



**POLITECNICO  
DI TORINO**

POLITECNICO DI TORINO

Master Degree course in Communications and Computer Networks Engineering

Master Degree Thesis

# **Building a RAG system for Kiruna's Urban Transformation Data**

## **Supervisors**

Prof. Riccardo COPPOLA  
Prof. Mesut DINLER

## **Candidate**

Burak KARGACI

ACADEMIC YEAR 2025-2026

# Acknowledgements

I would like to express my sincere gratitude to my supervisors, Prof. Riccardo Coppola, Prof. Mesut Dinler and Prof. Valeria Federighi for their guidance, support, and valuable feedback throughout the development of this thesis. Their insights and suggestions were essential in shaping the direction of this research. I would also like to thank the Department of Communications and Computer Networks Engineering at Politecnico di Torino for providing the academic environment and resources necessary to carry out this work. Finally, I would like to thank my family and friends for their encouragement and support during my studies.

## Abstract

This thesis presents the design and implementation of a retrieval-augmented generation (RAG) system to organize and access scattered information about the city of Kiruna in northern Sweden. Because of long-term mining activity, Kiruna is being gradually relocated. Buildings, roads, and entire districts are being moved or rebuilt. As a result, a large amount of documentation has been produced over time. This includes planning reports, maps, images, municipal records, and research papers. These materials exist in many different places and formats, which makes them difficult to search and connect. General large language models are not trained on this local data, so they cannot reliably answer detailed questions about Kiruna's transformation.

The thesis begins with a systematic investigation of the types of documents that exist in the Kiruna case. Sources from municipal websites, academic publications, and publicly available online resources were reviewed. Based on this analysis, a sample dataset of Kiruna-related documents was collected and organized. Different document categories were identified, and a classification structure was defined to describe their roles within the documentation of the city's transformation. In parallel, a study of metadata standards and metadata types was conducted to determine which approaches are most suitable for different kinds of materials, such as textual reports, spatial data, and visual records. The work evaluates how metadata can support consistency, traceability, and long-term usability of heterogeneous urban documentation.

The goal of this thesis is to build a system that can retrieve Kiruna-specific information and generate accurate answers based on verified sources. To do this, a RAG pipeline was developed in Python using a curated dataset. The system combines semantic search with language-model generation so that responses are based on retrieved documents rather than general knowledge.

The retrieval process works in several stages. First, document abstracts are embedded using a transformer-based model and compared with user queries through cosine similarity. The most relevant documents are then processed at page level. Pages are split into overlapping text chunks to preserve context. Embeddings are generated in batches to improve speed. Similarity scores are calculated for each chunk, and the most relevant sections are selected using a sliding-window approach. Irrelevant pages, such as tables of contents, are filtered out. The selected context is then passed to a local instruction-tuned language model, which generates an answer grounded in the retrieved material.

The results show that this approach improves access to fragmented documentation and allows users to ask detailed questions about Kiruna's planning, relocation, and urban change. The system demonstrates how retrieval-based methods can support research on cities undergoing long-term transformation. It also provides a foundation for future work on metadata structuring and digital archival systems for urban and heritage research.

# Contents

<b>1</b>	<b>Introduction</b>	5
<b>2</b>	<b>Background</b>	9
2.1	Digital Heritage	9
2.1.1	Introduction to Digital Heritage	9
2.1.2	Definition of Digital Heritage	10
2.1.3	Types of Digital Heritage Data	10
2.1.4	Technologies for Digital Heritage	11
2.1.5	Data Management in Digital Heritage	12
2.1.6	Digital Heritage Platforms	13
2.1.7	Use Cases of Digital Heritage	13
2.1.8	Challenges in Digital Heritage	14
2.1.9	Overview and Key Insights	14
2.2	LLM in Digital Heritage	15
2.2.1	Introduction to Large Language Models in Digital Heritage	15
2.2.2	Large Language Models (LLMs)	16
2.2.3	LLMs in Digital Heritage	17
2.2.4	Technical Integration	18
2.2.5	Use Cases	19
2.2.6	Challenges and Risks	20
2.2.7	Overview and Key Insights	21
2.3	Retrieval-Augmented Generation	21
2.3.1	Introduction to Retrieval-Augmented Generation	21
2.3.2	Concept of Retrieval-Augmented Generation	22
2.3.3	Architecture of RAG Systems	23
2.3.4	Technical Components	24
2.3.5	Applications of Retrieval-Augmented Generation	25
2.3.6	Advantages of Retrieval-Augmented Generation	26
2.3.7	Challenges and Limitations	27
2.3.8	Overview and Key Insights	28
<b>3</b>	<b>Methodology</b>	29
3.1	Data Model	29
3.1.1	Definition of the Data Model	29

3.1.2	Use-Case Model of the System . . . . .	32
3.2	Creation of Document Dataset . . . . .	34
3.2.1	Example of Dublin Core Metadata Implementation . . . . .	35
3.3	RAG Pipeline Architecture and Execution . . . . .	37
3.3.1	Document Indexing . . . . .	39
3.3.2	Query Processing and Document Retrieval . . . . .	39
3.3.3	Page Segmentation and Chunk Embedding . . . . .	39
3.3.4	Page Scoring and Filtering . . . . .	39
3.3.5	Window-Based Page Selection . . . . .	40
3.3.6	Context Construction . . . . .	40
3.3.7	Prompt Construction and Response Generation . . . . .	40
3.3.8	Summary . . . . .	41
3.4	Implementation Challenges and Design Decisions . . . . .	41
3.4.1	Memory Limitations and Context Size . . . . .	41
3.4.2	Irrelevant Section Retrieval and TOC Filtering . . . . .	42
3.4.3	Embedding Model Selection for Document Retrieval . . . . .	42
3.4.4	Multilingual Retrieval Challenges . . . . .	42
3.4.5	Matching Relevant Document Sections . . . . .	43
3.4.6	Model Quantization and Performance Trade-offs . . . . .	43
3.4.7	Summary of Design Decisions . . . . .	43
<b>4</b>	<b>Evaluation</b> . . . . .	<b>47</b>
4.1	Evaluation Methodology . . . . .	47
4.2	Definition of Test Documents . . . . .	48
4.3	Definition of Test Questions . . . . .	50
4.4	Response Quality Evaluation . . . . .	52
4.4.1	Qualitative Analysis of Generated Responses . . . . .	52
4.4.2	Questionnaire Results . . . . .	55
4.4.3	Additional Questions . . . . .	68
4.4.4	Discussion of Questionnaire Results . . . . .	71
<b>5</b>	<b>Conclusion</b> . . . . .	<b>73</b>
	<b>Bibliography</b> . . . . .	<b>75</b>



# Chapter 1

## Introduction

Kiruna is a city in northern Sweden and it is going through a major urban transformation. This change is caused by long-term mining activity under the ground. Over time, the mining has made parts of the land unstable. Because of this, large areas of the city can no longer stay where they are. Many parts of the urban area are being moved, demolished, or rebuilt in new locations. This process does not affect only a few buildings. It affects the structure of the whole city. Housing areas, public buildings, roads, and cultural sites are all involved. Some places are slowly disappearing while new districts are being planned and built. The relocation will continue for many years and will happen in different stages. These stages include planning, demolition, construction, and documentation.

As the city changes, its physical form changes as well. Streets, buildings, and neighborhoods are not always in their original locations anymore. At the same time, older layers of the city become harder to trace. Once a building is removed, its physical presence is gone. What remains are documents, images, maps, and memories. For this reason, documentation becomes very important. It helps people understand how Kiruna developed in the past. It also helps explain how the city is changing today. In the future, it will help researchers study the transformation and understand its long-term effects. Without systematic documentation, it will be difficult to connect the past city with the new one that is being built.

The transformation of Kiruna produces a very large amount of material. There are historical archives that describe earlier planning and development. There are also new planning documents that explain relocation strategies and future construction. Maps and geographical data show how the boundaries of the city and the use of land are changing. Photographs record buildings before they are demolished and new areas as they appear. Reports and municipal records describe decisions and technical processes. Heritage collections preserve cultural and social information about the city. All these materials together form a rich record of Kiruna's transformation.

However, these materials are not stored in one place. They are produced by many different organizations. These include municipal authorities, planning offices, cultural institutions, and research groups. The materials are also stored in different formats. Some are paper documents kept in archives. Others are digital files. Some are stored in map databases. Each source contains useful information. Still, they are often separated

from each other. They follow different classification systems and different standards. Because of this, it is hard to connect them. Information about Kiruna is spread across institutions and formats. This makes it difficult to create a clear and complete overview of how the city is changing.

This situation creates an important research problem. When a city changes on a large scale, information about it often becomes fragmented. In Kiruna, data about the city exists in different places and in different forms. Researchers who want to study the city must search in many archives and databases. They may need to look for maps in one place, reports in another, and images somewhere else. Even when they find the materials, it can be hard to compare them. Metadata is not always consistent. Some documents are incomplete. Some are easy to access, but others are not. Many materials are digital, but some still exist only in physical form. Some planning documents were created for short-term use and were not fully preserved.

As relocation continues, there is also a risk of losing knowledge. When buildings are demolished, their physical context disappears. Without good documentation, it becomes harder to understand how those places looked and how they were used. Over time, it may become difficult to reconstruct the original urban structure. The lack of a unified system for organizing and linking materials makes this problem worse. Without such a system, it is hard to document the transformation in a clear and accessible way. For urban historians, planners, and heritage researchers, this limits the ability to study Kiruna's development over time. It also makes it harder to preserve the memory of the city.

The need for structured documentation is especially urgent because physical traces of the past are being removed. When a building is demolished or a district is relocated, it cannot be studied in the same way again. If documents and images are not connected, the link between past and present becomes weak. This affects both historical research and future planning analysis. To understand how the city evolved, it is important to see how decisions were made and how different parts of the city were connected. Some archives already contain valuable information. However, they are often built for administrative use. They are not always designed for long-term research or for linking different types of materials. Maps, reports, and photographs may exist in separate systems. Because there is no integrated digital infrastructure, it is difficult to connect them. Over time, knowledge can become scattered and harder to interpret.

Digital heritage systems offer one possible response to this situation. Such systems aim to collect and organize different types of material in a structured way. They can bring together documents, images, and spatial data. They also help people search for and compare information. By using shared standards and metadata, digital archives can support long-term access and analysis. Researchers can follow connections between sources and trace changes over time. However, the growing amount of digital data also creates new challenges. Collections are large and complex. They include many formats and come from different institutions. Traditional archival methods may not be enough to manage this complexity.

Recent developments in artificial intelligence suggest additional ways to support archival work. Large Language Models can help interpret texts and generate descriptions. They

can also help identify links between documents. Retrieval-based approaches make it possible to search across structured datasets and retrieve relevant information. These tools do not replace archival practices. Instead, they can support the organization and use of complex collections. They are especially useful in situations where data is heterogeneous and constantly growing. In a case like Kiruna, where transformation is ongoing and documentation is expanding, such tools may help connect different sources and make them easier to use.

Within this context, this thesis focuses on the problem of fragmented urban documentation in Kiruna. The aim is to explore how heterogeneous materials about the city can be organized within a structured digital framework. The work looks at how maps, planning documents, images, and archival records can be connected. It also seeks to define criteria for classifying, describing, linking, and validating data. The thesis examines how a retrieval-based approach could support structured access to this information. The goal is to support research on Kiruna's transformation and long-term development. The proposed framework is not meant to replace existing archives. Instead, it aims to connect them and make them more accessible and useful over time.

The scope of the thesis is limited. It focuses on data organization and digital structure. It does not attempt to simulate the future of the city or evaluate planning policies. It also does not present a complete technical implementation of a system. Instead, the work defines methods for structuring and linking information. The goal is to support documentation and research during a period of large-scale urban change. By focusing on the problem of fragmented knowledge, the thesis contributes to discussions on digital heritage and urban transformation. It highlights the importance of preserving and connecting information so that the history and evolution of Kiruna can be understood in the future.



## Chapter 2

# Background

### 2.1 Digital Heritage

#### 2.1.1 Introduction to Digital Heritage

Cultural heritage has always been an important part of human society. It shapes identity, shared memory, and knowledge. In the past, heritage was mostly linked to physical objects and traditions. These included monuments, buildings, artworks, and cultural practices passed from one generation to another. Today, digital technology is changing how heritage is preserved and accessed. New tools allow cultural materials to be recorded, stored, and shared in digital form. They also make it easier for people to study and interpret these materials.

Over time, heritage has moved beyond a small academic field. It has become part of everyday cultural life. Public interest in heritage has grown strongly in recent decades. As noted in heritage studies, “heritage has burgeoned over the past quarter of a century from a small elite preoccupation into a major popular crusade” [18]. This growth has increased the demand for preservation. It has also created new expectations. People now expect heritage to be accessible online. They want to explore collections, images, and archives through digital platforms.

The idea of digital heritage connects several disciplines. It stands between computer science, digital humanities, and heritage studies. Digital technologies are now used to document, preserve, manage, and present cultural resources. Advances in computing and networks allow institutions to create digital copies of objects and sites. They also allow the creation of new cultural works that exist only in digital form. Because of these changes, the meaning of heritage is expanding. It is no longer limited to physical artifacts.

Digital heritage includes more than scanned copies of objects. It also includes materials that were created digitally from the beginning. These are known as born-digital materials. Examples include digital art, virtual environments, and digital records. Such materials may never exist in physical form. They still represent important cultural information. Their preservation requires new methods and tools. The growing variety of digital formats makes digital heritage work more complex. It also makes the field strongly interdisciplinary.

Today, digital heritage is an important research area. It combines technical innovation

with cultural and social concerns. Many professionals must work together. These include computer scientists, archivists, museum staff, historians, and policy makers. Their shared goal is to make sure cultural resources remain accessible and meaningful over long periods of time.

### 2.1.2 Definition of Digital Heritage

The idea of digital heritage builds on earlier ideas of cultural heritage. Cultural heritage includes both physical and non-physical elements. These can be monuments, buildings, traditions, or practices. They have historical, artistic, or social value. International frameworks describe heritage as sites and objects with “outstanding universal value from the point of view of history, art or science” [34]. These definitions help explain what should be preserved and why it matters.

Digital heritage extends this idea. It focuses on how digital technology supports preservation and access. Digital heritage includes materials that have been digitized from physical objects. It also includes materials created directly in digital environments. Examples include digital photos of artifacts, 3D models of sites, and digital archives. It also includes multimedia records of traditions and events. Websites, social media content, and digital artworks can also be part of digital heritage.

The growth of digital heritage is closely linked to computing technology. Digital tools have changed traditional preservation practices. Recent research explains that “cultural heritage preservation has entered a transformative era with the integration of advanced computing technologies” [9]. These tools affect how heritage is recorded, stored, restored, analyzed, and presented. They make it possible to manage larger collections and share them widely.

Some digital heritage exists only in digital form. These materials may have no physical version. Still, they represent important cultural knowledge. Their preservation requires new frameworks and methods. Research in digital preservation notes that “many forms of digital information cannot be expressed in traditional hard-copy or analog media” [30]. Examples include interactive web pages, geographic systems, and virtual reality models. These examples show how the scope of heritage is expanding. Digital heritage also improves access. Digital platforms allow institutions to share collections with global audiences. People can access materials remotely. This supports education and public engagement. It also reduces geographic barriers.

Digital heritage can therefore be defined as a collection of digital resources that represent and preserve cultural heritage. These resources may come from digitized objects or from digital-only environments. This definition highlights both technological and cultural aspects.

### 2.1.3 Types of Digital Heritage Data

Digital heritage data can be grouped into several categories. These categories depend on how the data is created and used. One category is digitized heritage. This includes digital copies of physical objects. Examples include scanned manuscripts, photos of artworks,

and 3D scans of archaeological sites. These digital versions help preserve fragile originals. They also allow wider access and support research and education.

Another category is born-digital heritage. These materials are created directly in digital form. Examples include digital art, digital architecture, video games, and online cultural content. These materials can be difficult to preserve. They often depend on specific software or hardware. Digital preservation research notes that digital objects come in many forms and that “there is no single definition or model of a digital document that would be valid in all cases” [30]. This diversity makes preservation more complex.

Metadata is another important type of digital heritage data. Metadata describes digital objects. It gives information about origin, context, format, and usage rights. Metadata helps organize and retrieve digital materials. Without good metadata, digital objects can become difficult to understand or access over time.

Digital heritage data also includes multimedia. This can be audio, video, or interactive content. These formats create richer representations of heritage. Virtual museum systems often combine text, images, audio, and interaction. Recent research notes that technologies such as artificial intelligence and extended reality are “transforming how cultural knowledge is preserved and experienced” [21]. These tools allow new forms of engagement.

Geospatial data is also important. Maps, spatial data, and sensor data help document heritage sites and landscapes. Combining different data types allows more complete representations of heritage.

The wide variety of digital heritage data shows that the field is interdisciplinary. Managing these data requires skills from computer science, information science, and heritage studies. It also requires standards for interoperability and preservation.

#### 2.1.4 Technologies for Digital Heritage

Digital heritage relies on many technologies. These technologies support the collection, processing, storage, and sharing of data. Digitization tools are among the most basic. These include scanning, photography, and 3D imaging. They create digital copies of objects and sites. High-resolution digital records support research, conservation, and access.

Data processing tools are also important. These include image processing, machine learning, and data analysis. They help interpret digital heritage data. They can support tasks such as object recognition and restoration. New computing technologies are playing a larger role. Artificial intelligence, virtual reality, and augmented reality are being used in heritage projects. They create interactive and immersive experiences. Research notes that computing technologies now support many stages of heritage work, from acquisition to presentation [9].

Digital repositories and information systems store digital heritage data. They manage metadata and support long-term preservation. Virtual environments and digital twins allow users to explore digital reconstructions of sites. These tools improve access and understanding.

Digital heritage also depends on broader digital infrastructure. Cloud computing and high-speed networks allow large amounts of data to be stored and shared. They support

collaboration between institutions and allow global access to collections.

As digital technology continues to develop, its role in heritage will grow. It will create new opportunities and new challenges for preservation and research.

### 2.1.5 Data Management in Digital Heritage

Data management is a key part of digital heritage work. Digital heritage projects create large amounts of data. These data come in many formats. They can include images, 3D models, text files, audio recordings, and interactive media. All of these materials must be stored and organized carefully. They must also be preserved so they can be used in the future.

Digital preservation is not simple. Digital objects depend on technology that changes quickly. File formats, software, and storage systems can become outdated. This creates risks for long-term access. Research in digital preservation points out several challenges. These include policy issues, institutional responsibilities, legal questions, and metadata needs [30]. Because of these issues, institutions must work together. They must plan for long-term preservation from the beginning.

A central task in digital preservation is deciding what must be preserved. Digital files may change format over time. Storage systems may also change. Even so, the main informational content must remain accessible. Preservation strategies often focus on keeping essential characteristics. These include structure, meaning, and context. It is not always necessary to preserve every technical detail. Instead, the goal is to keep the content usable and understandable.

Metadata is very important in this process. Metadata describes digital objects and their context. It explains where they come from and how they should be used. It also helps organize and retrieve files. Without metadata, digital collections can quickly become difficult to navigate. Over time, files without proper description may lose meaning.

Sustainability is another major concern. Digital heritage data must remain accessible for many years. Sometimes it must last longer than the systems that store it. This means that data must be updated and migrated over time. Formats may need to be converted. Storage systems may need to be replaced. Long-term preservation often requires open standards and backup systems. Regular checks are also needed to make sure data remains intact. Legal and ethical issues must also be considered. Digital heritage often includes materials with copyright restrictions. Some materials may also be culturally sensitive. Institutions must decide how to share such materials responsibly. They must balance access with respect for ownership and cultural values.

Accessibility is another key part of data management. Digital collections should be usable by many different people. This includes researchers, students, and the general public. Systems should provide clear interfaces and search tools. They should also work across different devices and platforms. Good data management supports both preservation and access.

### 2.1.6 Digital Heritage Platforms

Digital heritage platforms provide the systems used to store and present digital collections. These platforms include digital repositories, online databases, and virtual museums. They allow users to access cultural materials from different locations. In many cases, they provide global access.

Digital repositories are common in heritage work. They store digital objects and their metadata. They also support long-term preservation. Many repositories follow shared standards. These standards help different institutions exchange data. This allows collections to be connected and shared.

