https://doi.org/10.1186/s40494-024-01231-3

Heritage Science

RESEARCH

Open Access

Nanjing Yunjin intelligent question-answering system based on knowledge graphs and retrieval augmented generation technology

Liang Xu[1*], Lu Lu[2], Minglu Liu[3], Chengxuan Song[4] and Lizhen Wu[1]

Abstract

Nanjing Yunjin, a traditional Chinese silk weaving craft, is celebrated globally for its unique local characteristics and exquisite workmanship, forming an integral part of the world's intangible cultural heritage. However, with the advancement of information technology, the experiential knowledge of the Nanjing Yunjin production process is predominantly stored in text format. As a highly specialized and vertical domain, this information is not readily convert into usable data. Previous studies on a knowledge graph-based Nanjing Yunjin Question-Answering System have partially addressed this issue. However, knowledge graphs need to be constantly updated and rely on predefined entities and relationship types. Faced with ambiguous or complex natural language problems, knowledge graph information retrieval faces some challenges. Therefore, this study proposes a Nanjing Yunjin Question-Answering System that integrates Knowledge Graphs and Retrieval Augmented Generation techniques. In this system, the ROBERTA model is first utilized to vectorize Nanjing Yunjin textual information, delving deep into textual semantics to unveil its profound cultural connotations. Additionally, the FAISS vector database is employed for efficient storage and retrieval of Nanjing Yunjin information, achieving a deep semantic match between questions and answers. Ultimately, related retrieval results are fed into the Large Language Model for enhanced generation, aiming for more accurate text generation outcomes and improving the interpretability and logic of the Question-Answering System. This research merges technologies like text embedding, vectorized retrieval, and natural language generation, aiming to overcome the limitations of knowledge graphs-based Question-Answering System in terms of graph updating, dependency on predefined types, and semantic understanding. System implementation and testing have shown that the Nanjing Yunjin Intelligent Question-Answering System, constructed on the basis of Knowledge Graphs and Retrieval Augmented Generation, possesses a broader knowledge base that considers context, resolving issues of polysemy, vague language, and sentence ambiguity, and efficiently and accurately generates answers to natural language queries. This significantly facilitates the retrieval and utilization of Yunjin knowledge, providing a paradigm for constructing Question-Answering System for other intangible cultural heritages, and holds substantial theoretical and practical significance for the deep exploration and discovery of the knowledge structure of human intangible heritage, promoting cultural inheritance and protection.

Keywords Knowledge graphs, Retrieval augmented generation, Question-Answering System, Nanjing Yunjin, Vector retrieval

*Correspondence:
Liang Xu
xuliang@hdu.edu.cn
Full list of author information is available at the end of the article

Introduction

Nanjing Yunjin, a distinguished representative of Chinese silk weaving craftsmanship, is acclaimed as the final milestone in ancient Chinese silk weaving techniques. In 2009, the Yunjin weaving technique was officially inscribed in the UNESCO's Representative List of the Intangible Cultural Heritage (ICH) of Humanity.

Nanjing Yunjin is typically based on the complex and colorful Shu brocade, employing various intricate embroidery skills along with gold and silver threads to create vivid, splendidly ornate patterns and colors. This not only reflects the unique aesthetics of Nanjing Yunjin but also highlights its profound historical and cultural significance and exceptional craftsmanship. With the development of the internet, digital libraries, technical documents, etc., most information is provided in unstructured language form [1]. The complex production processes, diverse classifications, unique weaving techniques, and rich pattern connotations of Nanjing Yunjin are mostly documented in text. Analyzing this information to obtain corresponding thematic content, efficiently storing unstructured text data, and facilitating the management and utilization of specialized domain knowledge is crucial for understanding the historical origins of Nanjing Yunjin and for the inheritance and preservation of the weaving technique [2, 3].

Previous research [4] achieved high-quality, accurate answers in handling questions related to the field of Nanjing Yunjin using a knowledge graphs (KGs)-based Question-Answering System (Q&A system). However, this system also has its drawbacks and limitations.

Firstly, the Q&A system relies on the completeness of the KG, necessitating a precise, comprehensive, and structured KG as a foundation [5]. Most relationships in KGs tend to reflect only static interactions between entities, failing to represent dynamic activities and state changes of related entities. This limitation hinders KGs embedding models from effectively learning rich and comprehensive entity representations [6]. Inadequacies or inaccuracies in the KGs can adversely affect the performance of the Q&A system [7]. Secondly, the system depends on predefined entity and relationship types [8]. A primary constraint of KGs-based Q&A systems is their ability to answer only questions related to entities and relationships already modeled in the KGs. Existing KGs embeddings typically consider only direct relationships between entities, neglecting higher-order structural relationships [9]. If a query involves entities or relationships outside the KGs, the system cannot directly provide answers. Thirdly, natural language ambiguity and polysemy pose challenges, especially when user queries are vague or complex, making information retrieval (IR) based on KGs challenging [10]. Finally, there is a lack of reasoning capability. While humans can infer correct answers from multiple sources, traditional KGs usually rely on matching features of entities and relationships for independent reasoning on individual question–answer pairs, lacking deep understanding and inferential abilities to address complex questions requiring reasoning. This limitation presents significant constraints when dealing with complex problems and inferential queries.

In response to these issues, recent advancements in large semantic models have the potential to address problems inherent in KGs-based question-answering.

Large language models (LLM) can answer some questions beyond the KGs, including speculation and interpretation of unknown entities or relationships, due to their broad knowledge base acquired from extensive corpora and zero-shot inference capabilities on new questions [11]. This inferential ability to handle unknown information can alleviate the KG's reliance on known entities and relationships. LLM are particularly noteworthy as they generate fluent natural language texts and dialogues, achieving a quality and smoothness comparable to human creations [12]. With mechanisms for learning deep semantic representations, these models enhance understanding of complex contexts and reasoning capabilities, effectively handling contextual information and resolving semantic ambiguities caused by polysemy and vague language expressions, thereby excelling in answering questions requiring deep understanding and logical reasoning.

To address the limitations of KGs-based Q&A systems, this study proposes a Nanjing Yunjin Intelligent Q&A system that combines KGs and Retrieval Augmented Generation (RAG) techniques. Building on the KGs-based Q&A system from prior research, the RAG question-answering module is added. Initially, Yunjin texts are vectorized to extract multifaceted knowledge about production, classification, weaving techniques, and pattern connotations. This is followed by retrieval and augmented generation steps to reorganize the extracted semantic information into natural language answers, further promoting the exploration and utilization of Nanjing Yunjin culture, ensuring its protection and inheritance. Specifically, First, the ROBERTA model quantifies Nanjing Yunjin texts, yielding word vectors; second, the vectorized data are stored in the Facebook AI Similarity Search (FAISS) vector database, creating vector indexes and employing efficient, accurate approximate nearest neighbor search algorithms for retrieval; finally, retrieval results and user queries are fed into the Large Language Model Meta AI (LLAMA) model for enhanced generation of more accurate final answers. The RAG method integrates retrieval models and LLM, extracting semantic features of questions and answers and recalling results based on semantic relevance between texts, then feeding retrieval results into a large language model for augmented generation to achieve intelligent question-answering on Nanjing Yunjin-related knowledge.

The innovations and contributions of this paper are summarized as follows:

1. Semantic information is extracted using a vector model. The ROBERTA model is utilized to convert textual data into vector form, enhancing the identification of Nanjing Yunjin text features and improving the accuracy of text analysis.
2. The use of the FAISS database provides an efficient way to store and index a large amount of vector data. By calculating the similarity between the features of questions to be retrieved and the feature library, the top-k similar retrieval results are obtained, achieving a deep match between Nanjing Yunjin questions and answers.
3. Introduction of LLM. Relevant retrieval results and user queries are passed as background information to the LLAMA model for augmented generation, aiming to obtain more accurate text generation results. By combining retrieval models and LLM, semantic match retrieval and natural language generation are organically integrated, enhancing the interpretability and credibility of the Q&A system.
4. To improve the development and utilization of Nanjing Yunjin information and effectively preserve and protect this ICH, a Q&A system based on KGs and RAG technology is developed and implemented. The system integrates KGs and RAG technology to recognize user-input questions, answering through the KGs module if related entities and relationship types exist, and intelligently answering through the RAG module otherwise.

The structure of the remaining parts of this paper is as follows: The "Related work" section reviews the research status of Q&A systems and RAG technology. The "Methodology" section introduces the algorithms related to this study, the construction of the Yunjin knowledge base, and the design and implementation of the retrieval and augmented generation modules. The "Results and discussion" section discusses the implementation of the Q&A system, providing a detailed introduction and operational examples for each module in the Model View Controller (MVC) architecture. The "Conclusion" section summarizes the work of this study and analyzes the content and direction of future research.

Related work

This chapter delves into the scholarly examination of Q&A systems and the RAG technology, delineating the distinctions and enhancements of this study in comparison to existing research.

Question-Answering Systems in the intangible cultural heritage

Q&A systems are designed to furnish users with accurate answers to their inquiries promptly [13]. These systems are regarded as an advanced form of IR [14], allowing users to pose questions in natural language, with the system searching or filtering answers from a pool of candidate documents [15]. The evolution of Q&A systems is inextricably linked with advancements in AI and NLP, aiming to provide intelligent solutions for queries expressed in natural language, thus signifying a significant leap in IR technology [16].

Historically, Q&A systems were predominantly specialized, utilizing rule-based templates to process narrow and structured data sets to answer domain-specific questions. From the Turing test [17] to the inception of Eliza [18], and the development of the LUNAR system [19], the progression of Q&A systems has been evident. With the advancement of information and internet technologies, Q&A systems have evolved from specialized to general-purpose frameworks, albeit confined to natural language interactions. In recent years, with the advancement of AI, intelligent dialogue systems have garnered attention, such as OpenAI's latest release of GPT-4 [20], a model based on the Transformer architecture, which supports not only text input but also image-based question answering, marking the transition of Q&A systems into the era of intelligent interactive querying.

Q&A systems have expanded their applicability to the domain of ICH, enabling dynamic human–computer interaction. For instance, Wang and others have developed an interactive system based on a philosophical texts knowledge base, integrating natural language processing (NLP) and IR technologies to autonomously perform partial answer searching tasks [21]. Zhao constructed a specialized database focused on Mongolian stringed instruments, designing a detailed metadata structure and employing MySQL for the database, with retrieval and querying facilitated by the Lucene framework [22]. Sperli proposed an interactive cultural heritage framework, allowing visitors to access the system through a dialogue agent based on the Seq2Seq model [23]. These relational database-based ICH Q&A Systems have somewhat enhanced the efficiency of queries and utilization. However, the typically unstructured nature of data hinders semantic associations, often resulting in sluggish responses when handling large volumes of data and complex queries [24].

KGs, by offering semantically rich structured data representations, have facilitated a paradigm shift in the architecture of Q&A Systems. Within the domain of ICH, numerous studies have integrated KGs into Q&A Systems to enhance the performance of question-answering tasks. For instance, Liu et al. employed the BiLSTM-CRF model for named entity recognition to construct a KG and Q&A System for the Liao dynasty historical and cultural domain [25]. Liu developed a deep learning model based on Bidirectional Encoder Representations from Transformers (BERT) to recognize the intent and entities/attributes of input questions, constructing a mineral KG for querying and returning answers [26]. Aurpa utilized a Transformer-based deep neural network model for accurate and rapid retrieval of answers in Bengali language reading comprehension [27]. Suissa proposed a genealogical information system based on a deep neural network (DNN) with self-attention networks to answer genealogy-related questions [28]. These KG-based ICH Q&A Systems are capable of providing efficient and precise answers, yet they face limitations in graph updating, reliance on predefined types, and semantic understanding.

Recently, LLM have attracted considerable attention for their learning mechanisms, ability to reason and interpret, and potential to address issues inherent in KG-based question answering. OpenAI's ChatGPT marks the advent of a new era for LLM [29]. Since the introduction of GPT1.0, the GPT series models have significantly increased in parameter size, with GPT3.0 reaching 175 billion parameters [30]. LLM have made rapid progress and have been applied across various domains [31–33]. Despite this, challenges remain in the authenticity of the content generated by LLM. Retrieval augmentation has been proven to enhance the performance of Q&A Systems and reduce the occurrence of hallucinations produced by large models.

