A Retrieval-Augmented Generation System for Accurate and Contextual Historical Analysis : AI-Agent for the Annals of the Joseon Dynasty

Jeong Ha Lee
AI-Robotics, KIST School,
University of Science and Technology(UST),
Seoul, Republic of Korea
wjdgk1029@gmail.com

Ghazanfar Ali
Intelligence and Interaction Research Center,
Korea Institute of science and technology(KIST),
Seoul, Republic of Korea
ali@kist.re.kr

Jae-In Hwang
Intelligence and Interaction Research Center,
Korea Institute of science and technology(KIST),
Seoul, Republic of Korea
hji@kist.re.kr

Abstract

In this paper, we propose an AI-agent that integrates a large language model(LLM) with a Retrieval-Augmented Generation(RAG) system to deliver reliable historical information from the Annals of the Joseon Dynasty through both objective facts and contextual analysis, achieving significant performance improvements over existing models. In order for an AI-agent using the Annals of the Joseon Dynasty to deliver reliable historical information, clear source citations and systematic analysis are essential. The Annals, an official record spanning 472 years (1392–1897), offer a dense, chronological account of daily events and state administration that shaped Korea's cultural, political, and social foundations. We propose integrating a LLM with a RAG system to generate highly accurate responses based on this extensive dataset. This approach provides both objective information about historical figures and events from specific periods and subjective contextual analysis of the era, helping users gain a broader understanding. Our experiments demonstrate improvements of approximately 23 to 50 points on a 100-point scale compared to the GPT-4o and OpenAI AI-Assistant v2 models.

Keywords: AI-agent, Large Language Model, Retrieval Augmented Generation, Historical Analysise

1 Introduction

The Joseon Dynasty (1392–1910) governed the Korean Peninsula for about 500 years, main-

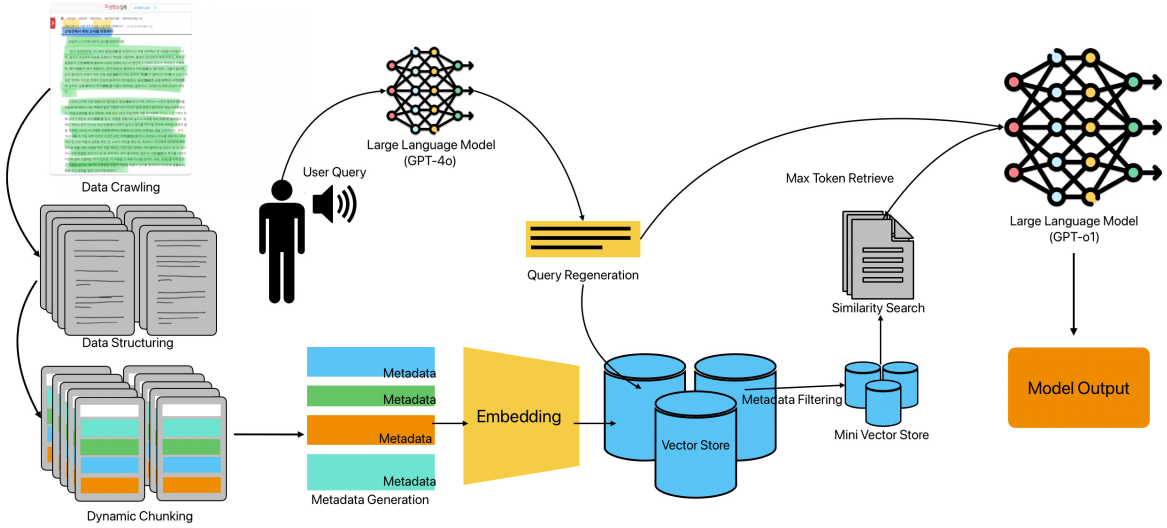

Figure 1: This figure illustrates the overall pipeline of our model, showing both the data preprocessing steps and how it operates through a large language model (LLM). By utilizing a vector store built on data from the Annals of the Joseon Dynasty, the system generates accurate responses to user queries.

taining a centralized administration rooted in Confucian ideals. During this era, a thorough record-keeping culture flourished, documenting a wide range of state affairs, from administration and diplomacy to military operations. A prime example is the Annals of the Joseon Dynasty, which meticulously documented the reigns and events of successive kings. Today, it is recognized as a UNESCO Memory of the World for its significant historical value[14, 13]. When developing an AI-agent based on historical information, delivering reliable information without distortion is important.[20]. When an AI-agent generates an answer based on large historical data with a vast amount of information, the accuracy and reliability of the answer can be increased by presenting the source of the answer and the basis of the analyzed content[7]. In this annual style of historical data, it is difficult to comprehensively grasp the information of a specific subject or person because the events are arranged in chronological order, and the same event may be described differently depending on the recorder or point of view, so its reliability should be reviewed. Accordingly, we propose a novel approach that integrates LLM with a proposed RAG system to generate highly accurate answers by utilizing the extensive chronological records found in the Annals of the Joseon Dynasty particularly the Sejong Annals.