Virtual museums are another type of platform. They use digital tools to create interactive experiences. Users can explore digital collections online. Some systems use artificial intelligence, extended reality, or digital twins. These tools allow users to explore heritage sites in new ways. Recent research notes that new technologies support “persistent, immersive, and interactive environments” for cultural exploration [21]. These environments allow users to learn through interaction.

Digital platforms also support collaboration. Museums, archives, and libraries can share data through common systems. Shared platforms make collections more visible. They also support research and education. Researchers can access materials from multiple institutions in one place. New technologies are often integrated into these platforms. Artificial intelligence can help with image recognition and metadata creation. Recommendation systems can suggest related content. Virtual and augmented reality can create immersive tours. Cloud computing supports storage and data processing.

Usability is very important for digital platforms. They should be easy to use. They should also support different types of users. Some users are experts, while others are not. Platforms should include multilingual interfaces and accessible design. They should also work well on different devices.

### 2.1.7 Use Cases of Digital Heritage

Digital heritage has many practical uses. One common use is the digitization of cultural collections. Museums and archives create digital copies of objects and documents. These copies can be shared online. This allows people around the world to access them. It supports research, education, and public interest.

Another use is virtual reconstruction. Digital tools can recreate heritage sites. These reconstructions can be used for research or tourism. They can also recreate sites that no longer exist. Virtual environments allow people to explore places that are far away or difficult to reach. They can also show how sites looked in the past. Digital heritage is widely used in education. Digital platforms provide access to cultural materials for students. Learners can explore collections and take virtual tours. Interactive tools can improve understanding of history and culture. These experiences can support both formal and informal learning.

Researchers also use digital heritage data. Digital images, 3D models, and archives can support research in many fields. These include archaeology, history, and art history. Digital tools allow detailed analysis. They also support collaboration between institutions

and disciplines.

Digital heritage supports community participation. Communities can contribute to documentation of their own heritage. They can share local knowledge and memories. This helps preserve cultural identity. It also promotes diversity and inclusion.

Digital heritage can also support tourism and local economies. Online platforms can promote heritage sites. They can attract visitors and raise awareness. Virtual experiences can also provide access when physical visits are not possible.

### 2.1.8 Challenges in Digital Heritage

Despite its benefits, digital heritage faces several challenges. One major challenge is technological change. Digital formats and systems become outdated. Files may become unreadable over time. Research notes issues such as digital obsolescence, accessibility, copyright, and data security [24]. These issues make long-term preservation difficult.

Another challenge is sustainability. Digital heritage projects require funding and technical support. Maintaining digital collections can be expensive. Institutions must plan for long-term management. Without stable funding, digital projects may not survive.

Legal and ethical issues are also complex. Questions of ownership and access must be addressed. Some materials may have cultural restrictions. Others may be protected by copyright. Institutions must handle these issues carefully.

Data quality is another concern. Digital copies may not fully represent physical objects. Important details may be lost. Ensuring accuracy and authenticity is essential. Researchers must trust the digital data they use.

Interoperability is also a problem. Digital heritage data is stored in many systems. These systems may not communicate well. Standards and protocols are needed. Without them, collections remain isolated.

Access and inclusion are also important issues. Digital technologies can improve access. However, not everyone has equal access to technology. Digital inequality can limit participation. Projects must consider accessibility and digital literacy.

### 2.1.9 Overview and Key Insights

Digital heritage is an important development in cultural preservation. It connects traditional heritage work with digital technology. It supports documentation, preservation, and access. Digital heritage includes both digitized objects and born-digital materials. It uses tools such as digitization, data systems, artificial intelligence, and virtual reality.

Digital technology has changed how heritage is managed and experienced. It allows wider access to collections. It supports research and education. It also creates new ways to engage with culture. At the same time, digital heritage faces challenges. These include preservation, sustainability, and ethical concerns.

As technology continues to change, digital heritage will remain important. It will require cooperation across disciplines. Computer scientists, heritage professionals, and communities must work together. Their goal is to keep cultural resources accessible and meaningful for future generations.

## 2.2 LLM in Digital Heritage

### 2.2.1 Introduction to Large Language Models in Digital Heritage

The digital transformation of cultural heritage has changed how historical and cultural materials are preserved, accessed, and understood. Museums, archives, and libraries now rely heavily on digital technologies to manage large collections. These collections include documents, images, and many types of multimedia content. Digitization has made cultural heritage easier to access for a wider audience. At the same time, it has created new challenges. Institutions must now deal with problems related to data organization, interpretation, and user interaction. Cultural heritage datasets are often very large. They can also be heterogeneous and multilingual. Because of this complexity, advanced computational tools are needed to support analysis and retrieval.

Artificial intelligence has become an important technology for addressing these issues. Natural language processing methods allow automatic analysis of textual heritage data. Such data includes archival records, exhibition descriptions, and academic texts. As digital collections continue to grow, the need for tools that can process large amounts of information has also increased. These tools must handle both text and multimedia data. Within this context, Large Language Models have gained attention as powerful AI systems. They are seen as tools that can support heritage preservation, interpretation, and access.

Large Language Models are now used in several cultural heritage applications. Research shows that they are becoming more common in tasks such as writing descriptions of historical monuments, translating ancient texts, and creating educational materials. One study notes that “large language models (LLMs) become increasingly prevalent in tasks related to cultural heritage, such as generating descriptions of historical monuments, translating ancient texts, and creating educational content” [4]. These capabilities suggest that LLMs can support both professionals and the public. They can help generate descriptive content, answer questions, and support interactive exploration of digital collections.

However, the use of LLMs also raises concerns. The content produced by these models may affect how users understand history and culture. If the generated information is inaccurate, it may lead to misunderstanding or misinterpretation. Studies have shown that generated heritage texts may “distort or misrepresent important cultural values, historical context, or cultural nuances” [4]. This means that the use of LLMs in heritage contexts must be examined carefully.

The integration of LLMs into digital heritage environments therefore creates both opportunities and risks. On one side, LLMs offer new ways to analyze and present cultural information. They can support access and engagement with heritage collections. On the other side, their limitations must be considered. Issues such as accuracy, bias, and cultural sensitivity are especially important in heritage contexts. For this reason, understanding how LLMs work is necessary before they can be used responsibly in digital heritage systems.

### 2.2.2 Large Language Models (LLMs)

Large Language Models are artificial intelligence systems designed to process and generate natural language. They are trained on very large datasets that contain text from many domains. During training, the model learns patterns in language. It learns how words are used together and how meaning is formed in context. This allows the model to predict and generate coherent text. Unlike earlier rule-based systems, LLMs do not rely on fixed rules. Instead, they use deep neural networks that learn statistical relationships between words and contexts. LLMs are general-purpose models. They can perform many tasks without being programmed separately for each task. These tasks include summarizing texts, translating between languages, answering questions, and generating new content. In digital humanities research, LLMs are seen as flexible tools for working with large textual datasets. They make it easier to analyze and explore large collections. Research notes that LLMs provide “significant advantages in scalability and accessibility” for researchers working with large cultural corpora [19]. This scalability is especially important for cultural heritage collections, which often include millions of documents and records. LLMs are also developing into multimodal systems. Traditional language models focused only on text. Newer models can process images, audio, and text together. Multimodal LLMs combine these data types and analyze them in relation to each other. Research on intangible cultural heritage shows that LLMs demonstrate “robust multimodal processing capabilities and sophisticated semantic comprehension” [12]. This means they can analyze visual and textual heritage materials at the same time. For example, they can interpret images of artifacts while also processing related textual descriptions.

Most modern LLMs are built on the Transformer architecture. This architecture uses attention mechanisms to model relationships between words in a sequence. Instead of reading text strictly from left to right, the model looks at all words in relation to each other. It evaluates how each word connects to others in the sentence. This allows the model to capture context and meaning more effectively. The attention mechanism is important because it helps the model identify relevant information across long text passages. This is useful for cultural heritage texts, which often include long and complex sentences. Historical documents may contain references to events, people, and places across different sections. The Transformer architecture allows the model to capture these connections. Another advantage of the Transformer architecture is that it supports large-scale training. Models can be trained with billions of parameters. These parameters represent learned relationships within the data. The large scale allows the model to capture complex semantic patterns. In cultural heritage contexts, this helps when analyzing texts that contain historical language, references, and variations in style. LLMs can identify relationships between historical events, places, and individuals mentioned in archival records.

LLMs are usually trained using self-supervised learning. During training, the model learns to predict missing words or the next word in a sentence. It does this by analyzing large amounts of text. Through this process, the model learns grammar, vocabulary, and contextual relationships. Once training is complete, the model can perform many language tasks. LLMs can generate new text, summarize long documents, translate between languages, and answer questions. In cultural heritage contexts, these abilities support many applications. LLMs can generate descriptions of artifacts and historical

objects. They can provide contextual explanations for users. They can also support interactive systems that respond to user questions. Some research describes LLMs as digital storytelling tools. For example, studies on museum systems describe LLMs as systems that can function as “a digital storytelling machine that can be trained and guided to act as a museum guide” [33]. In this role, LLMs can provide information about exhibits and historical contexts. They can adapt responses based on user interests. This allows more personalized experiences for visitors. LLMs can also help extract information from historical texts. They can identify names of people, places, and events. They can also identify relationships between them. Research in digital humanities shows that multimodal LLMs can serve as “reliable tools for historical information extraction within Digital Humanities” [19]. This means they can support automated analysis of large archival datasets.

Even though LLMs have strong capabilities, they also have limitations. One important limitation is accuracy. LLMs generate text based on patterns learned from data. They do not verify facts in real time. Because of this, they can produce incorrect or misleading information. In cultural heritage contexts, incorrect information can affect how history is understood. It can also affect how cultural values are represented. Another limitation is cultural misalignment. Generated texts may include simplified or inaccurate representations of cultural contexts. Research shows that many generated heritage texts include misalignments. One study reports that “over 65% of the generated texts exhibit notable cultural misalignments” [4]. These misalignments can affect how cultural values and traditions are represented. Bias is also a concern. Training data may contain dominant cultural perspectives. This can influence model outputs. Some cultures or viewpoints may be underrepresented. This can lead to biased interpretations. There are also concerns about transparency. Users may not know what data was used to train the model. Research in digital humanities highlights concerns about “the opacity of training data, the privacy of user data and the bias hidden in the model structure” [19]. These limitations show that LLMs must be used carefully in cultural heritage systems. Their outputs should be evaluated. Human oversight is important. Responsible deployment requires attention to accuracy, bias, and cultural sensitivity.

### 2.2.3 LLMs in Digital Heritage

One of the most important uses of Large Language Models in digital heritage is the analysis of archival texts. Archives contain very large collections of documents. Many of these documents are only partly digitized. Some are digitized but not fully annotated. Manual analysis of such collections takes a long time. It also requires specialized knowledge. Researchers must read and interpret many documents. This makes the process slow and difficult. LLMs can support automated analysis of archival texts. They can process large amounts of text and extract useful information. For example, they can identify themes, names, places, and events in documents. They can also identify relationships between these elements. This allows researchers to explore connections across collections. Research on archival heritage shows that AI-based processing can help “discover descriptive and topical keywords of previously digitized archival records” [32]. This improves searchability. It also helps users find relevant materials more easily. LLMs can

classify documents and group them by topic. They can also generate summaries of long documents. These summaries help users understand content quickly. This is useful when working with large archives. LLMs can also help restore historical texts. Many digitized archives rely on optical character recognition. OCR systems often produce errors. Old documents may have unusual fonts or damage. This makes OCR less accurate. LLM-based pipelines can help correct these errors. They can improve text quality. Studies show that using LLMs with historical archives opens new possibilities for restoration and analysis [31]. This improves usability of digitized texts.

Metadata enrichment is another important use of LLMs in digital heritage. Many heritage records contain limited metadata. Some records include only basic descriptions. Others may have incomplete information. This makes it harder to search and retrieve materials. LLMs can help generate additional metadata. They can create descriptions, keywords, and contextual notes. These additions improve discoverability. Users can find materials more easily. They can also understand the context of items better. Research on cultural heritage search systems shows that LLM-based methods can improve discoverability of artifacts. They do this by generating contextual interpretations and linking related items [35]. LLM-driven search systems can analyze exhibit descriptions and metadata. They can then suggest related items. This helps users explore collections in new ways.

Cultural heritage collections often include multiple languages. Historical documents may exist in different languages or scripts. Some collections include translations. Others do not. This can create barriers for users. LLMs can support multilingual access. They can translate texts into different languages. They can also generate explanations. This allows more people to access collections. It supports global audiences and cross-cultural research. Multilingual processing also helps preserve linguistic diversity. It allows smaller language communities to access heritage materials. However, translation accuracy is important. Cultural context must be preserved. This remains an area of ongoing research. LLMs can also function as digital assistants in heritage systems. These assistants can answer questions. They can provide explanations. They can guide users through collections. Conversational interfaces allow users to interact with collections in natural language. Research on cultural heritage websites shows that integrating LLMs can provide “assisted question-answering functions to answer information needs that the system’s knowledge base fails to cover” [7]. This means that LLMs can help fill gaps in information. They can respond to user queries even when information is not directly stored in the system. LLM assistants can be used in websites, mobile applications, and virtual museums. They can provide personalized information. They can suggest related materials. They can support educational experiences. This makes digital heritage more interactive and accessible.

#### **2.2.4 Technical Integration**

Integrating Large Language Models into digital heritage systems requires careful design. LLMs are rarely used alone. They are usually part of larger systems. These systems connect data sources, processing modules, and user interfaces. The goal is to make sure that model outputs are based on reliable heritage data.

A typical LLM system starts with data ingestion. Text data is collected from digital repositories, archives, or museum databases. This may include catalog descriptions, archival records, or digitized texts. Before being processed by the model, the data is cleaned and prepared. This can include normalization and segmentation. In many systems, retrieval mechanisms are used. These mechanisms provide the model with relevant context. This helps reduce incorrect outputs. It also improves accuracy.

LLMs often interact with digital repositories and metadata systems. Instead of generating responses only from internal knowledge, the system retrieves relevant records. These records are then given to the model as context. This hybrid approach improves reliability. It also makes it easier to trace information sources. Research shows that combining LLMs with structured metadata can “expand metadata and enhance access to the past” [32]. This allows users to explore archives in more meaningful ways.

Knowledge bases and ontologies also play an important role. Cultural heritage information is often organized using controlled vocabularies. When LLMs are connected to these structures, they can generate responses that follow established terminology. Research shows that semantic integration improves consistency and reduces ambiguity [35].

Domain adaptation is another key step. General LLMs are trained on broad data. They may not include specialized heritage knowledge. Fine-tuning models on museum or archive texts improves relevance. Research shows that fine-tuned models produce more accurate results in heritage systems [33].

User interfaces are also important. Visual dashboards and conversational interfaces allow users to interact with LLM outputs. Combining LLM processing with visualization helps users explore data. Research shows that such systems support “augmented exploration of digitized archival heritage” [32]. Interfaces help users evaluate AI-generated content.

### 2.2.5 Use Cases

Museums are among the first institutions to use Large Language Models in practice. LLMs are used to improve visitor experiences. They can provide interactive guides and personalized recommendations. They can also explain exhibits in simple language. Visitors can ask questions in natural language and receive answers. The system can adapt responses based on visitor interests. Research on museum applications describes LLMs as systems that can act as “interactive guide[s], providing personalized information to visitors about artworks and historical contexts” [33]. These systems can be used inside museums or online. In virtual museums, LLMs can guide users through collections. They can explain objects and historical narratives. This creates more engaging and immersive experiences. LLMs also help museum staff. They can assist with writing exhibition texts and educational content. They can help draft descriptions and labels. Human experts still review the content. However, LLMs reduce the time needed for writing and editing. They also support new storytelling approaches.

Digital libraries and archives benefit from LLMs mainly through improved access. LLMs can analyze large document collections. They can extract key topics and relationships. They can generate summaries that help users understand content quickly. This is important for archives that contain millions of documents. Research shows that

LLM-based processing can help “unlock digitized— not born-digital—archival heritage data created before the advent of AI” [32]. This means that existing archives can be explored in new ways. LLMs can connect information across collections. They can reveal links between people, places, and events. LLMs also help improve digitized text quality. Many historical documents were digitized using OCR. OCR often produces errors. LLM-based pipelines can correct these errors. Studies show that such pipelines can improve readability and usability of historical texts [31]. This makes archives easier to use for research.

LLMs are also used as research tools. Researchers in digital humanities use them to analyze large text collections. LLMs can identify patterns and trends. They can support topic analysis and entity recognition. They can compare texts across time periods. LLMs also support search systems. Research shows that LLM-based search tools improve discoverability. They use both metadata and generated interpretations [35]. Hybrid systems combine traditional search with LLM analysis. This allows more flexible exploration of collections.

LLM-based assistants provide conversational access to collections. Users can ask questions and receive answers. They can also receive suggestions for related content. Research shows that integrating LLMs into heritage websites enables “advanced information exploration support” [7]. These systems go beyond static content. These assistants are helpful for non-expert users. They make complex information easier to access. They support learning through exploration. However, careful design is needed. Systems must remain transparent and accurate.

### 2.2.6 Challenges and Risks

Despite their benefits, LLMs introduce risks in digital heritage contexts. One major issue is hallucination. LLMs may generate information that sounds correct but is wrong. In heritage systems, this can lead to misinformation. It can also distort historical narratives.

Cultural misalignment is another issue. Generated content may not represent cultural values correctly. Research reports that many outputs include cultural inaccuracies. Studies note that “cultural value misalignments may exist in generated texts” [4]. These may include misrepresentation of historical facts or oversimplification of cultural narratives. Such issues are sensitive in heritage contexts.

Bias is closely linked to cultural misalignment. Training data may overrepresent certain perspectives. This can influence outputs. Minority cultures may be underrepresented. Researchers emphasize the need for culturally aware evaluation and human oversight [12].

Data quality also affects results. Heritage datasets often contain incomplete metadata. If input data is poor, output may also be poor. High-quality data and transparent processing are necessary.

Another concern is user trust. Users may assume AI outputs are always correct. Research stresses that users need “careful warnings when interacting with AI-processed data” [32]. Transparency is essential. Systems should show sources and limitations.

### 2.2.7 Overview and Key Insights

Large Language Models represent a major development for digital heritage. They provide tools for text analysis, metadata enrichment, multilingual access, and interactive systems. They help process large collections. They support new ways to access and explore heritage.

At the same time, LLMs introduce risks. These include inaccuracies, bias, and cultural misalignment. Hallucinated information can distort history. Responsible integration requires careful design. It also requires human oversight and domain adaptation.

For research in computer science and digital humanities, understanding LLMs is essential. They should be seen as support tools, not authoritative sources. When used carefully, they can improve access and engagement. They can support next-generation digital heritage systems while respecting historical accuracy and cultural integrity.

## 2.3 Retrieval-Augmented Generation

### 2.3.1 Introduction to Retrieval-Augmented Generation

Large language models (LLMs) are now a central part of modern artificial intelligence and natural language processing. These models are trained on very large collections of text. Because of this training, they can generate text that is coherent and appropriate for many different contexts. Recent progress in transformer-based architectures and large-scale training has greatly improved their performance. Today, LLMs can perform tasks such as question answering, summarization, and dialogue generation at a high level. However, even though they are powerful, they still face practical challenges. Recent research points out that “large language models (LLMs) demonstrate powerful capabilities, but they still face challenges in practical applications, such as hallucinations, slow knowledge updates, and lack of transparency in answers” [10].

Despite their success, standalone LLMs have several important limitations. One well-known problem is hallucination. This happens when a model produces text that sounds correct but is actually wrong. The output may be fluent and convincing, yet factually incorrect. Another limitation is that LLMs often rely on outdated knowledge. After training is finished, their parameters remain fixed. This makes it difficult for them to incorporate new information. In addition, they may not contain enough knowledge about specific domains. Specialized or technical topics are sometimes missing from training data. Research has shown that LLMs “often fabricate facts and lack knowledge when dealing with specific domains or highly specialized queries” [10]. These issues reduce trust in real-world use. They are especially problematic in areas like medicine, law, or scientific research, where accuracy is critical.

A further challenge concerns how knowledge is stored inside LLMs. Most of their knowledge is encoded in the model parameters. This means the information cannot easily be updated without retraining or fine-tuning the model. As one study explains, “purely parameterized language models store their world knowledge... in the parameters of the model” [10]. This design makes knowledge difficult to change or expand. Retraining very large models takes a lot of time and computational resources. Because of this, it is hard

for them to keep up with fast-changing information.