Therefore,this study enhances the KG-based approach by incorporating a RAG module to construct a Nanjing Yunjin Q&A System. Compared to traditional KG Q&A Systems, this study focuses on undefined entity types or relationships, and can enhance the generation of natural language answers by storing and retrieving unstructured information vectors. Compared to Q&A Systems based solely on LLM, this study retrieves substantiated data from knowledge bases to augment generation by large models, thereby mitigating the extent of hallucinations by the models. Through a collaborative approach involving various methods, this research achieves deep exploration, analysis, widespread dissemination, and utilization of knowledge on Nanjing Yunjin.

Research related to retrieval-augmented generation

RAG technology enhances the generation of more accurate and relevant natural language answers by integrating content retrieved from external knowledge bases with prompts for LLM [34]. Scholars have conducted research in indexing, retrieval, and generation to improve the quality of RAG.

Indexing involves the cleaning, extraction, transformation into plain text, segmentation, and indexing of data. Embedding models play a crucial role in this process by converting text segments into vector representations. Word embedding techniques have evolved from statistical models to neural network-based models. Early statistical methods primarily included one-hot encoding [35], bag of words [36], n-gram models [37], and Term Frequency-Inverse Document Frequency (TF-IDF) [38], which mainly characterized the relationships between words based on their frequency statistics in the text. However, with the advent of deep learning, neural network-based word embedding algorithms such as Word2Vec [39] and GloVe [40] have been widely adopted. These methods capture contextual information and train models by predicting target word vectors, thereby generating semantically rich word embeddings. Due to the static and unified nature of the generated distributed word vectors, Recurrent Neural Networks (RNNs) have been widely applied [41, 42]. The latest development is the Transformer model, such as BERT, which are deep learning-based pretrained models capable of processing bidirectional context information simultaneously, more effectively addressing the problem of polysemy. The ROBERTA embedding model used in this study is an optimization and refinement of BERT, incorporating larger model parameters and more training data, with improvements also made in training methodologies.

Retrieval refers to the process of querying and computing the text segments with the highest similarity based on user needs. In terms of text preprocessing before retrieval, Ashmawy proposed a method for text normalization, utilizing a transformer-based sequence-to-sequence (Seq2Seq) model to construct a machine translation model from multilingual characters to words [43]. In terms of retrieval model optimization, keyword retrieval has gradually evolved into semantic retrieval, reflecting the continuous development and progress of text retrieval technology [44]. Traditional text IR mainly encompasses the Boolean model, vector space model, and probabilistic model, representing complex retrieval processes through simple logical operators, but these can only reflect whether retrieval results are related without indicating the degree of relevance. The distributional hypothesis posits that the semantic proximity between words can be judged by the similarity of their contexts.

Notable models include the Deep Structured Semantic Models (DSSM) and the Deep Relevance Matching Model (DRMM). For example, Xie used DSSM to calculate the semantic relevance between questions and predicates among candidates, combining semantic-level and lexical-level scores to rank candidates [45]. Huang and others introduced Convolutional Neural Networks (CNN) on top of DSSM to better capture the interaction information of local features [46]. Palangi proposed the LSTM-DRMM model, which uses a Recurrent Neural Network (RNN) with Long Short-Term Memory (LSTM) units to sequentially capture each word in a sentence, while leveraging the probabilistic pooling layer of DRMM to capture the complex matching relationship between queries and documents and embedding it into semantic vectors [47].

With the recent rise of LLM, vector retrieval has garnered attention. Unlike traditional keyword-based text retrieval methods, vector retrieval uses embedding technologies to convert text content into feature vector representations with deep semantic information, thereby revealing the implicit connections and semantic similarities between texts [48]. This transformation not only significantly reduces the data storage requirements but also greatly improves retrieval efficiency and accuracy, and is widely applied in IR, data mining, machine learning, and other fields [49, 50]. Common vector retrieval methods include Nearest Neighbor Search (NNS) [51], K-Nearest Neighbor Search (K-NNS) [52], and Approximate Nearest Neighbor Search (ANNS) [53].In recent years, with the continuous increase in the size and dimensionality of data, ANNS has emerged as a popular method for fast and efficient retrieval from large-scale datasets. FAISS, an efficient open-source library, offers various implementations of ANNS to accommodate tasks with varying scales and performance requirements [54]. This study utilizes the FAISS vector database to extract semantic features through deep neural networks, aiming to map textual information to continuous vector representations in a high-dimensional feature space. Then, through ANNS, it quickly locates the vectors most similar to the query sample, thereby achieving precise retrieval and effective utilization of various information resources related to Nanjing Yunjin.

Generation refers to the process of inputting questions and retrieved documents as prompts into a LLM to enhance the generation of natural language answers. Numerous studies have optimized generators, such as Guo, who introduced Retrieval-Augmented Language for Literature Ambiguity generation, leveraging external knowledge sources to enhance the background explanation of biomedical texts [55]. Siriwardhana proposed the RAG-end2end model, which, through joint training of the retrieval and generation components, adapts to specific domain knowledge bases and has been improved for domain adaptability in open-domain question-answering tasks [56].

These studies on RAG provide a theoretical and practical foundation for this research. Through experimental comparisons, this study employs the advanced embedding model ROBERTA and the fast and efficient vector retrieval method. In the generation phase, it uses the currently highest-performing large language model, LLAMA, marking the first application of RAG technology in the field of Nanjing Yunjin. This aims to enhance the accuracy and effectiveness of the Q&A System.

Methodology

Models and algorithms used

TRANSFORMER

Textual materials related to Nanjing Yunjin exhibit highly discrete and fragmented characteristics, posing challenges to the effective extraction of textual features. Transforming these complex and fragmented texts from a deep semantic perspective to adapt them for algorithmic processing and understanding is particularly necessary. There are many methods to convert text content into word vectors, such as one-hot, word2vec, doc2vec, etc. These methods provide an effective pathway for text quantification but often neglect contextual semantic information. The TRANSFORMER model overcomes the shortcomings of existing text quantification methods, constructed using unsupervised training on large volumes of unlabelled corpus data. Introduced by Google in 2017, Transformer is a neural network architecture widely used in NLP, text classification, Q&A systems, and more. It consists of several encoders and decoders [57], each composed of multiple layers of identical blocks. These blocks are stacked together to form the overall architecture of the Transformer, as shown in Fig. 1. The ROBERTA vectorization model used in this paper employs 24 layers of transformer encoders, with a vector dimension of 1024. It first preprocesses the Nanjing Yunjin text data through tokenization, vocabulary mapping, etc., converting it into a sequence of word indices. Word Embedding is used to represent the input Nanjing Yunjin sentences as word vector expressions, with Positional Encoding capturing the positional encodings of these word vectors. This information is then fed into a multi-head self-attention mechanism layer, capturing global contextual information for each element in the input sequence through the relationships between the Query, Key, and Value vectors. Following the multi-head self-attention mechanism, there is a feed-forward neural network layer that performs the same operation on the vector of each

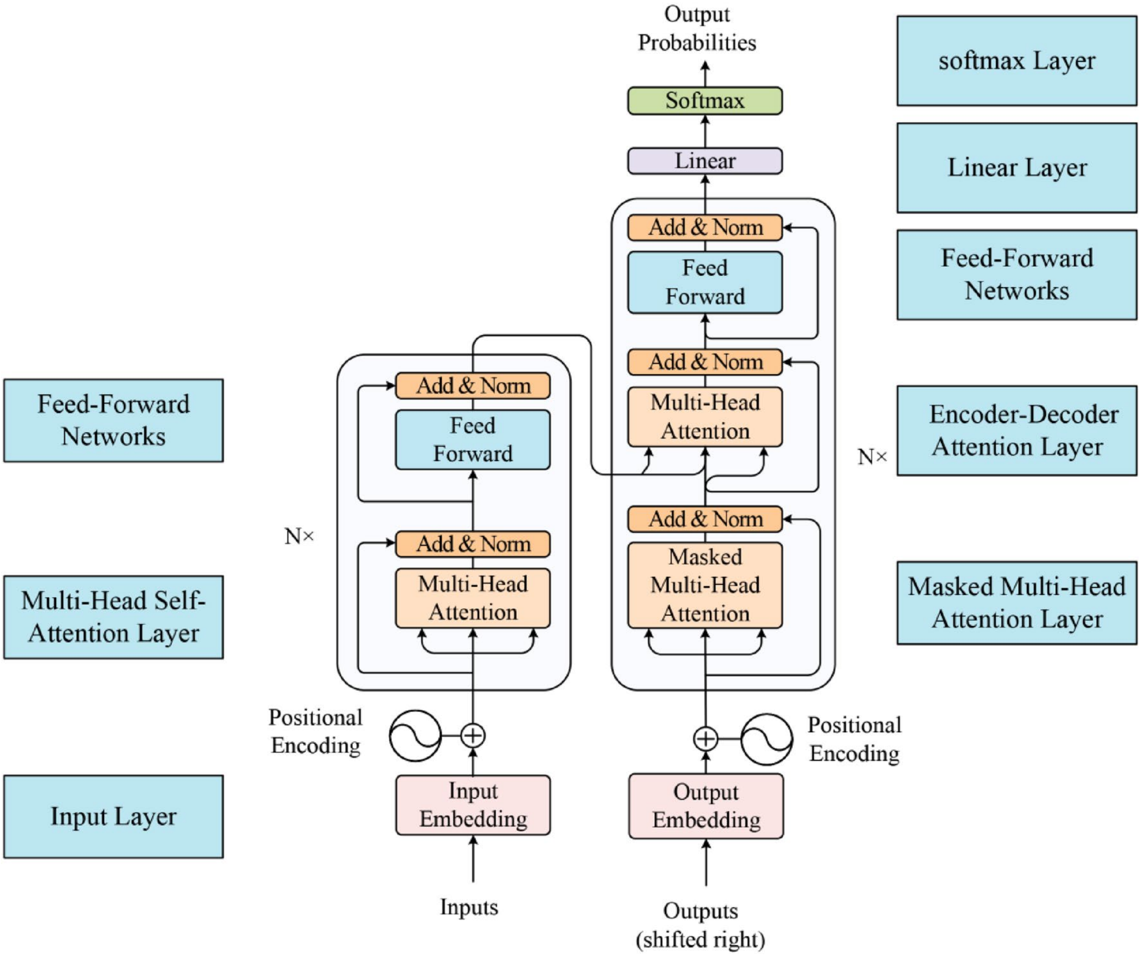


Fig. 1 Structure of transformer

position, including two linear transformations and a ReLU activation output. Each layer employs Residual Connections to address the problem of gradient vanishing in deep learning. Additionally, each layer's output undergoes Layer Normalization to stabilize the training process and enhance the model's generalizability. In the decoding phase, the masking operation of the Multi-Head Attention layer is used to enhance the memory of contextual information, thereby obtaining more global information. Subsequently, an encoder-decoder attention mechanism is introduced, allowing the decoder to access more source language information from the encoder's output, followed by a feed-forward neural network layer, residual connections, and layer normalization. After passing through multiple layers of the decoder structure, the final hidden state is projected onto a logits vector equal to the size of the vocabulary. This is followed by the application of the Softmax function to obtain a probability distribution, producing a vector of the same dimension as the input sequence as the output.

Euclidean distance

Euclidean distance is a mathematical measure of linear distance between two points in multidimensional space [58]. It quantifies the absolute distance between points in space, reflecting the absolute differences in individual numerical characteristics. In a two-dimensional space, for two points A(×1, y1) and B(×2, y2), the Euclidean distance is defined as the length of the straight line segment between these two points, formulated as:

$$\mathrm{d(A,B)} = \sqrt{(\mathrm{x2}-\mathrm{x1})^2 + (\mathrm{y2}-\mathrm{y1})^2} \tag{1}$$

Extended to an n-dimensional space, for two points A(×1, y1, z1,..., n1) and B(×2, y2, z2,..., n2), their Euclidean distance is calculated using the formula:

$$\mathrm{d(A, B)} = \sqrt{\sum_{\mathrm{i=1}}^{\mathrm{n}} (\mathrm{B_i} - \mathrm{A_i})^2} \tag{2}$$

Here, A and B represent two n-dimensional vectors, and and are their respective coordinates in the ith dimension.

Simply put, Euclidean distance is the square root of the sum of squared differences in coordinates across dimensions. It reflects the straight-line distance between two points in Euclidean space. This study employs the calculation of Euclidean distance between two vectors to measure their similarity, thus establishing the FAISS indexFlatL2 index.

