

Index-Aware Knowledge Grounding of Retrieval-Augmented Generation in Conversational Search for Archival Diplomatics

1st Qihong Zhou

*Department of Electrical and Computer Engineering
University of British Columbia
Vancouver, Canada
qihong.zhou@ubc.ca*

2nd Binming Li

*Computer Science
Simon Fraser University
Vancouver, Canada
Binming.Li@outlook.com*

3rd Victoria Lemieux

*School of Information
University of British Columbia
Vancouver, Canada
v.lemieux@ubc.ca*

Abstract—This paper discusses a novel index-aware method of semantically grounding chunking in the preprocessing phase of a conversational search pipeline. The paper outlines the novel index-aware chunking strategy, explains the setup for an experimental evaluation, and concludes with a discussion of the experimental results. The results indicate that using index-aware knowledge grounding in the conversational search pipeline can help reduce computational costs, processing resource demands, hallucinations, and the precision of answers.

Index Terms—conversational search, retrieval-augmented generation, computational archival science

I. INTRODUCTION

Conversational search is an AI-powered information retrieval method that allows users to ask questions in natural language in a manner that simulates human conversation, rather than relying on specific one-off queries based on keywords. The most common example of this type of search is the now ubiquitous chatbot [1]. Conversational search uses natural language processing to aid information retrieval and digital interactions that enables users to engage with digital platforms using unstructured, conversational queries. Large Language Models (LLMs) have revolutionized conversational search by demonstrating outstanding capacity to retrieve information in response to natural language queries. However, these LLMs come with high computational demands, require large amounts of data, and result in considerable energy consumption, all of which can lead to high operational costs and environmental concerns [2].

To address these concerns, in 2021, Meta AI introduced the concept of Retrieval-Augmented Generation (RAG) [3]. RAG enhances the capabilities of LLMs by connecting them to external knowledge sources at inference time. In this way, an LLM does not need to rely only on the information already stored within the parameters of a pre-trained model. Instead, a retrieval system is used to find relevant document passages from a knowledge-base external to the model and uses this retrieved context to generate more accurate and up-to-date responses. Using RAG enables LLMs to operate with less demand for computational resources and data, since it

means that it is no longer necessary to fit an exhaustive representation of all world knowledge directly into the model's parameters. The RAG approach also allows LLMs to operate more effectively in specialized knowledge domains, such as archival diplomatics, since the LLM is able to access domain-specific knowledge to augment the pre-existing information in its model.

Despite these advantages, the RAG approach is not perfect. LLMs have input limitations as defined by their context windows, that is, the maximum number of tokens (roughly words or sub-word units) that can fit into the model at once (including both the prompt and the generated response). Typical input limits for LLMs are 4K–32K tokens for widely used commercial models and 128K–200K tokens for cutting-edge ones [4]. Traditional RAG methods chunk external documents in order to segment a knowledge-base into smaller, fixed-size segments representative of the corpus of external documents. While this approach helps overcome input limitations, it often fragments context, resulting in incomplete retrieval and diminished precision of query results [5].

In the remainder of this paper, we provide the background and related research connected to our approach. In section three, we discuss our approach in detail and the experimental methodology. In section four, we discuss the results of our experiment, and in section five we present our lessons learned.

II. BACKGROUND & RELATED RESEARCH

A. Conversational Search Pipeline Using RAG

Typically, conversational search using RAG entails conducting the following steps:

- 1) **Ingestion:** In this step, documents are parsed, datafied (i.e., converted to markdown, json or other computable format, and broken into manageable chunks).
- 2) **Indexing:** These chunks are then converted into vector embeddings and stored in a specialized vector database.
- 3) **Retrieval:** When a user asks a question, it is also converted into an embedding to search for the most relevant chunks. This step may also involve query rewriting to

improve retrieval results by (e.g., adding appropriate words, and removing non-relevant ones to a query.

- 4) **Ranking:** A ranker model re-orders the retrieval chunks to prioritize the most relevant information.
- 5) **Generation:** The top-ranked chunks and the original question are fed to an LLM to generate a coherent, context-aware answer.

Research on improving the conversational search pipeline using RAG has focused on all steps in this pipeline. For example, [6] explores methods to mitigate limitations in query understanding by enriching user inputs with structured common sense knowledge and the potential of instruction-tuned language to automatically reformulate queries and improve the effectiveness of conversational information retrieval. Despite wide-ranging research on how to improve the conversational search pipeline with RAG, a mostly overlooked aspect of the pipeline remains the process of chunking during ingestion of RAG input data.