To address these problems, researchers have explored ways to connect language models with external knowledge sources. External retrieval mechanisms have become more important as LLMs are used in knowledge-intensive tasks. When models can retrieve information from outside sources, they can access up-to-date data. They can also check facts and provide references. This improves reliability. These ideas have led to the development of retrieval-augmented generation (RAG). RAG combines text generation with information retrieval. As described in one survey, “RAG refers to the retrieval of relevant information from external knowledge bases before answering questions with LLMs” [10].

The integration of retrieval systems with generative models marks an important change in how AI systems are designed. Instead of relying only on internal model knowledge, modern systems combine internal parameters with external data. This hybrid approach has become a key direction in research on reliable and trustworthy AI.

### 2.3.2 Concept of Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) is a framework that combines information retrieval methods with generative language models. The main idea is simple. The system first retrieves relevant documents or data from external sources. It then uses this information as context when generating a response. In this way, the model can access knowledge that is not stored in its internal parameters. One foundational paper explains that RAG models “combine pre-trained parametric and non-parametric memory for language generation” [16].

In traditional language models, knowledge is stored inside the model parameters. This is called parametric knowledge. It is learned during training and encoded in the neural network weights. In contrast, RAG introduces non-parametric knowledge. This knowledge is stored outside the model, usually in external databases or document collections. According to research, “parametric knowledge is acquired through training LLMs and stored in the neural network weights,” while “non-parametric knowledge... resides in external knowledge sources such as vector databases” [10]. By combining these two types of knowledge, RAG systems can generate text that is both fluent and factually grounded.

The core idea of RAG is to retrieve information at the moment when a question is asked. Instead of relying only on what the model learned during training, the system searches external sources and retrieves relevant content. This content is then used during generation. This process improves accuracy and helps the system provide updated information. It also avoids the need for retraining the model every time new knowledge appears. As one study notes, “by supplying LLMs with retrieved relevant factual information, the hallucination problem can be alleviated to some extent” [37].

The RAG approach builds on earlier work that combined retrieval methods with neural models. The term “retrieval-augmented generation” was introduced in research that linked a neural retriever with a sequence-to-sequence generator. In these systems, documents are represented as dense vectors. The system retrieves the most relevant documents and then generates text based on them. The original RAG model included “a pre-trained seq2seq model” as parametric memory and “a dense vector index” as non-parametric memory [16]. This hybrid design allows the system to use external information

while still generating natural language responses.

Over time, RAG methods have developed further. Early systems followed a simple retrieve-and-generate pipeline. They retrieved documents once and then generated an answer. Newer systems use more dynamic retrieval strategies. Some systems retrieve information multiple times during the generation process. This allows the model to refine its answer step by step. Research explains that “retrieval augmented LMs commonly use a retrieve-and-generate setup where they retrieve documents based on the user’s input, and then generate a complete answer conditioning on the retrieved documents” [13]. This shows how retrieval and generation work together.

RAG has quickly become one of the most important approaches for improving the reliability of language models. By separating knowledge storage from text generation, it allows systems to update information more easily. It also supports domain-specific knowledge and flexible adaptation. As one survey concludes, RAG “effectively combines the parameterized knowledge of LLMs with non-parameterized external knowledge bases” [10]. This combination helps address several limitations of standalone models. Because of this, RAG has become a central idea in modern AI system design.

### 2.3.3 Architecture of RAG Systems

A typical RAG system is made of several components that work together as a pipeline. These components include document collections, embedding models, vector databases, retrieval modules, and generation modules. Each part has a specific role. The architecture is modular. This means that different implementations can be used for each component depending on the application.

RAG systems depend on external knowledge sources. These sources may include document collections, knowledge bases, or structured databases. They contain the information that the system retrieves when answering a query. In many cases, documents are divided into smaller sections or chunks. This makes retrieval faster and more precise. Each chunk is indexed using embeddings or other retrieval methods. External knowledge provides non-parametric memory for the system. This memory is separate from the language model. It can be updated without retraining the model. Research notes that non-parametric knowledge “empowers LLMs to access and leverage the latest or domain-specific information” [10]. Because of this separation, RAG systems can stay up to date as new data is added.

To support retrieval, documents and user queries are converted into vector representations. These vectors capture semantic meaning. Similarity between vectors can then be measured. Dense retrieval methods are often used. In these methods, both queries and documents are encoded into high-dimensional vectors. Vector databases store these embeddings. They support fast similarity search. When a user sends a query, the system converts the query into a vector. It then searches for the most similar document vectors. This search often uses techniques such as approximate nearest neighbor search or maximum inner product search. Foundational research explains that the retriever “provides latent documents conditioned on the input,” and the generator then uses them to produce the final output [16].

The retrieval module selects relevant documents based on the user query. This process

may include several steps. These steps can include query processing, document retrieval, reranking, and filtering. A common workflow retrieves candidate documents first. Then it ranks them by relevance. Finally, it selects the top results. These results are passed to the generation module as context. Research describes a typical workflow as including steps such as “query classification... retrieval... reranking... repacking... [and] summarization” [36]. Each step can be implemented differently depending on the system. Some systems use dense retrieval. Others use sparse or hybrid retrieval methods.

The generation module is usually a large language model. It produces the final response. It takes both the user query and the retrieved documents as input. The retrieved documents provide context and factual grounding. In many systems, retrieved text is added to the query to form a prompt. This prompt is then given to the language model. The model generates an answer based on the combined input. This allows the model to use external knowledge directly when producing text.

A typical RAG pipeline follows several steps. First, the system receives a user query. Second, it converts the query into a vector representation. Third, it retrieves relevant documents from the external database. Fourth, it combines the retrieved documents with the query. Finally, it generates a response using the language model. This pipeline connects retrieval and generation in one system. Survey literature explains that “given an input query, the retriever identifies relevant data sources, and the retrieved information interacts with the generator to improve the generation process” [39]. Because the design is modular, different components can be replaced or improved. This makes the architecture flexible and suitable for many applications.

### 2.3.4 Technical Components

Dense retrieval is a central part of many RAG systems. It uses neural embedding models to represent queries and documents in the same vector space. Each text is converted into a numerical vector. Similarity between vectors is then used to find relevant documents. One widely used method is Dense Passage Retrieval (DPR). DPR encodes text into dense vectors so that similar texts are close to each other in vector space. Dense retrieval improves semantic matching. It works better than simple keyword search in many cases. Traditional keyword methods rely on exact word matches. Dense retrieval looks at meaning instead of exact words. This allows the system to find documents that are related in meaning even if they use different wording. This is useful for natural language queries and complex information needs.

After embeddings are created, similarity search is used to retrieve relevant documents. The system compares the query vector with document vectors. It then selects the most similar ones. Common similarity measures include cosine similarity and inner product. Vector databases are designed to perform this search efficiently. They can handle large numbers of vectors. To make search faster, many systems use approximate nearest neighbor methods. These methods allow quick retrieval even from large datasets. Similarity search is very important because retrieval quality affects generation quality. If the system retrieves irrelevant documents, the model may generate incorrect answers. Because of this, improving retrieval accuracy is an active research area.

After retrieval, the system builds a prompt for the language model. This prompt

combines the user query with the retrieved documents. The prompt is then passed to the language model. The model generates a response based on this input. The way retrieved documents are included in the prompt matters. Some systems simply concatenate the documents. Others summarize or filter them first. The length of the prompt is also important. Language models have limits on how much text they can process at once. If too much text is included, some information may be lost. Good prompt design ensures that the most relevant information is included while staying within limits.

The integration between retrieval and generation can be done in different ways. In simple systems, retrieval happens once before generation. The retrieved documents are then used to generate the final answer. In more advanced systems, retrieval may happen several times. The model may request more information while generating its response. This is called iterative or active retrieval. Research shows that “augmenting LMs with retrieval components that look up relevant information from external knowledge resources is a promising direction to address hallucination” [13]. Some systems also use joint training. In these cases, the retriever and generator are trained together. In other systems, they are trained separately and combined later. The best approach depends on resources and application needs.

### 2.3.5 Applications of Retrieval-Augmented Generation

Retrieval-Augmented Generation has been used in many areas of artificial intelligence, natural language processing, and information retrieval. Its main strength is that it combines external knowledge with generative models. Because of this, it works well in tasks where accurate and up-to-date information is important. It is especially useful in knowledge-intensive applications.

One of the most common uses of RAG is open-domain question answering. In these systems, a user asks a question. The system retrieves relevant documents from large collections. It then uses these documents to generate an answer. Early research showed that RAG models “set the state of the art on three open domain QA tasks” [16]. This shows that hybrid systems work well for tasks that require factual knowledge. In question answering, RAG improves both accuracy and interpretability. Retrieved documents provide evidence for the generated answer. Users can check these documents to verify the information. This increases trust in the system. Research notes that RAG can generate “more specific, diverse and factual language than a state-of-the-art parametric-only seq2seq baseline” [16]. RAG is also used in domain-specific question answering. Examples include medical and scientific systems. In medical settings, RAG can provide clinicians with updated information from trusted sources. One study explains that providing models with relevant documents from reliable collections can reduce hallucinations and outdated information [38]. This makes RAG useful in areas where accuracy is critical.

RAG is increasingly used in conversational AI and knowledge assistants. Chatbots and virtual assistants often need to answer questions using large knowledge bases. RAG allows them to retrieve information dynamically and generate context-aware responses. Modern conversational systems benefit from RAG because they can access updated information without retraining. This is useful in enterprise settings where knowledge changes often. Organizations can connect language models to internal documents and databases. This

allows assistants to provide accurate answers. RAG also improves transparency. When systems show retrieved documents along with generated answers, users can verify the information. This increases trust and accountability.

RAG is widely used in enterprise search. Many organizations store large collections of internal documents. These include reports, manuals, emails, and technical documentation. Traditional search methods rely on keywords. They may miss relevant documents that use different wording. RAG improves search by using semantic retrieval and generation. A RAG system can find relevant documents and then generate summaries or answers. This allows employees to find information quickly. Engineering-focused research explains that “a RAG system involves finding documents that semantically match a query and then passing the documents to a large language model... to extract the right answer” [2]. This supports decision-making and knowledge management.

RAG is also used in scientific research. Researchers often need to search large collections of academic papers. RAG systems can retrieve relevant articles and generate summaries. This helps with literature reviews and information discovery. Scientific RAG systems can combine structured and unstructured data. For example, they may retrieve information from databases and research papers at the same time. This allows researchers to access more complete information. Recent work shows that RAG has already been used in scientific and clinical question-answering systems [38]. The combination of retrieval and generation makes it useful for complex research tasks.

### 2.3.6 Advantages of Retrieval-Augmented Generation

RAG has several advantages when compared to standalone language models. These advantages relate to accuracy, updates, transparency, and flexibility. Because RAG combines retrieval with generation, it can address many limitations of models that rely only on internal parameters.

One of the main advantages of RAG is better factual accuracy. The system retrieves documents from external sources and uses them during generation. This provides factual grounding for the response. It reduces the chance of hallucination. Survey research notes that RAG “has been demonstrated to significantly enhance answer accuracy” and reduce hallucinations, especially in knowledge-intensive tasks [10]. When a model can access external documents, it can generate responses that are more reliable. Retrieved documents provide context and evidence. This allows the system to produce outputs that are more factually correct. In contrast, models that rely only on internal parameters may generate incorrect information when they lack knowledge.

Another important advantage is the ability to update knowledge easily. External databases can be updated without retraining the language model. This allows the system to include new information quickly. Research explains that non-parametric knowledge stored in external sources is “easily updatable” and can include domain-specific or recent data [39]. This feature is useful in areas where information changes quickly. Examples include news, finance, and healthcare. Instead of retraining a large model, developers can update the external knowledge base. The system can then retrieve new information during inference.

RAG also improves transparency. Because the system retrieves documents, users can

inspect the sources used for generation. This allows them to verify the information. It also increases trust in the system. Research highlights that RAG “improves the transparency of LLMs by grounding their reasoning on the retrieved documents” [38]. By linking generated responses to external sources, RAG systems provide some level of explanation. Users can see where the information comes from. This is important in applications that require justification or traceability.

RAG systems are modular. This means that different components can be changed or updated separately. Developers can replace retrievers, update databases, or improve the generator without redesigning the whole system. This flexibility allows systems to be adapted to different domains. The modular design also supports scalability. New data sources can be added easily. Retrieval strategies can be improved. Because of this, RAG can be used in many different applications. It works well in systems that require both flexibility and reliability.

### 2.3.7 Challenges and Limitations

Even though RAG offers many advantages, it also introduces several challenges. These challenges relate to retrieval quality, system performance, evaluation, and complexity. Because RAG combines several components, problems in one part can affect the whole system.

The quality of retrieved documents is very important. If the system retrieves incorrect or irrelevant documents, the generated answer may also be incorrect. External knowledge sources can sometimes include unreliable information. Research notes that retrieved data may contain “noise information even fake news,” which can mislead the model during generation [6]. Because of this, retrieval systems must be designed carefully. Filtering and reranking methods are often used to improve document selection. However, retrieval errors still remain a major challenge. Improving retrieval accuracy is an active research area.

RAG systems require more computation than standalone models. The system must compute embeddings, search for documents, and generate responses. These steps increase latency. In real-time applications, slower responses may affect user experience. The pipeline can include several stages, such as query processing, retrieval, reranking, and generation. Each stage adds processing time. Some studies note that RAG systems “still suffer from their complex implementation and prolonged response times” [36]. Finding a balance between speed and accuracy is important.

Evaluating RAG systems is more difficult than evaluating standalone language models. Both retrieval and generation must be assessed. Traditional metrics for language models focus on text quality. Retrieval metrics focus on document relevance. RAG requires both types of evaluation. Recent research emphasizes that RAG evaluation must consider multiple factors. These include how well the system integrates information and how robust it is to noise. Benchmark studies mention challenges in evaluating “noise robustness, negative rejection, information integration, and counterfactual robustness” in RAG systems [6]. Because of this, evaluation methods are still developing.

RAG systems are more complex than standalone models. They include multiple components such as retrievers, vector databases, and language models. Each component

must be configured and maintained. This increases engineering effort. Developers must make decisions about document chunking, embedding models, and retrieval strategies. One study notes that implementing RAG requires many design choices about indexing and retrieval methods [36]. This complexity can make deployment and maintenance more difficult.

### 2.3.8 Overview and Key Insights

Retrieval-Augmented Generation is an important development in modern artificial intelligence. By combining retrieval methods with generative language models, RAG addresses several key limitations of standalone models. It allows systems to access external knowledge, improve accuracy, and provide updated information.

The RAG approach integrates two types of knowledge. Parametric knowledge is stored in the model parameters. Non-parametric knowledge is stored in external sources such as databases. This hybrid design allows models to generate responses that are grounded in evidence. Foundational research explains that RAG models “combine pre-trained parametric and non-parametric memory for language generation” [16].

The architecture of RAG systems includes several components. These include document collections, embedding models, vector databases, retrieval modules, and generation modules. These components form a pipeline. The pipeline retrieves relevant documents and generates responses based on them. The modular design allows flexibility and scalability.

RAG has been used in many applications. These include question answering, conversational systems, enterprise search, and scientific research. Its main advantages include improved accuracy, updatable knowledge, and better transparency. At the same time, it introduces challenges related to retrieval quality, latency, evaluation, and complexity.

Overall, RAG has become a key paradigm for improving the reliability and usefulness of large language models. By integrating retrieval with generation, it allows AI systems to produce responses that are more accurate, trustworthy, and context-aware. As research continues, RAG is expected to play an increasingly important role in knowledge-intensive AI systems.

# Chapter 3

## Methodology

### 3.1 Data Model

#### 3.1.1 Definition of the Data Model

This section introduces the data model that supports the retrieval-augmented generation (RAG) system developed in this thesis. The goal of the model is to organize different types of documentation related to the city of Kiruna. It also aims to support reliable retrieval and clear answer generation. The model was designed after an initial analysis of the Kiruna case. This analysis reviewed the types of documents, metadata structures, and user needs identified in the example application at the start of the project. The final structure reflects both the characteristics of the collected materials and the functional needs of a RAG-based question-answering system.

The main goal of the data model is to connect three layers. These layers are users and interactions, document-based knowledge, and retrieval processes. The model must store documents and metadata in a consistent way. It must support semantic retrieval at a detailed level. It must also show clearly which sources are used for each generated answer. For this reason, the class structure includes entities for users, documents, knowledge segments, retrieval records, and extended metadata. The model is designed to be flexible. It should support different document types and allow future extensions.

The user layer includes two main roles. These roles are residents and urban planners. Both roles represent possible users of the system, but they reflect different needs. Residents may search for information about the relocation of Kiruna. They may look for planning decisions or changes in infrastructure. Urban planners represent professional users. They may consult documents and also contribute new documentation. Each user entity includes basic authentication attributes. These include name, email, and password hash. Assigning roles allows the system to separate public access from professional functions. Both user types can send queries to the system. However, the distinction between roles supports future features such as role-based access control and document management tasks.

User interaction with the system is represented by the Chat Interaction class. Each interaction represents one question-and-answer exchange. The class stores a unique identifier for the interaction. It also stores the user's question and the generated answer.

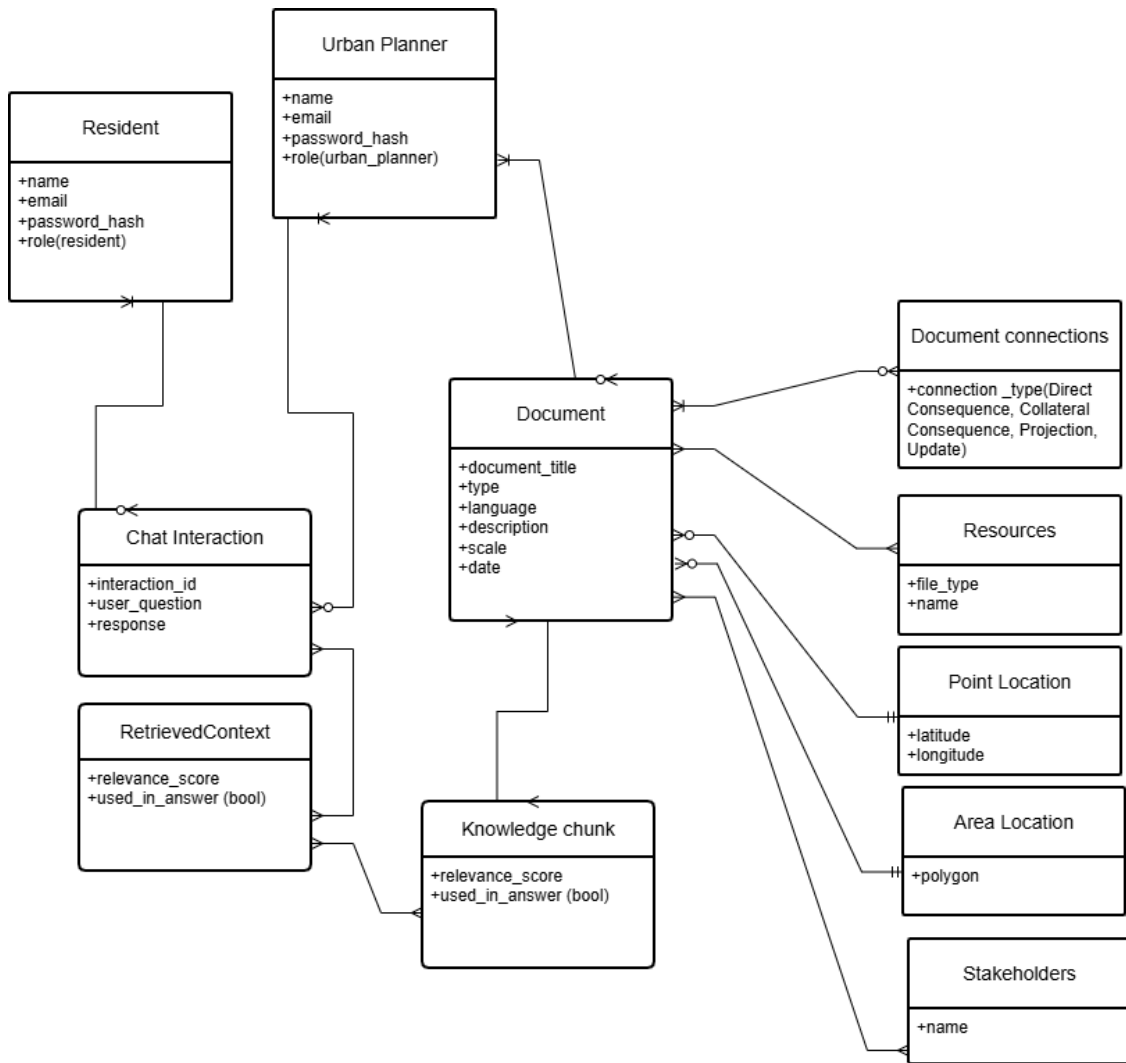


Figure 3.1. Entity-relationship diagram

This structure allows the system to keep a history of queries and responses. Keeping this history is important for evaluation, reproducibility, and debugging. By linking each interaction to a user, the system can also study usage patterns. It can later support features such as conversation history or follow-up questions.