Our method is designed not only to provide objective, fact based historical information regarding figures, events, and social contexts queried by users, but also to offer in-depth analysis and interpretation of the circumstances of the time. To achieve this, the study introduces two primary approaches:

- Objective Information Providing: When users request clear and precise details about historical events or figures from a specific period, the AI-agent provides objective and detailed information based on the corresponding historical records.

- Subjective Contextual Analysis and Interpretation: Drawing on related historical records, the system provides comprehensive analysis and interpretation of the sociocultural context of the era, thereby enabling users to gain a broader understanding of the period. In particular, it retrieves the information needed for the model's answers to generate answers and shows their sources together, allowing accurate analysis and reliable answers to be generated.

Our system uses an integrated framework that includes *dynamic chunking, metadata filtering,*

query regeneration, and *Max Token Retrieve* techniques. This framework overcomes the limitations of traditional fixed-token chunking methods[6] by preserving the full context and source information, ensuring that no critical details are omitted. Evaluation results based on the GPT-assisted benchmark question set, we evaluate our model with GPT-4o, ai-assistant v2 models. and in three evaluation categories, on average, improving answer accuracy by 23 to over 50 points in comparison with each model, especially in psychedelic prevention and evidence-based reasoning capabilities.

In the following sections, we first present an overview of related work, detailing previous approaches to processing historical records with LLM and RAG, and introducing the Annals of the Joseon Dynasty along with its format and prior studies. Next, we describe the system implementation, including our *dynamic chunking, metadata filtering, query regeneration*, and *Max Token Retrieve* methods. We then evaluate the system's performance against established benchmarks, and finally, discuss the implications, limitations, and potential future directions of our work.

2 Related Work

Research on efficiently retrieving information from large-scale historical texts has become a major topic[15]. Traditional keyword-based search faces limitations in dealing with massive archives, leading to the adoption of vector embedding techniques for semantic retrieval, as well as metadata tagging systems[1, 22]. For instance, the U.S. National Archives and Records Administration (NARA) has launched a pilot project that leverages AI-based *semantic search* to analyze users' intent and context, thereby connecting them to relevant records more effectively [17]. Because semantic search goes beyond string matching to discover semantically similar or contextually linked documents, it holds the potential to present users with precisely the information they need, even in extensive archives, and reveal new historical insights [17]. Moreover, comprehensive *metadata tagging* such as labeling entities (people, places), time periods, and subjects has been implemented in multiple archives, including the digitized version of the Annals of the Joseon Dynasty, to assist with data discovery and filter information by, for example, date or named entities [21].

Retrieval-Augmented Generation (RAG), introduced around 2020, integrates knowledge retrieval with large language models to produce more trustworthy, evidence-based answers[10]. Specifically, RAG retrieves relevant documents for a user query and then generates an answer grounded in that retrieved context, thereby alleviating the risk of hallucination and improving factual accuracy[10, 6]. The European Parliament Archives recently deployed a RAG-based question-answering system that indexes more than 100,000 archival documents. This system allows users to ask questions interactively, returning a concise generated answer underpinned by relevant archival records [5]. This approach has shown the potential to let non-specialists easily navigate vast historical collections by engaging in conversational interactions.

In parallel, various text-mining and NLP techniques have been applied to the Annals of the Joseon Dynasty, a canonical example of large-scale historical texts in Korean studies. Bak and Oh (2015)[2] introduced a corpus extracted from the Annals and conducted statistical analysis to categorize the decision-making patterns of different monarchs. Their results revealed intriguing patterns, such as deposed or tyrannical kings tending to exhibit more unilateral behaviors, demonstrating how text-mining can unveil new interpretive insights [2]. Subsequent studies expanded on this work by employing techniques like dynamic word embeddings to observe linguistic changes over time within the Annals [8]. Moreover, Yoo et al. (2022)[21] constructed HUE, a dataset and pretrained model specialized in reading Hanja texts, which are the classical Chinese characters used in historical Korean records. By integrating additional metadata such as dates and personal names, they achieved improved performance in tasks like summarization and named-entity recognition. Such research underscores the necessity of specialized NLP approaches for large historical corpora written in older forms of language. Recent developments show that *conversational AI* can make historical knowledge