Cosine similarity

Cosine similarity is a method that uses the cosine value of the angle between vectors to measure differences between entities and quantify their similarity in multidimensional space [59]. Unlike Euclidean distance, it mainly distinguishes differences in direction and is insensitive to absolute values.

For two non-zero vectors A = (a1, a2, ..., an) and B = (b1, b2, ..., bn), the cosine similarity between them is calculated as:

$$\text{cosine similarity}(A, B) = \cos(\theta) = \frac{A \bullet B}{||\mathrm{A}||||\mathrm{B}||} = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} {a_i}^2}\sqrt{\sum_{i=1}^{n} {b_i}^2}} \tag{3}$$

where "·" denotes the dot product of the vectors, "||||" indicates the magnitude (or length) of a vector, and θ is the angle between the two vectors.

The range of cosine similarity is from 0 to 1, with values closer to 1 indicating an angle closer to 0 degrees, meaning that the two vectors are more similar. A value of 0 indicates no similarity between the two vectors.

In fields such as text mining, IR, and recommendation systems, cosine similarity is commonly used to compare the degree of similarity between high-dimensional feature vectors of documents, user interests, etc. In this study, cosine similarity is employed from a semantic perspective to measure the similarity between user queries and document data.

Building a Yunjin knowledge base

In this study, Nanjing Yunjin data is mapped to an efficient vector space, extracting key information and features. The vectorized data can be stored in a vector database. A vector database is a novel type of database management system, facilitating the storage, indexing, and retrieval of high-dimensional vector data. This study selects the FAISS vector library as the storage and retrieval medium for the Nanjing Yunjin knowledge base. FAISS is an efficient vector database and similarity search library, capable of searching billions of vectors, making it the most mature approximate nearest neighbor retrieval library to date. It offers advantages such as fast retrieval speed, the ability to understand search semantics, and precise ranking of recall results.

Data collection

The primary data used in this study mainly comes from official internal materials provided by the Nanjing Yunjin Museum and Nanjing Yunjin Research Institute. Documents related to the ICH inheritors mostly come from official data on the Chinese ICH website. Information on weaving materials, machinery, and other related aspects was mainly obtained through on-site inspections and collections, and some academic papers and book materials related to Nanjing Yunjin were also selected as data sources.

Data preprocessing

Preprocessing of data from different sources and structures includes removing useless symbols, deleting data without textual content, eliminating duplicate data and advertisements. Regular expressions are used to delete useless symbols, including links and irrelevant characters on web pages. Data without textual content, some of which are purely images or have little text, are also deleted. Embedded duplicate content and irrelevant advertisements in web pages created by forwarding or quoting are also removed.

Language processing

The original data on Nanjing Yunjin is predominantly in Chinese, encompassing aspects such as production processes, weaving techniques, descriptions of artworks, along with relevant cultural connotations and technical explanations. Given the Q&A System's objective to cater to a broad audience, including non-native Chinese speakers, thereby facilitating widespread dissemination and utilization of knowledge on Nanjing Yunjin, this study employed professional human translation services to convert all data from Chinese to English. This approach ensured the accuracy and cultural sensitivity of the translations while also allowing for meticulous adjustments to terminology and phrasing to preserve the integrity of the original cultural meanings. To further guarantee the accuracy and consistency of the translations, a review team composed of linguists and experts in the field of

Nanjing Yunjin was established to meticulously re-examine the translated content.

Data segmentation

The first stage of building a knowledge base is data segmentation. The original Yunjin text needs to be divided into segments of n words (typically 100 to 200), splitting the original data. The process involves reading file content, removing newline characters, skipping empty lines, splitting each line into words, and adding them to the all_words list. Using list comprehension, the words in all_words are divided into segments of 100 words, forming a list of paragraphs. Each paragraph's words are joined with spaces and written into the result file, with each paragraph occupying one line. The segmented data serve as the corpus for vectorization.

Text vectorization

Processed and segmented data are converted into vectors. Vectorization transforms data into high-dimensional numerical vectors, enabling it to participate in subsequent indexing, retrieval, and sorting operations. Common vectorization methods include one-hot, word2vec, doc2vec, etc. These methods provide an effective pathway for text quantification but often fall short in capturing context semantic relationships. This study employs the ROBERTA model for contextual semantic extraction of input text and generation of word vector representations. The model uses 24 layers of transformer encoders with a vector dimension of 1024. It undergoes unsupervised training on large unlabelled corpora, capturing contextual semantic information, thereby addressing polysemy and vague expressions in Nanjing Yunjin text. For the usage scenarios of the Nanjing Yunjin corpus knowledge base, the vectorized results are stored in the FAISS database.

Comparison of vectorization models

To validate the superiority of the ROBERTA vector model in text feature extraction, three sets of vectorization models are selected for comparison, as shown in Table 1. The comparison models include m3e-base, BERT-Large-Cased, and BGE. The m3e-base model uses 12 layers of transformer encoders for contextual semantic extraction of input text, with a vector dimension of 768. The BERT-Large-Cased model also employs the Transformer architecture, containing 24 layers of Transformer encoder layers to learn semantic and contextual information in input text and generate word vector representations, with a dimension of 1024. BGE, short for BAAI General Embedding, is a general embedding model by Beijing Academy of AI, capable of mapping any text to a low-dimensional dense vector. The model version used in this experiment is bge-large-en-v1.5, with a generated word vector dimension of 1024.

Table 1 Comparison of retrieval results

Model name	Target segment retrieved	Ranking of target segment	Similarity to target segment
m3e-base	Yes	Top 1	0.9439394
roberta-large	Yes	Top 1	0.9982275
bge-large-en-v1.5	Yes	Top 1	0.9223862
bert-large-cased	Yes	Top 1	0.8572579

In this experiment, the collected corpus is divided into segments of 100 words, totaling 1846 segments. Under the same hardware conditions, the FAISS vector index files generated by the three models with a vector dimension of 1024 are all 7.21 MB.

Taking the user input "Yunxia is like brocade: the past and present life of Yunjin" as an example, the corpus contains the target segment "Yunxia is like brocade: the past and present life of Yunjin. In the aesthetic taste of the Chinese people, no one can avoid Jiangnan. The south of the Yangtze River is full of mist and rain, and the rouge beauty is elegant and gorgeous, with a drunken tenderness. This kind of taste, mixed with a philosophical outlook and scholar-bureaucrat temperament of 'looking back on time will become empty, and taking pleasure in time and leisure' has been written into every page of the history of Jiangnan people. Nanjing Yunjin, a luxury product from a thousand years ago."

As can be seen, under the condition of English corpus, using the roberta-large model for vectorization, the recall result in the FAISS retrieval stage shows a similarity of 99.82% with the target segment, indicating the best recall effect.

Fine-tuning of the ROBERTA model

In comparative experiments, using ROBERTA as the vectorization model for RAG experiments yielded significant results but failed to fully encapsulate the linguistic nuances of the Nanjing Yunjin, a specific domain of ICH. This shortfall led to a scenario where the text segments recalled in the RAG phase had high similarity but insufficient differentiation. Therefore, we fine-tuned the ROBERTA model to better adapt to the context and characteristics of the Nanjing Yunjin domain, thereby enhancing the text understanding capabilities in this field.

(1) Dataset introduction. For the fine-tuning process, we prepared 200 text segments related to Nanjing Yunjin. Each text segment was rewritten to maintain semantic integrity, generating five similar text segments as positive samples, labeled as 1. Addi-

tionally, other texts not belonging to the same similarity topic were randomly extracted from the text corpus as negative samples, labeled as 0. The total training dataset comprised 1540 entries, with a validation dataset of 192 entries and a test dataset of 193 entries.

(2) Model construction. In this experiment, Roberta was fine-tuned using a classification task approach with the constructed dataset of similar queries related to Nanjing Yunjin. During the model training phase, the stacked Transformer encoders within the Roberta model captured the context of sentences more effectively. The output from the Transformer encoders was then fed into a Dropout layer, which randomly deactivated a portion of the neurons to prevent overfitting during training. This was followed by a fully connected layer, where the scores for each category were transformed using the exponential function and then normalized to obtain the probabilities for each category. The difference between the predicted results and the actual labels was measured using the Cross-Entropy Loss function. Finally, the gradient was calculated using the backpropagation algorithm, and the model parameters were updated using the gradient descent algorithm to minimize the loss function, achieving the goal of fine-tuning the model.

(3) Hyperparameter settings. During the text input phase, the maximum sentence truncation length was set to 256, with 16 sentences per training batch. At the word vector representation stage, the ROBERTA-LARGE pre-trained model was used, with each layer having a vector dimension of 1024. The semantic encoding stage utilized 24 layers of Transformer encoders. During the model training phase, the dropout rate was set at 0.1, the learning rate at 5e-5, and the training duration was 20 rounds.

(4) Evaluation metrics. The experiment utilized Precision (P), Recall (R), and F1 Score (F1) to assess the model's performance. The best evaluation results from the validation dataset were recorded during the model training process. After training concluded, the model with the highest F1 score was loaded to evaluate the test dataset's performance as shown in Table 2.

(5) Experimental results. To demonstrate the effects of fine-tuning, we continue with the example of user input "Yunxia is like brocade: the past and present life of Yunjin. In the aesthetic taste of the Chinese people, no one can avoid Jiangnan." Using the fine-tuned ROBERTA model, through the FAISS vector database, the top 5 most relevant text segments were retrieved from the Nanjing Yunjin corpus as background material for the RAG stage. The experimental results are shown in Table 3.

Table 2 Test results

Dataset	Precision	Recall	F1 Score
Validation dataset	99.35%	99.50%	99.42%
Test dataset	98.71%	99.00%	98.84%

It can be observed that, after fine-tuning, the model is better able to capture the distinctions in the Nanjing Yunjin data. The similarity of the recalled text segments no longer clusters around 0.99 + but instead presents a ranking where higher similarities correspond to higher ranks, and lower similarities correspond to lower ranks. By doing so, it is possible to set a similarity threshold to filter out irrelevant information as much as possible before the RAG generation phase.

Design and implementation of the retrieval module

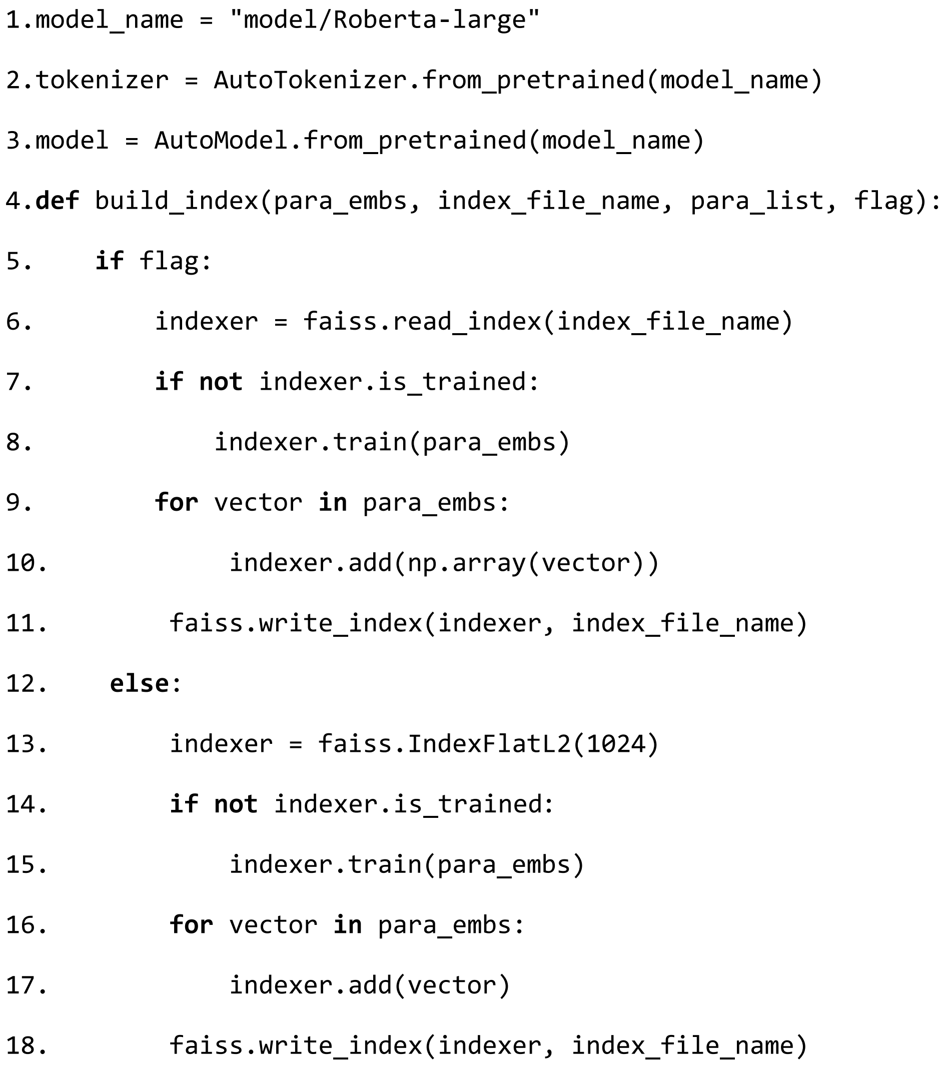
Establishing FAISS Index

A FAISS indexFlatL2 index type is established. This is an index method based on Euclidean distance, measuring the similarity between two vectors by calculating their Euclidean distance. After creating the index, the Yunjin text embeddings, fine-tuned to 1024 dimensions, are added to the index. This allows the index to store the embedding vector of each text and its unique identifier within the index. The constructed index is saved to disk for future research and applications, allowing for quick loading and use when needed.

