

Article

An Intelligent Docent System with a Small Large Language Model (sLLM) Based on Retrieval-Augmented Generation (RAG)

Taemoon Jung  and Inwhee Joe * 

Department of Computer Science, University of Hanyang, Seoul 04763, Republic of Korea; moiiom@hanyang.ac.kr

* Correspondence: iwjoe@hanyang.ac.kr; Tel.: +82-2-2220-1088

Abstract

This study designed and empirically evaluated a method to enhance information accessibility for museum and art gallery visitors using a small Large Language Model (sLLM) based on the Retrieval-Augmented Generation (RAG) framework. Over 199,000 exhibition descriptions were collected and refined, and a question-answering dataset consisting of 102,000 pairs reflecting user personas was constructed to develop DocentGemma, a domain-optimized language model. This model was fine-tuned through Low-Rank Adaptation (LoRA) based on Google's Gemma2-9B and integrated with FAISS and OpenSearch-based document retrieval systems within the LangChain framework. Performance evaluation was conducted using a dedicated Q&A benchmark for the docent domain, comparing the model against five commercial and open-source LLMs (including GPT-3.5 Turbo, LLaMA3.3-70B, and Gemma2-9B). DocentGemma achieved an accuracy of 85.55% and a perplexity of 3.78, demonstrating competitive performance in language generation and response accuracy within the domain-specific context. To enhance retrieval relevance, a Spatio-Contextual Retriever (SC-Retriever) was introduced, which combines semantic similarity and spatial proximity based on the user's query and location. An ablation study confirmed that integrating both modalities improved retrieval quality, with the SC-Retriever achieving a recall@1 of 53.45% and a Mean Reciprocal Rank (MRR) of 68.12, representing a 17.5 20% gain in search accuracy compared to baseline models such as GTE and SpatialNN. System performance was further validated through field deployment at three major exhibition venues in Seoul (the Seoul History Museum, the Hwan-ki Museum, and the Hanseong Baekje Museum). A user test involving 110 participants indicated high response credibility and an average satisfaction score of 4.24. To ensure accessibility, the system supports various output formats, including multilingual speech and subtitles. This work illustrates a practical application of integrating LLM-based conversational capabilities into traditional docent services and suggests potential for further development toward location-aware interactive systems and AI-driven cultural content services.



Academic Editors: Francisco De Arriba-Pérez and Silvia García-Méndez

Received: 25 July 2025

Revised: 22 August 2025

Accepted: 22 August 2025

Published: 27 August 2025

Citation: Jung, T.; Joe, I. An Intelligent Docent System with a Small Large Language Model (sLLM) Based on Retrieval-Augmented Generation (RAG). *Appl. Sci.* **2025**, *15*, 9398. <https://doi.org/10.3390/app15179398>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: small Large Language Model (sLLM); Retrieval-Augmented Generation (RAG); Intelligent Docent; Museum; LangChain; LoRA; HuggingFace; Natural Language Processing model (NLP); Speech-to-Text (STT); Ultra-Wideband (UWB); Facebook AI Similarity Search (FAISS)

1. Introduction

1.1. Research Background

Museums and art galleries are spaces that fulfill a public mission of conveying cultural arts and historical knowledge, and interpretation services that help visitors deeply

understand the exhibits have become an essential element. Traditional docent services have performed this role, but have limitations in encompassing all visitors due to issues such as staff shortages, language barriers, and a lack of consideration of information-vulnerable groups [1]. In particular, visually and hearing-impaired individuals and foreign visitors often struggle to access sufficient information through existing audio guides or text-based interpretation services, leading to disparities in their exhibition-viewing experiences [2].

As a result, there has been an increase in attempts to implement real-time interactive commentary systems using artificial intelligence-based natural language processing technology, particularly Large Language Models (LLMs) [3]. However, general-purpose LLMs are not optimized for specialized domains such as docent services, and their massive computational resource requirements make them impractical for application in actual mobile or kiosk environments [4]. This study aims to overcome these limitations by utilizing a small Large Language Model (sLLM) specialized for museum exhibition commentary and designing an intelligent docent system that integrates various functions such as indoor location-based exhibit recognition, multilingual support, and TTS output. Additionally, the system was pilot-tested in a museum to collect user feedback and validate its effectiveness and scalability through quantitative and qualitative analysis. This study aims to contribute to improving information accessibility in the metaverse environment and presenting a new paradigm for enjoying cultural content through intelligent docent services.

This paper is structured as follows: Section 1 presents the research background and objectives, Section 2 reviews related studies, and Section 3 presents the methodology of this study. Section 4 covers the preparation process and results for performance evaluation, as well as pilot cases, and Section 5 discusses the conclusions and future research directions.

1.2. Research Purpose

The main purpose of this study is to design an RAG (Retrieval-Augmented Generation)-based intelligent docent system to overcome the limitations of information accessibility and explanatory power in museum and art gallery exhibition environments and to verify its usability and effectiveness through quantitative and qualitative evaluations. Additionally, beyond the simple application of LLMs, this study aims to empirically analyze how an RAG system specialized for docent services can function in actual viewing environments and explore the applicability of language generation technology in the field of cultural content.

The specific research objectives are as follows:

(1) Design and implementation of an RAG-based question-answering structure:

Embed museum domain commentary data and design an RAG structure that combines similar document search and response generation to generate explanatory responses to natural language queries in real time.

(2) Development of a docent domain-specific sLLM:

Build and train a small-scale LLM (sLLM) optimized for museum commentary rather than a general-purpose LLM to ensure high response accuracy and fluency for exhibition content [5].

(3) Generation of user-customized responses by query type:

Appropriate document combinations and prompt structures are applied to various query types, such as exhibition descriptions, artist-related questions, and historical background, to construct explanations that meet user requirements.

(4) Evaluate usability and information delivery through empirical experiments:

Apply the developed system to actual museum sites and conduct quantitative and qualitative evaluations focusing on user satisfaction, comprehension, and response reliability to verify the effectiveness of the technology and derive directions for improvement [6].

2. Background & Related Works

In this section, we review related technologies and previous studies to establish the theoretical basis for intelligent docent systems. First, we examine trends in LLMs and their lightweight version, sLLMs, and discuss their applicability for real-time inference. Next, we analyze RAG structures and domain-specific application cases and examine the possibility of combining image- and video-based information through multimodal RAG. Additionally, we introduce vector search technologies based on FAISS and OpenSearch, as well as precise search methods based on morphological analysis. Finally, we derive the relevance to this study by presenting examples of component integration and automation using the LangChain framework.

2.1. Large Language Model (LLM) Research Cases

Large Language Models (LLMs) are artificial intelligence technologies that learn from large amounts of text data to achieve human-level natural language understanding and generation capabilities. Various architectures, such as GPT-3.5, BERT, and T5, have been commercialized [7–9]. These models contain over a billion parameters and demonstrate outstanding performance in various Natural Language Processing (NLP) tasks, including conversational AI, summarization, translation, and question-answering [10].

2.1.1. The Rise of Multimodal LLM (MLLM)

LLMs are evolving into MLLMs, which understand and generate various modalities such as images, audio, and video beyond text. Google's PaLM 2 attempts to mimic human multisensory cognitive abilities by simultaneously processing text and image information to respond to complex questions or generate image captions [11]. Additionally, there are increasing cases where text and images are combined for applications such as visual question-answering and image description generation. The intelligent docent system in this study plans to provide visual information through AR glasses and guide robots, and it is expected that integration with MLLMs will enable richer interactions, such as directly understanding the visual features of exhibits and reflecting them in explanations.

2.1.2. Combination with Agentic AI

Research is actively underway to develop LLMs beyond simple question-and-answer systems into agentic AI that can utilize external tools and autonomously perform complex tasks [12]. Research cases have been reported where the LLM accesses various external tools such as calendar apps, email, and web browsers to process complex user requests (e.g., "Change tomorrow's meeting to 10 a.m. and send a summary of the meeting minutes to the participants") [13]. This can contribute to providing more intelligent services, such as docent systems that actively search for exhibition-related information based on visitors' questions, link with route guidance systems to suggest optimal viewing routes, and even perform museum event reservations and shop guidance.

2.1.3. Reinforcement Learning-Based Alignment

Research on alignment, which controls LLM outputs to align with human intentions and ethical standards, is actively underway, with Reinforcement Learning from Human Feedback (RLHF) emerging as a key technology [14]. RLHF involves humans assigning preferences to multiple responses generated by the LLM and using these preferences as reward signals to train the model, thereby reducing harmful or biased responses and encouraging the generation of beneficial and truthful responses [15]. For example, in domains such as medicine and law, where high reliability and ethical standards are required, research cases measuring and mitigating LLM bias have been presented [16].

2.2. Small Large Language Model (sLLM) Research Cases

LLMs require large-scale computational resources (GPUs/TPUs), high memory capacity, and long processing times, making them unsuitable for mobile environments or device-centric applications that require real-time responsiveness and lightweight performance [17]. To overcome these limitations, small Large Language Models (sLLMs) have emerged, focusing on reducing the number of parameters and improving response speed and resource efficiency [18]. Notable examples include Alpaca, Vicuna, Mistral, and Google's Gemma2-9B, which was adopted in this study. These models feature lightweight structures compared to LLMs through technologies such as Knowledge Distillation, Quantization, and LoRA (Low-Rank Adaptation). In particular, despite being a relatively small-scale model, Gemma2-9B has secured competitiveness in terms of sentence generation fluency and semantic consistency, and its Korean language processing performance is also excellent, making it an appropriate base model for the docent domain [19,20].

2.2.1. On-Device AI and Edge Computing

The trend toward lightweight sLLMs is reducing dependence on cloud servers and accelerating the development of on-device AI technology that runs AI models directly on edge devices such as smartphones and AR glasses. Examples include lightweight translation models running on smartphone apps and AI assistant services that process speech in real time on edge devices. Such on-device AI offers the advantages of maximizing real-time response speed, reducing network latency, and strengthening privacy protection.

2.2.2. RAG-Based sLLM Optimization

Compared to LLMs, sLLMs have less pre-trained knowledge, which can lead to more frequent hallucination issues. Therefore, research is actively underway to improve performance by combining sLLMs with RAG (Retrieval-Augmented Generation) techniques that utilize external knowledge bases [21]. Examples of its application span various industries, such as combining sLLMs and RAG in medical chatbots to search for the latest medical information and provide accurate answers, or in legal consultation systems to reference legislation data and present relevant precedents. To minimize latency and performance degradation at each stage of the RAG pipeline (document chunking, embedding, search, and generation), techniques such as prompt compression, caching, and efficient indexing strategies are being actively developed to minimize latency and performance degradation at each stage of the RAG pipeline (document chunking, embedding, search, and generation and joint optimization). One effective hyperparameter optimization technique is the RGHL (Rapid Genetic Exploration with Random Direction Hill-Climbing Linear Exploitation) algorithm, which combines global genetic exploration with surrogate-based local exploitation. RGHL HPO enables the efficient tuning of key parameters such as chunk size, embedding dimensions, and retrieval depth—crucial for optimizing RAG pipelines in resource-constrained environments. The algorithm has demonstrated strong performance in high-dimensional, low-resource search spaces, making it well-suited for edge-deployed sLLMs that demand both scalability and fast inference [22].

2.2.3. Hybrid Search

To improve search accuracy in RAG systems, hybrid search, which combines traditional keyword-based search (e.g., BM25) and vector similarity search, is gaining attention [23]. This contributes to improving search accuracy by considering not only semantic similarity but also keyword matching [24]. Research in this area includes studies on recommending more suitable products by simultaneously utilizing product name keywords and sentence vectors indicating the user's purchase intent when searching product catalogs in response to user queries on e-commerce platforms.

2.3. Retrieval-Augmented Generation (RAG) Research Cases

Retrieval-Augmented Generation (RAG) is a structure proposed to solve the problem of large language models generating incorrect information (hallucinations) about knowledge not included in the training data, integrating the document retrieval and generation processes. It works by first searching for relevant documents from an external knowledge base or vector database in response to a question, then incorporating them into the input of the generation model to generate more accurate and contextually appropriate responses. The RAG structure primarily consists of two components. The Retriever selects documents that are semantically similar to the input query, using embedding-based search techniques such as Dense Passage Retrieval (DPR), BM25, GTR, and GTE [25]. The Generator is a language model that generates responses based on the selected documents, and most often uses a Transformer-based decoder model [26]. In this study, we adopted a structure that combines FAISS-based vector search and Amazon OpenSearch in the document retrieval stage, and designed a Spatio-Contextual Retriever (SC-Retriever) that reflects the exhibition location information to improve the context and accuracy of the docent system. Recent research trends show that structures are becoming more sophisticated, such as Multi-hop RAG for reasoning between multiple documents, Contextual RAG including prompt optimization, and Spatio-RAG for spatio-temporal information expansion. This study reflects these trends and applies them to domain-specific applications.

2.3.1. Multi-Modal RAG

Research on Multi-modal RAG, which goes beyond text RAG to search for various modalities such as images and videos and provide them to the LLM, is emerging. This demonstrates cases where relevant images or video clips are searched and presented together when generating answers to text queries, thereby enriching the information [27]. In the docent system described in this paper, actual images of exhibits and related video materials are searched and utilized in explanations, enabling the provision of more abundant and visual information to visitors.

2.3.2. Self-Correcting RAG

Self-correcting RAG techniques, which enable LLMs to evaluate the reliability of their own responses and perform additional searches to improve responses in cases of uncertainty, are being researched. Such self-correcting RAG can be utilized in fields such as finance and law to enable models to recognize information gaps on their own and search for additional materials to generate more reliable answers to complex questions.

2.3.3. Continual Learning RAG

Continual learning RAG methodologies, which continuously reflect new information in LLMs in environments where knowledge changes frequently, are becoming increasingly important. Research is underway to quickly reflect changed knowledge in RAG modules and update them when LLMs generate responses based on the latest news articles or real-time data feeds.