more approachable to the public. Pataranutaporn et al. (2023)[16] presented the notion of *Living Memories*, wherein AI-generated characters are trained on extensive historical sources, allowing users to interact with a *digital persona* of a historical figure. Empirical studies suggest that engaging in conversation with these AI agents can enhance learning engagement more effectively than merely reading a text. Similar lines of research exist outside the historical domain, such as Kim et al. (2024)[9], who explored ways to improve patient understanding of surgical procedures by incorporating facial expressions and gestures in animated physician characters, illustrating how nonverbal communication and storytelling techniques can deepen comprehension. These examples highlight how systems designed to deliver large-scale domain knowledge should not only provide correct information but also present it in a manner that users find engaging and comprehensible. traditional RAG remained confined to a simple retrieve–generate framework, recent work has proposed various pipeline enhancements to boost accuracy and reliability on long and dense texts. First, Mix-of-Granularity dynamically selects the optimal chunk size per query—rather than using fixed-size chunks—thereby minimizing context loss while improving retrieval efficiency[24]. Next, query-focused rewriting approaches such as RQ-RAG[4] and Rewrite–Retrieve–Read[12], which explicitly reformulate and decompose the user's initial query before retrieval, have demonstrated improved retrieval precision using these refined queries. Beyond that, the DRAGIN framework[18] performs multi-stage retrieval during the generation process itself, enabling iterative information enrichment in complex, interactive historical QA tasks. To mitigate hallucination caused by mismatches between retrieval and generation, methods for adjusting the balance between internal and external knowledge sources and integrating automatic verification steps have been introduced, further enhancing the traceability and trustworthiness of responses[19, 23]. These pipeline-enhancement strategies offer theoretical and empirical foundations for the design choices in this study—namely, dynamic chunking, metadata filtering, and query regeneration.

3 System Architecture

3.1 System Framwork

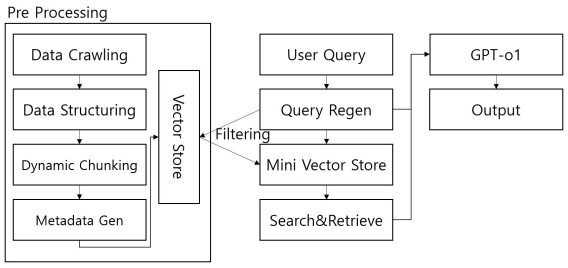

Figure 2: An overview of our integrated RAG-based system for delivering reliable historical information from the Sejong Annals. The pipeline comprises four stages: data preprocessing, query regeneration, metadata filtering, and RAG.

In Fig 2 at data preprocessing, Sejong Annals data was web-crawled and consolidated into a single source. To avoid issues with fixed token-based chunking that split articles and context, we adopted *dynamic chunking* to keep each article intact. However, this generated large chunks that sometimes mixed topics. To resolve this, we parsed date information (year, month, day) from each article and incorporated it into the vector store.

Since users often provide vague queries, the *query regeneration* module uses the GPT API to refine input queries by adding event details and approximate dates, greatly improving search precision[11, 3]. In *metadata filtering*, date details are extracted from the refined query so that only relevant chunks are selected—using exact matches for complete dates or aggregating chunks for partial or absent date info.

In the RAG stage, a similarity search is performed on the filtered vector store. Instead of retrieving a fixed number of chunks, we collect chunks until the maximum token threshold we set is reached. This maximum token threshold is determined based on the model's input token limit. Through a system prompt, the model selects only the necessary information to generate an accurate final answer.

By combining *dynamic chunking, metadata filtering*, *query regeneration, and Max Token Retrieve*, our system forms a comprehensive

framework for extracting and delivering detailed historical records aligned with user intent.

3.2 Data Pre-processing

Our model is developed to provide users with reliable historical information by utilizing the Sejong Annal section of the Annals of the Joseon Dynasty. The Sejong Annal is a historical record documented daily from 1418 to 1450, consisting of a total of 163 volumes and containing an extensive dataset of 3,392,418 words. During the data collection phase, a web crawling technique was applied to the official website of the Annals of the Joseon Dynasty to selectively extract records specifically from the reign of King Sejong. The extracted data was then consolidated into a single source for further processing.

A particularly crucial step in data preprocessing was text chunking. Typically, when processing textual data, chunking is performed using fixed number of token. However, this approach artificially segments the text into fixed lengths, often resulting in a single article being split across multiple chunks or disrupting the contextual flow. Consequently, it becomes difficult to clearly present the source of the information in the model-generated responses, limiting the user's ability to fully grasp the overall context of an article.

In this problem, we adopted a *dynamic chunking* method. Specifically, each article within the Annals of the Joseon Dynasty was treated as an independent unit, ensuring that the entire content of a given article was preserved within a single chunk. This approach allows the vector store to retain the original structure and context of the articles when retrieving chunks, enabling the model to generate responses that clearly reference complete source articles. As a result, the generated responses include comprehensive source information, significantly enhancing the system's reliability and transparency.

However, several issues were discovered in the total of 30,949 chunks generated through the aforementioned preprocessing steps. First, each chunk was too large, resulting in the mixing of content on various topics within a single chunk. This led to a decrease in the precision of the information contained in each chunk during vector searches. Since the embedding model converts the input text into a single vector, a large chunk size causes too much information to be condensed into one vector, which in turn dilutes important keywords or key concepts.

In particular, when queries include specific dates, the matching accuracy significantly deteriorated, frequently resulting in information from the wrong period being retrieved instead of the specific period requested by the user.

To resolve these issues, date information was parsed from each article to extract three pieces of metadata year, month, and day, which were then incorporated during the construction of the vector store. This enabled precise, date-based matching during searches and, ultimately, significantly enhanced the overall accuracy and reliability of the vector search.