To build the index, firstly, the pretrained model and tokenizer are loaded. During the initial construction of the index file, it needs to go through both the Train and Add processes. Subsequent additions of new vectors to the index file can be implemented through the Add operation for incremental index building. The core code is as follows:

Table 3 Experimental result

Text segment ID	Similarity before fine-tuning	Similarity after fine-tuning
1	0.9982275	0.9989113
2	0.9971823	0.9972273
3	0.9969448	0.9762033
4	0.9968793	0.9520621
5	0.9968697	0.8896874
6	0.9982275	0.9989113



```
model_name = "model/Roberta-large"
tokenizer = AutoTokenizer.from_pretrained(model_name)
model = AutoModel.from_pretrained(model_name)
def build_index(para_embs, index_file_name, para_list, flag):
    if flag:
        indexer = faiss.read_index(index_file_name)
        if not indexer.is_trained:
            indexer.train(para_embs)
        for vector in para_embs:
            indexer.add(np.array(vector))
        faiss.write_index(indexer, index_file_name)
    else:
        indexer = faiss.IndexFlatL2(1024)
        if not indexer.is_trained:
            indexer.train(para_embs)
        for vector in para_embs:
            indexer.add(vector)
        faiss.write_index(indexer, index_file_name)
```

FAISS retrieval

When performing the recall operation, the input query vector needs to be preprocessed first. The ROBERTA text embedding model is used to vectorize the user's input question. Then, after obtaining the input semantic vector, the FAISS index is used to retrieve similar texts from the knowledge base. The method used for calculating the similarity between vectors is cosine similarity, determining whether the two vectors are zero. If so, the similarity is directly returned as 0. A similarity threshold is set, and if the similarity in the retrieval results is greater than this threshold, it is considered related to the user input.

The key code for retrieval is as follows:

```
class FaissTool():
    def __init__(self, text_filename, index_filename):
        self.engine = faiss.read_index(index_filename)
        self.id2text = []
        for line in open(text_filename,'r',encoding='utf-8'):
            self.id2text.append(line.strip())
    def cosine_similarity(self,vector1, vector2):
        dot_product = np.dot(vector1, vector2)
        norm_vector1 = np.linalg.norm(vector1)
        norm_vector2 = np.linalg.norm(vector2)
        if norm_vector1 == 0 or norm_vector2 == 0:
            return 0
        similarity = dot_product / (norm_vector1 * norm_vector2)
        return similarity
    def search(self, q_embs, topk=5):
        res_dist, res_pid = self.engine.search(q_embs, topk)
        result_list = []
        sim = []
        for i in range(topk):
            index = int(res_pid[0][i])
            vector = self.engine.reconstruct(index)
            q = q_embs.flatten()
            sim.append(self.cosine_similarity(q ,vector))
            result_list.append(self.id2text[index])
        return zip(result_list,sim)
```

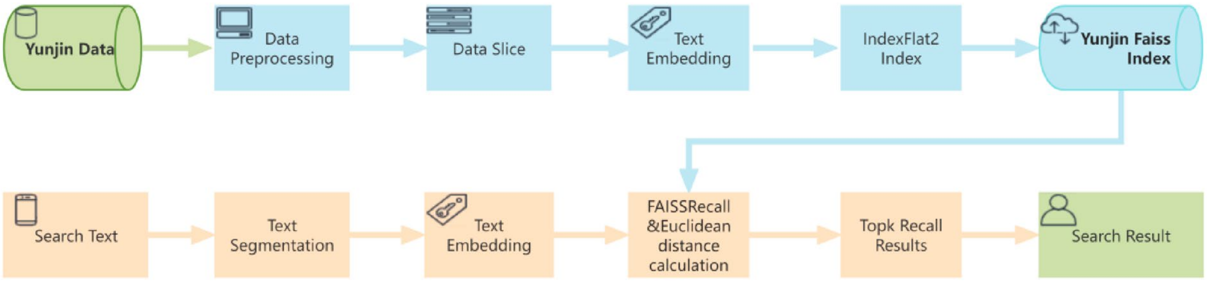


Fig. 2 FAISS index creation and result recall process

Result output

The k most similar texts to the search text vector are retrieved as output results. Specifically, they are sorted in descending order by cosine value, and the k texts with the highest cosine values, i.e., highest similarity, are selected as recall results. The value of k can be set according to practical needs, and the retrieval results should be provided as background information to the large model for generating question–answer results. The FAISS index creation and retrieval process is as shown in Fig. 2.

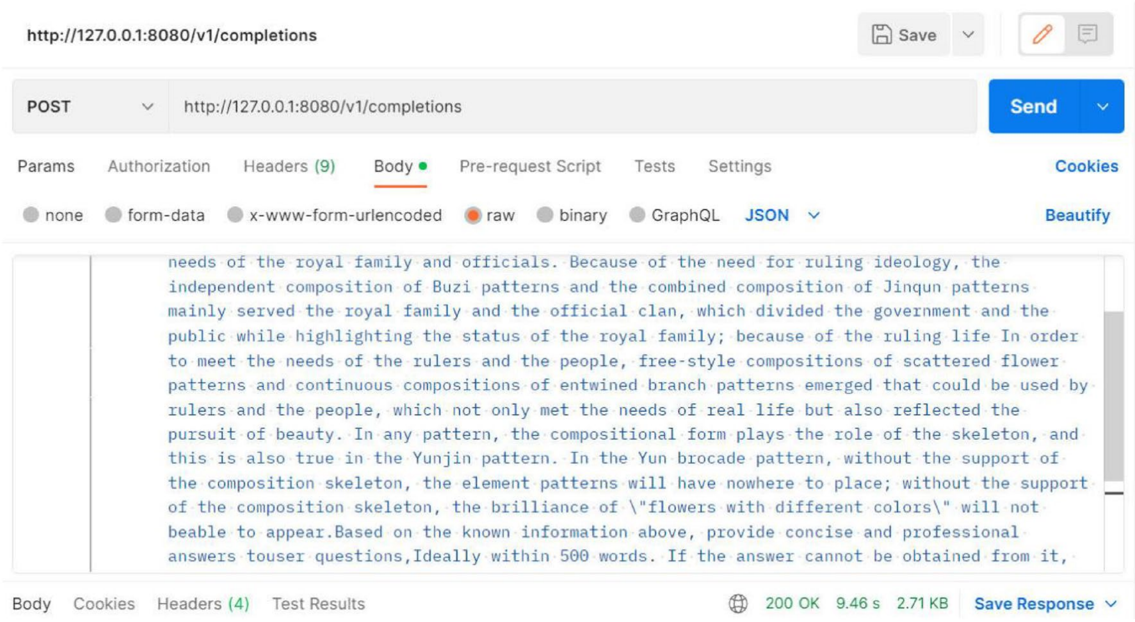
Design and implementation of the augmented generation module

Feeding the information provided by FAISS retrieval results into a large language model can enhance the generation of more natural and accurate answers, to some extent mitigating illusions often encountered with large models. In this study, the chosen large language model is LLAMA, the latest open-source large-scale language model released by Meta, with parameters ranging from 7 to 70 billion. It underwent deep pretraining and fine-tuning using 1.4 trillion tokens of training data, achieving exceptional text generation capabilities. It excels in common sense reasoning, question-answering, mathematical reasoning, code generation, and language understanding.

Deployment of the large language model

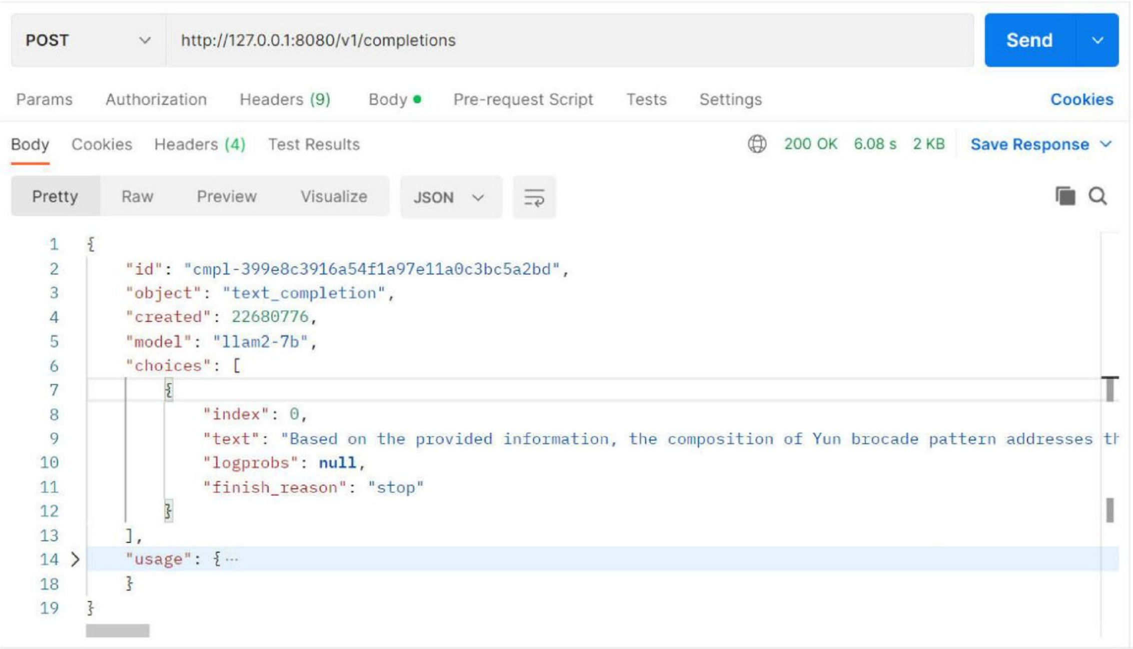
In this experiment, LLAMA-2-7b-chat was used. The downloaded model weights are provided in the official LLAMA2 format. To facilitate usage, they need to be converted to the Hugging Face official API and data structure, allowing the model to be fine-tuned, evaluated, generated, and deployed using various tools provided by Hugging Face.

The LLAMA2-7b-chat-hf large model requires a current memory of > = 28 GB, and a single 3090 graphics



```
needs of the royal family and officials. Because of the need for ruling ideology, the
independent composition of Buzi patterns and the combined composition of Jinqun patterns
mainly served the royal family and the official clan, which divided the government and the
public while highlighting the status of the royal family; because of the ruling life In order
to meet the needs of the rulers and the people, free-style compositions of scattered flower
patterns and continuous compositions of entwined branch patterns emerged that could be used by
rulers and the people, which not only met the needs of real life but also reflected the
pursuit of beauty. In any pattern, the compositional form plays the role of the skeleton, and
this is also true in the Yunjin pattern. In the Yun brocade pattern, without the support of
the composition skeleton, the element patterns will have nowhere to place; without the support
of the composition skeleton, the brilliance of \"flowers with different colors\" will not
beable to appear.Based on the known information above, provide concise and professional
answers touser questions,Ideally within 500 words. If the answer cannot be obtained from it,
```

Fig. 3 Request example



```
{
    "id": "cmpl-399e8c3916a54f1a97e11a0c3bc5a2bd",
    "object": "text_completion",
    "created": 22680776,
    "model": "llam2-7b",
    "choices": [
        {
            "index": 0,
            "text": "Based on the provided information, the composition of Yun brocade pattern addresses th
            "logprobs": null,
            "finish_reason": "stop"
        }
    ],
    "usage": {···
    }
}
```

Fig. 4 Request results

card has a memory of 24 GB. Therefore, two 3090 graphics cards were used in this experiment to jointly complete the inference tasks. To accelerate model inference, an inference framework for large models, vLLM, needs to be installed. Developed at the University of California, Berkeley, vLLM equipped with PagedAttention redefines the latest technology level for LLM services. Its throughput is 24 times higher than HuggingFace Transformers, without any need for changing the model architecture. vLLM also has the capability to deploy models on servers. Once successfully deployed, it can provide question-answering services via an API.

