https://doi.org/10.1038/s40494-026-02384-z

Visual information identification and Q&A of intangible cultural heritage inheritors by using enhanced Graph-Retrieval framework

Check for updates

Runzhou Wang[1] ✉, Xinsheng Zhang[2], Qilei Liu[1], Jinqi Su[1], Yuezhou Zhang[3] & Yulong Ma[2]

Visual business cards provide a digital medium for presenting information on intangible cultural heritage (ICH) inheritors. Accurately recognizing these cards allows the extraction of key details that, when combined with large language models, support semantic understanding and contextual reasoning to generate enriched cultural content. This study introduces a Graph-Retrieval framework that integrates graph-based methods with retrieval-augmented generation for visual information recognition and ICH-related question answering. A dataset of Chinese ICH inheritors' visual cards was built, covering 10 information types. To enhance extraction robustness, graph feature enhancement is applied through semantic recognition, random node masking, and edge deletion, while positional attention captures spatial relationships. A complementary Loop-RAG strategy dynamically integrates external knowledge with inner-outer loop retrieval. Experiments show the Graph-Retrieval framework surpasses benchmarks on multiple datasets, achieving a macro-average F1 of 0.928, with ablation studies validating feature enhancements. Loop-RAG also excels in report generation, question answering and few-shot conditions.

Intangible Cultural Heritage (ICH) includes social practices, performing arts, festivals, traditional knowledge and skills, as well as related tools, objects, handicrafts, and cultural spaces. It is a living form of culture that is continuously preserved and renewed through intergenerational transmission within communities[1,2]. As a carrier of historical memory and cultural identity, ICH reflects the cultural accumulation of a country or region and plays an important role in cultural continuity and dissemination[3]. ICH inheritors are the primary holders of relevant knowledge and skills. The systematic recording and dissemination of their information are fundamental to both ICH protection and public understanding[4]. With the increasing application of digital technologies in the cultural sector, ICH is gradually moving toward digitalization and platform-based dissemination. In this process, the digital storage and recognition of information related to ICH inheritors have become key components of ICH digitization efforts[5].

Among different digital carriers, text remains the most basic form of information representation. Visual business cards further complement textual information through layout and visual cues, enabling more intuitive presentation[6]. Visual business cards of ICH inheritors typically contain personal information, project names, and regional attributes, which facilitate information access and public awareness when shared on online platforms and social media[7]. However, accurately extracting such information remains challenging due to heterogeneous data sources and highly diverse layout styles.

Visual Information Extraction (VIE) offers a technical solution for converting document images into structured information. Existing studies have demonstrated that jointly modeling textual content, spatial layout, and visual features enables effective recognition of key document fields[8]. As an extension of traditional text information extraction, VIE explicitly considers the joint semantic structure formed by text and layout in document images[9,10], and has been widely applied to structured document scenarios such as invoices and web pages[11–13]. Introducing VIE into ICH inheritor visual business cards can therefore support the automatic extraction of core information, including names, project titles, and regions. At the same time, the rapid development of Large Language Models (LLMs) has provided new opportunities for understanding and querying ICH-related information. Models such as ChatGPT, Llama, and ChatGLM have demonstrated strong capabilities in semantic understanding and contextual reasoning across a wide range of language processing tasks[14,15]. In the context of ICH dissemination, LLMs can interpret domain-specific content and enable on-demand information access through question-answering interactions, thereby enhancing user engagement and experience[16,17].

[1]School of Economics and Management, Xi'an University of Posts and Telecommunications, Xi'an, China. [2]School of Management, Xi'an University of Architecture and Technology, Xi'an, China. [3]Institute of Flexible Electronics, Northwestern Polytechnical University, Xi'an, China. ✉e-mail: wrz@xupt.edu.cn

From a methodological perspective, VIE research has evolved from rule-based approaches to grid-based methods, deep learning models, and graph-based representations. Early methods relied primarily on handcrafted rules and template matching to parse document structures[18–20]. While effective for fixed-format documents, these approaches were highly sensitive to layout variations. Grid-based methods, such as Chargrid, BERTgrid, and their extensions, improved document structure modeling through two-dimensional representations[21–24], but still suffered performance degradation in complex layouts or low-quality images. End-to-end deep learning methods further enhanced feature representation and robustness under sparse or noisy conditions[25–29]. For example, the BROS model encodes two-dimensional positional information to better handle unordered documents[30], and subsequent studies introduced retrieval mechanisms and layout-aware architectures to improve semantic modeling[31,32]. However, these methods mainly focus on local feature learning and have limited capacity to capture global structural relationships.