B. Current Research on the Chunking Process

As previously mentioned, chunking is necessary to overcome the limitations of context windows in LLMs. By chunking external documents into smaller, fixed-size segments the number of data tokens to be fed into an LLM can be reduced [5]. However, traditional methods often fail to create chunks that capture sufficient semantic meaning, as they do not account for underlying textual structure and semantics, resulting in incomplete or inaccurate retrieval in highly specialized domains.

Several researchers have aimed at overcoming this shortfall. [5] investigated the effects of twelve different approaches to chunking of three different types: divided into length-based, structure-based, and semantic-based approaches. These approaches were implemented and applied to a corpus of documents. Results of the experiment confirmed that the choice of the chunking technique significantly influences the quality of the retrieved documents. According to [5], structure-based strategies achieved the best outcomes, whereas simpler length-based approaches suffered from information fragmentation. The optimal chunking strategy appeared to be context-dependent, varying with document structure and question type. Another paper [7] explored two advanced techniques: 1) late chunking and 2) contextual retrieval to preserve global context, finding that late chunking offers a more computationally efficient solution by leveraging the natural capabilities of embedding models, while contextual retrieval, with its reliance on LLMs for context augmentation and re-ranking, incurs higher computational expenses. In yet another recent piece of research, [8] propose a novel framework that enhances RAG by integrating hierarchical text segmentation and clustering to generate more meaningful and semantically coherent chunks. Nevertheless, chunking that produces accurate, context-sensitive results for highly specialized domains, such as archival diplomatics, remains an open area of research. In the following section, we present our approach to chunking—Index-Aware Knowledge Grounding RAG.

III. METHODOLOGY

A. Domain Context

Archival diplomatics is a specialized knowledge domain that focuses on the critical analysis of documents, particularly documentary heritage, to determine their authenticity, reliability, and context of creation. It studies the genesis, form, transmission, and relationship of archival documents to the facts they represent and their creators, with the goal of identifying, evaluating, and communicating their true nature (i.e., whether authentic or inauthentic) [9]. The discipline originated in the 17th century with the French Benedictine monk Dom Jean Mabillon, who developed a methodology to authenticate medieval records and detect forgeries, notably in his work *De re diplomatica libri VI* [10]. The scope of diplomatics has expanded beyond medieval documents to include modern and digital records, reflecting its adaptability to contemporary archival science [11]. Archival diplomatics is also increasingly applied in digital forensics, where its principles help maintain the long-term authenticity of digital records through technical and system migrations [12].

B. Index-Aware Knowledge Grounding RAG Dataset Overview, OCR and Text Extraction

In order to capture the archival diplomatics knowledge domain for purposes of generating our index-aware RAG, we form our corpus from Luciana Duranti's six part series of articles [9], [13], [14], [15], [16], [17], representing the traditional foundation of archival diplomatics. To this corpus, we added the InterPARES Terminology database [18], comprised of terms that extend the foundations of archival diplomatics into the digital realm.

The input to our system consists of a set of PDF files representing different sections of the same article or document. Each part is stored as an independent PDF file, and one of these files serves as an index table that lists terms and their corresponding locations across the documents. In our case study, the index primarily includes domain-specific terms related to archival diplomatics. However, the same process can be generalized to other domains, such as legal, medical, or technical documentation, without modification to the pipeline logic. No image pre-processing, such as denoising or layout correction, is performed prior to OCR, as the system relies on the OCR engine's internal processing capabilities. The pipeline also does not depend on the number of datasets or sources; as long as the file names match the entries in the index table, the system can correctly associate terms with their respective documents and page numbers.

The first step in our Index-Aware Knowledge Grounding RAG involves extracting text from all provided PDF files. A hybrid extraction approach is implemented using PyMuPDF (fitz) as the primary tool and Tesseract OCR via the pytesseract library as a fallback mechanism. For each PDF, the system first attempts to extract the text layer directly through PyMuPDF. If no text layer exists or if extraction fails, the pages are converted into images using pdf2image, and OCR is applied