Inference of the large language model

Following the requirements of the large language model, the IDs retrieved by FAISS are indexed to the text content, and then prompt words are concatenated. The constructed prompt words are passed to the large language model for inference. The model generates one or multiple possible answer candidates based on the input information, reflecting a deep understanding and interpretation of the query question.

Taking the question "How does the composition of Yunjin patterns meet the specific needs and requirements of the royal family and officials?" as an example, a POST request is sent to the server, as shown in Fig. 3.

The IDs retrieved by FAISS are indexed to the text content, and prompt words are concatenated through the code:

Known information:Due to the particularity of the Yun brocade pattern specifically serving the royal family and officials, its composition was formed around the needs of the royal family and officials. Because of the need for ruling ideology, the independent composition of Buzi patterns and the combined composition of Jinqun patterns mainly served the royal family and the official clan, which divided the government and the public while highlighting the status of the royal family; because of the ruling life In order to meet the needs of the rulers and the people, free-style compositions of scattered flower patterns and continuous compositions of entwined branch patterns emerged that could be used by rulers and the people, which not only met- the needs of real life but also reflected the pursuit of beauty. In any pattern, the compositional form plays the role of the skeleton, and this is also true in the Yunjin pattern. In the Yun brocade pattern, without the support of the composition skeleton, the element patterns will have nowhere to place;

Based on the known information above, provide concise and professional answers touser ques tions,Ideally within 500 words. If the answer cannot be obtained from it, please say "cannot answer the question based on known information" or "did not provide enough relevant information, and it is not allowed to add fabricated elements to the answer.

The user question is:

What specific needs and requirements of the royal family and officials does the composition of Nanjing Yunjin pattern address?

In the prompt template, < s > , [INST], < < SYS > > , < < / SYS > > , [/INST], and [/s] are all special markers designated by LLAMA to distinguish prompt words and user problems.

Post-processing and presentation of answers

The answer candidates generated by the large language model undergo post-processing, including the removal of irrelevant information, correction of grammatical errors, and format adjustments, to make them more natural and understandable. Then, based on certain evaluation criteria, the answer candidates are scored, and the best answer is selected. Finally, the chosen optimal answer is presented to the user in natural language form to meet their information needs regarding Yunjin.

Following the example request, the results obtained are as follows, shown in Fig. 4.

As a symbol of royalty: Yunjin patterns are specifically designed for the royal family and officials, highlighting their status and power. The intricate design of the patterns reflects the wealth and luxury associated with the royal court.

Symbolism and ceremony: Yunjin patterns have rich symbolic meanings, with each element and color representing specific aspects of royal life and ceremonies. For example, "scattered flowers" in the patterns signify the court's prosperity and flourishing, while "entwined branches" symbolize the unity and stability of the monarch's rule. The design of the patterns reflects the belief that a monarch's power and authority originate from the natural world and the universe.

Representing political thought: Yunjin patterns are closely linked to the court's political ideology. The design of the patterns echoes Confucian principles of harmony, balance, and hierarchy, which are core to the imperial system. The "different colored flowers" in the patterns represent the various levels and positions within the government, while "entwined branches" symbolize the nation's unity and stability.

Meeting practical needs: Besides symbolic and ceremonial functions, Yunjin patterns also satisfy the actual needs of the royal family. The complex design of the patterns allows for a highly variable and flexible arrangement, suitable for various garments and textiles. This flexibility makes the patterns multifunctional and practical, suitable for real-life occasions, such as ceremonial and court clothing.

Reflecting aesthetic tastes: Yunjin patterns are not only functional and symbolic but also aesthetically pleasing. The combination of bright colors and complex patterns creates a visually striking design, reflecting the thoughtful aesthetics of the artisans who created them. The beauty of the patterns mirrors the cultural and artistic values of the time, as well as the thoughtful craftsmanship involved in their creation.

As can be seen, by obtaining background information relevant to the question from the knowledge base through FAISS retrieval and providing it to LLAMA2, the large model successfully generates answers relevant to the question from the background information. By querying substantiated data from the knowledge base before generating responses and submitting it to the large model, illusions of the large model can be somewhat mitigated. Continuously querying external sources ensures information remains up-to-date without frequent retraining of the large language model, reducing costs and addressing issues of real-time responsiveness and timeliness in large models. At the same time, vectorizing data through deep neural networks is particularly suited for unstructured data that KGs struggle to handle. Employing various methods to work in tandem on data processing, the study maximizes the extraction of information from Nanjing Yunjin-related data, providing users with higher quality IR services.

RAG evaluation

In the fields of NLP and IR, the effectiveness of RAG techniques has been validated through extensive practice. However, to rigorously assess content accuracy and quality control, this study designed a qualitative evaluation experiment for RAG, specifically focusing on the theme of Nanjing Yunjin. The experiment involved a set of 100 questions related to Nanjing Yunjin. Initially, researchers manually compiled standard answers and input them into the RAG system for automatic generation of answers and their corresponding contexts. The generated results were evaluated based on the following five metrics:

Faithfulness: Measures the consistency of the generated answer with the facts provided in the given context.

Answer relevancy: Assesses the relevance of the generated answer to the user's question.

Context precision: Measures the degree of match between the retrieved context information and the user's question.

Context recall: Evaluates the completeness of the context information retrieved from external knowledge sources.

Table 4 Partial RAG evaluation results

Question	Generated Content	Context	True answer in context	F	AR	CP	CR	CRY
For what occasions was Nanjing Yunjin mainly used in ancient times?	Nanjing Yunjin was mainly used in ancient times for royal court clothing and as a special product for the royal family during the Yuan, Ming, and Qing dynasties	…… this shows the status of Yun brocade in the royal ruling class. ……	Nanjing Yunjin was mainly used for royal clothing in ancient times and was a symbol of the royal family	1	1	1	1	1
Where is Nanjing Yunjin produced?	Nanjing Yunjin is produced in Nanjing, Jiangsu Province, China	……many comrades recalled and expressed their gratitude to this artist who made outstanding contributions to the protection, inheritance and development of Nanjing', 'respectively.……	Nanjing Yunjin is produced in Nanjing, Jiangsu Province, China	1	1	1	1	1
What are the main raw materials of Nanjing Yunjin?	Silk is the main raw material for Nanjing Yunjin. The material is soft to the touch, delicate, smooth, and wear-resistant. In addition to silk, gold and silver threads, and peacock feather threads are also commonly used in Yunjin production	Third step is the preparation of raw materials (such as gold and silver thread, peacock thread) Feather thread and silk raw materials that require nearly ten processes for processing and dyeing)……	Mainly mulberry silk, sometimes gold and silver thread, peacock feathers and other special materials are also used	1	1	1	1	1
When was Nanjing Yunjin Research Institute established?	The Nanjing Yunjin Research Institute was established on December 5, 1957	In March 1957, the Yunjin Research Institute was formally established. ……	1957	1	1	1	1	1
What does Bufu mean?	The information does not provide enough context or information to answer this question in detail, so I cannot provide a definitive answer	The inheritance and dissemination trend of Nanjing Yunjin in the new era ……	Bufu also refers to the official uniforms worn by ancient Chinese officials. Because of the distinctive Buzi pattern on the clothing, it is called Bufu	0.5	0	0	0.5	0

In the table, F denotes Faithfulness, AR denotes Answer Relevancy, CP denotes Context Precision, CR denotes Context Recall, and CRY denotes Context Relevancy

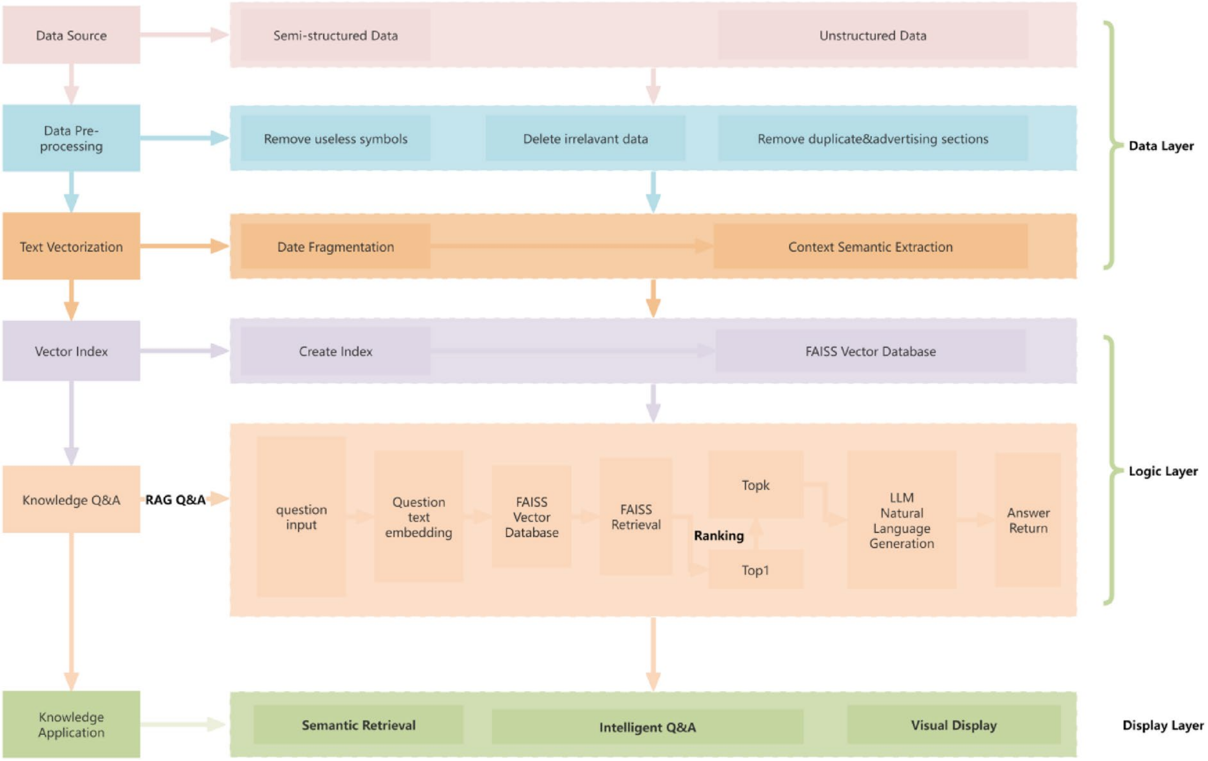


Fig. 5 System architecture diagram

Context relevancy: Assesses the relevance of the retrieved context information to the user's question.

During the evaluation process, each metric was assigned a quantified scoring standard, where high accuracy answers received a score of 1, average ones received 0.5, and completely irrelevant ones scored 0. To ensure the reliability and validity of the evaluation process, authoritative experts in the field of Nanjing Yunjin were invited to participate in manual judgment, assigning scores to each metric based on rigorous evaluation criteria. The partial evaluation results are summarized in the Table 4.

Overall, this study, through a scientifically rigorous method, has conducted a detailed and comprehensive qualitative evaluation of the RAG system's performance from multiple key perspectives. For the set of 100 questions, the RAG evaluation yielded an average faithfulness of 0.76, answer relevancy of 0.76, context precision of 0.65, context recall of 0.75, and context relevancy of 0.65. With average scores exceeding 0.6, these results indicate that the RAG system can answer users' questions in a truthful and accurate manner. Not only do these outcomes help to deeply understand the applicability of the system within specific cultural knowledge domains, but they also provide empirical evidence and directions for future enhancements and optimizations of the RAG model.

Results and discussion

In response to the shortcomings in the Nanjing Yunjin KGs-based Q&A system, this study has developed a digital resource Q&A system for Nanjing Yunjin based on KGs and RAG technology.