3.3 Query Regeneration Module

In this study, we introduce a *query regeneration* technique to effectively retrieve information related to the user's query from the vector store and enhance retrieval accuracy. The *query regeneration* Module is designed to reconstruct the user's original query into a more detailed and enriched version by leveraging the GPT API, thereby improving the matching accuracy when searching for relevant information in the vector store. Specifically, the module employs prompt engineering techniques to augment the original query with additional detailed information about the mentioned events or topics, and to insert approximate date information when such details are missing.

The necessity for this module can be explained from three key perspectives:

3.3.1 Compensation for Inaccurate Queries

The queries provided by users are often vague or inaccurate. While experts in historical facts may formulate highly specific queries, general users or children might express their queries ambiguously or imprecisely. Answers generated from such inaccurate or distorted queries may not only deviate from the required information but may also risk producing entirely erroneous responses due to hallucination. The *query regeneration* Module addresses this issue by refining the query into a clearer and more specific

form. When user's query is received, it is passed to the GPT API. Through system prompting, the GPT model augments the query by adding approximate date information, clarifying ambiguous points, and appending relevant context. The regenerated query is then embedded and used to perform similarity search in the vector store. Due to the rich information compared to the user query and date information written to enable metadata filtering, the accuracy of the vector store search was improved.

3.3.2 Enhancement of Retrieval Efficiency

The RAG framework employed in this study embeds the user's query and compares its similarity with the data stored in the vector store to retrieve the most relevant information. For effective retrieval, the query must inherently contain sufficient detailed information; otherwise, it becomes challenging to match the precise information intended by the user. For instance, if the query "What are the main points of the educational or cultural policies enacted in 1420?" is submitted, the module augments it to: "Please describe the main aspects of the educational or cultural policies implemented under King Sejong's reign in 1420, particularly focusing on the establishment and role of Jiphyeonjeon, the promotion of scholarship, and specific initiatives related to cultural development for the populace." This augmentation enables more precise search results.

3.3.3 Support for Metadata-Based Filtering

We incorporate a *metadata filtering* approach by utilizing the date information associated with each article. During the search process, the date information contained in the user's query plays a crucial role in limiting the search to articles relevant to that date. However, not all users include date information in their queries; in such cases, the search must span the entire corpus, which can reduce accuracy. To mitigate this, the *query regeneration* Module automatically parses and supplements the query with relevant date information. For example, if a query such as "Tell me about the court's reaction when Hunminjeongeum was promulgated" is submitted, the module regenerates it as "Tell me about the court's reaction when Hunminjeongeum was promulgated in 1446," thus enhancing search accuracy by explicitly including the date. The *query regeneration* Module compensates for ambiguities in user queries and, by augmenting additional details, enhances the precision of vector store retrieval—thereby enabling the generation of more accurate and reliable responses.

3.4 Metadata Filtering

During the *dynamic chunking* process, each chunk is assigned three pieces of metadata—year, month, and day—based on the date information included in the article. This metadata is automatically generated by parsing the date data contained in each article. It later enables precise matching with user queries, effectively filtering out information related to specific dates from the vast historical records.

First, the system extracts date information from the user's query. If the query contains complete information for the year, month, and day, the system utilizes all three to select only the chunks that match exactly. For example, if the query "What was the educational policy on May 12, 1420?" is entered, the system filters only those chunks that have metadata corresponding to May 12, 1420, and builds a temporary vector store.

Conversely, if the query includes only the year and month or just the year, appropriate filtering is still performed. In such cases, the system aggregates all chunks corresponding to the specified month or year to create a temporary vector store. For instance, for the query "What was the educational policy in May 1420?", the system targets records for the entire month of May 1420; and for the query "What educational policies were implemented in 1420?", it filters the chunks relevant to the entire year of 1420.

If the user's query does not contain any date information, the *query regeneration* module activates to automatically supplement the query with appropriate date details. This module analyzes the content of the input query to identify related events or topics, extracts the approximate date information associated with the event, and then adds it to the query. Based on the supplemented date information, metadata matching

is performed, and a temporary vector store is constructed from the filtered chunks.

Finally, a similarity search is conducted within this temporary vector store. Because the search is performed only on the chunks selected through the filtering process, unnecessary information is excluded from the entire dataset, resulting in precise search results that align with the user's intent. In this way, *metadata filtering* plays a crucial role in focusing on the specific dates or periods desired by the user, thereby enabling effective search and delivery of information from the vast historical record data.

3.5 Retrieval Augmented Generation

Once the temporary vector store is constructed through the *metadata filtering* and *query regeneration* processes, a similarity search is performed based on the restructured user query. This search selects the article chunks that exhibit the highest similarity with the query. Existing RAG models typically use a fixed number of chunks to retrieve information. However, this fixed approach has clear limitations.