2.4. Vector Database (Vector DB) and FAISS Research Cases

The performance of a search engine in a RAG system is entirely determined by the accuracy of similarity calculations between document embeddings and search speed. Vector databases store document and query embeddings and return the most similar documents to the query vector, thereby playing a central role in RAG [28]. FAISS (Facebook AI Similarity Search) is a library developed for similarity search between high-dimensional vectors, supporting algorithms such as Inverted File Indexing (IVF), Product Quantization (PQ),

and HNSW. It maintains response speeds of under 100 ms even in large-scale vector search environments through GPU-based parallel processing [29]. In this study, we utilized FAISS's IVF-PQ structure to index over 200,000 exhibition descriptions and question-answering data, minimizing real-time search response latency. Additionally, we integrated Amazon OpenSearch, a distributed vector search platform, to ensure scalability and operational stability, and incorporated the Nori morphological analyzer into the preprocessing stage to enhance search precision for Korean-language data.

2.4.1. Large-Scale Image and Video Search

FAISS is used to store millions or tens of millions of image and video feature vectors and quickly find content similar to a specific image. For example, it can be applied to systems that recommend products similar to images uploaded by users on online shopping malls or find clips similar to specific scenes on media content platforms [30].

2.4.2. Recommendation Systems

FAISS is used to convert user profiles or item features into vectors, store them in a vector database, and provide personalized recommendations through similarity search [31]. Representative examples include recommending songs similar to a user's listening history in music streaming services or finding movies similar to a specific movie in movie recommendation services [32].

2.4.3. Anomaly Detection System

Research is underway to contribute to system security and stability by quickly detecting vectors that deviate from normal patterns (outliers) through FAISS by converting network traffic data or system logs into vectors [33].

2.5. LangChain-Based Integrated Framework Research Cases

LangChain is an open-source framework that allows various language model components to be modularized and connected, providing high scalability and flexibility in constructing LLM application systems. In particular, it allows for the chain-based management of components such as Retriever, Prompt Template, Memory, Tool, and Output Parser, making it advantageous [34,35] for structurally controlling complex RAG systems. Therefore, we built a RetrievalQA chain based on LangChain, which integrates user query processing and prompt template configuration with FAISS-based vector search and sLLM response generation. In addition, LangChain was used to implement features for operational automation, such as automatic embedding and the updating of documents within the system, automatic registration of prompts when new exhibits are added, and log-based analysis of user interaction records [35,36]. This structure functions as an effective framework in terms of maintenance in environments with short content cycles and frequent changes, such as museums and exhibition halls [36].

2.5.1. LLM-Based Chatbot Development

LangChain is widely used to develop LLM-based chatbots that process user queries and generate responses [34]. LangChain can be used to efficiently build internal chatbots that answer questions based on internal company documents, customer service chatbots that automatically handle customer inquiries, and expert chatbots that respond like experts on specific topics [37].

2.5.2. Code Generation and Automation

LangChain can be used to generate programming code or automate data analysis pipelines using LLMs. When a developer enters a request in natural language, LangChain

transmits it to the LLM to generate code and, if necessary, calls external tools to execute the code [37].

2.5.3. Data Analysis and Automatic Report Generation

LangChain can be used to analyze large datasets and build systems that automatically generate summary reports and dashboards based on the results. It can be applied to automatically generate monthly reports by analyzing market trend data or to summarize research data to create draft reports [38].

3. Research Methodology

This section describes the structure and main components of an intelligent docent system for information-vulnerable groups. The system is centered on a RAG-based SC-Retriever and an sLLM and a domain-specific vector database has been built for answering questions about exhibits. It also provides user-customized commentary through the LangChain framework and UWB-based location recognition.

3.1. Overview of the Overall Configuration

This intelligent docent system has a structure that can be divided into a development environment and a commercial environment, as shown in Figure 1. It integrates various functions, such as LLM-based question and answer, RAG document search, and UWB-based location information utilization, to define efficient data flow and functions between development and operation.

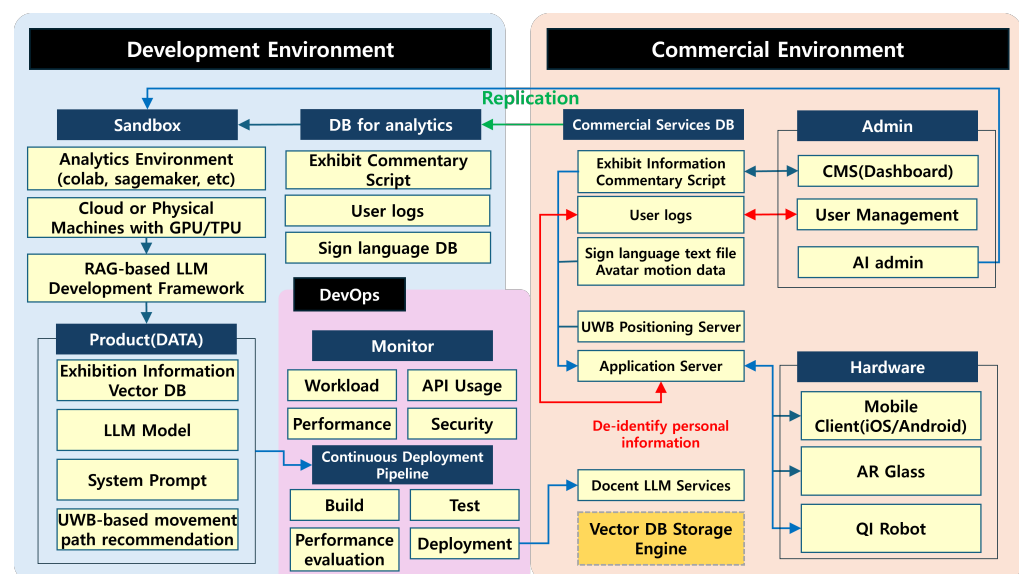


Figure 1. Intelligent docent system configuration diagram.

3.1.1. Development Environment

RAG-based LLM frameworks operate on cloud analytics platforms such as Colab or SageMaker, or on GPU/TPU-based physical equipment. This generates exhibition commentary vector databases, LLM models, system prompts, and other outputs. These outputs undergo automated testing, building, and deployment processes through DevOps and monitoring pipelines before being applied to commercial systems. In addition, experimental data (exhibition commentary, user logs) is managed in an analysis database and used for future retraining and evaluation.

- **Development Database:** This mainly manages source data necessary for research and learning, such as exhibition description texts, Q&A datasets, user persona data, and

pilot user logs, and is composed of a hybrid structure of a relational database (MySQL) and a vector database (OpenSearch) that stores embedded vectors. This database stores more than 199,000 exhibition description data and over 102,000 question–answer pairs in JSON format after pre-processing, vectorized using a multilingual sentence embedding model, and indexed through FAISS and Amazon OpenSearch. In addition, researchers can freely perform performance tuning, reflect new data, and configure the experimental environment, and replication from the development database to the commercial database is performed through an ETL pipeline (Extract-Transform-Load). Data from the development environment are transmitted in the background in periodic or trigger-based real-time, and exhibition content, Q&A, user logs, etc., are anonymized and preprocessed before being sorted and version-controlled in JSON format for transmission.

3.1.2. Commercial Environment

This refers to the operating system connected to terminals such as mobile apps, AR glasses, and guide robots used by visitors in actual museums, art galleries, and exhibition halls.

- **Commercial Database:** The commercial database is configured with a relational database (MySQL) and a vector database (OpenSearch) for real-time RAG search to store the latest exhibition information, user sessions, and logs through regular replication and synchronization with the development database, and to manage personalized data such as UWB-based location information, user language settings, and preferences. This database supports high-speed vector search and response generation to quickly respond to user queries and location-based search requests and is configured as a distributed cluster to provide high availability and stable response speeds. Additionally, service operators use commercial databases to perform user support, content update monitoring, and service log analysis tasks. User text or voice queries are vectorized through the Application Server and then quickly retrieved from the OpenSearch vector DB using the ANN (Approximate Nearest Neighbor) search method [39].
- The UWB Positioning Server tracks visitors' locations in real time to enable location-based content delivery [40].
- Docent LLM Services and Application Server search for relevant documents in the vector DB in response to visitors' queries and generate natural explanatory responses through LLM to deliver to users [41].
- Admin functions are responsible for system maintenance and content management, including the administrator dashboard, user management, and AI status monitoring.
- Hardware terminals are implemented as mobile apps (iOS/Android), augmented reality glasses (AR Glass), and exhibition guide robots (QI Robot), and support various output formats (text, voice).

User logs collected from commercial systems are anonymized and returned to the development environment, where they are used for model retraining and usability improvements. This design enables continuous system improvements based on real-time user feedback.

The intelligent docent system is based on the integrated operation of LLM and RAG technologies and consists of an input interface, information retrieval module (Retriever), generation module (sLLM), output rendering module, and CMS (Dashboard) for operation management. Users can submit text- or voice-based queries, and the system embeds them and searches for similar documents in the FAISS-based vector DB [29]. The retrieved documents are inserted into the prompt through LangChain, which has the structure

shown in Figure 2, and the docent-specific sLLM generates an appropriate explanatory response [42]. The generated response is output in text and voice formats according to the user's settings.

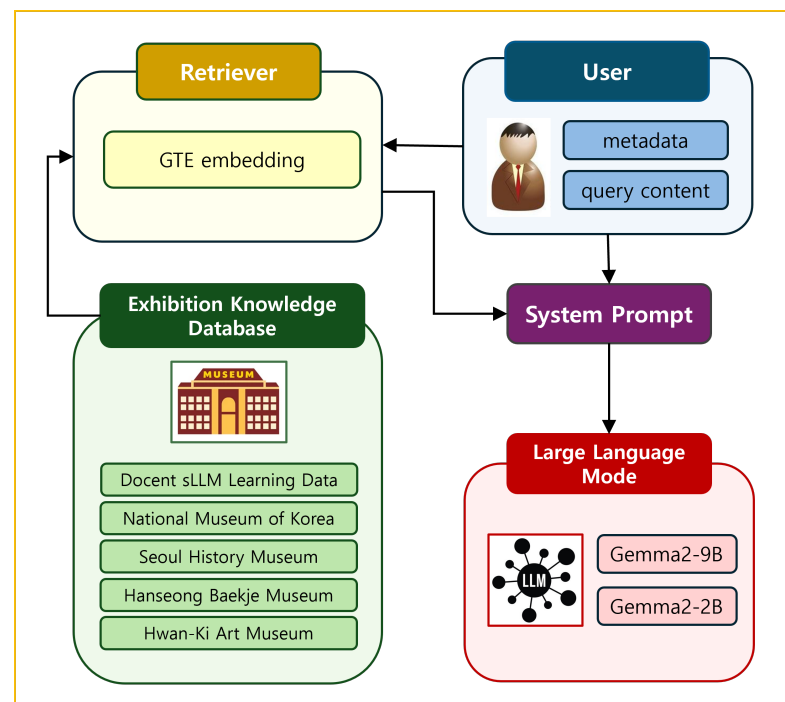


Figure 2. API deployment (LangChain).

3.2. Detailed Description of Applied Technologies

3.2.1. Information Search Module

User queries are vectorized using the Hugging Face's Sentence Transformer model, and the same embedding structure is applied to documents to enable consistent searching [43]. FAISS applies the IVF+HNSW algorithm to maintain an average search response speed of less than 150 ms, and parallel searches are performed in a multi-GPU environment. SC-Retriever improves search accuracy by reflecting spatial information such as user location and exhibition location as weights in addition to document semantic similarity.

3.2.2. Generation Module

LangChain's Prompt Template, RetrievalQAChain, and OutputParser modules were combined to construct a RAG flow based on user questions [44]. The generation model uses Gemma2-9B fine-tuned with LoRA in the AWS EC2 A100 environment, and the prompt configuration is subdivided according to literacy, query type, and whether or not to include document summaries, with the optimal template applied [45,46].

3.2.3. Ultra-Precise Indoor Positioning Technology (UWB-Based)

To implement a service that precisely identifies the indoor location of museum visitors and automatically provides appropriate exhibition commentary information for that location, we have introduced Ultra-Wideband (UWB) indoor positioning technology [47]. This enables real-time, high-precision location tracking in spaces with a high density of exhibits, such as museums and art galleries, which cannot be achieved with simple BLE-based location recognition [48]. Figure 3 shows the interaction between visitors and the museum regarding this feature.

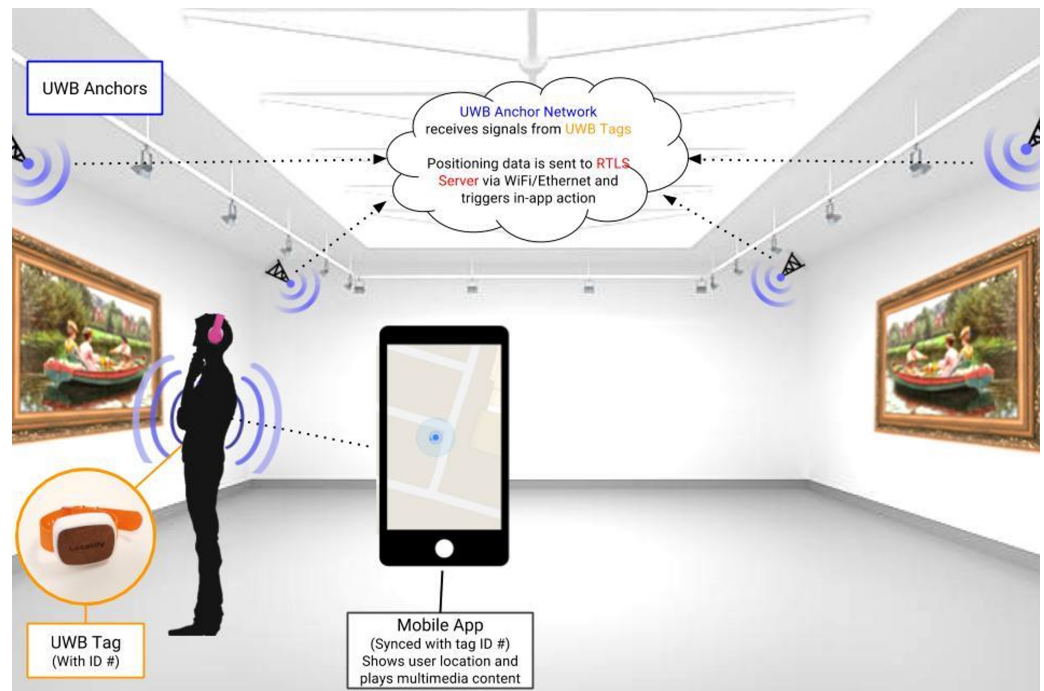


Figure 3. Providing situation-aware audio guides based on location tracking of visitors carrying UWB tags.