System development environment

The Q&A system developed in this project is platform-independent and can be run on common systems like Windows, Linux, and Mac. The system utilizes Python's Django web development framework. Python is a

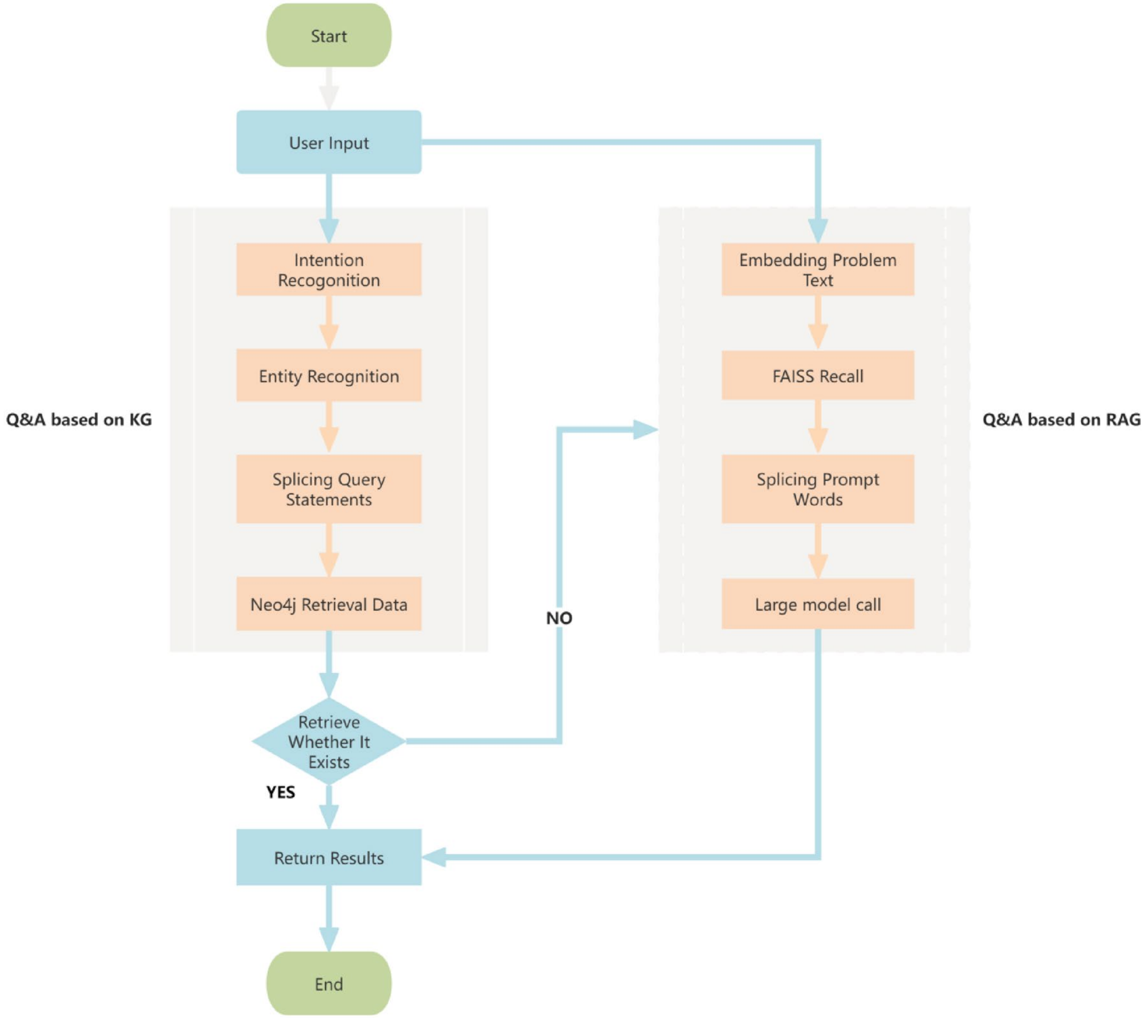


Fig. 6 Process diagram

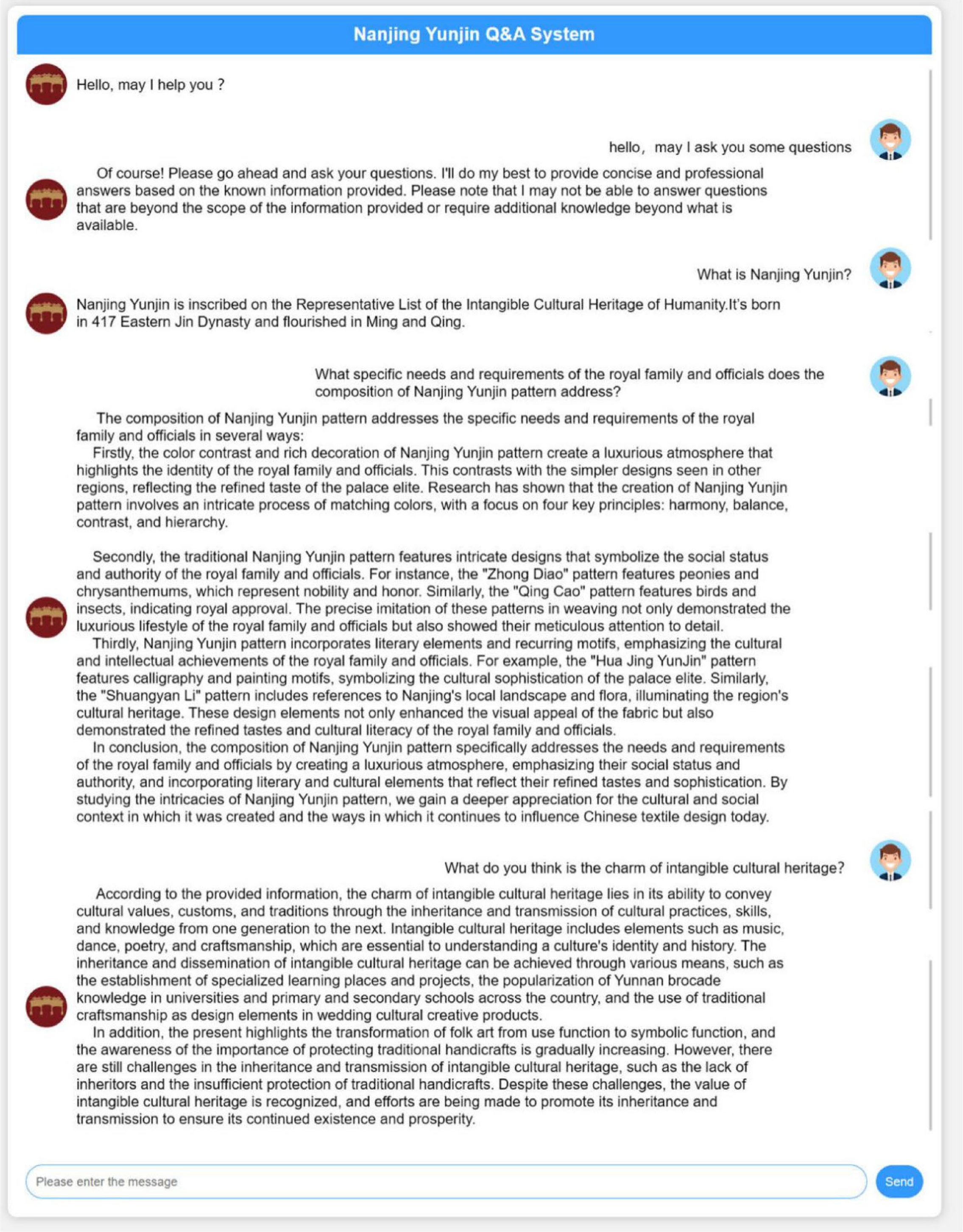
programming language known for its simplicity, ease of learning, and high code readability. It offers advantages such as simplicity, ease of use, and fast development speed, supporting multiple programming paradigms, and performs exceptionally in areas like big data processing, AI, and web development. Django is a comprehensive, large-scale open-source web design framework and is currently a commonly used application framework. The large language model employed is the LLAMA-2-13b-chat version, with a scale ranging from 7 to 70 billion parameters. The database storage utilizes the vector database FAISS, the most mature ANNS library currently available, providing interfaces for various programming languages and easily integrating into applications and platforms.

Overall system framework

This study constructs a Q&A system based on RAG technology on the foundation of the Nanjing Yunjin KGs, utilizing the MVC architecture. The system is divided into three layers: presentation, logic, and data [60]. The overall architecture of the system is shown in Fig. 5.

(1) Data layer: It primarily provides the data foundation for the entire Q&A system. Semi-structured or unstructured data related to Nanjing Yunjin, such as literature, papers, and relevant books, are obtained through official website crawling and manual selection. The data is then vectorized using the ROBERTA text embedding model, transforming it into high-dimensional numerical vectors for subsequent indexing, retrieval, and sorting operations.

(2) Logic layer: This layer is responsible for creating indexes, retrieval and sorting, and augmenting



Nanjing Yunjin Q&A System

Hello, may I help you ?

hello，may I ask you some questions

Of course! Please go ahead and ask your questions. I'll do my best to provide concise and professional answers based on the known information provided. Please note that I may not be able to answer questions that are beyond the scope of the information provided or require additional knowledge beyond what is available.

What is Nanjing Yunjin?

Nanjing Yunjin is inscribed on the Representative List of the Intangible Cultural Heritage of Humanity.It's born in 417 Eastern Jin Dynasty and flourished in Ming and Qing.

What specific needs and requirements of the royal family and officials does the composition of Nanjing Yunjin pattern address?

The composition of Nanjing Yunjin pattern addresses the specific needs and requirements of the royal family and officials in several ways:

Firstly, the color contrast and rich decoration of Nanjing Yunjin pattern create a luxurious atmosphere that highlights the identity of the royal family and officials. This contrasts with the simpler designs seen in other regions, reflecting the refined taste of the palace elite. Research has shown that the creation of Nanjing Yunjin pattern involves an intricate process of matching colors, with a focus on four key principles: harmony, balance, contrast, and hierarchy.

Secondly, the traditional Nanjing Yunjin pattern features intricate designs that symbolize the social status and authority of the royal family and officials. For instance, the "Zhong Diao" pattern features peonies and chrysanthemums, which represent nobility and honor. Similarly, the "Qing Cao" pattern features birds and insects, indicating royal approval. The precise imitation of these patterns in weaving not only demonstrated the luxurious lifestyle of the royal family and officials but also showed their meticulous attention to detail.

Thirdly, Nanjing Yunjin pattern incorporates literary elements and recurring motifs, emphasizing the cultural and intellectual achievements of the royal family and officials. For example, the "Hua Jing YunJin" pattern features calligraphy and painting motifs, symbolizing the cultural sophistication of the palace elite. Similarly, the "Shuangyan Li" pattern includes references to Nanjing's local landscape and flora, illuminating the region's cultural heritage. These design elements not only enhanced the visual appeal of the fabric but also demonstrated the refined tastes and cultural literacy of the royal family and officials.

In conclusion, the composition of Nanjing Yunjin pattern specifically addresses the needs and requirements of the royal family and officials by creating a luxurious atmosphere, emphasizing their social status and authority, and incorporating literary and cultural elements that reflect their refined tastes and sophistication. By studying the intricacies of Nanjing Yunjin pattern, we gain a deeper appreciation for the cultural and social context in which it was created and the ways in which it continues to influence Chinese textile design today.

What do you think is the charm of intangible cultural heritage?

According to the provided information, the charm of intangible cultural heritage lies in its ability to convey cultural values, customs, and traditions through the inheritance and transmission of cultural practices, skills, and knowledge from one generation to the next. Intangible cultural heritage includes elements such as music, dance, poetry, and craftsmanship, which are essential to understanding a culture's identity and history. The inheritance and dissemination of intangible cultural heritage can be achieved through various means, such as the establishment of specialized learning places and projects, the popularization of Yunnan brocade knowledge in universities and primary and secondary schools across the country, and the use of traditional craftsmanship as design elements in wedding cultural creative products.

In addition, the present highlights the transformation of folk art from use function to symbolic function, and the awareness of the importance of protecting traditional handicrafts is gradually increasing. However, there are still challenges in the inheritance and transmission of intangible cultural heritage, such as the lack of inheritors and the insufficient protection of traditional handicrafts. Despite these challenges, the value of intangible cultural heritage is recognized, and efforts are being made to promote its inheritance and transmission to ensure its continued existence and prosperity.