For example, assume a user submits a query regarding diplomatic matters during a specific period. If the diplomatic activities recorded in that period are divided into six separate chunks, but the system is configured to retrieve only five chunks, one important diplomatic record might be omitted. In such a case, the GPT model would generate an answer based on limited information, risking an inaccurate or incomplete response that fails to fully represent the entire diplomatic activity.

To solve this problem, we use the Max Token Retrive method, which takes into account the maximum input token value of the model, rather than the traditional fixed number of chunks. Instead of fixing the information to be delivered by the number of chunks, the information to be delivered can be delivered to the model as much as possible based on the tokens. This method can deliver a relatively large amount of chunks compared to the existing method. Also, due to the *dynamic chunking* described above, the size of each chunk is different, and the number of chunks to be delivered may vary depending on Query.

The extracted chunks are then passed to the

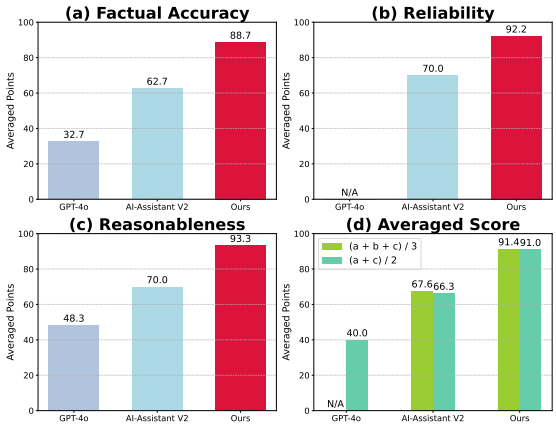

Table 1: Accuracy evaluation of the three models according to the criteria described in the "Evaluation" section. figure (a) shows Factual Accuracy, figure (b) shows Reliability, figure (c) shows Reasonableness, and figure (d), Averaged Score, shows the mean score across all evaluation metrics.

GPT-o1 model, which effectively supports long-context inputs. This model is optimized to understand lengthy contexts and, through a system prompt, selects only the necessary information from the delivered chunks for generating the final answer. Additionally, by setting a maximum token value in line with the GPT-o1 model's capabilities, the model can properly process all the information contained within the extended context.

Ultimately, in the RAG stage, sufficient chunks are secured based on the restructured user query and the dynamic token threshold. The GPT-o1 model then meticulously extracts the relevant information to generate accurate and reliable answers. This process plays a decisive role in enhancing the overall accuracy of information retrieval within the system, contributing to the delivery of rich and detailed historical information to the user.

Overall framework is described in Figure 1.

4 Evaluation

As shown in Table 1, we evaluated the accuracy of our model's answer generation using a benchmark set of questions and compared its performance with two other models: OpenAI's GPT-

4o and ai-assistant v2. GPT-4o is currently one of the most widely used general-purpose language models and is recognized as a model that exhibits high language understanding and generation capabilities. To clearly demonstrate the performance improvement of the RAG-based system proposed in this study, we selected GPT-4o as our baseline, as it produces strong results even without the intervention of a retrieval module. In addition, ai-assistant v2 is the latest RAG-integrated model officially provided by OpenAI; it features an architecture that generates accurate information from large external knowledge sources, representing a typical application of the RAG technique. These two models constitute the respective control groups of a non-RAG LLM and a standard RAG system, enabling us to compare how our proposed system overcomes the limitations of existing models through consistent evidence presentation and high accuracy. A total of 30 benchmark questions were generated based on GPT, and the corresponding ground truth was manually constructed through comprehensive searches of the Joseon Dynasty historical data. The set of questions includes 'erroneous questions' designed to probe the robustness of the models, such as queries about nonexistent events or inquiries that alter dates to reference different periods. These questions were incorporated to verify that the models not only provide accurate responses without hallucination, but also clearly indicate when requested information is unavailable. The evaluation was based on three criteria:

4.1 Factual Accuracy

This criterion assesses whether an answer includes all the necessary information to address the query. For example, if a query asks, "What diplomatic events occurred in October 1420 " and there are five documented diplomatic events, an answer must include all five events to be considered complete. A fully comprehensive answer receives 5 points, while answers missing some events are scored on a scale from 4 to 1. If the answer contains hallucinations or incorrect information, it is scored 0.

4.2 Reliability

This measure evaluates whether the answer clearly specifies its sources and provides concrete evidence for its claims. In the context of historical data, it is crucial that every piece of information is supported by verifiable references. Answers that present clear and complete source information are awarded 3 points; if some references are missing, the answer receives 2 points; if the provided references are only loosely related or overly broad, 1 point is given; and if no source is provided, the answer receives 0 points.

4.3 Reasonableness

This criterion examines whether the answer's interpretation is logical and sufficiently grounded in the evidence provided. An answer that is both logical and well-supported by evidence earns 2 points; if the interpretation is logical but lacks adequate supporting details or references, it earns 1 point; and if the answer is illogical or distorted, it scores 0.