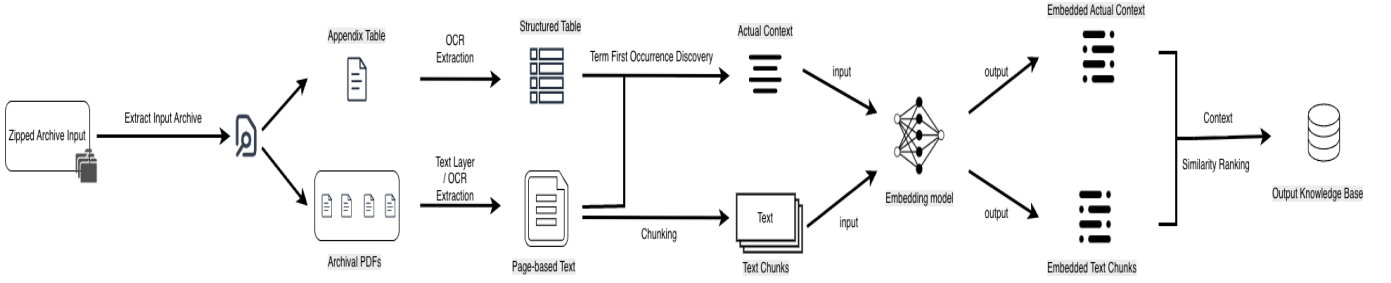


Fig. 1. Overall pipeline of Index-Aware RAG

to recover the textual content. This ensures that both digitally born and scanned PDFs are handled consistently.

The index table itself is also provided in PDF format, which requires additional pre-processing to convert it into a structured tabular representation. Since the table may not contain explicit borders or clear separation lines, the system applies OCR-based structure detection to identify rows and columns heuristically. The extracted data is organized into a tabular structure containing at least three key attributes: the term, the associated file name, and the page number where the term appears.

C. Chunking Strategy

1) *Parameters*: We use three integer parameters:

$$M > 0, \quad m > 0, \quad O \geq 0,$$

where M is the max tokens per chunk, m is the minimum target tokens, and O is the overlap (in words) between consecutive chunks.

2) *Token Counts and Notation*: For a chunk boundary context $C_k = (s_a, \dots, s_b)$, define the word count of a sentence s_i as

$$w_i = |\text{words}(s_i)|, \quad i \in \{a, \dots, b\}, \quad W_{a:b} = \sum_{i=a}^b w_i.$$

Here “words(·)” denotes the number of whitespace-delimited words (as in the implementation).

3) *Overlap Seed Selection*: Given an overlap target $O \geq 0$, define the minimal suffix index t by

$$t = \min \left\{ i \in [a, b] \mid \sum_{j=i}^b w_j \geq O \right\}.$$

Then the next-chunk seed is

$$\text{Seed}(C_k; O) = \begin{cases} \emptyset, & O = 0, \\ (s_t, \dots, s_b), & 0 < O \leq W_{a:b}, \\ (s_a, \dots, s_b), & O > W_{a:b}. \end{cases}$$

Its token count is

$$W(\text{Seed}(C_k; O)) = \begin{cases} 0, & O = 0, \\ \sum_{j=t}^b w_j, & 0 < O \leq W_{a:b}, \\ W_{a:b}, & O > W_{a:b}. \end{cases}$$

Minimality holds for $0 < O \leq W_{a:b}$:

$$\sum_{j=t}^b w_j \geq O \quad \text{and} \quad \forall i \in \{t+1, \dots, b\} : \sum_{j=i}^b w_j < O.$$

Equivalently, with suffix sums $S(i) = \sum_{j=i}^b w_j$ we have $t = \min\{i : S(i) \geq O\}$ (the implementation scans from b down to a until the threshold is reached, then restores the original order).

Bounds:

$$\min\{O, W_{a:b}\} \leq W(\text{Seed}(C_k; O)) \leq W_{a:b},$$

with equality $W(\text{Seed}) = W_{a:b}$ if $O \geq W_{a:b}$.

a) *Remark (when overlap applies)*.: Overlap is applied only after a *normal* emission (i.e., when the current chunk meets m and the next sentence would overflow M) and when $O > 0$. After a *forced* emission ($W(B) < m$) or an *oversize sentence* case ($w_j > M$), the next chunk starts empty (no seed).

4) *Algorithm (Pseudocode)*: The procedure converts per-file page text into sentence-aware chunks for retrieval.

5) *Index Table Parsing*: The appendix or index table is processed separately. Its tabular structure is recognized using OCR, and all identified cells are converted into structured records consisting of the term, file name, and page number. These entries serve as anchors for locating relevant text passages within the other PDFs. The extracted table is cached in memory for use in subsequent stages.