Fig. 7 Question-answering example

the generation of natural language answers. In this study, FAISS indexFlatL2 indexes based on Euclidean distance are created, which help quickly locate target data without having to compare each piece of data individually, thus enhancing retrieval speed and accuracy. Then, user inputs are vectorized, and cosine similarity is used to calculate the match between user input and text, sorting the cosine val-

Table 5 Stress test results

Label	HTTP request	Total
#Sample	2000	2000
Average value	11850 ms	11850 ms
Median	11300 ms	11300 ms
90%Line	15184 ms	15184 ms
95% Line	16443 ms	16443 ms
99%Line	43344 ms	43344 ms
Minimum value	2959 ms	2959 ms
Maximum value	46715 ms	46715 ms
Error%	0.00%	0.00%
Throughput	1.62718bps	1.62718bps

ues of user questions from high to low. The closer the cosine value is to 1, the more similar it is. Finally, the top k results with the highest similarity are extracted and submitted as background information to the large language model to augment the generation of more logical and readable results.

(3) Presentation layer: This layer primarily implements user interaction and page display. The presentation layer is the front-end page, mainly based on the Django framework to build a web-based Nanjing Yunjin Intelligent Q&A system, where users can ask questions. The presentation layer submits user data to the logic layer for processing, uses Python to connect and query the FAISS vector database, and ultimately returns the answer generated by the large language model back to the user.

System implementation and display

Upon receiving a user's query, the system first performs KGs-based question-answering. This involves entity recognition and intent identification to determine if the query contains specific entities and intents. If so, the KGs question-answering module is invoked to query knowledge from the local knowledge base. This approach is taken to maximize the use of high-quality local graph data and minimize illusions in the large model's responses. If the KGs cannot provide an answer, the FAISS vector database is used to retrieve the Topk segments of the Nanjing Yunjin corpus most relevant to the user's question. These segments, along with the user's question, are concatenated into prompt words and sent to the large model for question-answering. The process is shown in Fig. 6.

By accessing the Nanjing Yunjin Q&A system interface based on KGs and RAG technology through a browser, users are presented with a search box, send button, and answer display box.

The system supports three types of questions:

KGs-based Question-Answering. For example, for the first question, "What is Nanjing Yunjin?" the system, through named entity recognition, identifies 'Nanjing Yunjin' as a defined E1 CRM Entity type. Intent recognition then categorizes it as "Ask for definition", and the system finds the entity's corresponding properties and relationships in the triple store, such as the "Has Type" relationship between E1 CRM and E55 TYPE entities. The system then returns a natural language answer: 'Nanjing Yunjin is included on the Representative List of the ICH of Humanity.' This question involves entity and relationship types predefined in the KGs, enabling a direct search for answers within the KGs.

RAG Question-Answering. For example, the second question, "What specific needs and requirements of the royal family and officials does the composition of Nanjing Yunjin pattern address?" Since predefined relationship types in the KGs are limited, the system, unable to find a corresponding answer in the KGs, proceeds with retrieval-augmented generation. It recalls results through the vector database, inputs them as background information into the LLM model, and accurately generates a natural language text answer. Vectorization technology and LLM can to some extent compensate for deficiencies in KGs-based Q&A systems' reliance on predefined entities and relationships, untimely graph updates, and understanding of ambiguous or complex natural language.

Casual Conversation Question-Answering. For example, the third question, "What do you think is the charm of ICH?" The system still retrieves an answer. This type of question belongs to casual conversation and does not involve specific knowledge of Nanjing Yunjin. As there are no corresponding answers in the KGs and documents, and the large language model has certain logical and reasoning capabilities, it can infer the charm of ICH based on the relationship between Nanjing Yunjin and intangible heritage. It can provide a reasonable answer to such casual conversation questions, ensuring the system's question-answering is interesting and coherent, thereby enhancing the system's scalability and generalization capabilities. An example of the question-answering is shown in Fig. 7.

Testing with these three types of natural language questions demonstrates the feasibility of the system's process design and algorithm operation. By providing Nanjing Yunjin knowledge services to users through various KGs and RAG technology, the Q&A system achieves its intended functionality.

System testing

The constructed Nanjing Yunjin Intelligent Q&A system underwent practical effect testing, including system interface display, user sentence input, sentence query, and answer return, all of which achieved the expected results.

Stress testing

Stress testing is a method used to evaluate the performance and reliability of a system, network, or application under real load conditions. This test simulates a large number of concurrent user requests or high load situations to assess the response time, throughput, and resource utilization of the system under test. In this experiment, Apache JMeter was employed as the stress testing tool. The stress test simulated 100 users sending requests to the interface simultaneously, with a loop count of 20, totaling 2000 requests. The average response time was 11.85 s, the maximum response time was 46.715 s, and the minimum response time was 2.959 s, with an error rate of 0. The results of the stress test are shown in Table 5.

The stress test results indicate that the Nanjing Yunjin Q&A System based on the KG and RAG can still function normally with a concurrency of 100. It can withstand a certain number of concurrent requests, providing users with a fast and stable question-answering platform.

Performance evaluation

For this study, 100 questions related to Nanjing Yunjin were prepared, some targeting the KGs, some targeting documentation, and some casual conversation. Accuracy testing was conducted in the Q&A system. After multiple rounds of testing, the system operated well. Of the 40 KGs questions, the system provided objective and accurate responses to 35 questions, achieving an accuracy rate of 87.5%. Of the remaining 40 documentation questions, 32 were answered correctly, and 3 were not quite accurate, resulting in an accuracy rate of 80%. There were 18 casual conversation questions, of which the system accurately answered 12, resulting in an accuracy rate of 67%, improving the system's expandability. Overall, the system's answer accuracy rate was 79%, indicating that there is room for improvement in the model.

This study constructed a Q&A system for Nanjing Yunjin combining KGs and RAG technology and evaluated its accuracy, stability, and expandability. The system fundamentally meets the knowledge question-answering needs of the Nanjing Yunjin field. The Nanjing Yunjin Q&A system based on KGs and RAG technology serves as a window for users to utilize, protect, and inherit Yunjin culture and can be continuously improved and enhanced in future research.

Conclusion

This research developed and implemented an intelligent Q&A system for Nanjing Yunjin, integrating KGs and RAG technologies. The system enhances the organization, retrieval, protection, and application efficiency of Nanjing Yunjin information, particularly when dealing with challenges posed by unstructured natural language data. It employs text embedding models to manage scattered textual information and uses the Euclidean distance algorithm for computing the similarity between queries and text vectors, thereby facilitating the generation of more accurate and reliable textual answers through a large language model. The integration of RAG technology enables efficient and accurate natural language responses to user queries, enhancing the interpretability and credibility of the Q&A system.

The main research achievements include:

(1) Semantic information extraction using deep neural networks, with the ROBERTA vector model analyzing contextual semantic information to obtain text word vectors. Text embedding extracts semantic features from input documents, producing text vectors for Nanjing Yunjin, supporting user queries. Comparing with m3e-base, bge-large, and bert-large models, the experiments show that the use of the roberta-large model for vectorization, with a 99.82% similarity in the FAISS retrieval stage, outperforms other vectorization models.

(2) Construction of the FAISS database, providing an efficient way to store and retrieve a vast amount of vector data. An indexFlatL2 index is created based on Euclidean distance, and cosine similarity is used to calculate texts similar to user inputs. Sorting by cosine values from high to low, with values closer to 1 indicating greater similarity, recalls the top k most similar results for deep matching of Nanjing Yunjin queries and answers.

(3) Inputting relevant retrieval results and user queries into the large LLAMA model for enhanced generation, aiming for more accurate textual outcomes. The combination of text embedding and generative models realizes the organic integration of semantic match retrieval and natural language generation, enhancing the interpretability and credibility of the Q&A system.

(4) Development and implementation of a Q&A system based on KGs and RAG technology for effective utilization and protection of Nanjing Yunjin, an ICH. The system provides intelligent Q&A services for Nanjing Yunjin using KGs and RAG technology.

Despite significant progress in integrating neural IR with LLM for the Nanjing Yunjin Q&A system, several areas require further exploration and improvement:

(1) Optimization of data processing and knowledge sources: The current system overly relies on domain experts for text knowledge extraction, necessitating enhanced automated quality control. Additionally, the knowledge sources are relatively singular; future integration of multimodal data sources (like images and videos) will enrich the comprehensiveness and diversity of the Q&A system.

(2) Improvements in LLM and timeliness: LLM may generate content inconsistent with user input or existing world knowledge, calling for the incorporation of human feedback-driven reinforcement learning algorithms in future research. Continual updating of system documentation is also essential to maintain the timeliness and accuracy of the Q&A system.

(3) Advancements in data security and technical evaluation methods: Given data security concerns in large language model training, local data processing is recommended to ensure user privacy and data security. Moreover, the system's IR technology evaluation methods require further refinement, especially for the automatic parsing and assessment of answers generated by large models.

In summary, this research represents a significant step forward in the development of intelligent Q&A systems for Nanjing Yunjin, with implications not only for the heritage's preservation and protection but also as a key technological support for other ICHs. Future research will focus on addressing existing challenges to enhance the system's efficacy and application scope.

Abbreviations

ICH	Intangible cultural heritage
KGs	Knowledge graphs
Q&A system	Question-Answering System
RAG	Retrieval augmented generation
FAISS	Facebook AI Similarity Search
LLAMA	Large Language Model Meta AI
IR	Information retrieval
AI	Artificial Intelligence
NNS	Nearest neighbor search
K-NNS	K-nearest neighbor search
ANNS	Approximate nearest neighbor search
LLM	Large language models
RNNs	Recurrent neural networks
NLP	Natural language processing
LSTM	Long short-term memory
BERT	Bidirectional encoder representations from transformers
TF-IDF	Term frequency-inverse document frequency
MVC	Model view controller

Acknowledgements

This research was supported by experimental data created in the Nanjing Yun-Jing Institute. We also thank the ICH bearers of Nanjing YunJing for semantic annotation and data verification for this study. At the same time, we would like to express our deepest gratitude to the anonymous reviewers, whose comments improved this work tremendously.

Author contributions

Conceptualization: LX; Investigation: LX, LL, LZW; Methodology: LX; Software: LX, LML; Data preparation: LL, LX, CXS; Writing—original draft preparation: LX; Writing—review and editing: LX, LL. All authors read and approved the final manuscript.

Funding

This research was supported by the Jiangsu Provincial Social Science Foundation Project "Research on Knowledge Extraction and Organization of Nanjing Cloud Brocade Video Resources Based on Knowledge Meta" (Project No. 21TQC006). Jiangsu Provincial University Social Science Foundation Project "Research on the Implementation Path of the Living Heritage of Cloud Brocade Values in the DH Perspective" (Project No. 2021SJA0135).

Availability of data and materials

Data available on request from the authors.

Declarations

Competing interests

The authors claim there is no conflict of interest.

Author details

[1]Hangzhou Dianzi University, Lib, Hangzhou 310018, China. [2]Nanjing Forestry University, Nanjing 210037, China. [3]Unicom (Zhejiang) Industrial Internet Co. Ltd., Hangzhou 311103, China. [4]Nanjing University of Information Science and Technology, Nanjing 210044, China.

Received: 19 January 2024 Accepted: 1 April 2024

Published online: 09 April 2024

References

1. Sirsat SR, Chavan DV, Deshpande DSP. Mining knowledge from text repositories using information extraction: a review. Sadhana. 2014;39:53–62. https://doi.org/10.1007/s12046-013-0197-2.
2. Lu L, Liang X, Yuan G, Jing L, Wei C, Cheng C. A study on the construction of knowledge graph of Yunjin video resources under productive conservation. Herit Sci. 2023;11:83. https://doi.org/10.1186/s40494-023-00932-5.
3. Lu L, Li MT. Development of a virtual interactive system for Dahua Lou Loom based on knowledge ontology-driven technology. Herit Sci. 2023. https://doi.org/10.1186/s40494-023-01027-x.
4. Xu L, Lu L, Liu M. Construction and application of a knowledge graph-based question answering system for Nanjing Yunjin digital resources. Herit Sci. 2023;11:222. https://doi.org/10.1186/s40494-023-01068-2.
5. Zhang Y, Li Y, Wei X, Yang Y, Liu L, Murphey YL. Graph matching for knowledge graph alignment using edge-coloring propagation. Pattern Recogn. 2023;144: 109851. https://doi.org/10.1016/j.patcog.2023.109851.
6. Li D, Yan L, Zhang X, Jia W, Ma Z. EventKGE: event knowledge graph embedding with event causal transfer. Knowl Based Syst. 2023;278: 110917. https://doi.org/10.1016/j.knosys.2023.110917.
7. Wu T, Khan A, Yong M, Qi G, Wang M. Efficiently embedding dynamic knowledge graphs. Knowl Based Syst. 2022;250: 109124. https://doi.org/10.1016/j.knosys.2022.109124.
8. Seo S, Oh B, Jo E, Lee S, Lee D, Lee K-H, et al. Active learning for knowledge graph schema expansion. IEEE Trans Knowl Data Eng. 2022;34:5610–20. https://doi.org/10.1109/TKDE.2021.3070317.
9. Liu W, Cai H, Cheng X, Xie S, Yu Y, Dukehyzhang. Learning high-order structural and attribute information by knowledge graph attention

networks for enhancing knowledge graph embedding. Knowl Based Syst. 2022;250:109002. https://doi.org/10.1016/j.knosys.2022.109002.