By calculating the distance (Time of Flight, ToF) between UWB transceiver modules (tags and anchors), we estimate the user's location in 3D space with an accuracy of approximately 1.5 m to 3 m.

$$d_i = c \cdot \frac{(t_{rx,i} - t_{tx})}{2} \quad (1)$$

- d_i is the distance between the UWB tag and the i -th UWB anchor.
- c is the propagation speed (approximately 3×10^8 m/s).
- t_{tx} is the transmission time.
- $t_{rx,i}$ is the reception time.

Equation (1) is based on the Time-of-Flight (ToF) principle, which calculates the distance d_i between a transmitter and a receiver by measuring the propagation delay of an Ultra-Wideband (UWB) signal. The formula assumes that the signal travels at the speed of light c in free space and that the system is either time-synchronized or applies a two-way ranging mechanism to compensate for clock offset. This method is widely adopted in indoor positioning systems due to its high temporal resolution and robustness against multipath interference [49].

The user's location coordinates are mapped to exhibit IDs, and the system is configured to automatically call up the corresponding explanation content when the user reaches a specific exhibit [50]. This mapping is linked to the OpenSearch-based exhibit DB, and a hysteresis filter is applied to prevent unnecessary duplicate calls when switching content due to location changes [51].

Equation (2) indicates that the exhibition commentary content starts only when the distance is less than R_{enter} and ends only when the distance is greater than R_{exit} . This condition is commonly applied in location-based services using BLE or UWB and is effective in preventing duplicate content calls and maintaining consistency in the user experience [40].

Hysteresis filter (prevents duplicate content calls): A filter with different conditions set for entering and exiting within a certain distance from the exhibit.

$$\text{Enter if } d < R_{\text{enter}}, \quad \text{Exit if } d > R_{\text{exit}}, \quad R_{\text{exit}} > R_{\text{enter}} \quad (2)$$

- d is the distance between the current user location and the center of the exhibit.
- R_{enter} is the content trigger entry radius.
- R_{exit} is the content termination radius.
- UWB Tag: Acts as a transmitter worn or carried by the user [52]. (Center of Figure 4)
- UWB Anchor: A receiver fixed to the ceiling or wall of the exhibition room. At least 3–4 anchors are used to apply the triangulation method [53]. (Left side of Figure 4)
- UWB Anchor Network: Collects and calculates distance calculation results in real time to calculate the user's coordinates (x, y, z) [49].
- Mobile App: Provides customized explanations of exhibits based on location and plays related multimedia content via the app.

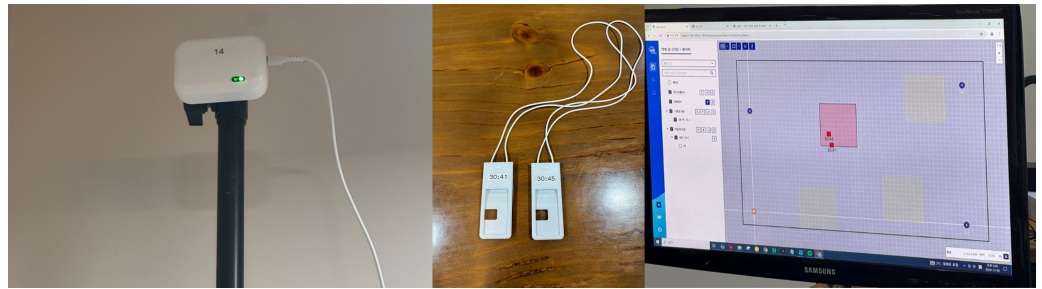


Figure 4. UWB Anchor, UWB Tag, Dedicated Program.

As shown in Figure 5, when a user registers a tag with a QR code, a unique ID is transmitted via mobile device and the real-time location is sent to the Geospace server. User entry and exit events are recorded in an AWS-based log database via the UWB server, and notifications are sent to mobile devices when events occur.

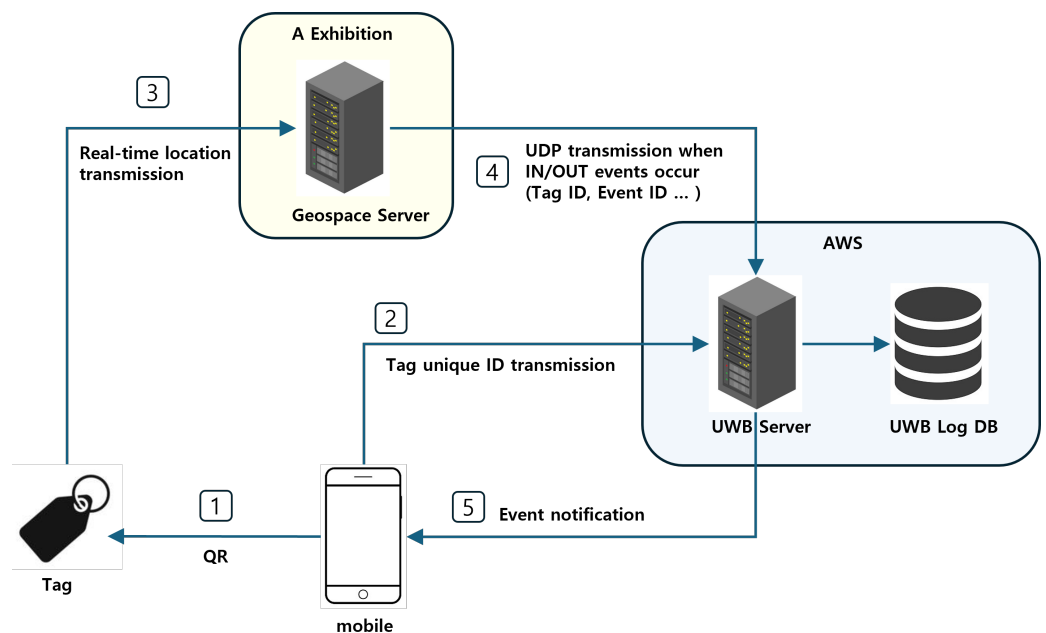


Figure 5. UWB positioning system configuration diagram.

3.2.4. Development of TTS and STT Voice Synthesis Audio Technology

Voice synthesis audio technology is a key component for improving the exhibition viewing experience for vulnerable groups such as the visually impaired. Exhibition commentary content was created using speech recognition (STT: Speech-to-Text) and voice synthesis (TTS: Text-to-Speech) technologies [54].

The STT technology was built based on Google Cloud STT API, and a noise removal preprocessing filter was applied to account for the reverberation and background noise unique to museums and art galleries [55]. This enabled the system to achieve over 90% speech recognition accuracy even in diverse user speech environments within museums [54]. By recognizing user speech in real time and converting it into text queries, the system was designed to enable a natural interface with the LLM-based docent system.

TTS technology is based on Naver Clova and OpenAI TTS API, and a custom dictionary was created to improve the pronunciation accuracy of proper nouns and technical terms that frequently appear in exhibition explanations [56,57]. This dictionary was compiled based on high-frequency nouns and cultural property names appearing in previously collected public commentaries, and reflects standard pronunciation and phonetic transcription information based on the results of TTS engine phonetic pattern analysis, as shown in Figure 6 [58]. To ensure natural listening experience in the response results, an algorithm was applied to automatically adjust word spacing, intonation, and tone in the TTS output, thereby achieving natural speech patterns similar to those of voice actors. This process contributed to reducing fatigue during long listening sessions for visually impaired individuals and improving their understanding of the commentary information [59].

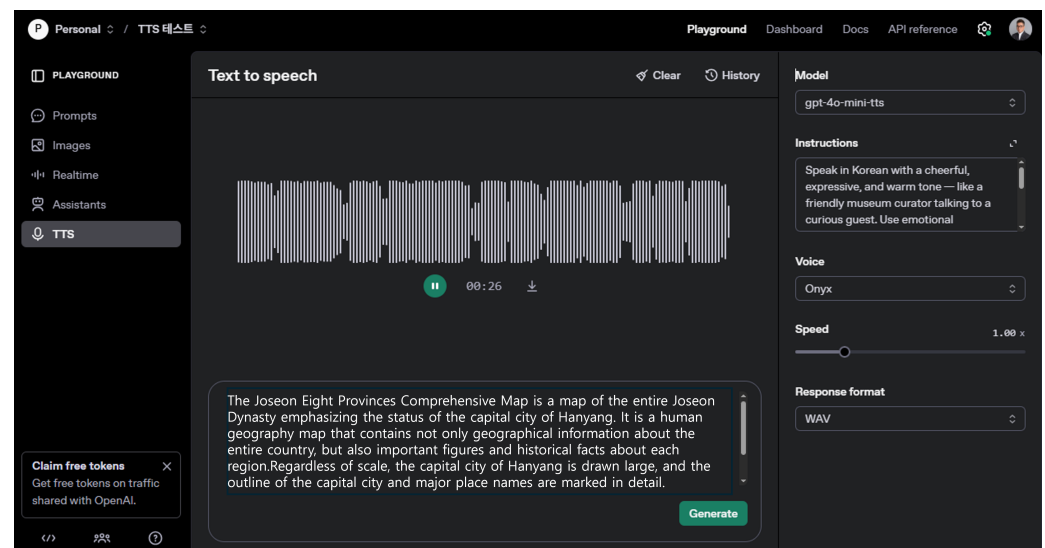


Figure 6. GPT-4o-mini-TTS Screen. (translated from Korean to English).

3.3. Building a Dataset

3.3.1. Refining Exhibition Explanations

The core of an intelligent docent system is to secure structured exhibition explanation data that include domain-specific professional information and to provide users with natural and accurate explanations based on these data. However, existing exhibition data are mainly composed of unstructured text, and the format and terminology also vary depending on the institution that holds the artifacts, making it difficult to apply directly to NLP (Natural Language Processing) models [60]. To address this issue, we collected over 100,000 exhibition descriptions for artifacts from public data such as the National Museum of Korea's exhibition items and museum websites [61]. To perform RAG from a standardized document database, we processed metadata and images into docent-style descriptions and refined all documents in the unprocessed public data, which were centered on specialized terminology and classical Chinese characters, into general-level descriptions through the process shown in Figure 7 [62].

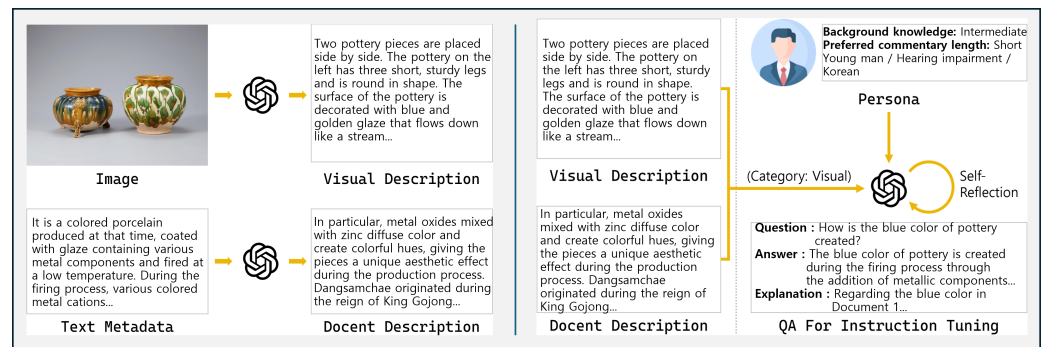


Figure 7. Data refinement and question-and-answer construction process for exhibition explanations (translated from Korean to English).

The refinement of exhibition commentary data was performed through a pipeline consisting of four stages. The first stage was text structuring, in which explanation text data provided by various institutions in HTML and PDF formats was automatically parsed to identify key items such as exhibition titles, artists, production periods, and descriptions, and these were stored in a standardized JSON format [63]. In this process, open source parser tools such as BeautifulSoup and PDFminer were used, and rule-based filters were also applied in parallel to organize documents with non-standard structures into consistent templates, as shown in Figure 8.

(Instruction) First, read the audience persona C and the docent's explanation D. Then, based on the audience persona C and the docent's explanation D, create a question-and-answer QA for the audience and an explanation E for the generated question-and-answer. At this time, the question Q should be able to be answered by the docent's explanation D. In addition, the question Q should be created to fit the audience persona C. For example, when the audience has little background knowledge about a specific work, difficult questions should be avoided, and questions should be asked in easy language for younger audiences. As a reference, when creating the question Q, do not refer to the work with a pronoun, but should have a complete meaning so that the intention can be understood from the question itself. In addition, the answer A should be appropriately consistent with the audience persona C. Finally, the explanation E should properly explain whether the generated QA satisfies the given conditions.

Figure 8. Prompt design that presents the direction and goals of response generation-1 (translated from Korean to English).