Using these evaluation metrics, we compared our model with GPT-4o and ai-assistant v2. Note that since GPT-4o does not employ a RAG framework, its reliability score was not considered. The evaluation results demonstrate that our model outperformed the other models across all metrics. In particular, our model exhibited strong resistance to hallucination and demonstrated exceptional ability in inferring and analyzing information based on provided sources. In contrast, although GPT-4o generated responses that appeared sophisticated, it was highly susceptible to hallucination, often responding confidently to erroneous queries. Similarly, the ai-assistant v2 model delivered relatively accurate information but frequently omitted essential details and scored lower in reasonableness compared to our model in Fig 3.

5 Discussion

This study implemented an AI-agent system based on the Sejong Annal section of the Annals of the Joseon Dynasty and proposed several techniques to enhance the reliability and transparency of historical information. The

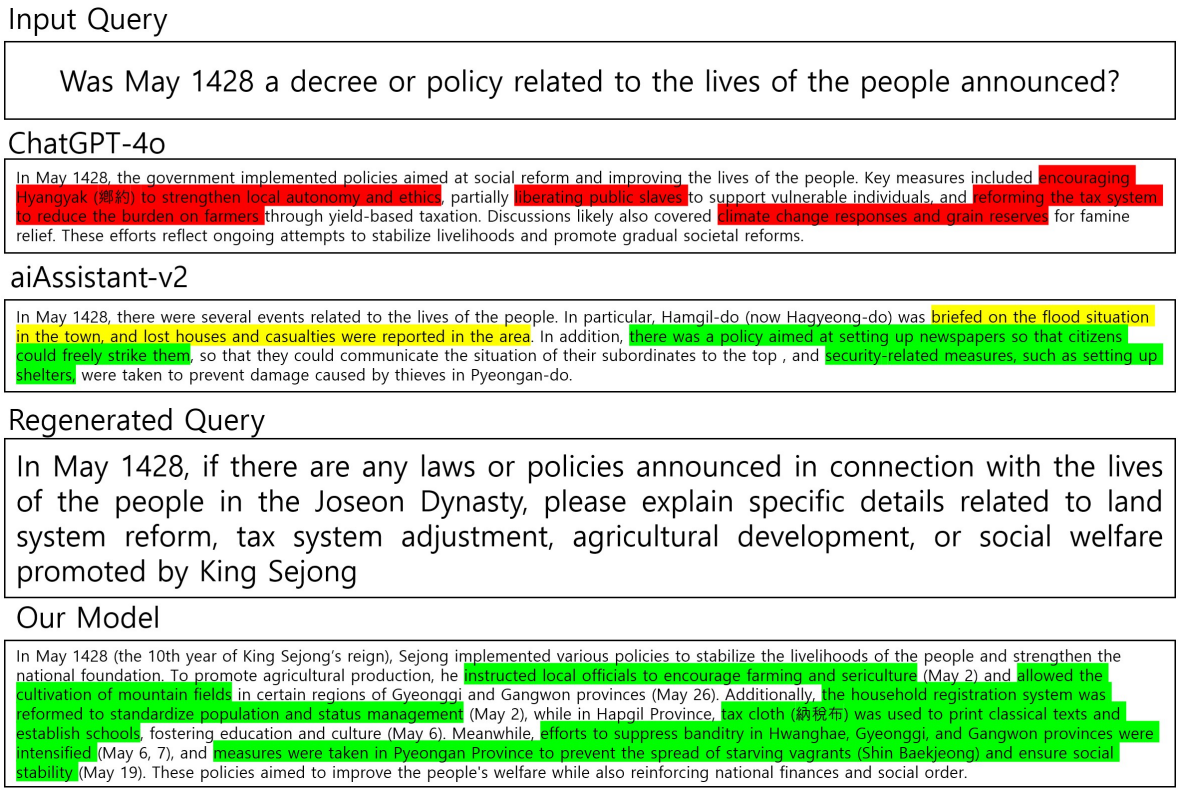

Figure 3: The figure above is part of the benchmark results we conducted. It shows the results of the chatGPT-4o model and the aiAssistant-v2 model for Input Query. Red highlighted section corresponds to hallucinations, yellow highlighted section corresponds to factual or less relevant information, and green highlighted section corresponds to factual and highly relevant information.Our model results through *query regeneration* produce accurate answers with the necessary information.

introduced *dynamic chunking, metadata filtering, query regeneration*, and *Max Token Retrieve* methods effectively overcome the limitations of traditional fixed-token chunking, significantly contributing to the model's ability to generate accurate answers that align with user intent without hallucination.

First, in the data preprocessing stage, the *dynamic chunking* approach—which preserves each article in its entirety—addressed the issue of contextual loss. This ensured that the chunks retrieved from the vector store maintained the full meaning of the original articles. Consequently, the model was able to generate responses that included clear source information, playing a crucial role in providing transparent and reliable information to users.

Furthermore, the *metadata filtering* technique assigns date information (year, month, and day) to each article, enabling the precise selection of information that exactly matches the user's query. When the user's query contains date details or is automatically supplemented with such information via the *query regeneration* module, a temporary vector store is constructed based on the filtered chunks. This process effectively limits the search scope, prevents the inclusion of unnecessary information, and enhances search accuracy.