At the center of the model is the Document class. This class represents the main knowledge objects in the system. Each document may be a report, map, research paper, municipal record, or similar source related to Kiruna. The class includes attributes such as title, type, language, description, scale, and date. These attributes were chosen based on earlier analysis of document types and metadata standards. Document type supports classification into planning reports, spatial data, images, and other categories. Language helps determine processing methods and user expectations. Scale and date are especially

important in the Kiruna context. Spatial extent and time sequence are key for understanding the relocation process. The Document class serves as the main reference point for traceability. It ensures that all retrieved content can be linked back to its original source.

To support semantic retrieval, documents are divided into smaller units. These units are called knowledge chunks. The Knowledge Chunk class represents these segments. Each chunk comes from a page or section of a document. It is used as the basic unit for embedding and similarity search. This structure matches the retrieval pipeline used in the system. Document pages are split into overlapping text segments to keep context. Each chunk is linked to its parent document. This allows the system to show source references when generating answers. The model also includes attributes for relevance and usage in answers. These attributes support evaluation and transparency in the retrieval process.

The Retrieved Context class links chat interactions with knowledge chunks. For each user query, the retrieval component selects the most relevant chunks from the document collection. These selected chunks are stored as retrieved context entries. Each entry includes a relevance score. It also includes an indicator showing whether the chunk was used in the final answer. This structure allows the system to record which pieces of information were retrieved and which were actually used to generate the response. Recording this information is important for transparency. It also supports evaluation of system performance. With this design, it is possible to show that answers are based on specific sources rather than general model knowledge.

In addition to textual content, the data model includes several classes for extended metadata. The Resources class represents files that are linked to a document. These files may include PDF reports, images, maps, or spatial datasets. This design allows one conceptual document to be connected to several digital resources.

The model also includes spatial metadata. Two classes represent spatial references: Point Location and Area Location. The Point Location class stores geographic coordinates for documents linked to specific places. This may include individual buildings or infrastructure points. The Area Location class stores polygon data. It represents larger areas such as districts or planning zones. This distinction reflects the spatial complexity of urban documentation. Some documents refer to single locations, while others refer to larger areas. Including both types of spatial reference supports future functions such as mapping or spatial filtering.

The Stakeholders class represents the actors involved in or mentioned by documents. These actors may include municipal authorities, mining companies, research institutions, or other organizations. Linking documents to stakeholders helps users understand the institutional context. It shows who is responsible for certain decisions or actions. This information can also be used for filtering documents by organization or responsibility.

The model also includes a Document Connections class. This class represents relationships between documents. Documents may be linked through updates, projections, or causal relationships. For example, one planning report may update or replace an earlier one. Another document may describe the results of a decision described in a previous report. In the Kiruna context, planning documents evolve over time. They influence each

other. Explicit connections between documents help reconstruct the sequence of decisions and their impacts.

Overall, the data model provides a structured way to represent heterogeneous urban documentation. It supports the functional needs of the RAG system. Documents and metadata are stored in a consistent format. Retrieval operates on clearly defined knowledge segments. Generated answers can be traced back to specific sources. This structure forms the basis for the system implementation described in later sections. It also allows systematic evaluation of retrieval and generation performance.

### 3.1.2 Use-Case Model of the System

This section explains the use-case model of the proposed system. The diagram shows how different user groups interact with the retrieval-augmented generation platform developed for the Kiruna case. The aim of the use-case model is to clarify how the system works. It also identifies user roles and the main tasks supported by the application. This section complements the data model described earlier. While the data model explains how information is stored, the use-case model shows how this information is accessed and used in practice. The system includes three main actors. These actors are Urban Planner, Resident, and Visitor. Each actor represents a different level of interaction with the platform. They reflect how the system may be used in both professional and public contexts. The Urban Planner represents expert users. These users manage and structure documentation. Residents represent local users. They search for information about relocation, planning decisions, and daily life changes. Visitors represent general users. They may explore information but do not add new content.

The first group of use cases focuses on account and access management. Users can register, log in, and log out of the system. Registration allows users to create an account. Logging in provides access to system features that require authentication. These functions are available to residents and urban planners. They make it possible to connect actions and queries to specific users. Logging out ends a session and supports basic security. Another basic function allows users to view documents stored in the system. This use case gives all actors access to available documentation. It serves as a general starting point for exploring the knowledge base.

The second group of use cases is specific to the Urban Planner role. These use cases focus on creating and organizing data. Urban planners can add new documents to the system. This reflects their responsibility for maintaining and updating the dataset. They can also define new areas and define point locations. These actions allow planners to assign spatial metadata to documents. For example, they can mark affected districts or individual buildings. Another function for planners is the creation of document connections. This allows them to link related documents and define relationships between them. These relationships may include updates, projections, or consequences. Together, these functions support the organization of heterogeneous documentation. They also help keep the knowledge base consistent and structured.

The third group of use cases focuses on information retrieval and question answering. These use cases are available to all actors. They represent the main purpose of the RAG application. Users can trace the relocation history of Kiruna. They can compare

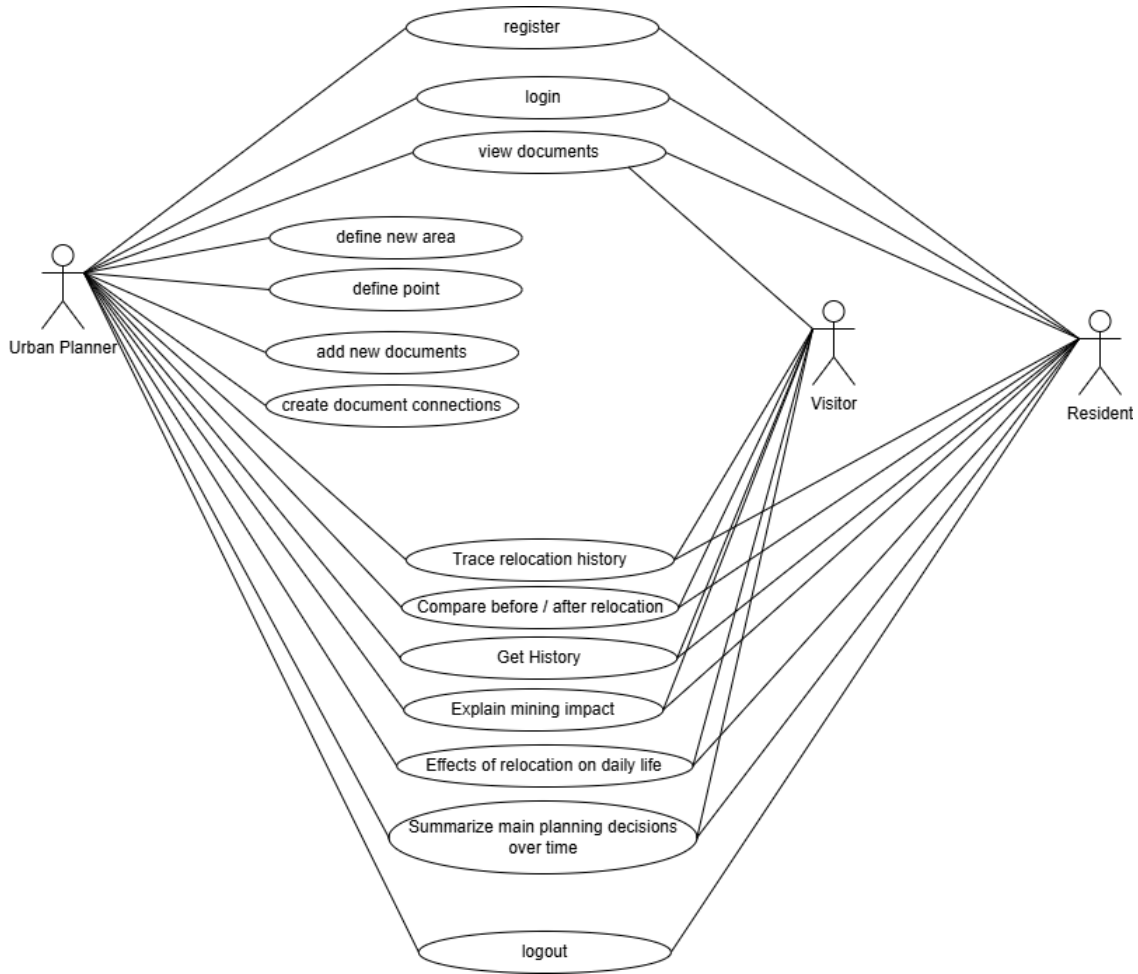


Figure 3.2. Use-case Diagram

conditions before and after relocation. They can retrieve historical information about planning decisions. They can also ask for explanations about the impact of mining. Another function allows users to explore how relocation affects daily life. Users can also request summaries of main planning decisions over time. These use cases reflect typical questions about the transformation of Kiruna. Each action activates the retrieval pipeline. The system searches for relevant documents and generates answers based on verified sources.

The use-case model shows that the system supports both content management and information exploration. Urban planners contribute to the structure and quality of the dataset. Residents and visitors mainly search for information and read results. However, all actors share access to core retrieval functions. This supports transparency and accessibility. The diagram also shows that the system is designed to answer complex questions. It focuses on thematic and historical questions rather than simple keyword searches. This supports the goal of providing contextual and reliable information about long-term urban

change.

In summary, the use-case diagram defines the functional scope of the system. It clarifies the roles of different users. It shows how document management, metadata definition, and retrieval-based question answering are connected in one platform. The model ensures that the technical components described in the methodology match realistic user scenarios. It also supports the overall aim of improving access to complex urban documentation.

## 3.2 Creation of Document Dataset

This section explains how the document dataset for the RAG system was created. It describes how documents were collected, structured, and prepared. The goal of this phase was to turn a heterogeneous set of materials about the relocation of Kiruna into a consistent knowledge base. This preparation stage builds on the data model defined earlier. It focuses on how documents were selected, described, and organized before being used in the retrieval pipeline.

The first step was to identify relevant sources. Documents were gathered from municipal publications, planning reports, academic papers, and publicly available archival material related to Kiruna's urban transformation. The final dataset used in the implementation contains fifteen PDF documents. These documents are written in different languages. Most are in English, five are in Swedish. This multilingual set reflects the international and institutional context of documentation about Kiruna. In addition to written reports, the early research phase also identified maps, images, and visual planning materials. These visual sources were not fully integrated into the current RAG pipeline. However, they influenced the design of the metadata model. They also showed the need for a flexible structure that can support different media types in future development.

Before building the retrieval system, a study of metadata standards and documentation practices was carried out. The purpose of this study was to understand how heterogeneous urban documents could be structured in a consistent and traceable way. Urban planning materials vary in format, scale, and authorship. Without structured descriptions, it is difficult to connect documents or understand their context. It is also difficult to maintain long-term usability. For this reason, several metadata standards were reviewed to see which ones were suitable for the Kiruna case.

Among the approaches examined, the Dublin Core metadata standard and the CIDOC Conceptual Reference Model (CIDOC CRM) were identified as the most relevant. These two standards address different but complementary aspects of documentation. Dublin Core provides a simple and widely used set of descriptive fields. CIDOC CRM provides a more complex conceptual structure. It focuses on relationships between cultural heritage entities.

Dublin Core is a general-purpose metadata standard used to describe digital and physical resources. It includes elements such as title, creator, subject, description, publisher, date, format, language, and identifier. These elements are flexible and can be used across different domains. In the Kiruna dataset, Dublin Core is useful for providing basic descriptive information. Fields such as title, date, language, and description can be applied

to reports, maps, and research papers. The simplicity of Dublin Core makes it suitable for creating a uniform template across different sources. It also supports interoperability and long-term access because it is widely used in digital libraries and archives. However, Dublin Core mainly describes individual resources. It does not fully represent complex relationships between documents, events, and actors.

CIDOC CRM is different. It is a conceptual ontology designed for cultural heritage documentation. It provides a structured way to represent entities such as documents, events, places, and actors. It also represents relationships between them over time. This approach is useful when understanding connections between objects and events is important. In the case of Kiruna, the relocation of the city involves many planning decisions, mining activities, and urban changes. Documents are linked through updates, projections, and consequences. CIDOC CRM allows these links to be represented clearly. For example, a planning report can be connected to a decision event, a location, and responsible stakeholders. This type of semantic structure supports long-term research and archival use. However, CIDOC CRM is complex and requires detailed modeling. For this prototype, a full implementation was not possible. Still, its concepts influenced the design of document connections and stakeholder relationships in the data model.

In addition to these two main standards, other metadata approaches were briefly reviewed. Spatial metadata standards were considered for describing geographic information such as maps and planning zones. Archival metadata schemas were also examined. These help with describing historical documents and institutional records. These approaches showed the importance of including attributes such as spatial extent, scale, and responsible organizations.

### 3.2.1 Example of Dublin Core Metadata Implementation

After reviewing several metadata standards, Dublin Core was selected as the primary schema for describing the collected documents. A set of example documents was then manually annotated using Dublin Core fields in order to test its suitability for urban planning and relocation-related materials.

The example shown in Table 3.1. illustrates the metadata description of an urban planning proposal related to the construction of a new hospital area in Kiruna. The document is represented as an image file and includes information about its creator, contributors, thematic focus, and spatial planning context. The title and description fields provide a concise summary of the document's content, while subject keywords support thematic categorization. The language and format fields describe technical characteristics relevant for processing and retrieval. The source and rights fields indicate the institutional origin and ownership of the material.

This example demonstrates how Dublin Core can be adapted to describe urban planning and heritage-related documentation in a consistent manner. Although the implemented RAG prototype primarily relies on textual summaries for retrieval, the metadata template provided a structured understanding of each document.

Although a structured metadata examples were prepared, the implemented RAG prototype focuses primarily on textual content summaries for retrieval. After evaluating different approaches, summaries were selected as the main representation for embedding



Figure 3.3. Development Plan for New Hospital Area

<b>FIELD</b>	<b>VALUE</b>
Title	Detailed Development Plan for New Hospital Area in Kiruna
Creator	Municipality of Kiruna
Contributor	Region Norrbotten; LKAB
Date	2025
Type	Urban planning proposal
Description	Proposal for a new hospital including parking and a helicopter landing pad, with preservation of industrial remains from the Tuolluvaara Mine
Subject	Healthcare infrastructure; urban planning; industrial heritage
Language	Swedish
Format	image/png
Source	Kiruna Municipality
Rights	© Kiruna Municipality

Table 3.1. Manually assigned Dublin Core metadata

and similarity search. This decision was made to simplify the pipeline and to focus on retrieval performance within the available time and technical constraints. Summaries provide a compact representation of each document’s main content and can be embedded efficiently. Nevertheless, the earlier metadata work remains relevant. It provides a conceptual structure for future system extensions and supports the long-term goal of building a more comprehensive archival and retrieval system.

The final dataset used in the RAG system consists of fifteen curated PDF documents related to the urban transformation and relocation of Kiruna. Each document was reviewed and summarized. These summaries were used as the basis for embedding and

retrieval. For implementation purposes, each document was assigned a simple identifier (paper0, paper1, etc.). The dataset includes academic articles, planning studies, journalistic texts, and student theses written in English, Swedish, and Italian. This multilingual and multi-format composition reflects the diversity of documentation produced around Kiruna’s relocation process. The documents included in the dataset are listed in the Table 3.2. The following summaries are examples that illustrate how individual documents were summarized and prepared for integration into the RAG pipeline.

**Summary of paper0** This paper examines how the concept of landscape shapes decision-making in the relocation of Kiruna, a mining town in northern Sweden. It frames landscape as a sociotechnical imaginary through which political, economic, environmental, and cultural values are negotiated. By analyzing the Kiruna 4-Ever masterplan and the Green/Blue Infrastructure project, the study identifies three competing narratives: a landscape of sacrifice dominated by extractive interests, a mono-functional technical landscape structured around mining systems, and a multi-functional landscape seeking social, ecological, and civic balance. The paper argues that planning documents and design tools leave material traces on the ground and play a critical role in mediating future urban trajectories amid ongoing extraction and uncertainty.

**Summary of paper1** This doctoral thesis investigates processes of heritagisation in built environments undergoing structural transformation, focusing on the urban relocation of Kiruna, Sweden. Conceptualising heritage as socially constructed and negotiated, the study follows how officially designated built heritage is reaffirmed, reinterpreted, or dismissed during the planning process between 2004 and 2015. Drawing on planning documents, media reporting, and observations, it shows how competing discourses of conservation, urban development, and architectural production shape decisions about historic buildings. The thesis introduces the concepts of re-heritagisation and de-heritagisation to describe how heritage meanings shift under conditions of urban change, and argues that heritage is continuously redefined through planning practice rather than preserved as a fixed category.

### 3.3 RAG Pipeline Architecture and Execution

This section explains the architecture and execution flow of the retrieval-augmented generation (RAG) pipeline developed in this thesis. Figure 3.4 illustrates the overall architecture of the system, including the interaction between the user, the Python RAG pipeline, the document dataset, and the language model. After describing the overall architecture, the detailed execution steps of the pipeline are shown in Figure 3.5, which illustrates the sequential processing stage. The pipeline supports access to heterogeneous documentation about the transformation of the Kiruna city. It is designed to retrieve information from a curated collection of texts and generate answers grounded in verified sources. By combining semantic retrieval with a local language model, the system allows users to ask questions about Kiruna and receive responses based on retrieved evidence rather than general model knowledge.

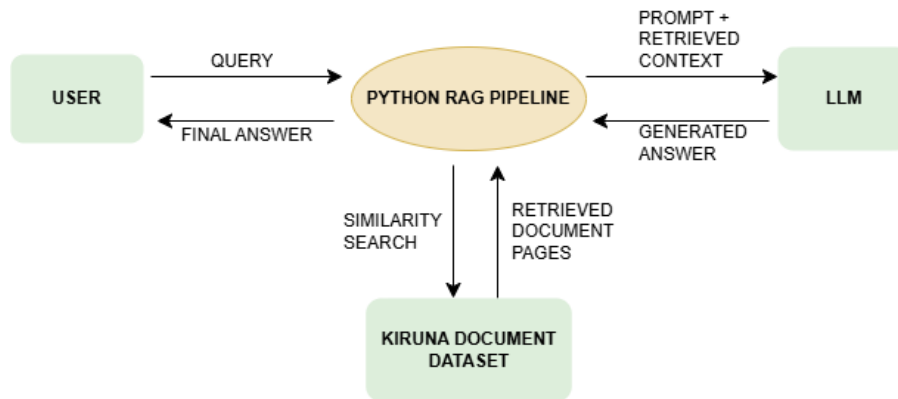


Figure 3.4. Architecture of the implemented RAG system.

The main goal of the pipeline is to find document segments that are most relevant to a user’s question. The implementation has several sequential stages. These stages include document indexing, query embedding, document retrieval, page-level ranking, context construction, and answer generation. Each stage performs a specific task and passes its output to the next step.