The second step involved refining Chinese characters and foreign words. Some explanatory texts contained a mixture of Chinese characters in accordance with ancient artifact notation practices, as well as numerous archaic expressions and Japanese-style foreign words. Accordingly, we adjusted the difficulty level and refined the terminology to aid the understanding of general visitors. For example, “朝鮮時代” was changed to “조선시대” by replacing the Chinese character notation with Hangul syllables. This process was handled by combining a predefined mapping table and GPT-based query filtering [64]. The third step was morphological analysis, which utilized the “Nori” plugin, a Korean text morphological analyzer provided by Amazon OpenSearch Service, to extract key word groups (nouns, verbs, adjectives, etc.) from the explanatory text. Nori is a Korean morphological analyzer optimized for Elasticsearch and OpenSearch environments, with a structure suitable for token-level segmentation for search indexing [65]. In this study, noun-centric keywords were extracted from the exhibition explanatory texts, and stop words and particles were removed to build a tokenization dictionary for document embedding and similarity-based search models [66]. The morphological analysis results were also used for question-answer generation, document retrieval, and LLM input prompt configuration. Finally, document

expression consistency and compression were improved through summary and rewriting tasks [67]. As shown in Figure 9, automatic summarization and style unification prompts were applied to the input explanatory texts, and the prompts were designed to focus on the core information (who, when, what, how, etc.) of each explanatory text during the summarization process.

Input: - Audience Persona C: An infant or toddler who is not able to actively engage with the artwork, but would like to see it explained as simply as possible. - Docent Commentary D: This artifact tapers down to the bottom of the [Meeting of the Gods]. This artifact, sometimes called a krater, is a vessel used in ancient Greece to mix water and wine, and it depicts several gods, centered on Omphalos, who was symbolically believed to be the place where all life on earth originated. Output: - Question Q: What is a krater a jar used for? - Answer A: Krater was used by the Greeks to mix wine and water for drinking. - Explanation E: Question Q is asking what Krater is used for in plain language from docent commentary D. Response A is also giving the correct answer to Question Q in plain language. Therefore, the generated question and answer QA is appropriate.	
--	--

Figure 9. Prompt design that presents the direction and goals of response generation-2 (translated from Korean to English).

The resulting sentences were all rewritten in a uniform declarative sentence structure. This step served as a key factor in determining the quality of question-and-answer generation and model training data preprocessing. During this process, we also classified the exhibits by type (paintings, crafts, sculptures, artifacts, etc.) and stored them as meta-data, ultimately securing approximately 199,000 pieces of structured text data. This data is presented in the form shown in Figure 10 and has been used for multilingual translation, question-answering dataset construction, and other purposes [68].

PS01001001005003200000000 • 문서1 (해설) ◦ 금동반가사유상은 한국 삼국 시대의 불교 예술품으로, 머리에 인동무늬가 있는 높은 보관을 쓰고 있는 것이 특징입니다. 매우 가늘게 표현된 상체는 옷을 걸치지 않았으며, 하체에는 치마를 걸쳐 그 치맛자락이 대좌를 덮고 있는 모습입니다. 머리 뒤쪽에 꽃이 가 있는 것으로 보아, 머리에서 나오는 빛을 표현한 광배를 달았음을 알 수 있습니다. 특히 이 반가사유상의 얼굴을 자세히 보면 코 밑에 수염이 있는데, 이는 현존하는 우리나라의 불상 가운데 유일하게 수염이 표현된 사례로 주목받고 있습니다. • 문서2 (시각) ◦ 이 이미지는 한 손을 턱에 댄 관음보살의 상반신 청동 조각상을 나타냅니다. 조각상은 머리에 장식에 있는 관을 쓰고 있으며, 그 위에는 정교한 무늬가 새겨져 있습니다. 얼굴은 부드러운 미소를 띠고 있어 평온한 느낌을 줍니다. 이 작품은 일반적으로 불교 조각에서 관세음보살, 즉 자비를 상징하는 보살로 여겨지며, 이 자세는 사유하는 모습을 표현한 것으로 해석될 수 있습니다. 조각상의 표면은 시간이 지나면서 자연스럽게 생성된 녹청이 있으며, 동양 고유의 청동 불상의 양식을 보여줍니다. 이러한 녹청(청록색의 녹)은 청동 조각의 진품성을 확인하는 또다른 요소이기도 합니다. 전체적인 디자인과 장식된 요소들은 아마도 특정 시대의 불교 예술 양식을 따르고 있을 것으로 추정됩니다. PS01001001011009499000000 • 문서1 (해설) ◦ 서안은 전통 한국 가옥에서 책상의 용도로 사용되었던 가구입니다. 이는 방바닥에 앉아서 생활하는 좌식 생활에 맞도록 높이가 낮으며, 책을 겨우 펼 수 있는 정도의 작은 크기였습니다. 서안은 사당방에서 손님을 마주 대할 때 주인의 위치를 정해주는 역할도 하였습니다. 또한,	PS01001001009001262000000 • 문서1 (해설) ◦ 청자 상감 연꽃 넝쿨 동자무늬 합은 고려시대에 제작된 청자 중 하나로, 신성함이나 생산, 풍요 등을 상징하는 동자 무늬가 특징적입니다. 이 합은 두껍기면에는 연꽃 넝쿨에 매달려 있고 있는 네 명의 동자가 상감 기법으로 장식되어 있으며, 문양의 구성이 매우 짜임새 있게 표현되어 있습니다. 동자들의 윤곽선은 흑상감으로, 동자들의 옷과 연방에는 흑점을 찍어 생동감을 더하였습니다. 유약의 빛깔이 맑고 밝게 표현되어 청자와 잘 어울려집니다. 두껍기면의 경사진 면과 밑으로 내려오는 입구 둘레 및 몸체의 옆면에는 각각 세로줄이 새겨져 꽃잎을 연상시킵니다. 두껍기장자리와 몸체의 턱 부분에는 유약을 닦아낸 후 모래 섞인 내화토 받침을 여러 곳에 놓아 포개어 구워졌습니다. 몸체의 밑바닥에는 유약을 닦지 않고 세 곳에 커다란 규격을 받쳐서 변조한 특징을 가지고 있습니다. • 문서2 (시각) ◦ 이 도자기는 고려 시대의 전형적인 상감청자 항합으로 추정됩니다. 원형의 합은 두껍기와 몸체로 구성되어 있으며, 상단에는 꽃무늬가 정교하게 새겨져 있습니다. 그릇의 표면에는 청자 유약이 균일하게 발라져 있으며, 전체적으로 은은한 청록색을 띠고 있습니다. 측면에는 수직으로 줄을 넣어 장식 효과를 주었습니다. 이러한 상감기법은 고려청자의 특징 중 하나로, 고유한 아름다움과 섬세한 공예 기술을 엿볼 수 있게 합니다. 이러한 종류의 상감청자는 주로 귀족층에서 의례용 또는 장식용으로 사용되었을 가능성이 있습니다. 고려 시대의 도자기 기술은 세계적으로 높은 평가를 받으며, 이 작품 역시 그러한 기술의 일환으로 보입니다. 이 도자기는 당시 사회에서 공예품이 어떤 역할을 했고, 미적 감각이 어떻게 발달했는지를 보여주는 중요한 유물로 여겨집니다.
---	--

Figure 10. An example of 200,000 questions and answers related to 100,000 artifacts written in Korean.

3.3.2. Building a Persona-Based Docent Dataset

To create input–output pairs suitable for sLLM learning, we collected conversational data that included various types of questions and response styles, as simple explanatory texts had limitations. We defined eight user types, including the hearing impaired, foreigners, children, and experts, and designed personas by setting variables such as language level, explanation length, preferences, voice/visual dependency, and question patterns [69]. After generating initial responses using GPT-4 and Claude3, we added a “self-reflection prompt” to evaluate the naturalness and appropriateness of the question–answer pairs, thereby improving the response content in a second round. Approximately 30,000 of the generated question–answer pairs were translated into English, Chinese, Japanese, and

Vietnamese, as shown in Figure 11. This process involved not only simple translation but also cultural context-appropriate expression adjustments and localization [70].

(함남단천 하여진평 응과가옥 기록)⁴² 질문: Khu vực được ghi chép trong tác phẩm này có ý nghĩa gì về cuộc sống và môi trường lúc bấy giờ? ⁴³ 답변: Tác phẩm này được ghi chép tại khu vực núi non Hamnam Dancheon Haejeonpyeong, nơi thể hiện hình ảnh điển hình của văn hóa nông thôn truyền thống Hàn Quốc, với nông nghiệp là nguồn sinh kế chính. Môi trường tự nhiên của khu vực này cho thấy mối quan hệ mật thiết giữa người dân nông thôn và thiên nhiên, điều này phản ánh rằng vào thời điểm đó, người dân (...) ⁴³	(청자 상감 연꽃 넝쿨 동자무늬 합)⁴⁴ 질문: この陶磁器の童子の模様は、高麗時代の社会的文脈の中でどのような意味を持っていたのでしょうか? ⁴⁴ 답변: 高麗時代の青磁において、童子の模様は神聖さ、生産性、そして豊かさを象徴していました。童子たちは蓮の臺にぶら下がって遊ぶ姿で表現され、象嵌技法によってより生き生きと細部まで描かれています。これらの文様は仏教の影響を受けたもので、当時の文化的・社会的背景を理解する (...) ⁴³
--	--

Figure 11. Examples of Vietnamese and Japanese datasets among multilingual languages.

In this manner, we built a dataset of over 200,000 questions and answers, with each sample containing metadata such as exhibit ID, persona type, difficulty level, language, and creation date.

3.4. Document Embedding and Vectorization

The exhibition commentary and Question-and-Answer (Q&A) data constructed in this study were embedded as high-dimensional vectors and converted into search indexes for use in document retrieval in the Retrieval-Augmented Generation (RAG) structure. The entire embedding and vector database construction was carried out through the following procedures.

3.4.1. Application of Multilingual Document Embedding Model

All text data were converted into 384-dimensional embedding vectors using the Hugging Face-based ‘sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2’ model. This model is a lightweight Transformer capable of calculating semantic similarity for multilingual expressions such as Korean, English, Chinese, Japanese, and Vietnamese, making it suitable for multilingual explanatory texts in the docent domain.

3.4.2. Similar Sentence Normalization and Duplicate Filtering

Exhibition explanations often contain repeated similar sentences or frequent stylistic variations, which can distort the similarity in cosine similarity-based searches. To prevent this, a similarity filtering normalization preprocessing was applied to consider sentences with a similarity of 0.98 or higher as duplicates and integrate them into a single representative sentence. The similarity between two different sentences is calculated using cosine similarity, and the filtering criteria are as follows.

$$\text{sim}(v_i, v_j) = \frac{v_i \cdot v_j}{\|v_i\| \cdot \|v_j\|} \geq \tau \quad (3)$$

- $\tau = 0.98$: Duplicate judgment threshold (merge if similarity is 0.98 or higher).
- When duplicates are found: $D_i \leftarrow \text{Representative}(D_i, D_j)$.

3.4.3. Building a Vector Search Infrastructure Based on OpenSearch

The vector index was built based on Amazon OpenSearch. OpenSearch is compatible with Elasticsearch and provides ANN (Approximate Nearest Neighbor) search functionality optimized for large-scale vector searches and high scalability. The indexed vectors are designed to respond to real-time search requests from the RAG Retriever, enabling precise document extraction for docent queries.

3.4.4. FAISS-Based Parallel Search Optimization

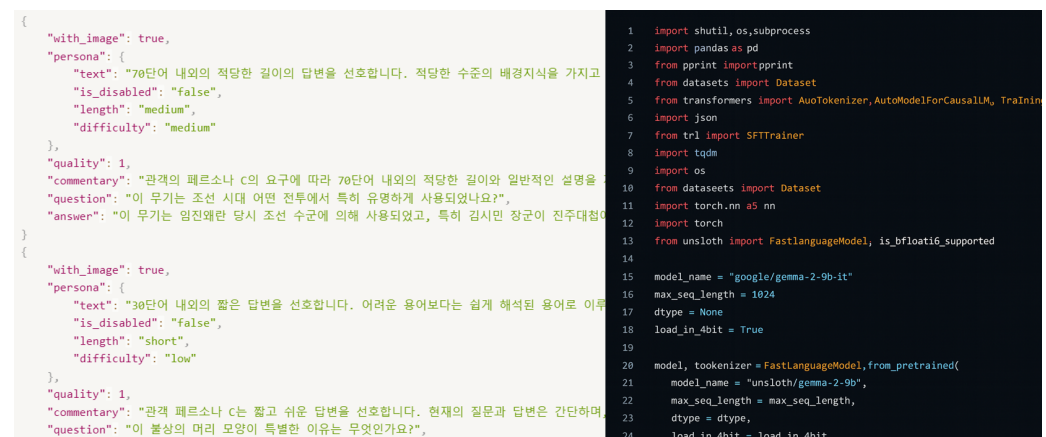
To ensure the response speed of the real-time question-answering system, the FAISS library was adopted for document similarity search. FAISS can perform high-speed similarity searches between high-dimensional embedding vectors using the built-in IVF-PQ (Index Flat Vector with Product Quantization) algorithm and maintains a stable response speed of less than 100 ms in a GPU environment. When new explanatory texts or exhibits are added, they are automatically vectorized and indexed through an ETL pipeline. This configuration minimizes the total latency from the moment a user's query is entered to the generation of a response, serving as the underlying technology that ensures the quality of the real-time docent service.

$$\text{Top-}k(v_Q) = \arg \max_{v_i \in \mathcal{I}} \text{sim}(v_Q, v_i) \quad (4)$$

- v_Q : the query embedding vector.
- $v_i \in \mathcal{I}$: document vectors in the index set $\mathcal{I}$.
- $\text{sim}(v_Q, v_i)$: the similarity score between v_Q and v_i (e.g., cosine similarity).
- $\arg \max$: returns the indices of the top k vectors with the highest similarity scores.