The *query regeneration* module mitigates the ambiguity and inaccuracy of user queries by reconstructing them into more detailed and specific questions. As a result, the retrieval process better reflects the user's intent, even in cases of erroneous or incomplete date information, thereby contributing to the extraction of accurate data.

In the RAG stage, instead of relying on a fixed number of chunks, the system dynamically collects sufficient chunks based on a pre-

determined token threshold. This approach ensures that all relevant information—for instance, in the case of diplomatic events where details may span multiple chunks—is captured without omission. Such an approach supports the GPT-o1 model's ability to effectively process long-context inputs, ultimately playing a decisive role in generating highly accurate answers.

The comparative evaluation results demonstrated that our system outperformed both GPT-4o and ai-assistant v2 across all evaluation criteria. In particular, our model showed remarkable robustness against hallucination and excelled in reasoning and analysis based on the provided evidence. In contrast, the comparison models exhibited distinct limitations: GPT-4o, despite generating sophisticated responses, was highly susceptible to hallucination, while ai-assistant v2 frequently omitted crucial details.

These findings indicate that in domains dealing with vast amounts of information and complex contexts, such as historical records, *dynamic chunking* and metadata-based retrieval techniques can significantly contribute to effective information extraction and response generation. At the same time, challenges such as information dilution during vector store construction and embedding, as well as chunk size adjustment issues, remain and warrant further research. Future work could involve the integration of various data sources, the utilization of multimodal information, and system enhancements based on real user evaluations.

In conclusion, this study proposed an integrated framework for the precise utilization of historical data like the Annals of the Joseon Dynasty, confirming that such an approach can substantially improve the reliability and accuracy of AI-agent implementations. This method is expected to serve as a valuable reference for the development of information retrieval and response generation systems in diverse domains.

6 Limitations and Future Work

Our study leverages automated database construction and pipeline automation effectively but faces limitations in real-time processing due to query regeneration and database search delays. Future work should optimize search algorithms, enhance distributed processing, refine query regeneration to improve date inference accuracy, adopt flexible date metadata filtering methods like fuzzy matching, and integrate multi-modal data for richer historical context.

We will extend the chatbot system presented in this paper into a more fully realized agent form. To achieve this, we plan to integrate our research on virtual agent systems, motion generation, and speech interaction to implement an agent capable of providing gesture, facial expression, and speech feedback appropriate for real conversational scenarios. Although these capabilities could not be applied in the current study, in our next research, we will incorporate the suggested concrete interaction examples to significantly enhance the system's immersion and usability.

This study is the first case of RAG based agent development using high density records which is the Annals of the Joseon Dynasty, but it has limitations in terms of technological novelty. In the future, we plan to apply the same approach to historical data from other periods and regions to verify versatility and expand the originality and scope of the study by upgrading the optimal search strategy and response generation technique for each literature.

The manual scoring approach used in this study carries the limitation of potential subjectivity in evaluators' judgments. In future work, we plan to introduce standardized evaluation guidelines based on expert review and a multi-evaluator validation process, tailored to the complex domain of historical information, to ensure objectivity and consistency. Through this, we aim to establish a more reliable and reproducible performance evaluation framework.

7 Conclusion

This study developed an AI-agent system based on the Sejong Annals of the Annals of the Joseon Dynasty, designed to extract accurate historical information aligned with user intent. Our approach integrates *dynamic chunking, metadata filtering, query regeneration*, and *Max Token Retreive* method to enhance accuracy and reliability.

Dynamic chunking preserved full context and

source information, addressing limitations of traditional chunking methods. *Metadata filtering* enabled precise retrieval for date-specific queries, while *query regeneration* refined ambiguous queries. The *Max token Retrieve* method delivered relatively large amounts of information in a model so that the information needed to answer could be delivered as much as possible. Comparative evaluations showed our system outperformed GPT-4o and ai-assistant v2 in factual accuracy, reliability, and reasoning, demonstrating robustness against hallucinations.

However, challenges remain. *query regeneration* may introduce incorrect date details, and *metadata filtering* struggles with discrepancies between event occurrence and mention dates. Response delays due to query processing also pose scalability concerns. Future improvements include refining *query regeneration*, enhancing temporal reasoning, and optimizing processing for real-time responses. Expanding to multimodal data could further enrich historical retrieval.

This framework showcases the potential for advanced historical information retrieval, with applications across various domains.

References

[1] Aman Ahluwalia, Bishwajit Sutradhar, Karishma Ghosh, Indrapal Yadav, Arpan Sheetal, and Prashant Patil. Hybrid semantic search: Unveiling user intent beyond keywords. *arXiv preprint arXiv:2408.09236*, 2024.

[2] JinYeong Bak and Alice Oh. Five centuries of monarchy in korea: Mining the text of the annals of the joseon dynasty. In *Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, pages 10–14, 2015.