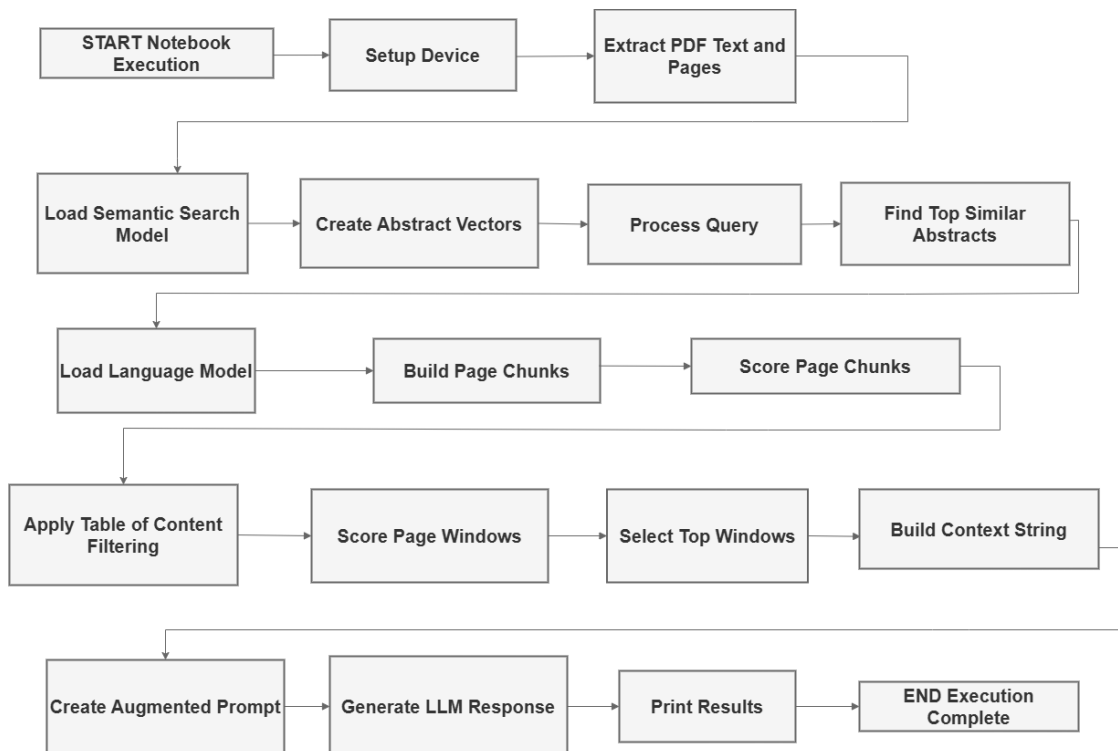


Figure 3.5. Execution pipeline of the implemented RAG system.

### 3.3.1 Document Indexing

The system works with a curated dataset of documents related to Kiruna. These documents were collected during the earlier document investigation phase of the thesis. They include planning reports, research papers, and other public texts describing the relocation and urban transformation of the city. Since the dataset contains long documents with different structures, preprocessing is needed to support efficient retrieval.

Each document is reviewed and a summary is extracted. Summaries provide a compact representation of the document's content. For this reason, they are used for first-level retrieval. All summaries are converted into vector embeddings using a transformer-based sentence-embedding model. These embeddings capture semantic similarity between texts. They allow comparison based on meaning rather than only keywords.

The abstract embeddings form an index. This index allows fast comparison between user queries and documents. Indexing is done once during system initialization. It provides the foundation for later retrieval steps.

### 3.3.2 Query Processing and Document Retrieval

When a user submits a question about Kiruna, the query is converted into an embedding vector. The same sentence-embedding model is used. Using a shared embedding space ensures that queries and documents can be compared consistently. The system calculates cosine similarity between the query embedding and all abstract embeddings. Documents with the highest similarity scores are considered most relevant. The indices of the top documents are returned. These documents are selected for deeper analysis. This step reduces the search space. It ensures that only relevant documents are processed further.

Abstract-level retrieval is efficient. However, it does not identify exact sections inside documents. For this reason, the pipeline continues with page-level analysis.

### 3.3.3 Page Segmentation and Chunk Embedding

The selected documents are processed page by page. Each page may contain large amounts of text. To support retrieval, each page is divided into smaller overlapping segments. In this implementation, pages are split into chunks of about 250 words. Some overlap is included between chunks. Overlap helps preserve meaning across boundaries. It ensures that important information is not lost when text is divided.

Each chunk is converted into an embedding vector using the same transformer model. Cosine similarity is then calculated between each chunk and the query embedding. For each page, the highest similarity score among its chunks becomes the page's relevance score. This method allows the system to detect relevant content even if only part of a page matches the query. Chunk embeddings are generated in batches. This improves efficiency when processing long planning reports and other documents related to Kiruna.

### 3.3.4 Page Scoring and Filtering

Similarity scores may still include irrelevant pages. For example, table-of-contents pages often contain repeated keywords but little useful explanation. To address this, the system

applies heuristic filtering. Pages that look like table-of-contents sections are given a score of zero. These pages often contain short lines, page numbers, and leader dots. Very short pages that lack meaningful text may also be excluded.

After filtering, each remaining page has a relevance score based on its best-matching chunk. However, important information in planning documents is often spread across multiple pages. Selecting only single pages may break the context. For this reason, the pipeline groups pages into larger sections.

### 3.3.5 Window-Based Page Selection

After page-level scores are calculated and filtered, the pipeline applies a window-based ranking method. Relevant information in documents is often spread across several pages. It is rarely contained on a single page. Selecting only the highest-scoring page can result in broken or incomplete context. This can reduce the quality of the generated answer.

To solve this, the system groups nearby pages into sliding windows. Each window contains a small number of consecutive pages. The relevance scores of the pages in each window are added together to produce a window score. All windows are then ranked based on their total scores. The highest-ranked windows are selected as the most relevant parts of the document.

This approach ensures that complete and coherent sections are retrieved instead of isolated fragments. It improves the continuity of the context given to the language model. It also increases the chance that full explanations are included in the final prompt. The selected pages are kept in their original order so that the logical structure of the document is preserved.

### 3.3.6 Context Construction

Once the most relevant windows are selected, the system builds the textual context that will be passed to the language model. The selected pages are combined into a single text. They are also labeled with metadata that shows the source document and page number. These labels support traceability. They allow users and researchers to identify where specific information comes from within the Kiruna documents.

The system also enforces a maximum context length. This ensures compatibility with the language model’s input limits. If the selected text is too long, it is shortened while keeping the highest-ranked content. This ensures that the most relevant information is included within the allowed context size. If no pages remain after filtering and ranking, the system uses a fallback method. In this case, a default part of the document collection is used as context. This fallback may be less precise. However, it ensures that the model still receives background information and can generate a useful answer.

### 3.3.7 Prompt Construction and Response Generation

The final stage of the pipeline is response generation. The constructed context is combined with the user’s query. A system instruction is also added. This instruction guides the model to produce answers that are grounded in the retrieved documents. It encourages the model to rely on provided sources and avoid unsupported claims.

The full prompt includes the system instruction, the user query, and the retrieved context. This prompt is tokenized and sent to a local instruction-tuned language model. The model generates a response based on the context. Only the newly generated tokens are returned to the user. The original prompt text is not included in the output. The result is an answer grounded in Kiruna-related documents from the dataset.

This design ensures that generated responses are based on verified sources. It allows users to ask detailed questions about the relocation of Kiruna, planning decisions, and urban transformation.

### 3.3.8 Summary

The RAG pipeline developed in this thesis combines semantic retrieval and language model generation. It improves access to fragmented documentation about Kiruna. The system first identifies relevant documents using abstract-level similarity. It then performs page-level and chunk-level analysis to find specific content. Irrelevant pages are filtered. Relevant pages are grouped into coherent windows. The selected context is passed to a language model. The model then generates responses grounded in retrieved material.

This architecture supports structured exploration of complex urban documentation. It shows how retrieval-based methods can support research on cities undergoing long-term transformation. The modular design also allows future extensions.

## 3.4 Implementation Challenges and Design Decisions

During the development of the retrieval-augmented generation (RAG) system for Kiruna-related documents, several technical challenges appeared. Most of these challenges were related to memory limits, retrieval accuracy, multilingual content, and model configuration. Solving these issues required repeated testing and adjustments to the system design. This section explains the main problems that occurred during implementation. It also describes the design decisions that were made to solve them.

### 3.4.1 Memory Limitations and Context Size

At an early stage, the full text of each selected PDF document was directly added to the prompt sent to the language model. The goal was to give the model as much context as possible. However, the documents in the Kiruna dataset are often long planning reports and research papers. When full documents were passed to the model, the system often produced CUDA out-of-memory errors. These errors occurred because the GPU did not have enough memory.

As an initial solution, a limit of 4000 characters was applied to each document before sending it to the model. This prevented memory overflow and allowed the system to generate responses without crashing. However, this solution created a new problem. Only the beginning of each document was processed. Relevant information later in the text was ignored. This reduced the quality and completeness of the answers, especially for long reports where key information appears in later sections.

To solve this issue, the retrieval process was redesigned. A chunk-based method was introduced. Instead of sending full documents to the model, each PDF was divided into smaller text segments called chunks. The system computed an embedding vector for the user query and for each chunk. Cosine similarity was then used to measure semantic similarity between the query and each chunk. Only the most relevant chunks were selected and sent to the language model.

This redesign greatly reduced memory usage and removed CUDA memory errors. It also allowed the system to consider all parts of long documents during retrieval. Only the most relevant sections were included in the final prompt. This improved both efficiency and answer quality.

### 3.4.2 Irrelevant Section Retrieval and TOC Filtering

After chunk-based retrieval was implemented, another issue appeared. Some queries were matched with irrelevant parts of documents. This happened especially with table-of-contents pages. These pages often contain repeated keywords and page numbers. Because of this, they sometimes received high similarity scores even though they contained little useful information.

To solve this problem, a filtering step was added. The system was designed to detect pages that look like tables of contents. Simple heuristic rules were used. These rules looked for patterns such as short lines, repeated numbers, and leader dots. Pages that matched these patterns were given a score of zero and excluded from retrieval. This filtering step improved the relevance of retrieved content. It reduced noise in the context passed to the language model.

### 3.4.3 Embedding Model Selection for Document Retrieval

Another important challenge was choosing the right embedding model. In early tests, the system used a general-purpose BERT model (bert-base-uncased) to create embeddings. However, this model is not designed for semantic search. Because of this, many queries were matched with the wrong documents. Around ninety percent of queries matched unrelated summaries. This caused inaccurate responses.

To improve retrieval accuracy, the embedding model was replaced with a model designed for semantic search tasks (sentence-transformers/multi-qa-mpnet-base-dot-v1). This change improved document-level matching. However, new issues appeared when working with multilingual documents.

### 3.4.4 Multilingual Retrieval Challenges

After improving document-level retrieval, new problems appeared because the Kiruna dataset is multilingual. Many documents in the collection are written in Swedish. At the same time, user queries are often written in English. When an English query was compared with embeddings created from Swedish documents, the system often failed to find the correct sections. Sometimes the correct document was selected at the abstract level, but the system could not find the most relevant passages inside the Swedish text.

To solve this issue, the embedding model was changed again. A multilingual sentence-transformer model (sentence-transformers/paraphrase-multilingual-mpnet-base-v2) was introduced. This model supports cross-lingual similarity. It allows a query in one language to match documents written in another language. After this change, the system was able to retrieve relevant passages from Swedish documents even when the query was written in English.

Additional testing showed another pattern. Response quality improved when the query language matched the language of the source document. For example, when Swedish queries were used with Swedish documents, the generated answers were more precise. They were also more accurate in context. This showed that language alignment between query and source text can improve results. However, the multilingual embedding model ensured that the system still worked when users asked questions in English.

### 3.4.5 Matching Relevant Document Sections

Even after selecting the correct document, the system initially had trouble finding the most relevant section inside the document. This problem was clear in long planning reports. Information about the same topic could appear across several pages. The chunk-based method helped, but further refinement was needed.

The final solution combined several techniques. Chunk-level similarity scoring was used. Table-of-contents filtering removed irrelevant sections. A sliding-window method grouped adjacent pages together. By combining these methods, the system was able to retrieve coherent sections instead of isolated sentences. This improved the quality of the context sent to the language model. It also reduced cases where answers were based on incomplete information.

### 3.4.6 Model Quantization and Performance Trade-offs

Another challenge was related to response time and computational efficiency. To reduce memory usage and speed up generation, an experiment was performed using 4-bit model quantization. The BitsAndBytes library was used to load the model in 4-bit precision. This reduced memory consumption and improved speed. However, testing showed that answer quality decreased when the model ran in 4-bit mode. Responses became less precise. In some cases, they were less coherent. Since the main goal of the system is to provide reliable information about Kiruna, answer quality was considered more important than small improvements in speed.

For this reason, the quantized configuration was not used in the final system. Instead, the model was loaded in 16-bit floating-point precision. This required more memory, but it produced better answers and more stable performance. The final design therefore prioritizes accuracy and reliability over minimal gains in speed.

### 3.4.7 Summary of Design Decisions

The development of the RAG system required several iterations. These iterations addressed challenges related to memory use, retrieval accuracy, multilingual content, and

model performance. The first approach of sending full documents to the model was replaced by a chunk-based retrieval strategy. This avoided memory overflow and improved scalability. Filtering methods were added to remove irrelevant sections such as tables of contents. Embedding models were tested and replaced to improve semantic matching and support multilingual retrieval. Model quantization was also tested, but it was not used in the final system because it reduced answer quality.

Together, these design decisions improved the reliability of the system. The final pipeline can retrieve relevant information from long and heterogeneous documents. It can also generate responses grounded in verified sources. The challenges encountered during development show the importance of testing and model selection when building retrieval-based systems for domain-specific and multilingual datasets.

ID	Document Title and Source
paper0	Which Landscape? Material Traces of an Integrated Design and Research Approach in Kiruna — Valeria Federighi; Taneha Kuzniecowa Bacchin; Kirthan Shekar
paper1	Adaptation and Relocation of Built Heritage: What Can We Learn from the Urban Transformations of Swedish Mining Towns? — Andrea Luciani; Jennie Sjöholm
paper2	Ideology, Environment and Forced Relocation: Kiruna – A Town on the Move — Bo Nilsson
paper3	Authenticity and Relocation of Built Heritage: The Urban Transformation of Kiruna, Sweden — Jennie Sjöholm
paper4	Heritagisation, Re-Heritagisation and De-Heritagisation of Built Environments: The Urban Transformation of Kiruna, Sweden — Jennie Sjöholm
paper5	Challenging the Set Mining Path: Agency and Diversification in the Case of Kiruna — The Extractive Industries and Society, Linda Stihl
paper6	Do Voluntary Corporate Actions Improve Cumulative Effects Assessment? Mining Companies’ Performance on Sami Lands — Rasmus Kløcker Larsen; Carl Österlin; Laura Guia
paper7	Memorialisation and its Denial: Slow Resistance through Derealisation in Kiruna, Sweden — Journal of Political Power, Eric Boyd
paper8	Smart(en)ing the Arctic City? The Cases of Kiruna and Malmberget in Sweden — Agatino Rizzo; Jennie Sjöholm; Andrea Luciani
paper9	Reframing Kiruna’s Relocation—Spatial Production or a Sustainable Transformation? — Aslı Tepecik Diş; Elahe Karimnia
paper10	How to Move a City: Mapping the Process of Kiruna Relocation — Giovanni Ceselli; Valeria Federighi
paper11	Konsten att flytta en stad – en studie av hur kommunikationerna i Kiruna påverkas av stadsomvandlingen — Linus Petersson
paper12	Att flytta en mönsterstad — Jennie Sjöholm
paper13	Uppdrag: Att flytta en stad – en studie av planeringsprocessen i samband med Kiruna stadsomvandling — Christin Gimberger; Karin Norberg
paper14	Ett nytt läge för Kiruna — Magnus Lindgren
paper15	Ett Kiruna i förändring – En studie om hur utflyttade Kirunabor upplever stadsflytten i Kiruna — Isak Kautto

Table 3.2. List of documents included in the Kiruna dataset



# Chapter 4

## Evaluation

### 4.1 Evaluation Methodology

The purpose of the evaluation is to determine whether the retrieval-augmented generation (RAG) system developed in this thesis improves the quality of answers to questions about the transformation of Kiruna. The system combines a document retrieval component with a language model. Therefore, the evaluation focuses on the quality and informativeness of the answers generated when domain-specific documents are included in the generation process. In particular, the evaluation compares responses produced by a language model without retrieval support and responses generated by the RAG system that retrieves relevant information from a curated document collection.

The evaluation is designed to represent a realistic information-seeking situation. Users who are interested in the relocation and urban transformation of Kiruna may ask questions about planning processes, historical developments, social effects, or the role of mining activities in shaping the city. The goal of the RAG system is to answer such questions using verified information retrieved from relevant documents. For this reason, the evaluation examines how well the system retrieves relevant text and how effectively it uses this information to produce informative and context-based answers.

The evaluation process consists of three main stages. First, a set of documents related to the relocation of Kiruna is selected and used as the knowledge base for the system. These documents include academic publications and research studies that discuss different aspects of Kiruna's transformation. The selected documents reflect the heterogeneous and fragmented nature of the available information about the city. The definition of this document collection is described in Section 4.2.

Second, a list of evaluation questions is created based on the content of the selected documents. These questions represent realistic queries that users might ask when searching for information about Kiruna's relocation and urban development. Each question corresponds to information that can be found in one or more documents in the dataset. This allows the correctness of the generated answers to be verified using the available sources. The development of these test questions is presented in Section 4.3.

In the third stage of the evaluation, answers are generated using two different system configurations. The first configuration uses the language model without retrieval

augmentation. In this case, the model produces responses based only on its pre-trained knowledge. The second configuration uses the RAG pipeline developed in this thesis. This pipeline retrieves relevant passages from the document collection and provides them as context to the language model before generating an answer. By comparing the responses produced by these two configurations, it is possible to examine the impact of document retrieval on the quality of the generated answers.

The quality of the responses is evaluated through human assessment. People who evaluate responses reviewed two answers for each question one of them is generated by the standard language model and the other one is generated by the RAG system. The answers are presented as two options, and evaluators choose the response they consider better. Their choice is based on criteria such as correctness, completeness, and relevance to the question.

## 4.2 Definition of Test Documents

The evaluation of a retrieval-augmented generation (RAG) system requires a clearly defined document corpus. This corpus should represent the knowledge domain addressed by the system. The goal of the developed system is to provide accurate information about the transformation and relocation of the city of Kiruna. For this reason, the evaluation must rely on documents that reflect the different types of knowledge produced about this process. The relocation of Kiruna has been widely discussed in academic research, urban planning studies, cultural heritage analysis, and socio-economic research. As a result, information about the city's transformation appears in many publications and institutional reports. This fragmentation makes the information difficult to access through standard search methods.

To simulate a realistic information retrieval scenario, a set of documents was selected from publicly available academic and institutional sources. These documents form the knowledge base used by the RAG system. They also provide the foundation for evaluating how effectively the system retrieves relevant information and generates accurate answers. The selected materials represent different perspectives on the relocation of Kiruna. These perspectives include urban planning, heritage studies, mining impacts, social change, and environmental governance. The diversity of topics reflects the complexity of Kiruna's transformation and creates an appropriate context for evaluating retrieval-based question answering.

The evaluation corpus consists of sixteen documents obtained from publicly accessible academic publications, conference proceedings, and university theses. These materials were selected because they contain detailed descriptions of the historical background of Kiruna and the processes behind its relocation. They also discuss planning strategies, social implications, and spatial transformation related to the city. Each document offers a specific perspective on the transformation process. At the same time, each provides contextual information that the system may retrieve when answering user questions. Table 3.2. summarizes the documents included in the evaluation dataset.

The selected corpus includes both English and Swedish publications. The English-language documents mainly consist of peer-reviewed journal articles and conference papers. These publications analyze different aspects of Kiruna’s relocation. They address topics such as the relocation of cultural heritage, urban transformation, planning processes, and the social and political effects of mining-driven spatial change. The Swedish-language documents mainly include academic theses and planning studies. These works often describe local processes, infrastructure changes, and community responses in greater detail. Including documents in both languages ensures that the dataset reflects the different ways knowledge about Kiruna has been produced and documented.

Another important feature of the dataset is the diversity of thematic perspectives represented in the documents. Several publications focus on heritage and cultural identity issues connected to the relocation of historical buildings and urban spaces. These works discuss how heritage values are preserved, transformed, or contested during the relocation process. Other studies examine the planning and governance aspects of the transformation. They analyze how municipal authorities, mining companies, and regional stakeholders coordinate the relocation process. Additional research explores broader socio-economic and environmental issues. These include the role of the mining industry, the effects on indigenous Sami lands, and the experiences of residents affected by the relocation.