6) *Term-Based Document and Page Retrieval*: Based on the extracted index, the pipeline identifies which document and page corresponds to each term. For each term, the associated PDF and page number are located and retrieved. Once the text of that page is available, the system searches for the first occurrence of the target term within the page and extracts the following one to two sentences as contextual information. These short passages form the query cache, which stores the

Algorithm 1 Sentence-Aware Chunking

```
1: Initialize empty chunk list  $\mathcal{C}$ .
2: for each file  $f$  do
3:   Build ordered sentences  $S_f = (s_1, \dots, s_{K_f})$  with token
   counts  $w_j$ .
4:    $B \leftarrow \emptyset$ ,  $W(B) \leftarrow 0$ .
5:   for  $j = 1$  to  $K_f$  do
6:      $w_j \leftarrow$  token count of  $s_j$ .
7:     if  $w_j = 0$  then
8:       continue
9:     end if
10:    if  $B \neq \emptyset$  and  $W(B) + w_j > M$  then
11:      if  $W(B) \geq m$  then
12:        Emit  $B$ ; add to  $\mathcal{C}$ .
13:      if  $O > 0$  then
14:        Build overlap seed from end of  $B$  until token
        sum  $\geq O$ .
15:         $B \leftarrow \text{Seed}(B)$ ;  $W(B) \leftarrow \sum w_i$  in seed.
16:      else
17:         $B \leftarrow \emptyset$ ;  $W(B) \leftarrow 0$ .
18:      end if
19:    else
20:      Emit  $B$ ; add to  $\mathcal{C}$ .
21:       $B \leftarrow \emptyset$ ;  $W(B) \leftarrow 0$ .
22:    end if
23:  end if
24:  if  $w_j > M$  then
25:    if  $B \neq \emptyset$  then
26:      Emit  $B$ ; add to  $\mathcal{C}$ ;  $B \leftarrow \emptyset$ ;  $W(B) \leftarrow 0$ .
27:    end if
28:    Emit single-sentence chunk  $\{s_j\}$  with  $w_j$ ; add to
     $\mathcal{C}$ .
29:    continue
30:  end if
31:  Append  $s_j$  to  $B$ ;  $W(B) \leftarrow W(B) + w_j$ .
32: end for
33: if  $B \neq \emptyset$  then
34:   Emit  $B$ ; add to  $\mathcal{C}$ .
35: end if
36: end for
37: return  $\mathcal{C}$ 
```

minimal text context needed for later retrieval steps. The query cache is maintained in memory during execution and saved locally as a JSON file (metadata.json) for persistence.

7) *Text Chunking*: After constructing the query cache, the rest of the PDFs are segmented into smaller text chunks. The chunking follows a fixed-length window with overlap between consecutive chunks. This ensures continuity of semantic meaning across chunk boundaries and maintains retrieval quality. Although the current implementation uses a fixed-length approach, the design allows for more sophisticated chunking strategies (e.g., semantic or dynamic chunking) in future extensions.

8) *Embedding and Similarity Computation*: Both the text chunks and query cache entries are embedded into vector representations using the SentenceTransformers model all-MiniLM-L6-v2. The embedding process supports batch processing with configurable batch size (default: 32) and can be executed on either CPU or GPU (CUDA-enabled). All embeddings are normalized to unit length, allowing cosine similarity to be computed efficiently using NumPy-based vector operations. No external vector database (e.g., FAISS or Pinecone) is required, which simplifies deployment and improves transparency in experimental evaluation.

9) *Knowledge Base Construction*: Finally, each query from the cache is used to retrieve the top three most relevant text chunks from the embedded corpus based on cosine similarity. These selected chunks together form the minimum knowledge base (KB) that represents the most relevant contextual information for the given query terms. This knowledge base can then be used as input for downstream tasks such as question answering, semantic search, or document summarization.

D. Experimental Set-up

To test the efficacy of our chunking strategy, we ran three experiments using two separate platforms offering conversational search, as described in the following subsections. The experiments entailed inputting the same set of standardized test prompts into each platform set-up to generate output results, and evaluating the accuracy of the output results against a pre-defined answer key.

1) *Experiments 1 & 3 Deployment Environment Details: Clio-X*: The first platform we used was Clio-X (<https://cliox.org>) [19], a Web3-powered, privacy-first AI orchestration platform that enables ethical application of AI using Compute-to-Data methods, blockchain-based access controls, and privacy-preserving machine learning. Clio-X supports AI/ML workflows that mask or obfuscate sensitive features (e.g. personal identifiers) before computation. The system ensures that analytic tasks like chatbots, clustering, topic modeling, or sentiment analysis can run without exposing underlying sensitive content. In the first experiment, we used the input data described above and deployed a conversational search pipeline on Clio-X using a typical RAG chunking strategy and a llama 3.2 3B model. In the third experiment, we deployed a conversational search pipeline on Clio-X using the same input data as for the first experiment, but this time using our Index-Aware Knowledge Grounding RAG chunking strategy with a llama 3.2 3B model.