[3] Sarah A Buchanan. Curation as public scholarship: museum archaeology in a seventeenth-century shipwreck exhibit. *Museum Worlds*, 4(1):155–166, 2016.

[4] Chi-Min Chan, Chunpu Xu, Ruibin Yuan, Hongyin Luo, Wei Xue, Yike Guo, and Jie Fu. Rq-rag: Learning to refine queries for retrieval augmented generation. *arXiv preprint arXiv:2404.00610*, 2024.

[5] European Parliament Archives Unit. Ask the archives: The ep archives unit launches its first generative ai tool. European Parliament News, 2023. Accessed: 2025-04-28.

[6] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2, 2023.

[7] Jeongyeon Hwang, Junyoung Park, Hyejin Park, Sangdon Park, and Jungseul Ok. Retrieval-augmented generation with estimation of source reliability. *arXiv preprint arXiv:2410.22954*, 2024.

[8] KyoHoon Jin, JeongA Wi, KyeongPil Kang, and YoungBin Kim. Korean historical documents analysis with improved dynamic word embedding. *Applied Sciences*, 10(21):7939, 2020.

[9] Hwang Youn Kim, Ghazanfar Ali, and Jae-In Hwang. Enhancing doctor-patient communication in surgical explanations: Designing effective facial expressions and gestures for animated physician characters. *Computer Animation and Virtual Worlds*, 35(3):e2236, 2024.

[10] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

[11] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM computing surveys*, 55(9):1–35, 2023.

[12] Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. Query rewriting in retrieval-augmented large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5303–5315, 2023.

[13] National Heritage Center. Unesco.

[14] National Institute of Korean History. The annals of the Joseon Dynasty.

[15] Vuong M Ngo, Gary Munnelly, Fabrizio Orlandi, Peter Crooks, Declan O'Sullivan, and Owen Conlan. A semantic search engine for historical handwritten document images. In *International Conference on Theory and Practice of Digital Libraries*, pages 60–65. Springer, 2021.

[16] Pat Pataranutaporn, Valdemar Danry, Lancelot Blanchard, Lavanay Thakral, Naoki Ohsugi, Pattie Maes, and Misha Sra. Living memories: Ai-generated characters as digital mementos. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pages 889–901, 2023.

[17] G. Shakir. National archives getting a big boost from ai to transform its search capabilities. FedScoop, Oct. 2024.

[18] Weihang Su, Yichen Tang, Qingyao Ai, Zhijing Wu, and Yiqun Liu. Dragin: Dynamic retrieval augmented generation based on the information needs of large language models. *arXiv preprint arXiv:2403.10081*, 2024.

[19] Zhongxiang Sun, Xiaoxue Zang, Kai Zheng, Yang Song, Jun Xu, Xiao Zhang, Weijie Yu, and Han Li. Redeep: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability. *arXiv preprint arXiv:2410.11414*, 2024.

[20] Georgios Trichopoulos, Markos Konstantakis, George Caridakis, Akrivi Katifori, and Myrto Koukouli. Crafting a museum guide using chatgpt4. *Big Data and Cognitive Computing*, 7(3):148, 2023.

[21] Haneul Yoo, Jiho Jin, Juhee Son, JinYeong Bak, Kyunghyun Cho, and Alice Oh. Hue: Pretrained model and dataset for understanding hanja documents of ancient korea. *arXiv preprint arXiv:2210.05112*, 2022.

[22] Hyunwoo Yoo, Minseok Kang, and Kyungwhan Oh. A semantic search model using word embedding, pos tagging, and named entity recognition. In *2018 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 1204–1209. IEEE, 2018.

[23] Wan Zhang and Jing Zhang. Hallucination mitigation for retrieval-augmented large language models: A review. *Mathematics*, 13(5):856, 2025.

[24] Zijie Zhong, Hanwen Liu, Xiaoya Cui, Xiaofan Zhang, and Zengchang Qin. Mix-of-granularity: Optimize the chunking granularity for retrieval-augmented generation. *arXiv preprint arXiv:2406.00456*, 2024.

8 Author Biography

Jeong Ha Lee is a Master student at AI-Robotics, University of Science and Technology. He completed his bachelor's degree in Art & Technology from Chung-Ang University in 2023. In the spring of 2024, he came to KIST as a UST integrated program student and started working in the MR lab. Now, he is interested in the integration of virtual humans and deep learning models.

Ghazanfar Ali completed his Ph.D. from UST in 2023. His research topic primarily focuses on infusing emotional Intelligence into virtual humans in AR/MR scenarios

by leveraging deep learning of human behavior. He did his B.E Computer Engineering from NUST – Pakistan in 2014. He is a former Faculty Member at a Pakistani University. In the Fall of 2017, he came to KIST as a UST student in the Integrative program and started working in the MR Lab.

Jae-In Hwang is a professor at KIST School and a principal research scientist in Center for Artificial Intelligence at Korea Institute of Science and Technology. He received the B.S., M.S., and Ph.D. degrees in computer science and engineering at Pohang University of Science and Technology, Korea, in 1998, 2000, and 2007, respectively.