Automatic embedding pipeline configuration: When new explanatory texts are added, we built an ETL pipeline that automates the process in the order of preprocessing, embedding, similarity filtering, and index registration. This allows the system to continuously update documents without manual intervention during maintenance and dynamically reflects changes in actual exhibits or additions to special exhibition content.

The code on the left side of Figure 12 is an example of structuring question-and-answer data according to visitor personas in JSON format. The code on the right loads the Small Language Model (sLLM) using the data and configures the environment for learning/inference. The two codes represent an integrated structure that connects data and models for customized docent response generation. In particular, it is designed so that lightweight LLM can operate effectively based on persona-based Q&A data.



```

{
  "with_image": true,
  "persona": {
    "text": "70년대 내외의 적당한 길이의 답변을 선호합니다. 적당한 수준의 배경지식을 가지고",
    "is_disabled": "false",
    "length": "medium",
    "difficulty": "medium"
  },
  "quality": 1,
  "commentary": "관객의 페르소나 c의 요구에 따라 70년대 내외의 적당한 길이와 일반적인 설명을",
  "question": "이 무기는 조선 시대 어떤 전투에서 특히 유명하게 사용되었나요?",
  "answer": "이 무기는 임진왜란 당시 조선 수군에 의해 사용되었고, 특히 감시된 장군이 진주대첩"
},
{
  "with_image": true,
  "persona": {
    "text": "30년대 내외의 짧은 답변을 선호합니다. 어려운 용어보다는 쉽게 해석된 용어로 이루어",
    "is_disabled": "false",
    "length": "short",
    "difficulty": "low"
  },
  "quality": 1,
  "commentary": "관객 페르소나 c는 짧고 쉬운 답변을 선호합니다. 현재의 질문과 답변은 간단하며",
  "question": "이 불상의 머리 모양이 특별한 이유는 무엇인가요?",
  "answer": ""
}

```

```

1 import shutil, os, subprocess
2 import pandas as pd
3 from pprint import pprint
4 from datasets import Dataset
5 from transformers import AutoTokenizer, AutoModelForCausalLM, TrainingArguments
6 import json
7 from trl import SFTTrainer
8 import tqdm
9 import os
10 from datasets import Dataset
11 import torch.nn as nn
12 import torch
13 from unsloth import FastLanguageModel, is_bfloat16_supported
14
15 model_name = "google/gemma-2-9b-it"
16 max_seq_length = 1024
17 dtype = None
18 load_in_4bit = True
19
20 model, tokenizer = FastLanguageModel.from_pretrained(
21     model_name = "unsloth/gemma-2-9b",
22     max_seq_length = max_seq_length,
23     dtype = dtype,
24     load_in_4bit = load_in_4bit,

```

Figure 12. Code example for docent sLLM model training and verification pipeline.

3.5. LangChain-Based RAG Integration Structure

In the process of embedding exhibition commentary and Q&A document data, we wrapped the SentenceTransformer-based embedding model with LangChain's Embeddings interface, and the vector search engine (VectoreStore) is integrated with the OpenSearch engine and called through the SimilaritySearch API within LangChain. This structure quickly searches approximately 200,000 multilingual document vector data items and delivers the top N similar documents to RAG. In the question-and-answer flow, the RetrievalQA chain is used. When a user query is entered, LangChain automatically calls relevant documents

through the vector searcher, uses these documents as context, and controls the LLM to construct prompts and generate responses. The prompt template applies a Few-shot Persona QA Template structure tailored to the docent domain, and the OutputParser function is used to reformat system output into various formats such as declarative sentences and multilingual summaries. LangChain provides modular connectivity between each component and prompt flow tracking (log traceability) functionality, thereby increasing the transparency of response generation for user queries and facilitating debugging and refinement processes. In addition, the Memory module was configured to continuously track the query–response context within the same user session, thereby laying the foundation for continuous query processing functionality.

3.6. sLLM Model Training Strategy

Various companies are continuously releasing sLLMs that can be used for research and commercial purposes, and we selected models by considering the balance of pros and cons in various aspects, such as response quality and speed, of individual models. We analyzed the latest model development trends, focusing on major small-scale commercial sLLMs and compared the performance of major models such as Upstage’s SOLAR, Mistral’s Nemo, Meta’s Llama3, Google’s Gemma2, and Alibaba’s Qwen2. In Figure 13, gemma-2b showed excellent performance, recording a minimum latency of approximately 0.03 seconds and a maximum value of approximately 35 tokens/second. However, considering the number of parameters, memory capacity, speed requirements, and Korean response quality, we ultimately selected the gemma2-9b model. While the publicly available Gemma2-9B model demonstrated high versatility and English response capabilities, it had inherent limitations such as insufficient Korean response capabilities, lack of understanding of the docent field, and lack of expertise in RAG documents and personality-based responses. To improve this, we performed supervised learning using LoRA-based training to optimize the model for the docent field while minimizing the impact on the model’s existing knowledge [71]. We experimentally demonstrate that LoRA-based fine-tuning is much more effective than prompt tuning in improving the quality of domain-specific automated response systems. Prompt tuning is limited in its performance to general queries and still has limitations in domain specialization and context adaptation [72]. On the other hand, model internalization through LoRA ensures both the accuracy and flexibility of responses in complex docent environments. A total of 250,000 document data points were collected, of which 100,000 were structured text from exhibition descriptions, and the remaining were structured docent scenario data in Q&A format. The docent data were standardized using a question–answer–source–person tag structure, and a total of 30,000 data points were manually reviewed and summarized through a rewriting process to create a high-quality training set. Fine-tuning was performed using the Hugging Face Transformers library.

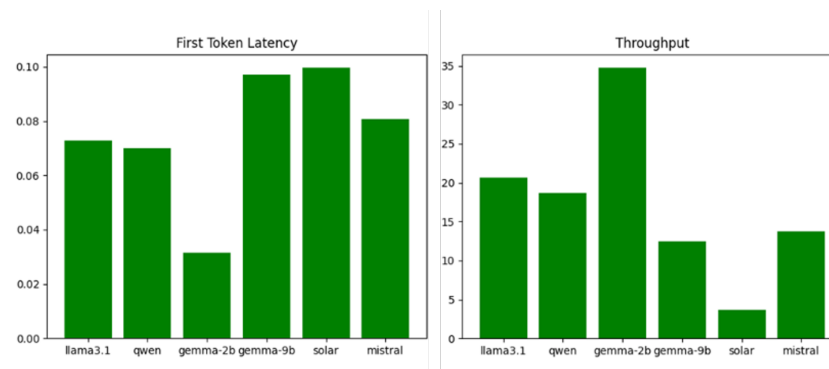


Figure 13. Comparison of response generation speeds for major small-scale commercial sLLMs.

Fine-tuning: LoRA

$$\theta_{FT} = \theta_{base} + \Delta_{LoRA} \quad (5)$$

- θ_{FT} : Final parameters of the model after fine-tuning (Fine-tuned Parameters).
- θ_{base} : Parameters of the pre-trained original LLM (Base Model Parameters).
- Δ_{LoRA} : Low-rank parameter changes learned using the LoRA method.

The training utilized two AWS EC2 A100 GPUs and achieved both lightness and accuracy through an LORA strategy that updated only about 0.5% of the total parameters. After fine-tuning, the accuracy increased to about 42.1%, which is about 4% higher than the baseline performance (38.4%) of the pre-trained Gemma2 model.

4. Experiments and Results

To quantitatively evaluate the performance of the docent LLM system proposed in this study, we designed a scenario-based evaluation system that simulates conditions similar to those in an actual museum environment. The evaluation scenario goes beyond simple question-and-answer tests to create a complex interactive environment that reflects the user's context (exhibition location, question type, preferred response style, etc.), with the aim of verifying the actual applicability of the service.

4.1. Hardware Specifications

The deep learning models for the docent LLM system, particularly the fine-tuning of the Gemma2-9B model and the optimization of the RAG pipeline, were developed and deployed on AWS EC2 A100 (p4dn.12xlarge) instances. This instance type was chosen to meet the demanding computational requirements of large language models and ensure efficient processing during both training and inference. The detailed hardware specifications are presented in Table 1.

Table 1. Hardware specifications for AWS EC2 A100 (p4dn.12xlarge) instance.

Specification	Details
Instance Type	p4dn.12xlarge
GPU	4 × NVIDIA A100 40 GB
GPU Memory	160 GB total
GPU Performance	Up to 0.5 PFLOPS (FP16 with sparsity)
vCPUs	48
RAM	576 GB
Network Bandwidth	Up to 200 Gbps (EFA)
NVMe Storage	4 × 1.8TB (7.2 TB total)

The 4 NVIDIA A100 GPUs with a total of 160 GB of memory were crucial for accelerating the parallel processing required for LLM fine-tuning and high-speed vector search operations within the FAISS framework. The substantial 576 GB of RAM and 7.2 TB of NVMe storage provided ample resources for handling large datasets. The high network bandwidth of up to 200 Gbps was also essential for efficient data transfer and communication within the distributed computing environment, ensuring minimal latency during model deployment and real-time interaction.

4.2. Background of Scenario Design

While general LLM evaluation criteria (MMLU, TruthfulQA, ELO, etc.) are mainly composed of general knowledge, mathematical thinking, and reasoning abilities, the museum docent domain has the following special characteristics. Visitors' questions span a variety of knowledge areas, such as exhibition descriptions, historical background, and artist intent,

and responses require not only simple correct answers but also natural language style and narrative depth. In addition, location information within the indoor space can affect the accuracy of question-and-answer interactions, and user preferences for question difficulty, length, and response format vary depending on user type. Accordingly, this study designed the following performance evaluation scenario, which includes user persona-based simulation and virtual museum environment configuration.

4.3. Scenario Components

Figure 14 shows the results of placing 500 artifact data sets in a virtual museum (including Joseon paintings, Buddhist artifacts, and everyday tools) and visualizing the interactions between the user's location and the exhibits. Each museum is organized based on a unique theme, and users move within the exhibition space (X1–X20) along UWB-based coordinates (Y1–Y20). User queries include their current location and exhibit ID, and the system automatically provides explanatory content based on semantic similarity and spatial proximity.

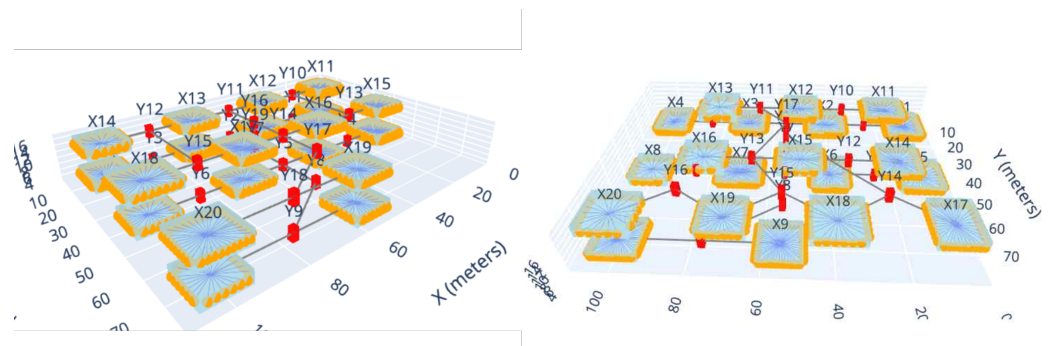


Figure 14. Examples of scenario samples from the performance evaluation dataset.

The simulation features three personas (elementary school students, general adults, and researchers), each designed to generate responses of varying length, tone, and information level. Elementary school students require short, simple sentences, general adults require medium-length responses that include background explanations, and researchers require multi-paragraph responses with a high level of expertise. Response quality was evaluated based on criteria such as the accuracy of information provided, the naturalness of the style, and the appropriateness of the location and persona reflection. By analyzing how customized responses are provided spatially based on the user's movement and queries, the accuracy of the recommendation system was verified. This experiment served as an empirical basis for verifying whether the SC-Retriever-based spatial-text fusion structure functions effectively in an actual service environment.

4.4. Evaluation Data Composition

A total of 2000 query–response samples were constructed based on scenarios, and each sample includes the following information: exhibition item ID and location, user persona, natural language question, correct answer document (gold reference), response from the model being evaluated, evaluator feedback (subjective and objective evaluations), etc. This scenario-based evaluation system provides a basis for comprehensively evaluating not only simple model performance comparisons but also the ability to generate responses and respond to context that can be actually applied in a docent environment.

4.5. Output Devices for Evaluation

Mobile docent app overview: The intelligent mobile docent app, which applies this technology, is designed to run on both iOS and Android platforms to improve visitors'

access to information and exhibition commentary experiences at cultural facilities such as museums and art galleries. It is based on a user-customized UX UI design that makes it easy to use not only for general visitors but also for information-vulnerable groups, including the visually impaired. Figure 15 shows screenshots of the app's features (Korean version). When launching the app, users can select a persona that suits them (general, visually impaired, foreigner, etc.), and the interface is configured according to the selected persona. For example, visually impaired users enter a voice-centric STT/TTS interface, while foreign users enter a language selection mode. Additionally, the app uses UWB (Ultra-Wideband) technology to precisely measure the user's indoor location and automatically provides relevant explanations when the user approaches an exhibit. Users can ask questions about exhibits via text or voice queries, and the app generates customized responses based on the user's background knowledge, language, and preferred explanation length using an RAG (Retrieval-Augmented Generation)-based small Large Language Model (sLLM). The generated explanations can be output in text or voice (TTS) format, and users can choose the format that best suits their environment. The system integrates map-based viewing flow control, tagging events, multilingual settings, voice feedback, and explanation difficulty adjustment functions. Administrators can analyze and update content and user logs through the CMS.

