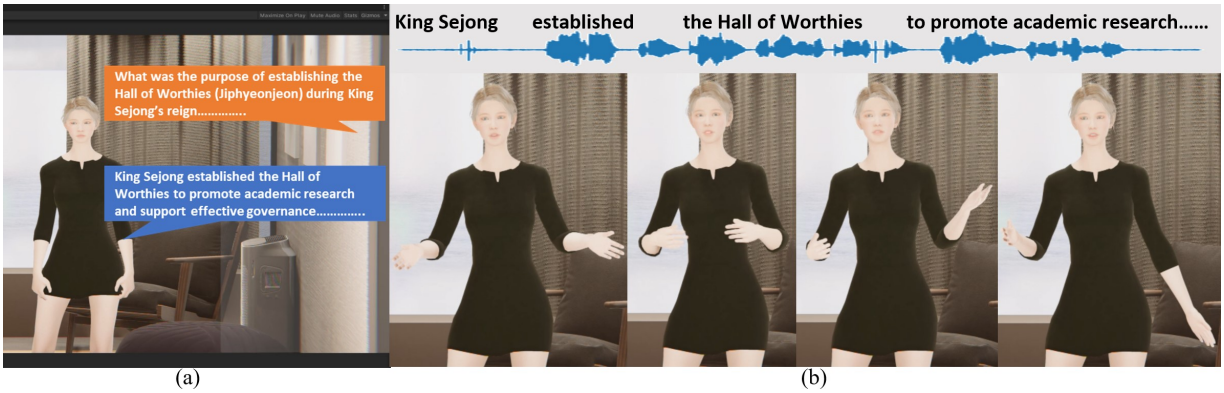
RAG based AI-Agent for Contextualized Analysis of High-Density Historical Records: Application to the Annals of the Joseon Dynasty



(a) (b)

Figure 1: (a) shows an example of our AI-Agent system interacting with a user's historical information query. The user inputs a question via voice, and the system responds with audio output. (b) shows the AI-Agent generating and displaying body animation synchronized with the context of its response.

Abstract

The field of digital heritage has increasingly focused on digitizing and reinterpreting historical materials to preserve and transmit cultural assets. Accurately conveying the content of historical records is critical for both academic research and education. Traditional digital archiving and retrieval systems have commonly relied on keyword-based searches or simple text matching, resulting in limited contextual understanding, insufficient source citation, and inadequate alignment with user intent. To address these challenges, this paper proposes a novel methodology that integrates a large language model (LLM) with a Retrieval-Augmented Generation (RAG) framework for high-density historical records such as the Annals of the Joseon Dynasty (49,646,667 characters, 1392-1910).Our system retrieves the most relevant historical sources, and delivers both objective facts and contextual analysis. Ultimately, we integrate this system with a Unity-based 3D AI-Agent, providing answers through synchronized voice output and body motion for an immersive interactive experience. By enabling natural, embodied interaction, this integration makes historical knowledge more accessible and engaging for users. Experimental results on a set of 30 benchmark questions demonstrate that our model outperforms both ChatGPT-4o and AI-Assistant v2 in Factual Accuracy, Reliability, and Reasonableness. This research expands the potential applications of digital heritage and lays the groundwork for broader integration with diverse historical data sources.

Index Terms: AI-Agent, LLM, Retrieval-Augmented Generation

1 Introduction

The Annals of the Joseon Dynasty, a UNESCO Memory of the World, is a comprehensive historical record that high-density documents over 500 years of Korean history[2]. This unparalleled collection provides daily, chronological accounts of the Joseon court, covering political, social, and cultural developments that shaped Korean society. Accurately conveying information from such historical sources is essential for academic research, cultural preservation, and public education. Ensuring factual correctness and providing clear references not only supports scholarly integrity but also helps users, both experts and the general public, understand complex historical contexts without misinterpretation or distortion. However, traditional digital archiving and retrieval systems have typically relied on keyword-based searches or simple text matching. These methods often fail to capture nuanced context, offer limited source transparency, and may not fully reflect the user's intent especially given the annal style, chronological format of the records, where the same event can appear in multiple contexts and sometimes even present contradictory details.

To address these limitations, we propose an advanced AI-Agent system that integrates a large language model (LLM) with a Retrieval-Augmented Generation (RAG) framework[1]. Our approach leverages the full depth of the Joseon Annals, particularly the Sejong Annals, using dynamic chunking, metadata filtering, and query regeneration to preserve context, enhance source accuracy, and deliver evidence-based answers. Notably, the system is integrated with a Unity-based AI-Agent (see Figure 1), offering users an immersive and interactive experience through synchronized voice and body animation. Our method achieves significant improvements in Factual Accuracy, Reliability, and Reasonableness compared to ChatGPT-4o and AI-Assistant v2, and sets the stage for broader applications in digital heritage and historical research.

2 Method

2.1 Data Preprocessing

We collect and unify records from the Sejong Annals through web crawling. Instead of breaking the text into fixed-size pieces, we use dynamic chunking so that each article remains whole. This preserves context and makes it easier to track sources, but can also lead to large chunks that sometimes mix different topics. To address this, we extract the year, month, and day data from each article and use these as metadata for more precise searches.