2) *Experiment 2 Deployment Environment Details: UBC NLP Hugging Face Space*: The second platform we used was the Hugging Face Space of the UBC Natural Language Processing Lab in which a conversational search pipeline had been set up (i.e., 'inter pares-chat'). Hugging Face Spaces is a platform within the Hugging Face ecosystem that allows artificial intelligence developers, researchers, and others to create, host, share, and deploy machine learning models and applications as interactive web demos [20]. For this experiment, an expanded input dataset was generated from a scrape

of the InterPARES Projects 1-4 websites. The pipeline was built using a typical approach to chunking with a backbone of OpenAI's recently released gpt-oss-20B model [21]

For each experiment, the following standard prompts were used a single time to test the conversational search information retrieval capabilities of each set up: 1) "Who/What are the 'persons' in archival diplomatic theory?", 2) "What is a 'writer' according to archival diplomatic theory?", 3) "Please explain what is an addressee in archival diplomatic theory", and 4) "In archival diplomatic theory, what is the difference between a 'fact' and an 'act'?"

IV. EXPERIMENTAL RESULTS

In this section, in the interest of brevity, we provide an overview of our experimental results for two of the experiments we ran.

A. Experiment 1: Standard Chunking Strategy on Clio-X with llama 3.2 3B model

1) *Test 1:* For Experiment 1, Test 1, we asked the Clio-X conversational search engine the question "Who/What are the 'persons' in archival diplomatic theory?" and received the following response:

Based on the knowledge base content, I can provide an accurate and relevant answer to your question. In archival diplomatic theory, the "persons" involved in records are primarily: 1. **Document creators**: Individuals who produce or sign documents, such as government officials, business owners, or individuals with a specific role or responsibility. 2. **Record keepers or custodians**: Those who manage, maintain, and preserve physical or digital copies of records over time. 3. **Appraisers or valuers**: Experts who evaluate the authenticity, value, or condition of documents for use in various contexts, such as appraisal, valuation, or litigation. These individuals play a crucial role in shaping the content, context, and meaning of archival records. The knowledge base highlights that diplomatics (a method of inquiry into archival materials) emphasizes the importance of considering the persons creating the records, their roles, and the concepts of function, competence, and responsibility when analyzing and managing archives. Source: [Source: New Uses Part6 records management.pdf] Diplomats as a Method of Inquiry...Gérard and Christiane Naud...the record forms convey and reveal content essential to the formation of the missing link between provenance and pertinence approaches. Source: [Source: New Uses Part6 records management.pdf] Diplomats as a Method of Inquiry...Gérard and Christiane Naud?! ...the record forms convey and reveal content essential to the formation of the missing link between provenance and pertinence approaches.

The response provided by the conversational search engine was inaccurate and incomplete. The correct response should

have included at least a reference to the terms used by diplomatics to designate the three necessary persons, which are the author, the addressee, and the writer [14]. In addition, the response should have indicated that the creator is the person in whose archival *fonds* the record exists, and that, with born digital records, a fifth person comes into play, the originator, the person from whose account the records is sent or by whose account the record is received (it maybe the same as the author or addressee, or different). In addition, the response should have mentioned other potential persons intervening in the creation of a record, such as witnesses and countersigners, noting that they do not participate in the creation of every record.

2) *Test 2:* For this test, we asked the question "In archival diplomatic theory, what is the difference between a 'fact' and an 'act'?". The returned response from the non-index-aware RAG in Clio-X was as follows:

Based on the knowledge base content, the difference between a 'fact' and an 'act' in archival diplomatic theory can be understood from the perspective of their relationship to documents. A fact refers to a specific, verifiable piece of information that is independent of any particular document or act. It is an objective reality that exists regardless of the documentation of it. In this sense, facts are often considered as the raw data that can be verified and documented in various ways. On the other hand, an act is a more complex entity that involves not only facts but also the context and circumstances surrounding their existence. An act is a single event or action that produces a particular document or set of documents. In this sense, acts are often considered as the means by which facts are created, recorded, and preserved. In terms of function, documents serve to present, verify, or document acts in some way. The act itself can be seen as the original fact, while the document serves as a record or presentation of that fact. This relationship is crucial in archival diplomatic theory, where understanding the distinction between facts and acts is essential for analyzing and interpreting historical documents. For example, consider a document from World War I that records an event, such as the signing of the armistice agreement. In this case, the document is a record of an act (the signing of the agreement), but it also contains information about a fact (the outcome of the war). The document serves to present and verify this fact, providing context and background information for understanding the significance of the event. In summary, in archival diplomatic theory, facts are objective pieces of information that exist independently of any particular document or act, while acts are single events or actions that produce documents. Understanding this distinction is essential for analyzing and interpreting historical documents.