From an information retrieval perspective, this thematic diversity is particularly important. The RAG system must be able to retrieve relevant information across multiple topics and perspectives. Some documents provide descriptive narratives about urban transformation and local experiences. Others present analytical discussions of policy frameworks, environmental governance, or spatial planning strategies. Because of this variation, the textual structures and terminology differ between documents. This diversity creates a realistic challenge for retrieval-augmented systems, which must identify relevant passages even when concepts and terminology vary.

The selected documents also differ in their length and structural organization. Some documents are relatively short research articles. Others are longer reports or academic theses that contain multiple sections, references, and methodological discussions. This variation is relevant for the evaluation process. The system must handle documents of different sizes and locate relevant information in different parts of the text. During pre-processing, the documents were first divided into pages. These pages were then split into smaller overlapping text chunks. This approach allows the RAG pipeline to perform semantic search more effectively. Instead of returning entire documents, the system retrieves specific passages that are relevant to a user’s question.

The selection of these sixteen documents does not aim to create a complete archive of all material related to the relocation of Kiruna. Instead, the objective is to construct a representative sample of the available knowledge sources. The selected documents capture different disciplinary perspectives and forms of documentation about the city’s transformation. By including documents from different disciplines, languages, and publication formats, the dataset reflects the fragmented nature of the available information. This fragmentation is one of the reasons why a retrieval-augmented system is needed.

Finally, the curated document set serves as the basis for generating the evaluation

questions used in the later stages of the evaluation process. Each question corresponds to information that can be found in one or more documents within the dataset. This design ensures that the evaluation focuses on questions with verifiable answers in the corpus. It also allows the performance of the retrieval-augmented system to be compared with systems that do not have access to the same document collection.

### 4.3 Definition of Test Questions

To evaluate the performance of the retrieval-augmented generation system, a set of evaluation questions was created using the document corpus described in the previous section. The goal of these questions is to reflect the kinds of information users may look for when trying to understand the relocation and urban transformation of Kiruna. Each question is based on information that appears in the collected documents. This makes it possible to check the answers produced by the system against the original sources.

The evaluation dataset includes thirty questions derived from the sixteen documents in the corpus. Each question is linked to a specific document where the relevant information can be found. This structure helps test whether the RAG system can retrieve information from the correct source and generate an answer based on that document. In comparison, the baseline language model must answer the same questions without access to these documents and relies only on its pre-trained knowledge.

The questions were written to represent different kinds of information found in the literature about Kiruna. They include conceptual explanations, interpretations of research findings, and descriptions of planning processes or social impacts related to the relocation of the city. This variety is important because the RAG system must work with documents that differ in topic, perspective, and writing style.

Several questions focus on cultural heritage and the preservation of historic buildings, which is an important theme in research on Kiruna’s transformation. Some questions examine how relocating historic buildings affects their authenticity and the surrounding historical environment. Others ask how planners decide which elements of Kiruna’s architectural heritage should be preserved or included in the new city centre. These questions are based on heritage-focused research, including studies by Sjöholm and related work on heritage management in Kiruna.

Another group of questions focuses on the social and cultural experiences of the relocation process. These questions look at how residents experience the changes in the city and how they deal with the loss of places that are demolished or relocated. For example, some questions ask why it may be difficult for people to publicly mourn the loss of the old city centre. Others explore how residents try to preserve memories of places that can no longer be physically commemorated. These topics appear in research on social responses to relocation, including studies on memorialisation, place attachment, and community identity.

A further group of questions addresses planning processes and governance. The relocation of Kiruna involves coordination between several actors, including the municipality, the mining company LKAB, and regional stakeholders. Some questions therefore examine how planning documents influence the design of the new city centre. Others consider

how planners balance industrial systems with public space or how time pressure and uncertainty influence planning decisions. These questions focus on the institutional and decision-making aspects of the relocation process.

The evaluation also includes questions about the economic and environmental aspects of the relocation. In particular, they focus on the role of the mining industry in Kiruna’s development. Mining is the main reason for the city’s relocation and strongly affects the region’s economy and infrastructure. For this reason, several questions examine how mining activities influence economic diversification, tourism development, environmental assessments, and relations with indigenous Sami communities. These questions test whether the system can retrieve information about broader socio-economic issues discussed in the literature.

Some questions also address technological and spatial planning topics, such as smart-city initiatives and the spatial development of the new Kiruna. These questions examine whether technological initiatives support sustainable urban development or mainly reinforce existing industrial structures. Other questions explore how maps, planning documents, and other tools help researchers and planners understand the ongoing relocation process.

The thirty questions therefore cover a wide range of topics related to the transformation of Kiruna. Table 4.1 shows how the evaluation questions are connected to the documents from which they were derived.

Question Numbers	Main Topic	Source Documents
Q1, Q10, Q23	Memory, mourning, and memorialisation	paper7
Q2, Q7, Q19, Q21	Authenticity and relocation of heritage buildings	paper3, paper4
Q3, Q17	Mapping and documentation of relocation processes	paper10
Q4, Q16	Smart city initiatives and sustainability debates	paper8
Q5, Q29	Place attachment and young resident’s perceptions	paper15
Q6, Q9	Mining economy and diversification	paper5
Q8	Heritage policy processes	paper1
Q11, Q12	Planning imaginaries and landscape design	paper0
Q13, Q24	Infrastructure and transport systems	paper11
Q14, Q15	Sustainability and transformation frameworks	paper9
Q18	Media narratives and public acceptance	paper2
Q22, Q20	Environmental and social impact assessments	paper6
Q25	Heritage selection in planning	paper12
Q26, Q30	Governance and planning coordination	paper13
Q27, Q28	Spatial planning and relocation decisions	paper14

Table 4.1. Relationship between the evaluation questions and the source documents

Designing the questions in this structured way helps ensure that the evaluation covers multiple aspects of Kiruna’s transformation while maintaining a clear link between

each question and the source material in the dataset. This is especially important when evaluating retrieval-augmented system. It makes it possible to check whether the system retrieves information from the correct document and whether the generated answer reflects the source content accurately.

Another important goal when designing the questions was to avoid very simple factual queries. The questions do not focus only on basic facts. Many of them require explanation or interpretation of issues discussed in the documents. For example, some questions examine the relationship between mining activities and economic diversification. Others ask how ideas about cultural heritage influence planning decisions. These types of questions better reflect the information that researchers or planners may look for when studying the relocation of Kiruna.

Overall, the evaluation questions are designed to test whether the system can retrieve relevant passages from the document corpus and produce informative answers that are specific to the Kiruna case. Because each question is linked to the selected literature, the quality of the answers can be checked against verifiable sources. The answers generated by the baseline language model and the RAG system are compared through the questionnaire described in the following section.

## 4.4 Response Quality Evaluation

### 4.4.1 Qualitative Analysis of Generated Responses

Before conducting the questionnaire-based evaluation, an initial qualitative analysis of the system outputs was carried out. The aim of this analysis was to observe how the retrieval-augmented generation (RAG) system performs in comparison with the language model operating without retrieval support. This step helped identify typical differences between the two configurations. It also provided an early indication of how access to retrieved documents influences the specificity and factual grounding of the generated answers.

To ensure consistent behavior of the language model during the experiments, a base instruction prompt was used in both configurations. This prompt defines general guidelines for generating helpful, honest, and concise responses. The full text of the base instruction prompt is shown below.

**Base Instruction Prompt:**

Provide helpful, respectful, and honest responses to the user’s questions. Avoid harmful, misleading, or unsupported information. Do not invent facts or present claims that cannot be supported. If the available information is insufficient, clearly indicate this instead of guessing. Produce concise answers written in plain text and avoid using bullet points.

In addition to the base instruction prompt, the RAG configuration included further instructions specifying how the model should use retrieved document excerpts when generating answers. These additional guidelines ensure that the generated responses remain grounded in the retrieved material. The complete RAG instruction prompt is shown below.

**RAG Instruction Prompt:**

Provide helpful, respectful, and honest responses to the user’s questions. Avoid harmful, misleading, or unsupported information. Do not invent facts or present claims that cannot be supported. If the available information is insufficient, clearly indicate this instead of guessing. Produce concise answers written in plain text and avoid using bullet points.

When retrieved document excerpts are provided, treat them as the primary source of information. Base the answer only on the retrieved passages and avoid relying on external knowledge that is not supported by them. The model may paraphrase or explain the retrieved text, but key claims must remain grounded in the excerpts. If a small logical step is required, clearly label it as an inference.

The configuration without retrieval augmentation received only the general instruction prompt. Since no document excerpts were provided, the model had to rely entirely on its internal pre-trained knowledge when generating answers. This setup makes it possible to compare responses based on general knowledge with responses grounded in the document corpus used by the RAG system.

To explore the behavior of both configurations, a subset of the evaluation questions was first tested manually. For each question, answers were generated using both the non-RAG configuration and the RAG pipeline. The responses were then examined with respect to their relevance, factual grounding, specificity, and consistency with the information contained in the source documents.

Several patterns emerged from this comparison. In many cases, answers produced without retrieval provided general explanations about mining towns, urban relocation, or sustainability challenges. These responses often contained reasonable background information. However, they rarely referred directly to the specific context of Kiruna. As a result, the answers sometimes appeared generic or only loosely connected to the research questions derived from the documents.

By contrast, the answers generated by the RAG system were more closely aligned with the content of the academic sources. Because the model received excerpts from relevant documents, the responses frequently included references to specific concepts, actors, or processes discussed in the literature. In addition, the structure and argumentation of these answers often reflected the analytical perspective presented in the original research papers.

One example illustrates this difference. A question asked why it is difficult for people to publicly remember and mourn the old Kiruna. The response generated without retrieval described the relocation of the city and discussed general themes such as displacement, nostalgia, and community disruption. Although these explanations are broadly plausible, the answer did not reflect the specific interpretation presented in the referenced study.

The RAG-generated answer, in contrast, followed the interpretation discussed in the corresponding academic article. It explained that public mourning is constrained by social and political dynamics connected to the mining industry. The answer also referred to a broader culture of silence surrounding the relocation process. In addition, it mentioned how the loss associated with the transformation of the city is often internalized within

the community and rarely acknowledged publicly. These elements correspond closely to the arguments developed in the source paper.

A similar difference appeared in responses to a question about smart-city initiatives in Kiruna and their relationship to sustainability and mining interests. The answer generated without retrieval presented a broad discussion of smart-city technologies, renewable energy, and environmental concerns. Although the response included generally plausible arguments, it also introduced inaccurate details and referred to organizations not mentioned in the original study. This suggests that the model attempted to construct a generic explanation rather than drawing on specific evidence.

The RAG-generated response focused instead on the initiatives described in the academic article. For instance, it referred to the establishment of the Kiruna Sustainability Centre and the testing of smart urban technologies. It also discussed tensions between sustainability ambitions and the economic priorities of the mining industry, which is a key theme in the paper. Because the answer was based on retrieved excerpts, it remained closely aligned with the arguments presented in the source material.

Across multiple test questions, this pattern appeared consistently. Answers generated without retrieval often relied on general knowledge about urban relocation, mining economies, or sustainability debates. While these responses were sometimes partly correct, they frequently lacked references to the specific case of Kiruna. In some cases, they also included assumptions that were not supported by the selected documents.

In contrast, answers produced by the RAG system generally showed stronger grounding in the document corpus. The retrieval step allowed the model to incorporate information directly from the academic literature. As a result, the responses were more closely connected to the Kiruna case. This grounding also reduced the occurrence of hallucinated information, since the model was encouraged to rely on the provided excerpts rather than generate unsupported claims.

Another difference was the level of detail in the responses. The answers generated by the RAG system often included more precise descriptions of social dynamics, planning conflicts, and governance processes related to the relocation of Kiruna. These details reflected the analytical approaches used in the source papers. In contrast, answers generated without retrieval were usually more general and rarely engaged with the conceptual discussions presented in the literature.

The qualitative analysis also showed some limitations of the RAG system. In some cases, the retrieved excerpts contained complex academic language or only parts of larger arguments. Because of this, some generated answers included incomplete explanations that needed additional context to be fully understood. This finding highlights the importance of careful document segmentation and retrieval strategies when working with academic texts.

Overall, the preliminary analysis suggests that the retrieval-augmented approach improves the specificity and factual grounding of the generated answers. Access to domain-specific documents allows the model to produce responses that better reflect the academic literature on the transformation of Kiruna. These observations led to the next stage of the evaluation, where the responses from both configurations were compared through a structured questionnaire study.

### 4.4.2 Questionnaire Results

To evaluate the perceived quality of the generated responses, a questionnaire was conducted. In this study, participants compared answers produced by two AI configurations: a standard language model and the retrieval-augmented generation (RAG) system developed in this thesis. For each question, participants were shown two responses. They were asked to choose the answer they considered better based on accuracy, relevance to Kiruna, and level of detail. The following subsections present the results for the evaluation questions included in the questionnaire. For each question, the two generated responses are presented. A graphical summary of participant preferences is also provided.

**Question 1** Why is it difficult for people to publicly remember and mourn the old Kiruna?

Response generated without RAG: The town of Kiruna in Sweden was relocated due to the discovery of iron ore deposits beneath the town, leading to concerns about subsidence and structural damage to buildings. This relocation has made it challenging for residents to maintain a strong sense of community and connection to the old town.

Response generated with RAG: The lack of public mourning in Kiruna is due to the forced relocation of the community, which has led to a culture of silence and denial. Residents are often discouraged from expressing negative feelings about the situation, and the focus is instead placed on the supposed benefits of the new development. This creates a celebratory atmosphere that does not fully acknowledge the loss and trauma experienced by the community.

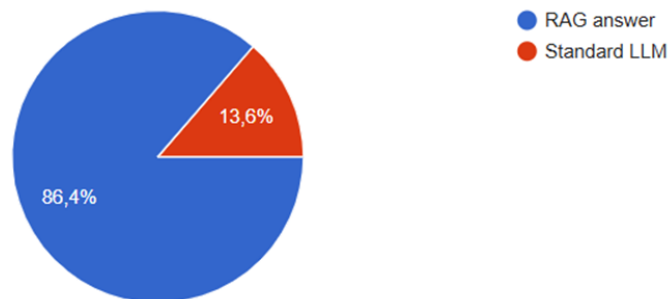


Figure 4.1. Distribution of responses for Question 1

Participants strongly preferred the RAG answer (86.4% vs 13.6%). The RAG response explains the social dynamics behind the lack of public mourning, including silence and denial within the community. In contrast, the standard LLM answer mainly describes the relocation process and emotional disruption in general terms, without addressing the deeper social mechanisms discussed in the source paper.

**Question 2** When historic buildings are relocated but their original surroundings disappear, how does this affect the overall historic environment?

Response generated without RAG: Relocating historic buildings can alter the surrounding historic environment by disrupting the original context and relationship between the building and its original setting, potentially changing the way it is perceived and understood by future generations.

Response generated with RAG: Relocating historic buildings can lead to the loss of surrounding context and authenticity, diminishing the overall historical environment, as the original surroundings often play a significant role in the building’s cultural significance.

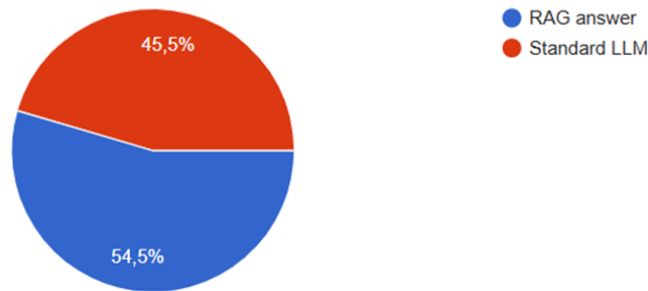


Figure 4.2. Distribution of responses for Question 2

Responses were relatively balanced (54.5% vs 45.5%). Both answers mention the loss of historical context when buildings are relocated. However, the RAG answer slightly emphasizes authenticity and cultural significance, which likely made it appear more closely aligned with heritage studies literature.

**Question 3** How can maps and documents help us understand how Kiruna is being moved over time? Response generated without RAG: Maps and documents from the 17th century show Kiruna as a small village, whereas modern maps and aerial photographs reveal its expansion and relocation due to iron ore mining, with the town now situated about 90 kilometers north of its original location.

Response generated with RAG: Maps and documents can help researchers studying Kiruna’s movement over time by allowing them to visualize the spatial relationships between documents and track changes in the city’s layout through a map interface. Additionally, documents can be filtered and sorted chronologically, enabling researchers to reconstruct a timeline of events and identify patterns and connections between documents.

Most participants preferred the RAG response (72.7% vs 27.3%). The RAG answer explains how maps and documents help analyze spatial and chronological changes in the relocation process. The non-RAG answer includes incorrect information about the town being moved 90 kilometers, which likely reduced its credibility.

**Question 4** Do smart-city initiatives in Kiruna genuinely support sustainability, or mainly reinforce mining interests?

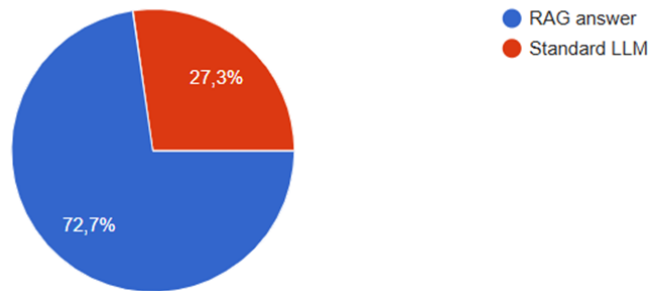


Figure 4.3. Distribution of responses for Question 3

Response generated without RAG: Kiruna’s smart city initiatives aim to improve the quality of life for residents while also reducing the environmental impact of mining operations, but some critics argue that they primarily serve to maintain and enhance the mining industry’s interests.

Response generated with RAG: The Kiruna City Sustainability Centre aims to support sustainability through various projects, including sustainable buildings, waste management, urban farming, energy systems, IT, flexible traffic, and integrated planning. However, the initiative is also driven by a desire to promote Kiruna’s image as a tourist destination and attract foreign investments, which may reinforce the interests of the mining industry.

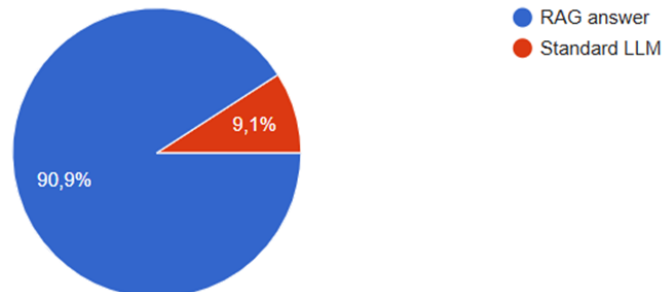


Figure 4.4. Distribution of responses for Question 4

The RAG answer was overwhelmingly preferred (90.9% vs 9.1%). It provides concrete examples such as the Kiruna Sustainability Centre and explains the tension between sustainability goals and mining interests. The non-RAG answer remains more general and does not reference specific initiatives described in the source paper.

**Question 5** How does Kiruna’s urban transformation affect place attachment and return intentions among young adults who grew up in Kiruna but moved away for studies?

Response generated without RAG: Research on Kiruna has shown that the ongoing urban transformation due to mining activities and relocation of residents can lead to

feelings of disconnection and loss of place attachment among those who have grown up there, potentially affecting their return intentions. This transformation can also create a sense of nostalgia and longing for the past, influencing their decision to return to Kiruna.

Response generated with RAG: Respondents who grew up in Kiruna but moved away for studies miss aspects of life in larger cities, such as cultural events, nightlife, and social meeting places. However, family-related security and local ties remain important factors motivating some of them to return to Kiruna.

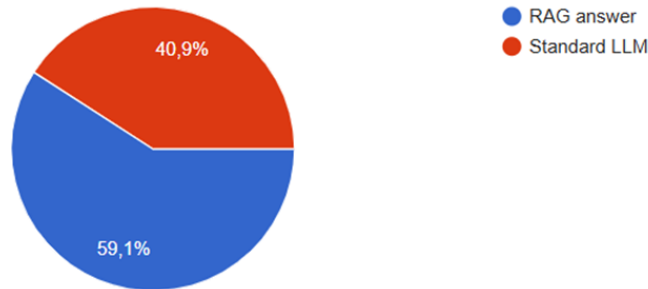


Figure 4.5. Distribution of responses for Question 5

The RAG response received slightly more support (59.1% vs 40.9%). It refers to specific factors affecting return intentions, such as family security and social opportunities. The non-RAG answer focuses more broadly on place attachment and nostalgia, which may have appeared less connected to the empirical findings of the study.