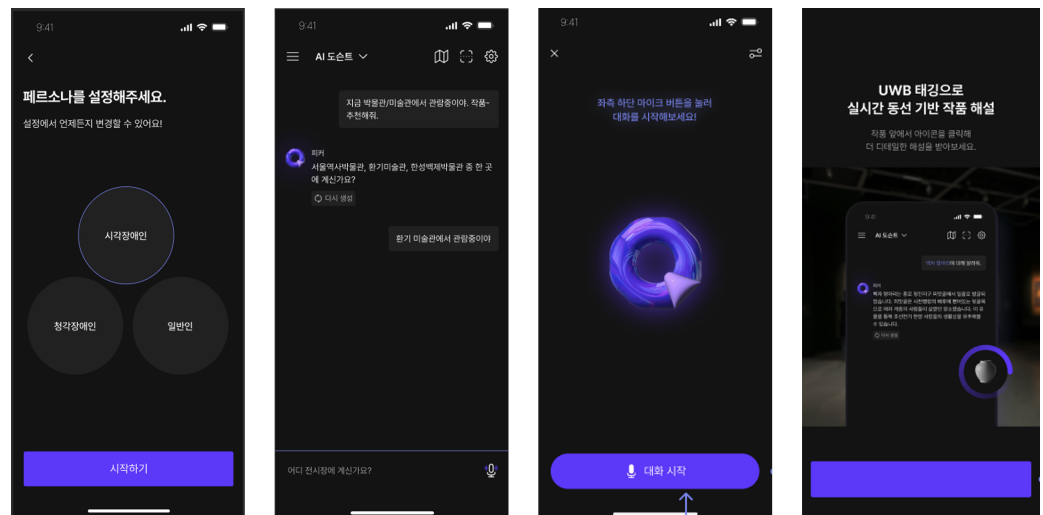


Figure 15. Persona settings, text chat, voice chat, UWB tagging. (Screenshot of the Korean version of the intelligent docent app).

4.6. Small LLM Performance Evaluation

To verify the performance of lightweight language models that can be used in the docent domain, we selected four representative open-source and commercial LLMs (Google's Gemma2, Meta's LLaMA3.1, Alibaba's Qwen2.5, and NVIDIA's Nemo) as comparison targets and named the results of this study DocentGemma. When tested using the same Docent Q&A set and explanation text dataset, DocentGemma achieved an accuracy of 41.25% after instruction tuning, as shown in Figure 16, demonstrating the best performance among individual sLLM models. This represents a 12.65%p difference in accuracy compared to Qwen2.5 (36.25%), Nemo (34.7%), LLaMA3.1 (30.4%), and Gemma2 (28.6%), which can be interpreted as the result of instruction tuning optimized for the docent domain. It produced improved results that reflected the user's location, query difficulty, and cultural context, which are important aspects of inferential thinking and contextual awareness.

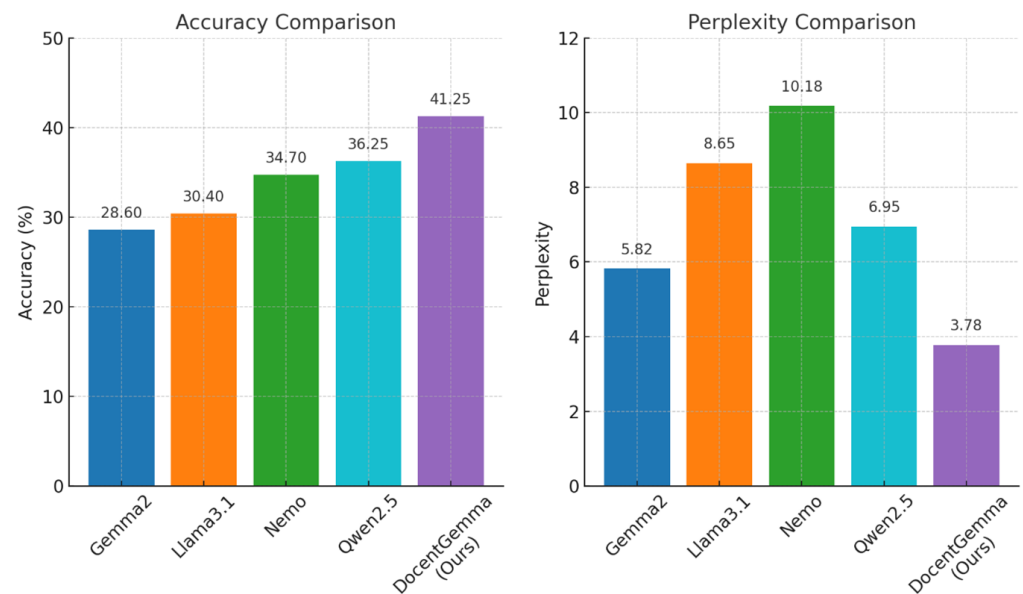


Figure 16. Comparison of answer generation performance among major small-scale commercial sLLMs.

DocentGemma recorded the lowest perplexity score of 3.78. This is significantly lower than Nemo 10.18, LLaMA3.1 8.65, Qwen2.5 6.95, and Gemma2 5.82. Perplexity is an indicator of how well the model's generated response sentences follow natural language context, with lower values indicating higher language generation quality. These performance results are not simply due to the number of model parameters or structural scale, but can be interpreted as the result of a customized instruction dataset, docent persona-based prompt structure, and response format control strategy. This is clearly shown in the results of the hallucination comparison test in Figure 17. In particular, since docent questions are mostly explanatory and leading questions rather than short-answer questions, the importance of the Perplexity metric is further emphasized in that contextual appropriateness and response fluency are directly linked to actual usability.

Question: Why doesn't the Mona Lisa have eyebrows?		
(Llama3.1)	(Qwen2.5)	(Gemma2)
In the case of the Mona Lisa, it was originally an animal with eyebrows, but over the centuries, Europeans studied many individuals academically to study their brains. In the process, some of the Mona Lisa's eyes were damaged, and later (omitted)	The painting called Mona Lisa was created by Italian artist Lucio Ferrazzi. By focusing on this energy and strengthening other elements, Ferrazzi demonstrates his artistic ability and unique style of expression. (omitted)	There are several possibilities for the Mona Lisa's competition: *Polishing process:* After the painting was completed, it may have been lost from the work due to various stability and environmental factors during the Middle Ages, especially those used by Leonardo da Vinci. (omitted)

Figure 17. Examples of hallucination phenomena in the docent domain among major small-scale commercial sLLMs (translated from Korean to English).

4.7. sLLM Model Training Strategy

Due to the nature of exhibition halls, exhibits are located at specific points in space, and users can freely move around the space and interact with the system. Therefore, we developed a Spatio-Contextual Retriever (SC-Retriever) that combines the similarity between the user's questions and documents with probabilities inferred from the user's current location and the spatial distribution of the exhibition to retrieve the correct document more accurately. The precision performance comparison evaluation of the Retriever module, which handles document search functionality, was conducted using a question-answering test set constructed based on a corpus of 500 docent commentary documents. Three comparison groups, GTE, GTE-Room, and SpatialNN, were applied, and they showed significant differences in search accuracy depending on how they reflected the user's exhibition location information and query context. In particular, SC-Retriever adopted a vector-combined search strategy that considers both meaning-based document similarity and space-based location similarity, going beyond simple text similarity calculations. This method is defined by the following weighted sum formula.

$$Sim_{total}(Q, D_i) = \alpha \cdot Sim_{text}(Q, D_i) + (1 - \alpha) \cdot Sim_{spatial}(U, P_i) \quad (6)$$

- Q : User query.
- D_i : i -th document (exhibit description, etc.).
- U : User's current location coordinates.
- P_i : i -th exhibit location coordinates.
- $\alpha \in [0, 1]$: Weighting coefficient between text and spatial similarity.
- $Sim_{text}(Q, D_i)$: Text-based semantic similarity (e.g., cosine similarity).
- $Sim_{spatial}(U, P_i)$: Spatial similarity (e.g., distance-based similarity).

Equation (6) defines the combined similarity score used by the SC-Retriever module to rank the relevant exhibit content based on both textual relevance and spatial proximity. The combination is a convex linear interpolation of two normalized similarity scores: Sim_{text} and $Sim_{spatial}$. This approach is commonly used in multimodal information retrieval tasks where multiple similarity dimensions are considered jointly [73].

As shown in Figure 18, SC-Retriever achieved a Recall@1 of 53.45%, demonstrating a performance 17.5–20% higher than GTE (32.7%), GTE-Room (35.95%), and SpatialNN (33.4%). It also recorded the highest MRR of 68.12, which indicates its ability to provide accurate responses within the top ranks for user queries. While existing methods were limited in accurately selecting documents for complex queries because they focused on only one piece of information, either semantic similarity (GTE) or location-based distance (SNN), SC-Retriever was able to respond more precisely to complex user queries by integrating this information.

SC-Retriever experimentally proved that its spatial context fusion search strategy, which goes beyond simple text-based search structures to reflect user location and exhibit context, is effective in the actual usage environment of docent systems, and secured structural expandability that can be applied to various location-based query response systems in the future.

Table 2 shows the change in Top-1 accuracy according to changes in the weighting coefficient α in the weighted sum-based similarity calculation formula. α is a parameter that adjusts the relative weight of text similarity and spatial similarity. According to the experimental results, when $\alpha = 0.0$, the accuracy was 32.54%, which is the lowest, reflecting only spatial similarity. On the other hand, even when $\alpha = 1.0$, which reflects only text similarity, the accuracy was 39.87%, which did not reach the optimal value. The highest accuracy was 41.25% when $\alpha = 0.7$, suggesting that a combination that places a high weight on text similarity while also reflecting spatial similarity to a certain extent is most effective. Accuracy showed a steady increase until α reached 0.7, followed by a slight decrease. This result indicates that the harmonious integration of the two similarity factors provides superior search performance compared to single-modal-based search [74].

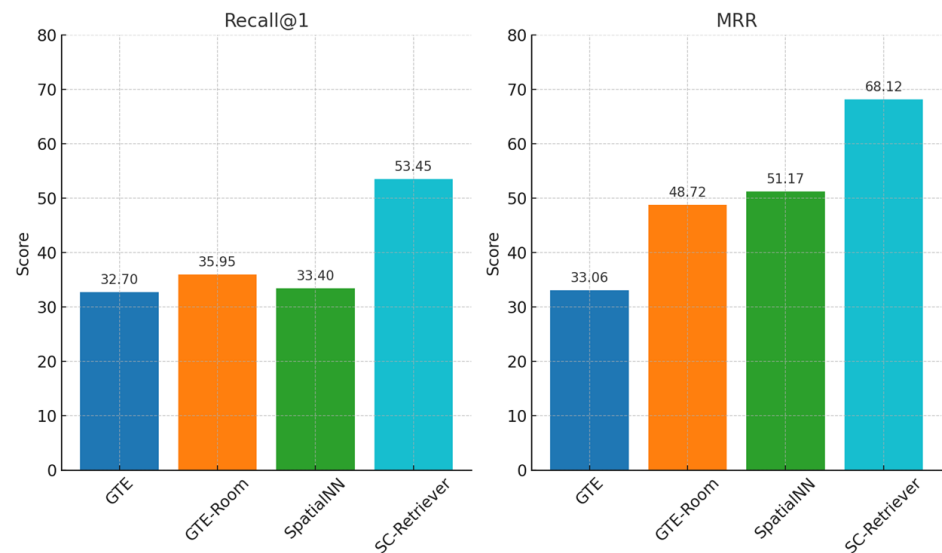


Figure 18. Comparison of the correct document selection performance of major Retriever methodologies.

Table 2. Top-1 Accuracy as a function of Alpha (α).

Alpha (α)	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Top-1 Accuracy (%)	32.54	33.72	35.30	36.23	37.74	38.76	40.83	41.25	41.08	40.54	39.87

4.8. Final RAG-Based LLM Model Performance Evaluation

As shown in Figure 19, the final performance evaluation results show that the DocentGemma-based RAG system proposed in this study demonstrated practical performance that surpassed existing major commercial and open-source language models in the docent domain-specific environment. Quantitatively, DocentGemma achieved the highest accuracy rate of 85.55% in correct answer selection, surpassing OpenAI's ChatGPT-3.5 Turbo (79.45%), Meta's LLaMA3.3-70B (70.45%), and Google's Gemma2-9B (63.7%), showing performance differences ranging from a minimum of 6.1 percentage points to a maximum of 31.85 percentage points. These results were achieved through the optimization of the sLLM based on Instruction-Tuning and the design of user persona-centric prompts, which played a key role. Additionally, the template structure combining the retrieved documents and user queries was layered to enable the generation of more consistent and highly coherent responses.

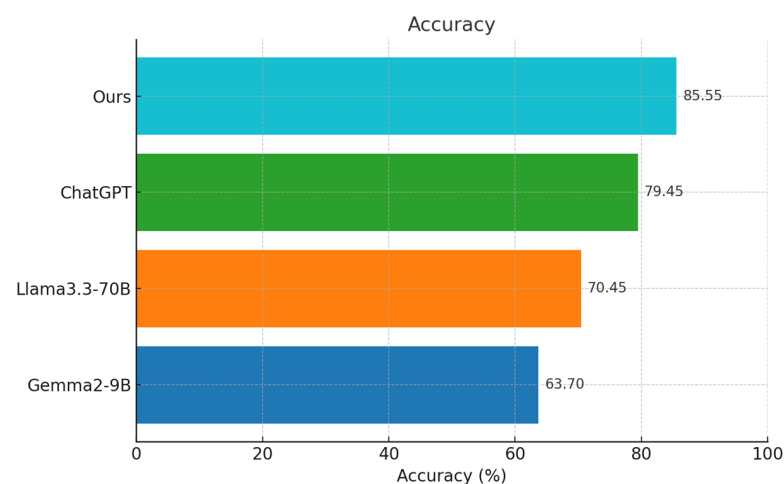


Figure 19. Development and evaluation of RAG-based docent sLLM.