Person	Role in the record-event relationship	Why they matter in diplomatic analysis
Creator / Producer (often a clerk, secretary, manager, or any office worker)	The individual or group that physically produces the record (writes, types, signs, etc.).	Diplomatics begins with the <i>record itself</i> and traces back to the event that produced it. The creator's identity, title, and place in the organization help determine the record's authenticity and provenance.
Owner / Custodian (department, agency, or institution that holds the record)	The entity that controls the record after its creation – it may be the same person who created it or a later administrative unit.	The owner's policies and procedures influence the record's format, preservation, and legal status. In diplomatic analysis, the chain of custody is crucial for establishing reliability.
User / Reader (researcher, archivist, lawyer, historian, etc.)	Anyone who later accesses, interprets, or cites the record.	The user's needs shape the way a record is presented and preserved. Diplomatic analysis must consider how the record will be understood by future users to ensure that its contextual integrity is maintained.
Subject of the Record (the person or entity about whom the record is written)	The individual or organization that is the focus of the record's content (e.g., a client, a project, a person's name).	In diplomatic scrutiny, the subject is a key element of the <i>record-event relationship</i> . The record's validity can hinge on whether it accurately represents the subject's actions or statements.
Archivist / Record-Keeper (the professional who processes, indexes, and preserves the record)	The person who applies archival principles, creates finding aids, and ensures the record's long-term accessibility.	The archivist's decisions—what to keep, how to describe, how to preserve—directly affect the record's evidentiary value. Diplomatic theory places the archivist at the intersection of the record and its contextual interpretation.

Fig. 2. Results of Experiment 2, Test 1

Although the general statement that the difference between a fact and an act in archival diplomatics rests upon their relationship to the document, a critical missing concept in the response was that an act always proceeds from a 'will' to produce it. Thus, we considered the response inaccurate, or, at the very least, imprecise.

B. Experiment 2: UBC NLP Hugging Face Space with Chatgpt 20B model

1) *Test 1*: For Experiment 2 Test 1, the UBC NLP Hugging Space conversational search engine was asked the same question and returned the response shown Figure 1.

Although a much larger model was used (i.e., gpt-oss-20B model vs. llama 3.2 3B model), the response was equally inaccurate and incomplete.

The results of the first two experiments indicated that we would either need to train the model to provide a more accurate and complete response, or fine tune the conversational pipeline to generate improved results. Since training a model requires human labeling of data and is resource intensive, we opted to fine-tune the conversational pipeline. After considering a number of options, we chose to focus on chunking due to the availability of a pre-existing index that we hypothesized could be leveraged to achieve more accurate and complete results.

2) *Test 2*: For this experiment, we prompted the UBC NLP Hugging Face conversational search engine with the same question as in Experiment 1, Test 2. The returned response is presented in a tabular format as indicated in figures 2 and 3.

The response was comprehensive, but also inaccurate in that it did not mention a 'will' as being essential to the formation of an act. Furthermore, it mentioned acts as always being juridically relevant, whereas, in fact, there can also

In archival-diplomatic theory the two terms refer to different parts of the record-event relationship:

Term	What it denotes	Why it matters in diplomacy
Act	A <i>juridically relevant action or event</i> that actually takes place (e.g., signing a contract, issuing a licence, a court ruling).	The act is the <i>cause</i> of the record. Diplomats ask whether the record genuinely records that act, because an act is what creates a legally or historically significant document.
Fact	The <i>content or reality</i> that the record is about – the information, <i>state of affairs</i> , or event described by the document (e.g., the date of the contract, the parties involved, the terms agreed).	The fact is what the record <i>represents</i> . Diplomats assess whether the fact recorded is true, complete, and faithfully reproduced, because a record is evidence of that fact.

Fig. 3. Results of Experiment 2, Test 2. First part of the response.

Key differences		
Aspect	Act	Fact
Nature	An <i>action</i> (the event that happens).	An <i>information item</i> (the reality that the document speaks about).
Legal relevance	Always juridically relevant – it creates a legal relationship or obligation.	May be juridically relevant or irrelevant; it can simply be a description of a <i>state of affairs</i> .
Role in record creation	The <i>trigger</i> – the act causes a record to be made.	The <i>subject</i> – what the record contains and what it purports to prove or describe.
Diplomatic focus	Verify that the record <i>actually</i> documents that act (authenticity, provenance).	Verify that the fact recorded is <i>true and complete</i> (historical authenticity, reliability).
Examples	Signing a deed, a court hearing, a government decree.	The date of the deed, the parties' names, the terms of the agreement.