To address this limitation, graph-based approaches have been increasingly adopted in VIE tasks. These methods explicitly model text units as nodes and represent spatial or semantic relationships through edges[33]. Prior studies indicate that graph-based models offer advantages in handling complex layouts and cross-style document scenarios[34–37]. Nevertheless, most existing approaches rely on static graph structures, which struggle to adapt to layout variations across platforms and usage contexts. This rigidity also restricts the model's ability to dynamically capture contextual relationships.

These limitations highlight the need for more flexible VIE frameworks that can incorporate contextual reasoning, handle incomplete information, and adapt to diverse document styles. Large language models provide promising solutions in this regard. Their strong semantic understanding, cross-domain generalization, and generative abilities make them particularly suitable for tasks involving unstructured or partially structured information[38]. Recent studies have explored the integration of LLMs into cultural heritage applications. Dai et al. combined diffusion models with LoRA to improve the recognition of paper-cutting art images[39]. Liu et al. applied LLMs to ancient book processing using prompt engineering and retrieval-augmented generation, enhancing semantic analysis[40]. Zhang et al. proposed ArchGPT as an intelligent agent for traditional building preservation, achieving high task satisfaction and response rates[41]. Xu et al. integrated knowledge graphs with retrieval-enhanced generation to support ICH question answering, improving reasoning coherence and interpretability[42]. Additional research has demonstrated the effectiveness of LLMs in cultural topic classification[43] and in domain-specific question answering when supported by retrieval and prompting strategies[44]. Despite these advances, significant challenges remain. Bouchra et al. reported that ChatGPT-based ontology construction for world heritage may fail when contextual cues are implicit, highlighting the necessity of external knowledge integration to improve reliability[45]. Ayash et al. showed that LLM performance declines markedly when addressing region-specific or highly specialized cultural questions, even for advanced models such as GPT-4 and DeepSeek[46]. These findings expose persistent issues, including hallucinations, limited understanding of domain-specific terminology, and the high computational cost of fine-tuning. Moreover, existing retrieval-augmented generation frameworks are largely static and lack mechanisms for dynamically optimizing knowledge selection during multi-round interactions. Therefore, how to effectively structure, update, and integrate external ICH knowledge into LLMs remains an open research question and a critical prerequisite for achieving accurate, reliable, and efficient ICH information extraction and question answering.

Although visual information extraction and large language models have advanced rapidly in recent years, significant challenges remain in applying these techniques to the recognition of ICH inheritors' visual business cards.

First, ICH visual documents suffer from strong heterogeneity and poor cross-scenario generalization. Most dissemination platforms lack standardized and large-scale datasets of inheritors' visual business cards, and existing datasets are limited in size and annotation consistency. Many platforms provide only basic fields, such as names and project titles, while omitting key information, including skill characteristics, representative works, and inheritance lineage, leading to insufficient semantic coverage. Moreover, existing methods often rely on fixed graph structures to model spatial relationships, which struggle to generalize across the highly diverse layouts of ICH visual business cards from different regions and platforms.

Second, the professional and context-dependent nature of ICH knowledge poses challenges for semantic modeling. ICH terminology is highly region-specific and context-sensitive, requiring fine-grained contextual understanding. For example, "horse riding (走马)" in Shaanxi shadow puppetry denotes a specific performance technique rather than a general concept. Without incorporating cultural and historical context, models may misinterpret such terms, causing semantic loss. Similar issues arise for many skill names when detached from their original contexts, reducing interpretability and practical value.

Third, large language models face limitations in generation reliability and domain adaptation. In the ICH domain, LLMs are prone to hallucinations when handling unfamiliar terminology or complex cultural backgrounds, which undermines information credibility. In addition, adapting LLMs to specialized ICH knowledge is costly due to their large parameter sizes and the need for high-quality domain corpora and substantial computational resources, limiting their practical deployment.

To address these challenges, this study proposes an enhanced Graph-Retrieval framework for information recognition and question answering on ICH inheritors' visual business cards. In the recognition stage, a graph structure enhancement method is introduced to improve cross-domain adaptability to heterogeneous visual business cards. In the question-answering stage, a Loop-RAG strategy is employed to mitigate hallucinations in large language models, enabling more accurate reasoning and semantic completion of ICH knowledge. The main contributions of this work are summarized as follows.