**Question 6** How does mining both enable and constrain economic diversification in Kiruna? Response generated without RAG: In Kiruna, Sweden, mining enables economic growth through job creation and revenue generation, but constrains diversification by dominating the local economy and limiting opportunities for other industries to develop.

Response generated with RAG: Mining both enables and constrains economic diversity in Kiruna by providing infrastructure benefits to tourism, but also competing with it for labor due to a small labor pool and hindering entrepreneurial activity by reinforcing a local economic structure that limits diversification.

Participants strongly favored the RAG answer (81.8% vs 18.2%). The RAG response explains concrete mechanisms such as labor competition and tourism infrastructure that affect diversification. The non-RAG answer provides a general explanation about mining dominance but lacks these specific structural details.

**Question 7** When buildings are relocated, how does this affect the overall integrity of Kiruna as a heritage town?

Response generated without RAG: The relocation of buildings in Kiruna, Sweden, due to mining activities poses a threat to the town's cultural heritage. The process can lead to the loss of traditional architecture, displacement of residents, and disruption of community identity, potentially altering Kiruna's unique character and historical essence.

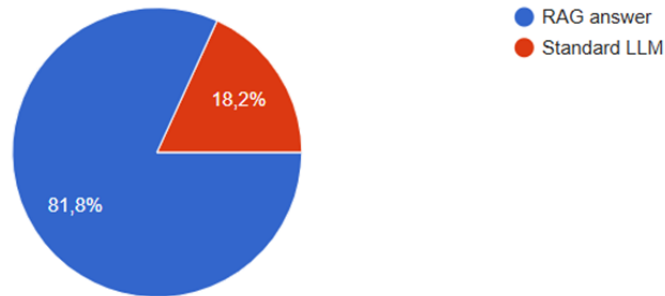


Figure 4.6. Distribution of responses for Question 6

Response generated with RAG: The relocation of historic structures in Kiruna affects the town’s heritage integrity by potentially losing parts of its authenticity due to changes in setting, while other heritage values may be maintained through restoration and reconstruction.

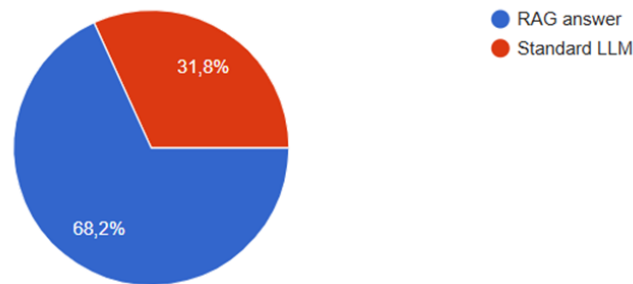


Figure 4.7. Distribution of responses for Question 7

The RAG answer was preferred by most participants (68.2% vs 31.8%). It directly discusses authenticity and heritage values when buildings are relocated. In comparison, the non-RAG response focuses more broadly on cultural heritage loss and community identity.

**Question 8** How have processes of heritagisation, re-heritagisation and de-heritagisation shaped decisions about buildings in Kiruna since 2004?

Response generated without RAG: In Kiruna, Sweden, processes of heritage conservation and urban planning have led to the relocation of the town due to copper mining activities causing subsidence, resulting in the decision to move the town by 2025.

Response generated with RAG: The processes of heritagisation (preservation and protection of cultural heritage) and re-heritagisation (revitalization and reuse of heritage sites) have been implemented in Kiruna since 2004 to address the impact of mining activities on the town’s historic environment. At the same time, de-heritagisation (demolition

or destruction of heritage structures) has also taken place, with several historic buildings being demolished to allow mining expansion, despite attempts to relocate or document them.

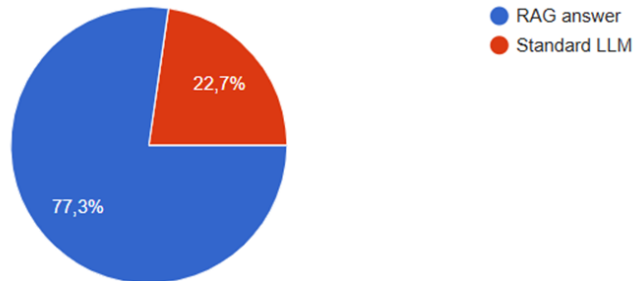


Figure 4.8. Distribution of responses for Question 8

The RAG response received clear preference (77.3% vs 22.7%). It explicitly explains the concepts of heritagisation, re-heritagisation, and de-heritagisation and connects them to planning decisions. The non-RAG answer contains incorrect information about copper mining and relocation timelines, which likely reduced its perceived accuracy.

**Question 9** How has tourism development in Kiruna emerged alongside iron-ore mining under strong structural constraints?

Response generated without RAG: Tourism development in Kiruna has been hindered by its remote location and harsh Arctic climate, but it has also been influenced by the mining industry, with many tourists visiting the area to see the mines and experience the unique culture of the Sami people.

Response generated with RAG: In Kiruna’s case, tourism development has largely emerged as a response to the constraints imposed by the iron-ore mining industry, which has dominated the local economy and created a highly specialized system in which the mine supplies services and labor to the local community.

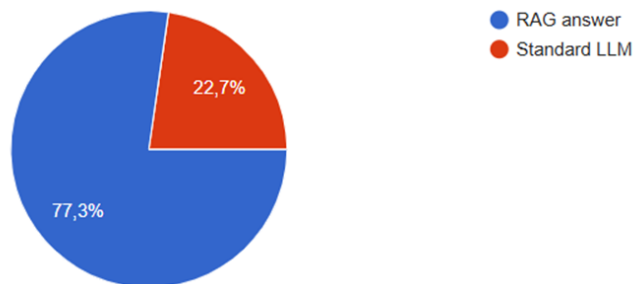


Figure 4.9. Distribution of responses for Question 9

Most participants selected the RAG answer (77.3% vs 22.7%). The RAG response explains tourism development as a structural outcome of a mining-dominated economy.

The non-RAG answer instead focuses on climate and tourism attractiveness, which is less relevant to the academic argument in the paper.

**Question 10** Do mining companies take extra voluntary steps, through corporate social responsibility, to improve their impact reports?

Response generated without RAG: Yes, many mining companies are taking extra steps to improve transparency and accountability through various initiatives, such as reporting on water and energy usage, rehabilitation of mined land, and community engagement. Some companies also publish sustainability reports and conduct regular audits.

Response generated with RAG: Mining companies may take additional voluntary steps to improve the impact reports, such as communicating with potentially affected Sami communities and including specific sections or chapters on the impacts of their operations on these communities.

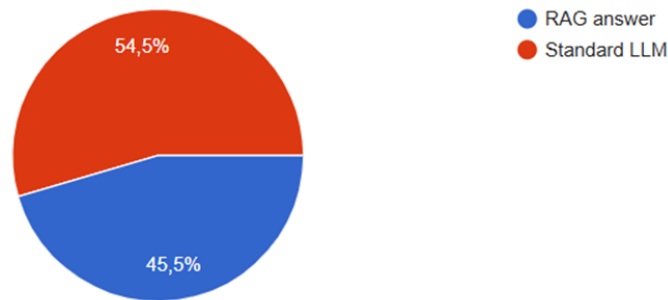


Figure 4.10. Distribution of responses for Question 10

The standard LLM answer was slightly preferred (54.5% vs 45.5%). The non-RAG response provides clearer examples of corporate social responsibility activities, making it appear more complete. The RAG answer focuses mainly on communication with Sami communities and therefore appears narrower in scope.

**Question 11** How do residents in Kiruna try to remember places that can no longer be publicly commemorated?

Response generated without RAG: Residents of Kiruna, Sweden, have had to adapt to the town's relocation due to mining activities, resulting in the original town being moved about 2 kilometers north. To cope with this, many residents use GPS and maps to navigate their surroundings, and some have also created new community centers and cultural institutions to preserve the history and identity of the relocated town.

Response generated with RAG: Residents in Kiruna try to cope with the memory of demolished buildings by finding alternative ways to pay their respects, such as through art projects like painting murals or holding workshops where locals can say goodbye to the buildings. This allows residents to maintain a sense of control and community engagement despite the loss of familiar places.

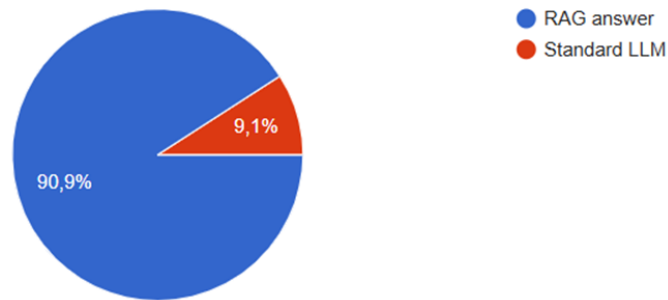


Figure 4.11. Distribution of responses for Question 11

Participants overwhelmingly preferred the RAG answer (90.9% vs 9.1%). The RAG response describes concrete practices such as murals and workshops used to remember demolished places. The non-RAG answer instead discusses relocation logistics and navigation, which are not directly related to the question.

**Question 12** How do planners balance mining systems and public space in the new Kiruna landscape?

Response generated without RAG: In the new urban planning for Kiruna, Sweden, there is a focus on balancing mining operations with public space by relocating residents and infrastructure away from the mining area and incorporating green spaces and recreational areas to mitigate the impact of mining activities. This approach aims to minimize environmental damage while preserving the town's character.

Response generated with RAG: The Kiruna mine operates as a mono-functional landscape primarily focused on extracting and processing iron ore, where spatial elements such as mining, logging, and logistics shape the structure of the surrounding environment.

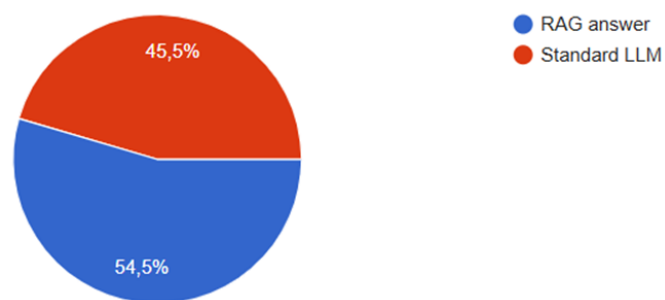


Figure 4.12. Distribution of responses for Question 12

Preferences were relatively balanced (54.5% vs 45.5%). The non-RAG answer explains planning strategies such as relocating infrastructure and creating green spaces. The RAG answer focuses on the mono-functional mining landscape, which may explain the

planning context but does not directly address the balance between public space and mining systems.

**Question 13** What challenges arise when redesigning transport and communication systems during Kiruna’s relocation?

Response generated without RAG: The relocation of Kiruna, a mining town in Sweden, poses significant challenges for redesigning its transport and communication systems due to the town’s location near the Arctic Circle, where the Earth’s magnetic field is weak and can affect compass-based navigation. This requires alternative navigation methods such as GPS and satellite-based systems, which may also be affected by weather conditions.

Response generated with RAG: Redesigning transport systems in Kiruna presents several challenges, including separating walking and cycling paths in densely built areas, planning snowmobile routes that intersect with new road plans, and adjusting bus routes to accommodate the relocated urban structure while aiming to reduce door-to-door travel time and encourage public transportation.

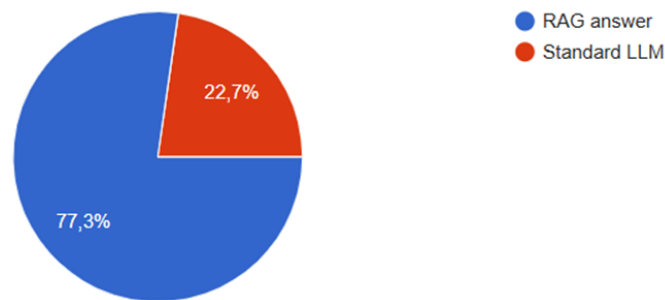


Figure 4.13. Distribution of responses for Question 13

The RAG response was strongly preferred (77.3% vs 22.7%). It describes specific challenges such as snowmobile routes, pedestrian infrastructure, and bus network adjustments. The non-RAG answer discusses magnetic fields and navigation systems, which are largely unrelated to the planning problem described in the paper.

**Question 14** How do planners decide what parts of Kiruna’s history should be carried into the new town centre?

Response generated without RAG: The planning process for Kiruna’s new town center involves analyzing the town’s history, cultural heritage, and community needs. Historians, archaeologists, and local stakeholders identify significant historical sites, landmarks, and cultural artifacts that should be preserved and incorporated into the design of the new development.

Response generated with RAG: Planners in Kiruna decide which parts of the town’s history should be carried into the new town centre by focusing on historically significant and culturally valuable areas, such as the original mining area, the railway district, and residential areas around the old town. These areas are considered of national interest

because of their unique character and the presence of notable buildings such as the Folkets Hus and the Kiruna Church.

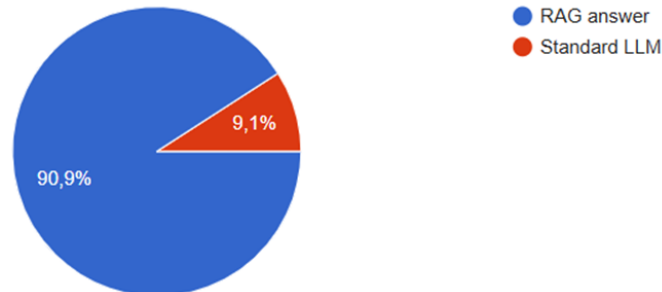


Figure 4.14. Distribution of responses for Question 14

Participants strongly preferred the RAG answer (90.9% vs 9.1%). The RAG response mentions specific historically significant areas and buildings in Kiruna. The non-RAG answer instead describes the planning process in general terms without referring to concrete examples from the case study.

**Question 15** How does time pressure influence planning decisions and coordination among stakeholders in Kiruna?

Response generated without RAG: In Kiruna, Sweden, time pressure can significantly influence planning decision-making among stakeholders due to the town's unique challenges, such as the relocation of infrastructure related to the expanding iron ore mine. This situation requires coordination among residents, mining companies, and government authorities, which can lead to rushed decisions and conflicts between stakeholders.

Response generated with RAG: Time pressure influences the planning process in Kiruna by creating tension between the two main stakeholders, the municipality and the mining company LKAB. For example, the municipality faced pressure to make decisions regarding infrastructure changes, such as the railway relocation, within strict deadlines. This urgency contributed to conflicts and difficulties in reaching consensus between the involved parties.

The RAG answer received strong support (81.8% vs 18.2%). It refers to specific conflicts between the municipality and the mining company LKAB and includes concrete planning deadlines. The non-RAG answer only provides a general description of coordination challenges.

**Question 16** Can smart city planning in Kiruna create real transformation, or does it mostly reinforce existing mining structures?

Response generated without RAG: Kiruna's smart city plan aims to transform the town by creating a more sustainable and resilient community, but it may also perpetuate the existing mining industry's influence, since the local economy is strongly dependent on the mining sector.

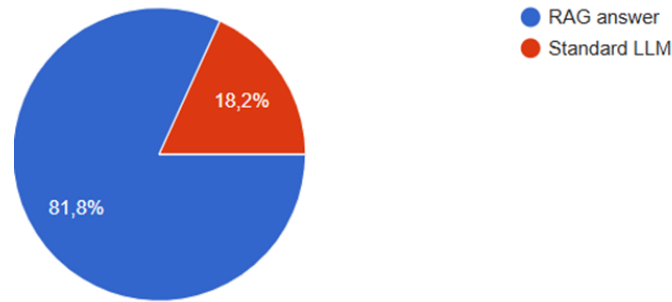


Figure 4.15. Distribution of responses for Question 15

Response generated with RAG: The Kiruna Smart City project aims to create transformation by establishing a sustainability centre and testing various smart city innovations. However, its success is influenced by the convergence and divergence of interests between the mining companies and the municipal government, which can lead to conflicts in planning decisions and reinforce existing resource extraction structures.

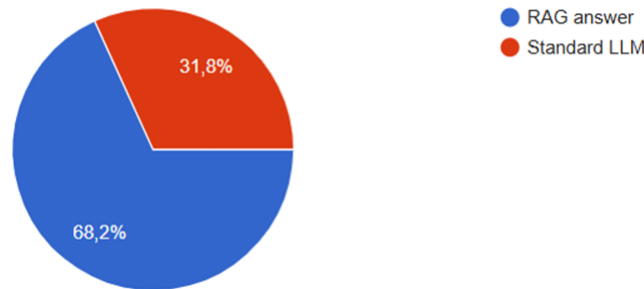


Figure 4.16. Distribution of responses for Question 16

The RAG response was preferred by most participants (68.2% vs 31.8%). It explains how competing interests between municipal authorities and mining companies influence smart city initiatives. The non-RAG answer describes sustainability goals more generally without addressing these governance dynamics.

**Question 17** Why is Kiruna’s relocation an evolving system rather than a single planning project?

Response generated without RAG: Kiruna’s relocation is considered an ongoing process due to the gradual expansion of the mine over time. As mining activities continue to affect the ground stability of the surrounding areas, the city must continuously adapt its planning decisions to ensure the safety of residents and infrastructure.

Response generated with RAG: Kiruna’s relocation is described as an evolving system because it involves multiple competing interests, including mining operations, cultural heritage preservation, and reindeer husbandry. The dominance of mining interests drives

the relocation process while also limiting available space and shaping the design of the new urban environment.

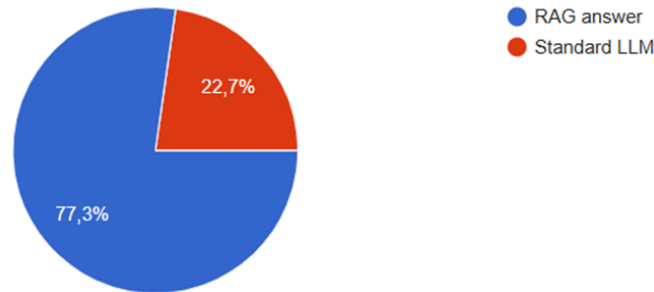


Figure 4.17. Distribution of responses for Question 17

Participants strongly favored the RAG answer (77.3% vs 22.7%). The RAG response highlights the interaction between mining, cultural heritage, and reindeer husbandry interests. This explanation better reflects the complexity of the relocation process than the simpler explanation provided in the non-RAG answer.

**Question 18** How do media narratives and municipal communication shape public acceptance of relocation in Kiruna?

Response generated without RAG: Media narratives can influence public perception by framing relocation as a necessary measure to address environmental and economic challenges. Municipal communication may also highlight the potential benefits of relocation, such as improved infrastructure and future development opportunities, in order to increase public acceptance among residents.

Response generated with RAG: Media narratives and municipal communication shape public opinion by presenting the relocation of Kiruna as necessary for the town’s survival. This framing creates a sense of inevitability and urgency while often downplaying negative impacts on residents, particularly those affecting indigenous Sami communities.

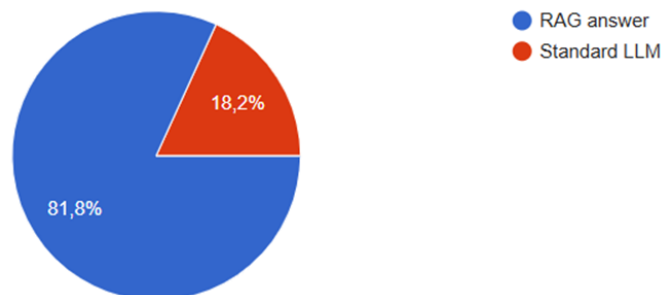


Figure 4.18. Distribution of responses for Question 18

The RAG response received strong preference (81.8% vs 18.2%). It explains how

relocation is framed as inevitable and how negative consequences are sometimes downplayed in communication. The non-RAG answer incorrectly frames relocation as related to climate change rather than mining expansion.