In practice

1. **Probative documents**: The record itself *is* the act (e.g., a signed contract).
2. **Supporting documents**: The record *supports* an act (e.g., a receipt that proves payment was made).
3. **Narrative documents**: The record *narrates* a fact that may be irrelevant to law (e.g., a diary entry).

Diplomatic analysis always asks: *Did the act happen?* (historical authenticity) and *Does the record faithfully represent the fact of that act?* (reliability). The two concepts are thus complementary but distinct.

Fig. 4. Results of Experiment 2, Test 2. Second part of the response.

be juridically irrelevant acts. Additionally, it indicates that a probative record is the act itself whereas a probative record memorializes an act and a dispositive record, which is not mentioned in the response at all, is the type of record that embodies the act itself.

C. Experiment 3: Index-Aware Knowledge Grounding RAG on Clio-X with llama 3.2 3B model

After developing our approach to the production of index-aware knowledge grounded RAG, as described above, we then ran our Experiment 3 to test the results of our new chunking strategy.

1) *Test 1*: We asked the same question as we had in the previous two experiments, and the conversational search engine returned the following results:

According to archival diplomatic theory, the "persons" involved in records refer to individuals or entities that concur in the formation of a document. These persons can include those who create, sign, direct, are written to, are concerned with, or have effect on the record (Source: IV and III). The identification of these persons is crucial as they are tightly linked to juridical conceptions and systems, and their subtle variations significantly influence diplomatic theory (Source: IV). In the context of archival science, special archival science applies archival theory to individual cases, focusing on how records

creators organized their memory. This discipline also highlights the importance of identifying persons in records, as they are central to understanding the document’s significance (Source: I).

Although this result was not inaccurate, it was imprecise in that it was lacking in detail. Nevertheless, we considered this to be a better response due to the lack of inaccuracy.

2) *Test 2*: Finally, we prompted the Clio-X conversational search engine using our revised index-aware knowledge grounded RAG with the same question and received the following response:

In archival diplomatic theory, a ‘fact’ refers to an objective piece of information or evidence, whereas an ‘act’ is the actual event or action that has occurred. According to the Archival Theory (no source specified), an act is a concrete expression of will, intention, or purpose, which leaves a record in a document. In contrast, a fact is the content of the recorded act. [Source: Archival Theory] This distinction highlights the difference between the object of investigation and the means by which it is established or disproved, as mentioned in the definition of ‘evidence’ provided earlier.

This response was more parsimonious than the response provided by the gpt-oss-20B model and was more accurate in that it mentioned the involvement of a will in the formation an act. Nevertheless, it was still inaccurate, in that facts do not need to be objective nor be evidence, and the response failed to capture that facts not only occur as the content of documents, but are also produced as a consequence of the creation of documents (e.g., a contract between parties creates a juridically relevant fact), and thus we considered it still to be incomplete.

V. LESSONS LEARNED

It is possible to characterize the results of our experiments as three distinct personas: The first persona is the untrained ‘young turk’ who offers a response based on general knowledge, but one which ends up being inaccurate or imprecise. The second persona is the ‘windbag’ that offers a long-winded response, but one that is still inaccurate. The final persona – based on our index-aware knowledge grounded RAG – can be characterized as the ‘Computational Archival Science student’, whose responses are more accurate, but who still has a lot to learn. In other words, our index-aware RAG still did not produce perfect results.

Despite the less than perfect results, this study indicates that it is not necessary to throw GPUs (i.e., more computational power) at a problem to achieve better results. In fact, in some cases, the results can end up being equally or more inaccurate and incomplete despite expending additional computing resources

Additionally, our results indicate that refinements to the conversational search pipeline, such as our adjustments to the chunking strategy, can achieve incremental improvements in

the accuracy and completeness of results without requiring time-consuming and labour-intensive model training.

Nevertheless, we must point to several limitations of our study. First, we acknowledge that the two experimental setups (i.e., Clio-X and the UBC NLP Hugging Face Space) utilized very different input data sets and deployment platforms, which, in addition to changes in the approach to chunking and the use of different models, could have influenced the results of our experiments. We were unable to control for these differences in our experimental design. In addition, we acknowledge that the conversational search pipeline includes many steps, as indicated in section 2, and, though we did not alter any of these other steps in our experiments, we did not specifically control for them in our experimental design. We also ran the test questions only once on each deployment platform, which, given that output from conversational search engines is non-deterministic, did not permit evaluation of the accuracy and completeness of output results over multiple tests. It may be the case, therefore, that different results would be achieved if the tests were run again.