(1) A dataset of 5237 visual business cards of Chinese ICH inheritors is constructed. Each card contains 10 categories of information, including project ID, project name, inheritor name, location, and project details, etc.
(2) A graph feature enhancement method is proposed for accurate ICH visual information extraction. The method integrates semantic feature extraction, random node masking, random edge deletion, and a positional attention mechanism. It achieves an F1 score of 0.928 on the proposed dataset and consistently exceeds 0.9 in F1 across five public benchmark datasets, demonstrating strong cross-scenario generalization.
(3) Construct an improved retrieval-augmented generation strategy (Loop-RAG). By combining external ICH knowledge, contextual prompts, and prompt engineering, and adding inner-outer loops to optimize the retrieval process, it effectively reduces the model's hallucination risk and significantly improves the accuracy and integrity of generation results. The BLEU, METEOR, and ROUGE-L values for the report generation task are 21.50, 22.10, and 23.80, respectively. The F1 value in the question-answering task can reach 0.941, and it still has a relatively excellent performance in the few-shot situation.

Methods

This study addresses the research challenges existing in the recognition and question answering of visual business cards of inheritors in the context of the vertical field of ICH, including insufficient visual business cards, inadequate representation of context semantics, and the problem of question answering illusion in LLMs. It proposes an ICH information recognition and question answering framework based on enhanced Graph-Retrieval. This framework not only achieves precise recognition of multi-category information in visual business cards but also enhances the credibility of text generation from LLMs in ICH scenarios through a Loop-RAG strategy. It is mainly divided into data acquisition and processing, graph feature enhancement, output module and Loop-RAG enhancement module. The specific process is shown in Fig. 1.

Firstly, to address the limited data size, lack of unified annotation standards, and significant layout variations in ICH visual business cards