10. Do P, Pham P. W-KG2Vec: a weighted text-enhanced meta-path-based knowledge graph embedding for similarity search. Neural Comput & Applic. 2021;33:16533–55. https://doi.org/10.1007/s00521-021-06252-8.
11. Zhang Q, Chen S, Fang M, Chen X. Joint reasoning with knowledge subgraphs for Multiple Choice Question Answering. Inf Process Manag. 2023;60: 103297. https://doi.org/10.1016/j.ipm.2023.103297.
12. Pavlick E. Symbols and grounding in large language models. Phil Trans R Soc A. 2023;381:20220041. https://doi.org/10.1098/rsta.2022.0041.
13. Sarrouti M, Ouatik El Alaoui S. SemBioNLQA: a semantic biomedical question answering system for retrieving exact and ideal answers to natural language questions. Artif Intell Med. 2020;102: 101767. https://doi.org/10.1016/j.artmed.2019.101767.
14. Cao Y, Cimino JJ, Ely J, Yu H. Automatically extracting information needs from complex clinical questions. J Biomed Inform. 2010;43:962–71. https://doi.org/10.1016/j.jbi.2010.07.007.
15. Tian D, Li M, Ren Q, Zhang X, Han S, Shen Y. Intelligent question answering method for construction safety hazard knowledge based on deep semantic mining. Autom Constr. 2023;145: 104670. https://doi.org/10.1016/j.autcon.2022.104670.
16. Ahmed S, Ahmad M, Swami BL, Ikram S. A review on plants extract mediated synthesis of silver nanoparticles for antimicrobial applications: a green expertise. J Adv Res. 2016;7:17–28. https://doi.org/10.1016/j.jare.2015.02.007.
17. Shieber S. The turing test: verbal behavior as the hallmark of intelligence. Cambridge: MIT Press; 2004.
18. Duggan GB. Applying psychology to understand relationships with technology: from ELIZA to interactive healthcare. Behav Inf Technol. 2016;35:536–47. https://doi.org/10.1080/0144929X.2016.1141320.
19. Woods WA. Progress in natural language understanding: an application to lunar geology. Proceedings of the National Computer Conference and Exposition. New York: ACM. 1973;441–450. https://doi.org/10.1145/1499586.1499695.
20. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI Chatbot for medicine. N Engl J Med. 2023;388:1233–9. https://doi.org/10.1056/NEJMsr2214184.
21. Wang X, Khoo ET, Nakatsu R, Cheok A. Interacting with traditional Chinese culture through natural language. ACM J Comput Cult Herit. 2014;7:18. https://doi.org/10.1145/2597183.
22. Zhao H. The database construction of intangible cultural heritage based on artificial intelligence. Math Probl Eng. 2022;2022:8576002. https://doi.org/10.1155/2022/8576002.
23. Sperli G. A cultural heritage framework using a Deep Learning based Chatbot for supporting tourist journey. Expert Syst Appl. 2021;183: 115277. https://doi.org/10.1016/j.eswa.2021.115277.
24. Yang Z, Wang Y, Gan J, Li H, Lei N. Design and Research of Intelligent Question-Answering (Q&A) system based on high school course knowledge graph. Mobile Netw Appl. 2021;26:1884–90. https://doi.org/10.1007/s11036-020-01726-w.
25. Liu S, Tan N, Yang H, Lukač N. An Intelligent Question Answering System of the Liao Dynasty based on knowledge graph. Int J Comput Intell Syst. 2021;14:170. https://doi.org/10.1007/s44196-021-00010-3.
26. Liu C, Ji X, Dong Y, He M, Yang M, Wang Y. Chinese mineral question and answering system based on knowledge graph. Expert Syst Appl. 2023;231: 120841. https://doi.org/10.1016/j.eswa.2023.120841.
27. Aurpa TT, Rifat RK, Ahmed MS, Anwar MM, Ali ABMS. Reading comprehension based question answering system in Bangla language with transformer-based learning. Heliyon. 2022;8: e11052. https://doi.org/10.1016/j.heliyon.2022.e11052.
28. Suissa O, Zhitomirsky-Geffet M, Elmalech A. Question answering with deep neural networks for semi-structured heterogeneous genealogical knowledge graphs. SW. 2022;14:209–37. https://doi.org/10.3233/SW-222925.
29. Bhattacharya K, Bhattacharya AS, Bhattacharya N, et al. ChatGPT in surgical practice—a new kid on the block. Indian J Surg. 2023. https://doi.org/10.1007/s12262-023-03727-x.
30. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. NIPS'20. 2020;1877–1901. https://doi.org/10.5555/3495724.3495883.
31. Tzachor A, Devare M, Richards C, Pypers P, Ghosh A, Koo J, et al. Large language models and agricultural extension services. Nat Food. 2023;4:941–8. https://doi.org/10.1038/s43016-023-00867-x.
32. Sorin V, Barash Y, Konen E, Klang E. Large language models for oncological applications. J Cancer Res Clin Oncol. 2023;149:9505–8. https://doi.org/10.1007/s00432-023-04824-w.
33. Eggmann F, Weiger R, Zitzmann NU, Blatz MB. Implications of large language models such as ChatGPT for dental medicine. J Esthet Restor Dent. 2023;35:1098–102. https://doi.org/10.1111/jerd.13046.
34. Luu RK, Buehler MJ. BioinspiredLLM: conversational large language model for the mechanics of biological and bio-inspired materials. Adv Sci. 2023. https://doi.org/10.1002/advs.202306724.
35. Rodríguez P, Bautista MA, Gonzàlez J, Escalera S. Beyond one-hot encoding: lower dimensional target embedding. Image Vis Comput. 2018;75:21–31. https://doi.org/10.1016/j.imavis.2018.04.004.
36. Yan D, Li K, Gu S, Yang L. Network-based bag-of-words model for text classification. IEEE Access. 2020;8:82641–52. https://doi.org/10.1109/ACCESS.2020.2991074.
37. Kikuchi M, Kawakami K, Watanabe K, Yoshida M, Umemura K. Unified likelihood ratio estimation for high- to zero-frequency N-Grams. IEICE Trans Fundam. 2021;E104.A:1059–74. https://doi.org/10.1587/transfun.2020EAP1088.
38. Zhang W, Yoshida T, Tang X. A comparative study of TF*IDF, LSI and multiwords for text classification. Expert Syst Appl. 2011;38:2758–65. https://doi.org/10.1016/j.eswa.2010.08.066.
39. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. NIPS'13.2013;3111–3119. https://doi.org/10.5555/2999792.2999959.
40. Pennington J, Socher R, Manning C. Glove:Global vectors for word representation.Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) .2014;1532–1543.https://aclanthology.org/D14-1162.
41. Huang L, Song Y. Intangible cultural heritage management using machine learning model: a case study of Northwest Folk Song Huaer. Sci Program. 2022. https://doi.org/10.1155/2022/1383520.
42. Chen Q, Zhao W, Wang Q, Zhao Y. The sustainable development of intangible cultural heritage with AI: Cantonese opera singing genre classification based on CoGCNet Model in China. Sustainability. 2022;14:2923. https://doi.org/10.3390/su14052923.
43. Ashmawy M, Fakhr MW, Maghraby FA. Lexical normalization using generative transformer model (LN-GTM). Int J Comput Intell Syst. 2023;16:183. https://doi.org/10.1007/s44196-023-00366-8.
44. Hambarde KA, Proença H. Information retrieval: recent advances and beyond. IEEE Access. 2023;11:76581–604. https://doi.org/10.1109/ACCESS.2023.3295776.
45. Xie Z, Zeng Z, Zhou G, Wang W. Topic enhanced deep structured semantic models for knowledge base question answering. Sci China Inf Sci. 2017;60: 110103. https://doi.org/10.1007/s11432-017-9136-x.
46. Huang P, He X, Gao J, Deng L, Acero A, Heck L. Learning deep structured semantic models for web search using clickthrough data. Proceedings of the 22nd ACM international conference on Information & Knowledge Management. New York, NY, USA: Association for Computing Machinery; 2013;2333–2338. https://doi.org/10.1145/2505515.2505665.
47. Palangi H, Deng L, Shen Y, Gao J, He X, Chen J, et al. Deep sentence embedding using long short-term memory networks: analysis and application to information retrieval. IEEE/ACM Trans Audio Speech Lang Process. 2016;24:694–707. https://doi.org/10.1109/TASLP.2016.2520371.
48. Iscen A, Furon T, Gripon V, Rabbat M, Jégou H. Memory vectors for similarity search in high-dimensional spaces. IEEE Trans Big Data. 2018;4:65–77. https://doi.org/10.1109/TBDATA.2017.2677964.
49. Hong W, Tang X, Meng J, Yuan J. Asymmetric mapping quantization for nearest neighbor search. IEEE Trans Pattern Anal Mach Intell. 2020;42:1783–90. https://doi.org/10.1109/TPAMI.2019.2925347.
50. Yuan GT, Lu L, Zhou XF. Feature selection using a sinusoidal sequence combined with mutual information. Eng Appl Artif Intell. 2023;126:107168. https://doi.org/10.1016/j.engappai.2023.107168.
51. Ozan EC, Kiranyaz S, Gabbouj M. K-subspaces quantization for approximate nearest neighbor search. IEEE Trans Knowl Data Eng. 2016;28:1722–33. https://doi.org/10.1109/TKDE.2016.2535287.

52. Miao X, Gao Y, Chen G, Zheng B, Cui H. Processing incomplete k nearest neighbor search. IEEE Trans Fuzzy Syst. 2016;24:1349–63. https://doi.org/10.1109/TFUZZ.2016.2516562.
53. Ferhatosmanoglu H, Tuncel E, Agrawal D, Abbadi AE. High dimensional nearest neighbor searching. Inf Syst. 2006;31:512–40. https://doi.org/10.1016/j.is.2005.01.001.
54. Johnson J, Douze M, Jegou H. Billion-scale similarity search with GPUs. IEEE Trans Big Data. 2021;7:535–47. https://doi.org/10.1109/TBDATA.2019.2921572.
55. Guo Y, Qiu W, Leroy G, Wang S, Cohen T. Retrieval augmentation of large language models for lay language generation. J Biom Inform. 2024;149: 104580. https://doi.org/10.1016/j.jbi.2023.104580.
56. Siriwardhana S, Weerasekera R, Wen E, Kaluarachchi T, Rana R, Nanayakkara S. Improving the domain adaptation of retrieval augmented generation (RAG) models for open domain question answering. Trans Assoc Comput Linguistics. 2023;11:1–17. https://doi.org/10.1162/tacl_a_00530.
57. Xiao H, Li L, Liu Q, Zhu X, Zhang Q. Transformers in medical image segmentation: a review. Biomed Signal Process Control. 2023;84: 104791. https://doi.org/10.1016/j.bspc.2023.104791.
58. Cardarilli GC, Di Nunzio L, Fazzolari R, Nannarelli A, Re M, Spanò S. N-dimensional approximation of Euclidean distance. IEEE Trans Circ Syst II Express Briefs. 2020;67:565–9. https://doi.org/10.1109/TCSII.2019.2919545.
59. Zhu S, Wu J, Xiong H, Xia G. Scaling up top-K cosine similarity search. Data Knowl Eng. 2011;70:60–83. https://doi.org/10.1016/j.datak.2010.08.004.
60. Sauter P, Vögler G, Specht G, Flor T. A model–view–controller extension for pervasive multi-client user interfaces. Pers Ubiquit Comput. 2005;9:100–7. https://doi.org/10.1007/s00779-004-0314-7.

Publisher's Note