, 2024. doi:10.1515/9780691258980.
[36] M. Makhortykh, V. Vziatysheva, M. Sydorova, Generative ai and contestation and instrumentalization of memory about the holocaust in ukraine, Eastern European Holocaust Studies 1 (2023) 349–355. doi:10.1515/eehs-2023-0054.
[37] M. Makhortykh, H. Mann, et al., Ai and the holocaust: rewriting history? the impact of artificial intelligence on understanding the holocaust, 2024. doi:10.54675/ZHJC6844.
[38] F. Finn, D. Khosrowi, C. Mello, M. Düring, D. Guido, Shaping history, responsibly: Seven principles to guide the design of archival ai assistants for cultural heritage collections, in: Proceedings of the 2026 ACM Conference on Fairness, Accountability, and Transparency, FAccT '26, Association for Computing Machinery, New York, NY, USA, 2026, p. 8300–8321. doi:10.1145/3805689.3812238.
[39] G. J. Nauta, W. van den Heuvel, S. Teunisse, Survey report on digitisation in european cultural heritage institutions 2015, Europeana, 2015. URL: https://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/ENUMERATE/deliverables/ev3-deliverable-d1.2-europeana-version1.1-public.pdf.
[40] W. Zou, R. Geng, B. Wang, J. Jia, Poisonedrag: Knowledge corruption attacks to retrieval-augmented generation of large language models, in: 34th USENIX Security Symposium (USENIX Security 25), 2025, pp. 3827–3844. URL: https://www.usenix.org/conference/usenixsecurity25/presentation/zou-poisonedrag.
[41] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, D. Mukhopadhyay, A survey on adversarial attacks and defences, CAAI Trans. Intell. Technol. 6 (2021) 25–45. doi:10.1049/CIT2.12028.
[42] A. G. Chowdhury, M. M. Islam, V. Kumar, F. H. Shezan, V. Kumar, V. Jain, A. Chadha, Breaking down the defenses: A comparative survey of attacks on large language models, CoRR abs/2403.04786 (2024). doi:10.48550/arXiv.2403.04786.
[43] L. Havens, M. Terras, B. Bach, B. Alex, Recalibrating Artificial Intelligence for Social Biases: Four New Priorities for Datasets, Models, and Metrics, Association for Computing Machinery, New York, NY, USA, 2026, p. 2253–2277. doi:10.1145/3805689.3812283.
[44] K. Soemer, H. Mihaljević, You frame it: How conceptual representations shape llm detection and reasoning about antisemitism, CoRR (2026). doi:10.48550/arXiv.2607.04945.
[45] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019, Association for Computational Linguistics, 2019, pp. 3980–3990. doi:10.18653/V1/D19-1410.
[46] M. Douze, A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, P. Mazaré, M. Lomeli, L. Hosseini, H. Jégou, The faiss library, IEEE Trans. Big Data 12 (2026) 346–361. doi:10.1109/TBDATA.2025.3618474.
[47] A. Steiner, R. Peeters, C. Bizer, MCP vs RAG vs nlweb vs HTML: A comparison of the effectiveness and efficiency of different agent interfaces to the web, in: H. Hacid, Y. Maarek, F. Bonchi, I. Guy, E. Yilmaz (Eds.), Proceedings of the ACM Web Conference 2026, WWW 2026, Dubai, United Arab Emirates, originally scheduled for April 13-17, 2026, rescheduled for June 29 - July 3, 2026, ACM, 2026, pp. 8493–8496. doi:10.1145/3774904.3792893.
[48] X. Zhao, M. Yang, K. Tian, H. Jiang, D. Guo, Y. Wang, J. Du, Performance of latest ai models, rag, and mcp on lung cancer-related questions, DIGITAL HEALTH 12 (2026). doi:10.1177/20552076261427503.
[49] A. Meroño-Peñuela, R. Hoekstra, grlc Makes GitHub Taste Like Linked Data APIs, in: The Semantic Web: ESWC 2016 Satellite Events, Heraklion, Crete, Greece, May 29 – June 2, 2016, Springer, 2016, pp. 342–353. doi:10.1007/978-3-319-47602-5_48.
[50] S. Es, J. James, L. E. Anke, S. Schockaert, Ragas: Automated evaluation of retrieval augmented generation, in: N. Aletras, O. D. Clercq (Eds.), Proceedings of the 18th Conference of the European

Chapter of the Association for Computational Linguistics, EACL 2024 - System Demonstrations, St. Julians, Malta, March 17-22, 2024, Association for Computational Linguistics, 2024, pp. 150–158. doi:10.18653/V1/2024.EACL-DEMO.16.

[51] J. Saad-Falcon, O. Khattab, C. Potts, M. Zaharia, ARES: an automated evaluation framework for retrieval-augmented generation systems, in: K. Duh, H. Gómez-Adorno, S. Bethard (Eds.), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024, Association for Computational Linguistics, 2024, pp. 338–354. doi:10.18653/V1/2024.NAACL-LONG.20.

[52] D. Ru, L. Qiu, X. Hu, T. Zhang, P. Shi, S. Chang, C. Jiayang, C. Wang, S. Sun, H. Li, Z. Zhang, B. Wang, J. Jiang, T. He, Z. Wang, P. Liu, Y. Zhang, Z. Zhang, Ragchecker: A fine-grained framework for diagnosing retrieval-augmented generation, in: A. Globersons, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. M. Tomczak, C. Zhang (Eds.), Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024, 2024. URL: http://papers.nips.cc/paper_files/paper/2024/hash/27245589131d17368cccdfa990cbf16e-Abstract-Datasets_and_Benchmarks_Track.html.

[53] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, P. Liang, Lost in the middle: How language models use long contexts, Trans. Assoc. Comput. Linguistics 12 (2024) 157–173. doi:10.1162/tacl_a_00638.

[54] M. A. de Luis Balaguer, V. Benara, R. L. de Freitas Cunha, R. de M. Estevão Filho, T. Hendry, D. Holstein, J. Marsman, N. Mecklenburg, S. Malvar, L. O. Nunes, R. Padilha, M. Sharp, B. Silva, S. Sharma, V. Aski, R. Chandra, RAG vs fine-tuning: Pipelines, tradeoffs, and a case study on agriculture, CoRR abs/2401.08406 (2024). doi:10.48550/arXiv.2401.08406.

[55] C. Wang, Z. Yang, S. Gao, C. Gao, T. Peng, H. Huang, Y. Deng, M. R. Lyu, RAG or fine-tuning? A comparative study on lcms-based code completion in industry, in: L. Montecchi, J. Li, D. Poshyvanyk, D. Zhang (Eds.), Proceedings of the 33rd ACM International Conference on the Foundations of Software Engineering, FSE Companion 2025, Clarion Hotel Trondheim, Trondheim, Norway, June 23-28, 2025, ACM, 2025, pp. 93–104. doi:10.1145/3696630.3728535.

[56] O. Ovadia, M. Brief, M. Mishaeli, O. Elisha, Fine-tuning or retrieval? comparing knowledge injection in llms, in: Y. Al-Onaizan, M. Bansal, Y. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024, Association for Computational Linguistics, 2024, pp. 237–250. doi:10.18653/V1/2024.EMNLP-MAIN.15.

[57] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Lora: Low-rank adaptation of large language models, in: The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022, 2022. URL: https://openreview.net/forum?id=nZeVKeeFYf9.

[58] T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer, Qlora: Efficient finetuning of quantized llms, in: A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, S. Levine (Eds.), Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023, 2023. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/1feb87871436031bdc0f2beaa62a049b-Paper-Conference.pdf.