4.9. Empirical Application Cases and User Response Analysis

The RAG-based docent system developed in this study was applied to three museums and art galleries in Seoul to review its feasibility and suitability for the field with the aim of improving information accessibility. The demonstration environment was deployed with multilingual support, including English, Chinese, Japanese, and Vietnamese, in addition to Korean, and demonstration evaluations were conducted with both general visitors and information-vulnerable groups.

4.10. Demonstration Locations and Deployment Environment

The system was installed at the Hwan-ki Art Museum, the Seoul History Museum, and the Hanseong Baekje Museum. As each institution has different spatial structures and exhibition characteristics, user location-based content provision functions (UWB-based positioning), multilingual explanations, and TTS output were set according to the environment. At the Hwan-ki Art Museum, the system was applied to modern art exhibitions, and many scene-interpretation-type questions arose to understand the artistic context. The Seoul History Museum provides content centered on historical timelines and artifacts, and time- and location-based search linkage functions (e.g., Joseon Dynasty late period lifestyle, requests for maps from that period, etc.) were demonstrated. The Hanseong Baekje Museum is a space centered on daily necessities and folk art exhibitions, and the effectiveness of UWB-based interaction, which provides automatic commentary when the user's location is close to the exhibition items, was confirmed.

4.11. User Composition and Participation Methods

A total of 110 visitors participated in the demonstration evaluation, including 14 visually impaired people, 14 hearing impaired people, 16 foreign visitors, and 21 elderly people (aged 60 or older). Users interacted with the system by viewing explanations for each exhibit through a mobile app or dedicated device and entering or speaking natural language questions. Questions and answers were provided in multiple languages in addition to Korean, and multimodal responses such as TTS and subtitles were also provided.

4.12. User Responses and Qualitative Evaluation Results

As shown in Table 3, according to the survey and interview results, the overall satisfaction rating was 4.24 out of 5 points, with particularly high scores in the areas of response clarity (4.12), expectation level (4.79), and appropriateness of technology and functionality (4.68). Users with limited access to information (such as those with visual or hearing impairments, or the elderly) showed particularly positive reactions to various output formats and requested additional support for subtitles. Meanwhile, foreign users generally expressed satisfaction with the professionalism of the commentary content and the quality of the translation, but suggested improvements in the lack of cultural context explanations and awkward translation expressions. In particular, Japanese and Chinese users pointed out the lack of cultural interpretation of specific traditional terms, and it was analyzed that the precision of multilingual localization should be improved in the future.

Table 3. Evaluation of an app based on various criteria.

Evaluation Items	Detail Items	Likert Scale	100-Point Scale	Average Scale
Digital Docent	Communication with the digital docent is easy.	4.21	84.1	4.12/82.4
	The digital docent's explanations match the facts.	4.47	89.4	
	The digital docent's content is rich.	4.52	90.4	
	The digital docent's explanations are accurately expressed in the local language.	4.25	84.9	
	The digital docent's feedback speed is fast.	3.61	72.1	
	There are typos or incorrect expressions in the digital docent's answers.	3.91	78.2	
	The digital docent's answers are satisfactory.	4.11	82.2	
	The answers to additional questions are satisfactory.	3.91	78.1	
Expectations	The actual user experience exceeded expectations.	4.83	96.5	4.79/95.8
	The functionality and service level exceeded expectations.	4.71	94.2	
	I expect the service to continue to improve in the future.	4.95	98.9	
	The app exceeded my expectations.	4.68	93.5	
Technology and Configuration	The app is easy to access.	4.89	97.8	4.68/93.5
	The app menu is well organized.	4.63	92.5	
	The app can be used stably.	4.93	98.5	
	Page loading does not take long.	4.26	85.2	
Design	The overall atmosphere and screen layout of the app are harmonious.	4.67	93.4	4.47/89.4
	The text and images (icons) are easy to read.	4.31	86.2	
	The app is easy to use at a glance.	4.63	92.5	
	The service menu layout and structure are consistent.	4.27	85.3	
Information	The app provides the latest information.	3.68	73.6	4.0/80.1
	The app provides a variety of information.	3.82	76.3	
	The information provided by the app is useful.	4.81	96.2	
	The information provided by the app is accurate.	3.71	74.2	
Voice Service	Voice recognition is accurate.	3.71	74.2	4.10/82.1
	Voice recognition speed is fast.	3.82	76.4	
	The content is accurately output in voice.	3.86	77.1	
	The speed of voice narration is appropriate.	4.40	87.9	
	The content of voice narration matches the facts.	3.96	79.2	
	There are no interruptions in the content of voice narration.	4.88	97.6	
Mobility	The service is easy to use while moving.	3.70	73.9	3.88/77.7
	Voice responses are not interrupted while moving.	3.93	78.5	
	Text is clearly displayed while moving.	4.28	85.6	
	The content of responses matches the artifacts.	2.98	59.6	
	It is easy to access the app while moving.	3.97	79.4	
	Page loading does not take long while moving.	4.23	84.5	
	The current location can be identified through the service while moving.	3.82	76.4	
	The app detects my location and responds appropriately.	4.39	87.8	
Expected Effects	The app has increased satisfaction with viewing the exhibition.	4.26	85.2	3.94/78.8
	It was easy to obtain new information through the app.	3.62	72.4	
Intention to Continue Using	I will continue to use the app in the future.	4.92	98.4	4.61/92.2
	I will recommend the app to my friends in the future.	4.86	97.2	
	I will speak positively about the app to my friends in the future.	4.82	96.3	
Overall Satisfaction	I am generally satisfied with the app.	4.24	84.8	4.24/84.8

4.13. Summary and Interpretation of Empirical Results

This system was applied to multiple users in an actual museum environment and received positive feedback in terms of information delivery, accessibility, and responsiveness. In particular, location recognition through UWB and user persona-based output methods were proven to be designed in line with actual user satisfaction. Furthermore, multilingual output, which was not implemented in existing systems, demonstrated new possibilities for exhibition guidance systems for information-vulnerable groups. However, UI friendliness and other aspects were identified as areas for future improvement. These empirical application results demonstrate that it is possible to develop a docent system that can be used immediately in the field, beyond the level of a laboratory prototype, and will be utilized as practical evidence for follow-up research and service expansion.

5. Conclusions & Future Research

This section summarizes the achievements of the proposed intelligent docent system and suggests future research directions. The system realized personalized exhibition commentary through a domain-specific RAG structure and an integrated framework based on sLLM and LangChain. Experiments proved high search accuracy and user satisfaction, and future goals include multilingual and sign language response enhancement, user-customized recommendations, and real-time processing optimization.

5.1. Conclusions

This study aimed to improve the exhibition viewing experience for users with low information accessibility in museum and art gallery exhibition environments, particularly information-vulnerable groups including visually and hearing impaired visitors and foreign visitors, by designing and implementing an intelligent docent system centered on a small Large Language Model (sLLM) based on RAG (Retrieval-Augmented Generation). This system integrates location-based exhibit identification, large-scale vectorization of explanatory texts, user-customized question-and-answer functionality, and multilingual and multimodal output capabilities to overcome the limitations of existing audio guides and QR-based explanation methods.

During the research process, we structured over 199,000 exhibition commentary data points and built a 102,000-pair question-and-answer dataset based on eight personas, including visually impaired, hearing impaired, and foreign visitors. The data were vectorized using a multilingual embedding model based on sentence-transformers and indexed in a distributed vector search infrastructure based on OpenSearch. Additionally, to improve the accuracy of document retrieval, we designed SC-Retriever, which combines location-based nearest neighbor search and semantic similarity-based search, and utilized it as a core module in the RAG architecture.

In quantitative comparison experiments, the DocentGemma model trained in this study recorded results such as Accuracy 85.55%, Perplexity 3.78, MRR 68.12%, and Recall@1 53.45%, demonstrating overall superior performance compared to existing public LLM and commercial models. In terms of qualitative evaluation, through empirical experiments conducted at the Hwan-ki Art Museum, the Seoul History Museum, and the Hanseong Baekje Museum, we obtained a high satisfaction rating of 4.24 on average from 110 users, including those from information-disadvantaged groups, and confirmed positive responses to core functions such as multilingual explanations, voice responses, and question-based explanations.

Overall, this study proposed a technical approach to solve the problem of information imbalance in exhibition viewing environments and demonstrated the applicability of RAG-based sLLM through the implementation and validation of a system at a level that can be applied in practice. In particular, by successfully building and verifying a user-customized natural language interface in the unique context of the docent domain, it showed the possibility of expansion into various real-time explanation services in fields such as cultural heritage, tourism, and education.

5.2. Limitations

First, although this study implemented the system based on a structure and data specialized for the docent domain, the learned docent dataset is biased toward certain exhibition themes, limiting its generalization performance. In particular, when specific traditional cultural terms or background knowledge about exhibition items are lacking, the information fidelity of the generated responses may be somewhat reduced. Second, in the case of multilingual support, sentence-level translation was implemented, but there

were language-specific differences in the quality of localization of cultural implications and traditional concepts. Japanese and Chinese users requested improvements in terms of unclear terminology and unnatural context.

5.3. Future Research

Future research will develop a domain-adaptive multilingual prompt structure that includes not only language translation but also cultural background knowledge and context interpretation, with the aim of improving the accuracy of question-answering for non-English-speaking users. In addition, refined voice mapping technology that enables real-time synchronization of TTS output through multimodal interface refinement and a situation-responsive voice commentary supplement system for the visually impaired will be added. Furthermore, a design for an online learning-based system improvement loop that collects field user queries and evaluation data and continuously reflects them in the learning data is expected to follow.

Author Contributions: Conceptualization, T.J. and I.J.; methodology, T.J.; software, T.J.; validation, I.J.; formal analysis, T.J.; investigation, T.J.; resources, I.J.; data curation, T.J.; writing—original draft preparation, T.J.; writing—review and editing, I.J.; visualization, T.J.; supervision, I.J.; project administration, I.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data that support the findings of this study are not available due to commercial restrictions.

Acknowledgments: The authors would like to thank the editor and the anonymous reviewers for their valuable comments and suggestions, which have significantly contributed to improving the quality of this manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Asakawa, S.; Guerreiro, J.; Ahmetovic, D.; Kitani, K.; Asakawa, C. The Present and Future of Museum Accessibility for People with Visual Impairments. In Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '18), Galway, Ireland, 22–24 October 2018; pp. 382–384. [\[CrossRef\]](#)
2. Jingyu, S.; Xinyi, L.; Zeyou, Z. Research on Museum Accessibility for the Visually Impaired. *SHS Web Conf.* **2023**, *163*, 02022. [\[CrossRef\]](#)
3. Mazzanti, P.; Ferracani, A.; Bertini, M.; Principi, F. Reshaping Museum Experiences with AI: The ReInHerit Toolkit. *Heritage* **2025**, *8*, 277. [\[CrossRef\]](#)
4. Qu, G.; Chen, Q.; Wei, W.; Lin, Z.; Chen, X.; Huang, K. Mobile Edge Intelligence for Large Language Models: A Contemporary Survey. *IEEE Commun. Surv. Tutor.* **2025**, Early Access. [\[CrossRef\]](#)
5. Trichopoulos, G.; Konstantakis, M.; Alexandridis, G.; Caridakis, G. Large Language Models as Recommendation Systems in Museums. *Electronics* **2023**, *12*, 3829. [\[CrossRef\]](#)
6. An, H.; Park, W.; Liu, P.; Park, S. Mobile-AI-Based Docent System: Navigation and Localization for Visually Impaired Gallery Visitors. *Appl. Sci.* **2025**, *15*, 5161. [\[CrossRef\]](#)
7. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2019**, arXiv:1810.04805. [\[CrossRef\]](#)
8. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *arXiv* **2023**, arXiv:1910.10683.
9. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models are Few-Shot Learners. *arXiv* **2020**, arXiv:2005.14165. [\[CrossRef\]](#)
10. OpenAI; Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; Almeida, D.; Altenschmidt, J.; Altman, S.; et al. GPT-4 Technical Report. *arXiv* **2024**, arXiv:2303.08774.
11. Driess, D.; Xia, F.; Sajjadi, M.S.M.; Lynch, C.; Chowdhery, A.; Ichter, B.; Wahid, A.; Tompson, J.; Vuong, Q.; Yu, T.; et al. PaLM-E: An Embodied Multimodal Language Model. *arXiv* **2023**, arXiv:2303.03378.