**Question 19** Can historic buildings in Kiruna still be considered authentic after they are moved to a new location?

Response generated without RAG: Moving historic buildings can alter their original context and surroundings, potentially affecting their authenticity. However, if the relocation preserves the building’s original structure and materials, it may still be considered an authentic representation of the original site.

Response generated with RAG: Historic buildings can still be considered authentic after relocation if their relationship to the surrounding landscape, the mining environment, and the broader historical context is maintained.

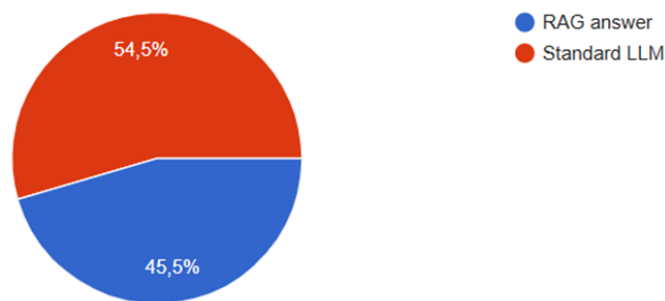


Figure 4.19. Distribution of responses for Question 19

The standard LLM answer was slightly preferred (54.5% vs 45.5%). The non-RAG response provides a clearer explanation of authenticity in relation to building relocation. The RAG answer focuses mainly on maintaining landscape relationships, which may have appeared less complete to participants.

**Question 20** Why are mining companies’ environmental and social impact assessments in Kiruna criticized as insufficient?

Response generated without RAG: Mining companies in Kiruna are criticized for conducting assessments that do not adequately consider the long-term effects of mining on local communities and the environment. These assessments are often seen as prioritizing economic benefits over social and environmental concerns, which can result in insufficient measures to mitigate risks and prevent harm.

Response generated with RAG: Voluntary corporate actions in mining companies’ environmental and socio-environmental impact assessments are criticized as inadequate because they often fail to provide a comprehensive understanding of the long-term consequences of mining operations on communities and the environment, resulting in what some researchers describe as “cosmetic improvements” rather than meaningful change. Most participants preferred the RAG answer (68.2% vs 31.8%). The RAG response high-

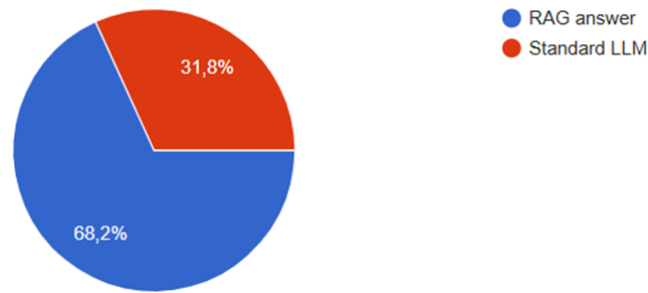


Figure 4.20. Distribution of responses for Question 20

lights criticism that voluntary corporate actions often lead only to “cosmetic improvements.” This argument closely reflects the critique presented in the academic source.

#### 4.4.3 Additional Questions

Although thirty evaluation questions were created when preparing the evaluation dataset, only the first twenty questions were used in the questionnaire study. This reduction was necessary to keep the survey at a reasonable length and to avoid participant fatigue. Each question required participants to read and compare two generated answers. The remaining questions and generated responses are documented in this section of the thesis.

**Question 21** Is it more important to keep historic buildings in their original place, or simply to save them somewhere else in Kiruna?

Response generated without RAG: Preserving historic buildings can be beneficial for cultural and historical significance, but relocating them can also help mitigate the effects of climate change and mining activities in sensitive areas like Kiruna.

Response generated with RAG: Preserving historic buildings where they are, rather than relocating them, seems to be the preferred approach among local interest groups, as they argue that many people support conserving historic buildings.

**Question 22** How does the concept of derealisation explain the lack of public memorialisation in Kiruna’s old town centre?

Response generated without RAG: Derealisation refers to a psychological phenomenon where an individual disconnects from reality, perceiving their surroundings as unreal or distant. In the context of Kiruna’s old town center, derealization may be applied to describe how the residents’ perception of their hometown has changed due to the ongoing mining-related relocation and urban renewal process, leading to a sense of disconnection from their past and physical environment.

Response generated with RAG: The concept of derealisation in this context refers to the emotional disconnection and detachment from the past, as evident in the reflections of informants who felt that the planned demolition of the historic city chambers did not acknowledge the significance of the building’s role in the community and its cultural

heritage. Despite its recognition as a national heritage site, the decision to demolish the building suggests a lack of meaningful commemoration, instead prioritizing the interests of the mining company and the new development plans.

**Question 23** How do planning documents influence what the new Kiruna will look like and feel like?

Response generated without RAG: Planning documents, such as master plans and zoning regulations, play a significant role in shaping the physical layout, density, and character of Kiruna, influencing factors like building heights, land use, transportation systems, and community amenities.

Response generated with RAG: Planning documents in Kiruna are likely to be influenced by sociotechnically constructed imaginaries that reflect the collective promise of a specific future, taking into account the technological, social, and economic interests and perspectives of various actors involved.

**Question 24** How has mining-related ground deformation affected infrastructure in Kiruna?

Response generated without RAG: Kiruna, a town in Sweden, is experiencing ground deformation due to mining activities, causing the town to slowly move about 1-2 cm per year. This has led to concerns about the structural integrity of buildings, roads, and other infrastructure.

Response generated with RAG: The construction of the new railway line in Kiruna has caused significant ground deformation, which has affected the infrastructure in the area. The new railway will pass through a sameby, specifically Laevasa, where one of its most important relocation areas is located, leading to potential negative consequences for the community.

**Question 25** The relocation of Kiruna is often described as sustainable, but how does the three-sphere transformation framework show gaps between plans and real experiences?

Response generated without RAG: Research on the Kiruna relocation project has identified gaps between the original plan's sustainability goals and the actual outcomes, including issues with waste management, social impacts on local communities, and inadequate consideration of long-term environmental consequences. These gaps highlight the need for more nuanced and context-specific approaches to sustainability planning.

Response generated with RAG: Despite the emphasis on involving residents in the planning through surveys, focus groups, and community engagement, the actual relocation process revealed significant gaps between the planned transformation and the actual outcomes. The relocation of cultural heritage buildings was only partially executed due to technical and financial limitations, indicating a disconnect between the planning phase and the implementation phase.

**Question 26** Why do technical and economic priorities seem stronger than social and cultural values in Kiruna's transformation?

Response generated without RAG: In Kiruna, Sweden, the mining industry has historically been the primary driver of economic growth, leading to significant investments in infrastructure and services catering to the needs of the mining sector. This has resulted in a strong focus on economic and technical priorities, overshadowing other values such as social cohesion and cultural heritage.

Response generated with RAG: The residents of Kiruna feel that their concerns and values were overlooked in the decision-making process, and that the relocation plan prioritized the interests of the mining corporation and the municipality over their own needs and emotions. In contrast, the history and experiences of the indigenous Sami community, who have been marginalized due to the mining industry's impact on their land and way of life, are largely absent from the planning process.

**Question 27** How does uncertainty about future mining impacts affect planning choices for the new centre?

Response generated without RAG: Uncertainty about future mineral reserves and prices can lead to caution in planning, potentially delaying investment and development decisions for the centre, as it may be difficult to justify significant expenditure on infrastructure and operations with uncertain returns. This can result in a more conservative approach to planning, prioritizing exploration and feasibility studies to better understand the potential risks and opportunities.

Response generated with RAG: Uncertainty about the future of mining in the area affects planning choices because it makes it difficult to determine the long-term viability of each alternative, which can lead to increased costs and risks for the city of Kiruna.

**Question 28** Why was the new centre of Kiruna built in its current eastward location instead of other possible sites?

Response generated without RAG: The new centre was built east of the original site due to concerns about subsidence caused by iron ore mining beneath the original location, which threatened the stability of the town.

Response generated with RAG: The new center of Kiruna was built in the eastward direction because it was deemed the most suitable location due to the market's preference for the area, which has already seen the establishment of businesses and shopping centers, such as KappAhls store. This location also offered easier conditions for building a grid of streets and houses, and was less prone to risks associated with the existing mining area, including sinkholes and environmental concerns.

**Question 29** Does the relocation make Kiruna more or less attractive for young adults?

Response generated without RAG: Kiruna's relocation due to mining activities may make it less attractive to young adults, as it requires frequent moves and can disrupt their established social networks and careers.

Response generated with RAG: Kiruna is less attractive to young adults because the lack of social meeting places and limited range of cultural events give a negative image of the city.

**Question 30** How do the municipality and LKAB cooperate or conflict during Kiruna’s relocation planning process?(about planning and governance)

Response generated without RAG: The municipality of Kiruna and L.K. AB (LKAB), the mining company, have a complex and ongoing relationship during the relocation planing process. The municipality has expressed concerns about the environmental and social impacts of the mine’s expansion, while LKB has emphasized the need for continued mining operations to support the local economy.

Response generated with RAG: LKAB and Kiruna Municipality’s cooperation during Kiruna’s relocation planning process can be described as conflictual rather than cooperative. The two main parties have different views on relocation alternatives, which creates disagreements and makes it difficult to reach consensus.

#### 4.4.4 Discussion of Questionnaire Results

The questionnaire results shown in Figure 4.21. a clear overall advantage for the retrieval-augmented generation (RAG) system. Across the twenty evaluated questions, participants preferred the RAG-generated responses in 18 cases (90%). The standard language model without retrieval was preferred in 2 cases (10%). In most questions, the RAG responses were considered more specific and informative because they included information retrieved from documents about Kiruna’s relocation. This allowed the answers to refer to concrete concepts, actors, and planning processes described in the source papers. In contrast, answers generated without RAG often relied on general knowledge. As a result, these responses tended to provide broader explanations that were less specific to the Kiruna case.

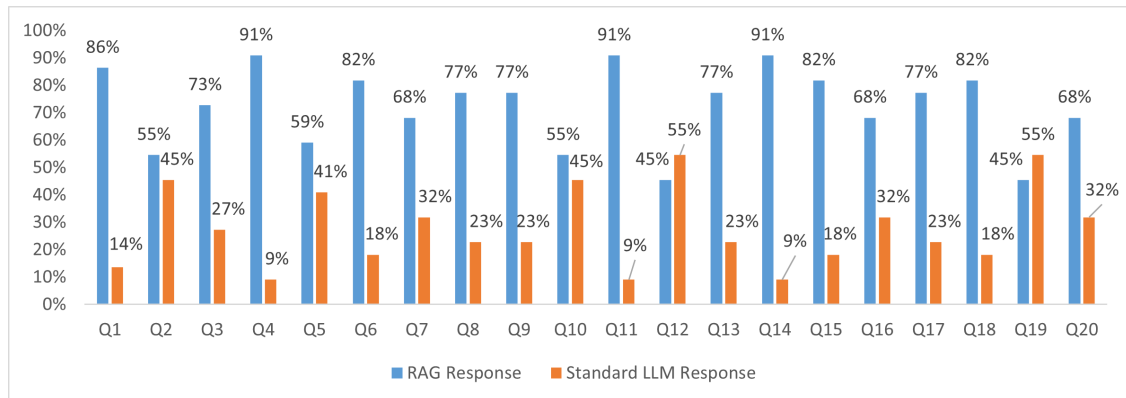


Figure 4.21. Comparison of participant preferences between RAG-generated responses and standard language model responses.

At the same time, the results also show some limitations of the RAG approach. In a few cases, the response generated by the standard language model was preferred because it provided a clearer or more general explanation of the topic. This can happen when the retrieval step selects passages that are only partly related to the question. It can also occur when the most relevant document is not retrieved effectively. In these situations, the

model may generate an answer based mainly on the retrieved excerpt instead of the most appropriate source. These observations highlight the importance of accurate document retrieval and ranking in RAG systems, since the quality of the generated answer strongly depends on the relevance of the retrieved context.

## Chapter 5

# Conclusion

This thesis examined the problem of fragmented urban documentation using the case of Kiruna, a city that is undergoing a long-term relocation due to mining activities. As the relocation continues, many materials have been produced by different institutions. These include planning documents, research publications, maps, and heritage records. The materials are stored across different platforms and formats. Because of this, they are difficult to access and analyze in a unified way. This thesis investigated how artificial intelligence techniques, especially retrieval-augmented generation (RAG), can support the organization and exploration of such heterogeneous data related to urban transformation.

The thesis first discussed the broader context of digital heritage and the role of artificial intelligence in managing complex cultural datasets. Large language models have strong capabilities for natural language understanding and text generation. However, their use in specialized domains is limited by problems such as outdated knowledge and hallucinated information. Retrieval-augmented generation addresses these limitations by combining language models with external document retrieval. By grounding answers in retrieved sources, RAG systems can generate responses that are more accurate and more specific to the context than those produced by standalone language models.

To explore this approach, a dataset of documents related to Kiruna’s urban transformation was collected and organized. The documents were analyzed and structured to support consistent indexing and retrieval. Based on this dataset, a RAG pipeline was implemented in Python. The system processes user queries through several stages. These stages include document embedding, semantic similarity search, page segmentation, chunk-level retrieval, and context construction. The retrieved passages are then given to a local instruction-tuned language model, which generates answers based on the retrieved content. This architecture allows the system to connect natural language questions with relevant parts of the document corpus.

The evaluation focused on the quality of generated answers to questions about Kiruna’s relocation and urban transformation. Two types of responses were compared. The first type was generated by a standard language model without retrieval. The second type was generated using the RAG pipeline. A qualitative analysis and a questionnaire study were conducted to evaluate which responses were seen as more accurate, specific, and

informative. The questionnaire results showed a clear overall advantage for the retrieval-augmented approach. In most cases, participants preferred the RAG-generated answers because they included more concrete details and reflected information taken from the document corpus.

At the same time, the evaluation also showed some limitations of the approach. The quality of the generated answer depends strongly on the quality of the retrieved context. If the retrieval component selects passages that are only partly related to the question, the generated answer may be incomplete or less accurate. Similar problems can occur if the most relevant document is not retrieved. In these situations, the standard language model may produce a more general explanation that appears clearer to readers. These observations show that the effectiveness of RAG systems depends not only on the language model but also on the design of the retrieval process. Important factors include document indexing, chunking strategies, and similarity ranking.

Overall, the results show that retrieval augmented generation can help improve access to complex and fragmented documentation about urban transformation processes. By connecting natural language questions to specific documents in a curated dataset, the system allows users to explore information about Kiruna in a more structured and accessible way. The implementation presented in this thesis is a prototype rather than a complete archival platform. However, it demonstrates how AI-assisted retrieval systems can support research on cities undergoing long-term spatial change. Future work could extend this approach by expanding the document dataset, improving retrieval strategies, and developing user interfaces that allow researchers and heritage professionals to interact more effectively with urban transformation archives.

# Bibliography

- [1] Adapting historic places to climate change. In Gemma Houston, Vanessa Glindmeier, and Carsten Hermann, editors, *Proceedings of the International Virtual Conference of the Project Adapt Northern Heritage*, 2020.
- [2] Barnett et al. Seven failure points when engineering a retrieval augmented generation system. 2024.
- [3] Eric Boyd. Memorialisation and its denial: Slow resistance through derealisation in kiruna, sweden. *Journal of Political Power*, 16(2):158–176, 2023.
- [4] F. Bu et al. An investigation into value misalignment in llm-generated texts for cultural heritage. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2025.
- [5] Giovanni Ceselli. How to move a city: Mapping the process of kiruna relocation. Master’s thesis, Politecnico di Torino, 2024.
- [6] Chen et al. Benchmarking large language models in retrieval-augmented generation. 2024.
- [7] A. G. Cossatin et al. Tell me more: Integrating llms in a cultural heritage website for advanced information exploration support. *Information Technology and Tourism*, 2025.
- [8] Valeria Federighi, Taneha Kuzniecowa Bacchin, and Kirthan Shekar. Which landscape? material traces of an integrated design and research approach in kiruna. In *Landscape as Horizon. Green Flashes in the Distance*. 2018.
- [9] X. Fu et al. Emerging computing technology for digital culture heritage preservation and inheritance: A literature review. *IEEE Transactions on Computational Social Systems*, 2025.
- [10] Gao et al. Retrieval-augmented generation for large language models: A survey. 2023.
- [11] Christin Gimberger and Karin Norberg. Uppdrag: Att flytta en stad – en studie av planeringsprocessen i samband med kiruna stadsomvandling. Master’s thesis, Uppsala University, 2006.
- [12] G. Han, J. Wang, and H. Li. Llm-driven safeguarding and inheritance of li brocade intangible cultural heritage. In *Proceedings of ICHSSR*, 2025.
- [13] Jiang et al. Active retrieval augmented generation. In *EMNLP*, 2023.
- [14] Isak Kautto. Ett kiruna i förändring: En studie om hur utflyttade kirunabor upplever stadsflytten i kiruna, 2024.
- [15] Rasmus Kløcker Larsen, Carl Österlin, and Laura Guia. Do voluntary corporate actions improve cumulative effects assessment? mining companies’ performance on

- sami lands. *The Extractive Industries and Society*, 5:375–383, 2018.
- [16] Lewis et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *NeurIPS*, 2020.
- [17] Magnus Lindgren. Ett nytt läge för kiruna, 2010.
- [18] David Lowenthal. *The Heritage Crusade and the Spoils of History*. Cambridge University Press, 1998.
- [19] W. Luo. Multimodal-llm as a reliable tool for information extraction from historical documents. Master’s thesis, Linnaeus University, 2025.
- [20] Bo Nilsson. Ideology, environment and forced relocation: Kiruna – a town on the move. *European Urban and Regional Studies*, 17(4):433–442, 2010.
- [21] L. Nisiotis et al. Enhancing digital heritage experiences: Evaluating fine-tuned llm integration within a cyber-physical-social virtual museum system. 2024.
- [22] Linus Petersson. Konsten att flytta en stad: en studie av hur kommunikationerna i kiruna påverkas av stadsomvandlingen, 2007.
- [23] Agatino Rizzo, Jennie Sjöholm, and Andrea Luciani. Smart(en)ing the arctic city? the cases of kiruna and malmberget in sweden. *European Planning Studies*, 32(1):59–77, 2024.
- [24] I. Siliutina et al. Cultural preservation and digital heritage: Challenges and opportunities. *Amazonia Investiga*, 2024.
- [25] Jennie Sjöholm. Att flytta en mönsterstad. *Fabrik & Bostad*, pages 24–43, 2015.
- [26] Jennie Sjöholm. *Heritagisation, Re-Heritagisation and De-Heritagisation of Built Environments: The Urban Transformation of Kiruna, Sweden*. PhD thesis, Luleå University of Technology, 2016.
- [27] Jennie Sjöholm. Authenticity and relocation of built heritage: The urban transformation of kiruna, sweden. *Journal of Cultural Heritage Management and Sustainable Development*, 7(2):110–128, 2017.
- [28] Linda Stihl. Challenging the set mining path: Agency and diversification in the case of kiruna. *The Extractive Industries and Society*, 11:101064, 2022.
- [29] Aslı Tepecik Diş and Elahe Karimnia. Reframing kiruna’s relocation—spatial production or a sustainable transformation? *Sustainability*, 13(7):3811, 2021.
- [30] Kenneth Thibodeau. Overview of technological approaches to digital preservation and challenges in coming years. In *The State of Digital Preservation: An International Perspective*.
- [31] A. Tomasino et al. Llm-based pipelines for the restoration of digitized historical printed archives. 2024.
- [32] G. M. Toth, R. Albrecht, and C. Pruski. Explainable ai, llm, and digitized archival cultural heritage: A case study of the grand ducal archive of the medici. *AI and Society*, 2025.
- [33] G. Trichopoulos. Large language models for cultural heritage. In *Proceedings of the ACM Greek SIGCHI Conference*, 2023.
- [34] UNESCO. Convention concerning the protection of the world cultural and natural heritage, 1972.
- [35] L. Vaštakas. Cultural heritage search with large language models. Master’s thesis, Linnaeus University, 2024.

- [36] Wang et al. Searching for best practices in retrieval-augmented generation. In *EMNLP*, 2024.
- [37] Wu et al. Retrieval-augmented generation for nlp: A survey. 2025.
- [38] Xiong et al. Benchmarking retrieval-augmented generation for medicine. In *ACL*, 2024.
- [39] Zhao et al. Retrieval-augmented generation for ai-generated content: A survey. 2026.