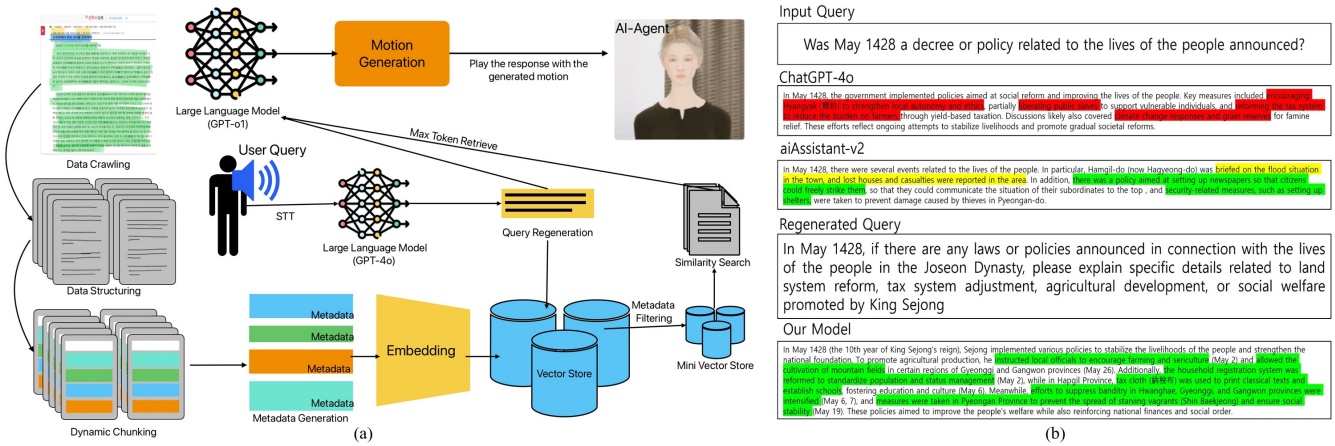


Figure 2: (a) shows the overall pipeline of our system, including data preprocessing, response generation, and AI-Agent operation. Using a vector store built from the Annals of the Joseon Dynasty, the system generates accurate responses to user queries and interacts through synchronized motion and audio. (b) presents part of our benchmark results, comparing the responses of ChatGPT-4o and AI-Assistant v2 for a given query. Red highlights indicate hallucinations, yellow indicates factual but less relevant information, and green shows factual and relevant information.

2.2 Query Regeneration

User queries are often imprecise or missing key details, such as exact dates or event descriptions. Such unclear user queries can make it difficult to retrieve highly relevant historical information from the vector store through similarity search. To solve this, we use the GPT API to regenerate the query, automatically adding missing context or date information. This makes the search more focused and increases the likelihood of retrieving relevant information.

2.3 Metadata Filtering

Our System uses the date information in the user query to narrow down the dataset (see Figure 2 (a)). If the query includes a specific date, only the records from that day are used. If it only includes a year or month, the system selects all records from that period. If no date is included, the query regeneration module provides it, ensuring that relevant records are always prioritized.

2.4 Retrieval Augmented Generation

Instead of setting a fixed number of chunks to retrieve, the system gathers as many relevant records as the model's token limit allows. This ensures that all necessary context is available for the model to generate accurate and thorough answers. The model reviews the retrieved records, extracts the key information, and delivers a clear, reliable, and well-referenced response.

2.5 Integrated In AI-Agent Interface

We integrated our system with a Unity-based AI-agent, enabling users to submit queries via voice and receive responses delivered with synchronized speech and motion. When a user's voice input is received, it is converted to text using a speech-to-text (STT). then, processes the text to generate an appropriate answer. Simultaneously, corresponding motion data are generated to match the response, and the final response is presented to the user through text-to-speech (TTS) synthesis along with the associated body and lip motion. It provides users with a immersive and interactive experience.

3 EXPERIMENTS

We evaluated our model's answer generation accuracy using a benchmark set of 30 questions, and compared its performance against ChatGPT-4o and AI-Assistant v2. The benchmark included

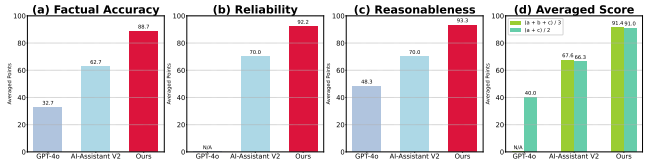


Table 1: Accuracy evaluation of the three models according to the criteria described in the "Experiments" section. figure (a) shows Factual Accuracy, figure (b) shows Reliability, figure (c) shows Reasonableness, and figure (d), Averaged Score, shows the mean score across all evaluation metrics.

not only standard historical questions but also 'erroneous questions', queries about nonexistent events or incorrect dates, to test each model's robustness and ability to avoid hallucination.

For evaluation, we used three criteria: Factual Accuracy (does the answer include all key facts and avoid errors or hallucination), Reliability (does it clearly cite sources and provide evidence), and Reasonableness (is the answer logical and grounded in evidence).

As shown in Table 1 and Figure 2 (b), our model consistently outperformed both ChatGPT-4o and AI-Assistant v2 across all metrics. Our approach was especially effective in resisting hallucinations and providing evidence-based, complete answers, while ChatGPT-4o, despite generating fluent responses, frequently hallucinated or responded confidently to false queries. AI-Assistant v2, while more reliable than ChatGPT-4o, often omitted key details and provided less comprehensive reasoning than our system.

4 CONCLUSION

Our AI-agent system effectively delivers accurate and reliable information from high-density historical data, helping users gain deeper insights into the history. This research shows the potential of AI for exploring complex historical records and lays the foundation for broader applications in cultural and educational fields.

REFERENCES

[1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020. 1

[2] National Institute of Korean History. The annals of the Joseon Dynasty. 1