12. Schick, T.; Dwivedi-Yu, J.; Dessì, R.; Raileanu, R.; Lomeli, M.; Zettlemoyer, L.; Cancedda, N.; Scialom, T. Toolformer: Language Models Can Teach Themselves to Use Tools. *arXiv* **2023**, arXiv:2302.04761. [\[CrossRef\]](#)
13. Yao, S.; Yu, D.; Zhao, J.; Shafran, I.; Griffiths, T.L.; Cao, Y.; Narasimhan, K. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. *arXiv* **2023**, arXiv:2305.10601. [\[CrossRef\]](#)
14. Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. Training language models to follow instructions with human feedback. *arXiv* **2022**, arXiv:2203.02155. [\[CrossRef\]](#)
15. Christiano, P.; Leike, J.; Brown, T.B.; Martic, M.; Legg, S.; Amodei, D. Deep reinforcement learning from human preferences. *arXiv* **2023**, arXiv:1706.03741. [\[CrossRef\]](#)
16. Bender, E.M.; Gebru, T.; McMillan-Major, A.; Shmitchell, S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21), Virtual Event, 3–10 March 2021; pp. 610–623. [\[CrossRef\]](#)
17. Guerrero, P.; Pan, Y.; Kashyap, S. Efficient Deployment of Vision-Language Models on Mobile Devices: A Case Study on OnePlus 13R. *arXiv* **2025**, arXiv:2507.08505. [\[CrossRef\]](#)
18. Jiao, X.; Yin, Y.; Shang, L.; Jiang, X.; Chen, X.; Li, L.; Wang, F.; Liu, Q. TinyBERT: Distilling BERT for Natural Language Understanding. *arXiv* **2020**, arXiv:1909.10351. [\[CrossRef\]](#)
19. Gemma Team; Mesnard, T.; Hardin, C.; Dadashi, R.; Bhupatiraju, S.; Pathak, S.; Sifre, L.; Rivière, M.; Kale, M.S.; Love, J.; et al. Gemma: Open Models Based on Gemini Research and Technology. *arXiv* **2024**, arXiv:2403.08295. [\[CrossRef\]](#)
20. Kim, J.-H.; Choi, Y.-S. Lightweight Pre-Trained Korean Language Model Based on Knowledge Distillation and Low-Rank Factorization. *Entropy* **2025**, *27*, 379. [\[CrossRef\]](#)
21. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-T.; Rocktäschel, T.; et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *arXiv* **2021**, arXiv:2005.11401.
22. Kim, C.; Joe, I. A Balanced Approach of Rapid Genetic Exploration and Surrogate Exploitation for Hyperparameter Optimization. *IEEE Access* **2024**, *12*, 192184–192194. [\[CrossRef\]](#)
23. Wahsheh, F.R.; Moaiad, Y.A.; El-Ebiary, Y.A.B.; Hamzah, W.M.A.F.W.; Yusoff, M.H.; Pandey, B. E-Commerce Product Retrieval Using Knowledge from GPT-4. In Proceedings of the 2023 International Conference on Computer Science and Emerging Technologies (CSET), Bangalore, India, 10–12 October 2023; pp. 1–8. [\[CrossRef\]](#)
24. Robertson, S.; Zaragoza, H. The Probabilistic Relevance Framework: BM25 and Beyond. *Found. Trends Inf. Retr.* **2009**, *3*, 333–389. [\[CrossRef\]](#)
25. Karpukhin, V.; Oğuz, B.; Min, S.; Lewis, P.; Wu, L.; Edunov, S.; Chen, D.; Yih, W.-T. Dense Passage Retrieval for Open-Domain Question Answering. *arXiv* **2020**, arXiv:2004.04906. [\[CrossRef\]](#)
26. Izacard, G.; Lewis, P.; Lomeli, M.; Hosseini, L.; Petroni, F.; Schick, T.; Dwivedi-Yu, J.; Joulin, A.; Riedel, S.; Grave, E. Atlas: Few-shot Learning with Retrieval Augmented Language Models. *arXiv* **2022**, arXiv:2208.03299. [\[CrossRef\]](#)
27. Lin, W.; Chen, J.; Mei, J.; Coca, A.; Byrne, B. Fine-grained Late-interaction Multi-modal Retrieval for Retrieval Augmented Visual Question Answering. *arXiv* **2023**, arXiv:2309.17133.
28. Johnson, J.; Douze, M.; Jégou, H. Billion-Scale Similarity Search with GPUs. *IEEE Trans. Big Data* **2021**, *7*, 535–547. [\[CrossRef\]](#)
29. Douze, M.; Guzhva, A.; Deng, C.; Johnson, J.; Szilvassy, G.; Mazaré, P.-E.; Lomeli, M.; Hosseini, L.; Jégou, H. The Faiss library. *arXiv* **2024**, arXiv:2401.08281. [\[CrossRef\]](#)
30. Khan, A.; S, N.C.; Gangodkar, D.R. An Overview Recent Trends and Challenges in Multi-Modal Image Retrieval Using Deep Learning. In Proceedings of the 2024 International Conference on Communication, Computing and Energy Efficient Technologies (I3CEET), Gautam Buddha Nagar, India, 20–21 September 2024; pp. 929–934. [\[CrossRef\]](#)
31. Mu, R. A Survey of Recommender Systems Based on Deep Learning. *IEEE Access* **2018**, *6*, 69009–69022. [\[CrossRef\]](#)
32. He, X.; Liao, L.; Zhang, H.; Nie, L.; Hu, X.; Chua, T.-S. Neural Collaborative Filtering. *arXiv* **2017**, arXiv:1708.05031. [\[CrossRef\]](#)
33. Pang, G.; Shen, C.; Cao, L.; Van Den Hengel, A. Deep Learning for Anomaly Detection: A Review. *ACM Comput. Surv.* **2021**, *54*, 1–38. [\[CrossRef\]](#)
34. Akkiraju, R.; Xu, A.; Bora, D.; Yu, T.; An, L.; Seth, V.; Shukla, A.; Gundecha, P.; Mehta, H.; Jha, A.; et al. FACTS About Building Retrieval Augmented Generation-based Chatbots. *arXiv* **2024**, arXiv:2407.07858. [\[CrossRef\]](#)
35. Topsakal, O.; Akinci, T.C. Creating Large Language Model Applications Utilizing LangChain: A Primer on Developing LLM Apps Fast. *Int. Conf. Appl. Eng. Nat. Sci.* **2023**, *1*, 1050–1056. [\[CrossRef\]](#)
36. Mavroudis, V. LangChain. *Preprints* **2024**, online. [\[CrossRef\]](#)
37. LangChain AI. Build a Customer Support Bot. LangGraph Tutorials. 2025. Available online: <https://langchain-ai.github.io/langgraph/tutorials/customer-support/customer-support/> (accessed on 18 August 2025).
38. Khachane, A.; Kshirsagar, A.; Takawale, P.; Charate, M. AUTOMATING DATA ANALYSIS WITH LANGCHAIN. *Int. Res. J. Mod. Eng. Technol. Sci.* **2024**, *6*, 11066–11070. [\[CrossRef\]](#)
39. Felicetti, A.; Niccolucci, F. Artificial Intelligence and Ontologies for the Management of Heritage Digital Twins Data. *Data* **2025**, *10*, 1. [\[CrossRef\]](#)

40. Al-Okby, M.F.R.; Junginger, S.; Roddelkopf, T.; Thurow, K. UWB-Based Real-Time Indoor Positioning Systems: A Comprehensive Review. *Appl. Sci.* **2024**, *14*, 11005. [CrossRef]
41. Qian, C.; Xie, Z.; Wang, Y.; Liu, W.; Zhu, K.; Xia, H.; Dang, Y.; Du, Z.; Chen, W.; Yang, C.; et al. Scaling Large Language Model-based Multi-Agent Collaboration. *arXiv* **2025**, arXiv:2406.07155.
42. Patriwala, A. LangChain: A Comprehensive Framework for Building LLM Applications (with Code). *Medium*. 2025. Available online: <https://medium.com/@patriwala/langchain-a-comprehensive-framework-for-building-llm-applications-e2800dba2753> (accessed on 18 August 2025).
43. Reimers, N.; Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *arXiv* **2019**, arXiv:1908.10084. [CrossRef]
44. LangChain. How-To Guides. LangChain Documentation. Available online: https://python.langchain.com/docs/how_to/ (accessed on 18 August 2025).
45. Google. Gemma: Introducing Google's New Family of Lightweight, State-of-the-art Open Models. 2024. Available online: <https://ai.google.dev/gemma> (accessed on 18 August 2025).
46. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv* **2021**, arXiv:2106.09685.
47. Yang, J.; Zhu, C. Research on UWB Indoor Positioning System Based on TOF Combined Residual Weighting. *Sensors* **2023**, *23*, 1455. [CrossRef]
48. Ma, W.; Fang, X.; Liang, L.; Du, J. Research on indoor positioning system algorithm based on UWB technology. *Meas. Sens.* **2024**, *33*, 101121. [CrossRef]
49. Zhao, W.; Goudar, A.; Tang, M.; Schoellig, A.P. Ultra-wideband Time Difference of Arrival Indoor Localization: From Sensor Placement to System Evaluation. *arXiv* **2024**, arXiv:2412.12427. [CrossRef]
50. Verde, D.; Romero, L.; Faria, P.M.; Paiva, S. Indoor Content Delivery Solution for a Museum Based on BLE Beacons. *Sensors* **2023**, *23*, 7403. [CrossRef]
51. Spachos, P.; Plataniotis, K.N. BLE Beacons for Indoor Positioning at an Interactive IoT-Based Smart Museum. *IEEE Syst. J.* **2020**, *14*, 3483–3493. [CrossRef]
52. Gong, M.; Li, Z.; Li, W. Research on ultra-wideband (UWB) indoor accurate positioning technology under signal interference. In Proceedings of the 2022 IEEE International Conferences on Internet of Things (iThings) and IEEE Green Computing & Communications (GreenCom) and IEEE Cyber, Physical & Social Computing (CPSCom) and IEEE Smart Data (SmartData) and IEEE Congress on Cybermatics (Cybermatics), Espoo, Finland, 22–25 August 2022; pp. 147–154. [CrossRef]
53. Krebs, S.; Herter, T. Ultra-Wideband Positioning System Based on ESP32 and DWM3000 Modules. *arXiv* **2024**, arXiv:2403.10194. [CrossRef]
54. Kingsbury, B.; Saon, G.; Mangu, L.; Padmanabhan, M.; Sarikaya, R. Robust speech recognition in noisy environments: The 2001 IBM SPINEevaluation system. In Proceedings of the 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Orlando, FL, USA, 13–17 May 2002; Volume 1, p. I-53–I-56. [CrossRef]
55. Google Cloud. Speech-to-Text. Available online: <https://cloud.google.com/speech-to-text/> (accessed on 18 August 2025).
56. Naver Cloud Platform. Clova Speech Synthesis. Available online: <https://guide.ncloud-docs.com/docs/clovaspeech-overview> (accessed on 18 August 2025).
57. OpenAI. Text to Speech. Available online: <https://platform.openai.com/docs/guides/text-to-speech> (accessed on 18 August 2025).
58. Raitio, T.; Latorre, J.; Davis, A.; Morrill, T.; Golipour, L. Improving the quality of neural TTS using long-form content and multi-speaker multi-style modeling. *arXiv* **2023**, arXiv:2212.10075.
59. Wood, S.G.; Moxley, J.H.; Tighe, E.L.; Wagner, R.K. Does Use of Text-to-Speech and Related Read-Aloud Tools Improve Reading Comprehension for Students with Reading Disabilities? A Meta-Analysis. *J. Learn. Disabil.* **2018**, *51*, 73–84. [CrossRef]
60. Leaman, R.; Khare, R.; Lu, Z. Challenges in clinical natural language processing for automated disorder normalization. *J. Biomed. Inform.* **2015**, *57*, 28–37. [CrossRef]
61. National Museum of Korea. Digital Collections. Available online: <https://www.museum.go.kr/ENG/contents/E0402000000.do?searchId=search&schM=list> (accessed on 18 August 2025).
62. Radeva, I.; Popchev, I.; Doukova, L.; Dimitrova, M. Web Application for Retrieval-Augmented Generation: Implementation and Testing. *Electronics* **2024**, *13*, 1361. [CrossRef]
63. Ozdemir, A.; Odaci, B.; Tanatar Baruh, L.; Varol, O.; Balcisoy, S. Enhancing Cultural Heritage Archive Analysis via Automated Entity Extraction and Graph-Based Representation Learning. *J. Comput. Cult. Herit.* **2025**. [CrossRef]
64. Zhang, J.; Cui, Y.; Wang, W.; Cheng, X. TrustDataFilter: Leveraging Trusted Knowledge Base Data for More Effective Filtering of Unknown Information. *arXiv* **2025**, arXiv:2502.15714.
65. Elastic. Nori Analysis Plugin. Available online: <https://www.elastic.co/guide/en/elasticsearch/plugins/current/analysis-nori.html> (accessed on 18 August 2025).

66. Li, J.; Bikakis, A. Towards a Semantics-Based Recommendation System for Cultural Heritage Collections. *Appl. Sci.* **2023**, *13*, 8907. [\[CrossRef\]](#)
67. Lee, J.; Cha, H.; Hwangbo, Y.; Cheon, W. Enhancing Large Language Model Reliability: Minimizing Hallucinations with Dual Retrieval-Augmented Generation Based on the Latest Diabetes Guidelines. *J. Pers. Med.* **2024**, *14*, 1131. [\[CrossRef\]](#)
68. Chen, D.; Sun, N.; Lee, J.-H.; Zou, C.; Jeon, W.-S. Digital Technology in Cultural Heritage: Construction and Evaluation Methods of AI-Based Ethnic Music Dataset. *Appl. Sci.* **2024**, *14*, 10811. [\[CrossRef\]](#)
69. Gîrbacia, F. An Analysis of Research Trends for Using Artificial Intelligence in Cultural Heritage. *Electronics* **2024**, *13*, 3738. [\[CrossRef\]](#)
70. Podsiadło, M.; Chahar, S. Text-to-Speech for Individuals with Vision Loss: A User Study. In Proceedings of the Interspeech 2016, San Francisco, CA, USA, 8–12 September 2016; pp. 347–351. [\[CrossRef\]](#)
71. Shin, D.; Park, J.; Kim, H. LoRA versus Prompt-Tuning: A Comparative Study in Domain-Specific Language Model Adaptation. *Electronics* **2024**, *6*, 56. [\[CrossRef\]](#)
72. Rasool, A.; Shahzad, M.I.; Aslam, H.; Chan, V.; Arshad, M.A. Emotion-Aware Embedding Fusion in Large Language Models (Flan-T5, Llama 2, DeepSeek-R1, and ChatGPT 4) for Intelligent Response Generation. *AI* **2025**, *6*, 56. [\[CrossRef\]](#)
73. Yu, D.; Bao, R.; Ning, R.; Peng, J.; Mai, G.; Zhao, L. Spatial-RAG: Spatial Retrieval Augmented Generation for Real-World Geospatial Reasoning Questions. *arXiv* **2025**, arXiv:2502.18470.
74. Theodoropoulos, G.S.; Nørvåg, K.; Doukeridis, C. Efficient Semantic Similarity Search over Spatio-Textual Data. In Proceedings of the 27th International Conference on Extending Database Technology (EDBT), Paestum, Italy, 25–28 March 2024; pp. 268–280.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.