Despite these limitations, we still see value in the directionality of our research, which is to pursue strategies of refining the conversational search pipeline as opposed to adding GPUs that allow for running larger models, or pre-training models to achieve better conversational search results. Our future research will focus on making additional adjustments to our conversational search pipeline, such as ranking of the relevancy of retrieved chunks.

In addition, although not a direct goal of this study, our research points to the value of experiments involving adjustments to the conversational search pipeline for the knowledge that these experiments generate about the behaviour of chatbots and other conversational search engines. This knowledge can contribute to frameworks aimed at greater transparency and accountability of AI.

VI. CONCLUSION

This paper has reported on a novel index-aware method of semantically grounding chunking in the preprocessing phase of a conversational search pipeline. The paper also reported on experimental evaluation of the efficacy of the strategy. Although the experiments had many limitations, the results indicate that using index-aware knowledge grounding in the conversational search pipeline may, indeed, help reduce computational costs, processing resource demands, hallucinations, and improve the precision of answers. Additionally, the results point to ways to test the behaviour of different strategies to adjust conversational search pipelines that could contribute to greater transparency, accountability and more ethical AI.

ACKNOWLEDGMENT

This research was supported by Canada’s Social Sciences and Humanities Research Council through the InterPARES Trust AI Project and Canada’s Natural Sciences and Engineering Research Council through Blockchain@UBC

REFERENCES

- [1] D. E. Staff, “Conversational search,” *DevXJ*, 2023.
- [2] E. Strubell, A. Ganesh, and A. McCallum, “Energy and policy considerations for modern deep learning research,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 09, 2020, pp. 13 693–13 696.
- [3] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, “Retrieval-augmented generation for knowledge-intensive nlp tasks,” *Advances in neural information processing systems*, vol. 33, pp. 9459–9474, 2020.
- [4] M. A. Ferrag, N. Tihanyi, and M. Debbah, “Reasoning beyond limits: Advances and open problems for llms,” *arXiv preprint arXiv:2503.22732*, 2025.
- [5] T. Moro and A. De Lorenzo, “Influence of chunking techniques on the performance of retrieval augmented generation,” Ph.D. dissertation, Università degli Studi di Trieste, 2025.
- [6] G. Rocchietti, “Efficient and effective methods for conversational search and summarization,” 2025.
- [7] C. Merola and J. Singh, “Reconstructing context: Evaluating advanced chunking strategies for retrieval-augmented generation,” in *International Workshop on Knowledge-Enhanced Information Retrieval*. Springer, 2025, pp. 3–18.
- [8] H.-T. Nguyen, T.-D. Nguyen, and V.-H. Nguyen, “Enhancing retrieval augmented generation with hierarchical text segmentation chunking,” in *International Symposium on Information and Communication Technology*. Springer, 2024, pp. 209–220.
- [9] L. Duranti, “Diplomatics: New uses for an old science, part i,” *Archivaria*, pp. 4–17, 1989.
- [10] V. L. Lemieux, *Searching for trust: blockchain technology in an age of disinformation*. Cambridge University Press, 2022.
- [11] L. Duranti and K. Thibodeau, “The concept of record in interactive, experiential and dynamic environments: the view of inter pares,” *Archival science*, vol. 6, no. 1, pp. 13–68, 2006.
- [12] L. Duranti, “From digital diplomatics to digital records forensics,” *Archivaria*, pp. 39–66, 2009.
- [13] —, “Diplomatics: New uses for an old science, part ii,” *Archivaria*, pp. 7–27, 1989.
- [14] —, “Diplomatics: New uses for an old science, part iii,” *Archivaria*, pp. 4–20, 1990.
- [15] —, “Diplomatics: New uses for an old science, part iv,” *Archivaria*, pp. 10–24, 1990.
- [16] —, “Diplomatics: New uses for an old science, part v,” *Archivaria*, pp. 6–23, 1991.
- [17] —, “Diplomatics: New uses for an old science, part vi,” *Archivaria*, pp. 4–24, 1991.
- [18] <https://interparestrustai.org/terminology>.
- [19] V. L. Lemieux, R. Gil, F. Molosiwa, Q. Zhou, B. Li, R. Garcia, L. D. L. T. Cubillo, and Z. Wang, “Clio-x: Aweb3 solution for privacy-preserving ai access to digital archives,” *arXiv preprint arXiv:2507.08853*, 2025.
- [20] <https://huggingface.co/spaces/UBC-NLP/inter pares-chat>.
- [21] OpenAI, “Introducing gpt-oss,” *OpenAI Blog*, pp. <https://openai.com/index/introducing-gpt-oss/>, 2025.