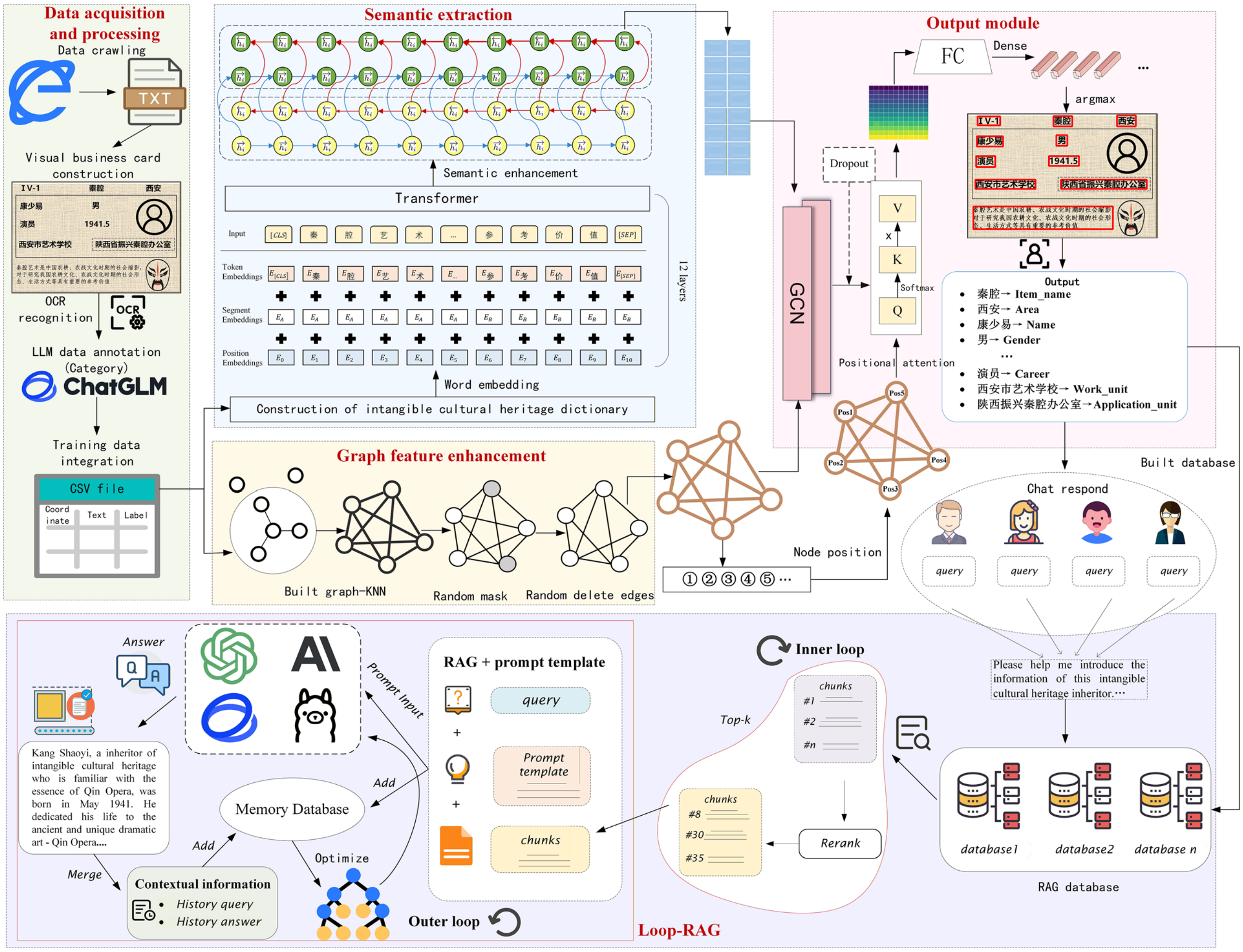


Fig. 1 | Graph-Retrieval model research framework.

across different regions and platforms, this study constructed a large-scale dataset of ICH inheritors' visual business cards in China. This dataset covers ten core elements, including project number, project name, inheritor name, skill characteristics, and work unit. Labels are standardized using a large language model to assist with annotation and correction.

Secondly, recognizing that traditional fixed graph structures struggle with heterogeneous document formats and layouts, this study introduces a graph-feature-enhanced modeling approach. Specifically, a semantic embedding method combining RoBERTa and BiLSTM captures text content and contextual information, while graph-structured modeling represents spatial dependencies among elements. Random node masking and edge deletion strategies improve robustness, and a positional attention mechanism explicitly encodes element locations. This approach enhances adaptability to diverse typographic structures, enabling robust recognition across scenarios.

Furthermore, to address the numerous specialized terms and context-sensitive expressions found in ICH information, this study integrated semantic embeddings, contextual features, and graph structure features into a multi-source fusion, training the model using a two-layer graph convolutional neural network (GNN). Through regularization using dropout and fully connected layers, the model accurately predicts the multi-category information in visual business cards and ensures that the extracted results remain consistent with the ICH knowledge system, thus avoiding semantic fragmentation caused by a lack of context.

Finally, to address the hallucination risks and high domain adaptation costs of large language models in ICH contexts, this study proposes an improved RAG strategy, Loop-RAG. Unlike traditional RAG methods, which rely solely on external knowledge retrieval and prompt design, Loop-RAG introduces a dynamic reasoning mechanism based on context engineering. Using a collaborative inner-outer loop structure enhances knowledge utilization efficiency and generation controllability in the ICH domain. In this mechanism, the inner loop handles immediate retrieval and decision-making. The inner loop is responsible for immediate retrieval and decision-making. That is, during the generation process, the model dynamically judges whether the current information is sufficient based on the input context and triggers targeted knowledge retrieval and reconstruction prompts. The outer loop enables long-term optimization through cross-session context accumulation. The model continuously records and analyzes retrieval paths, prompt combinations, and generation outcomes over multiple rounds, gradually learning more effective information selection strategies. By integrating inner-outer loop context engineering, Loop-RAG not only reduces hallucinations in individual tasks but also improves cross-scenario robustness.

Data acquisition and processing of ICH inheritors

The data for this study were primarily collected from the China intangible cultural heritage platform (https://www.ihchina.cn/representative). We have designed a three-stage annotation process: OCR extraction, LLMs automated annotation, and manual quality control. First, use PaddleOCR to identify the text content in the image and precisely extract the two-dimensional coordinates of this content. The OCR accuracy was evaluated on the training set, and the results showed that the character accuracy rate reached 95.2%. Subsequently, the LLMs were used to set a prompt to classify the text information extracted by OCR, totaling 10 types of tags, namely project, project name, region, inheritor's name, gender, occupation, date of birth, work unit, applying unit, and project description. Subsequently, a team composed of five experienced annotators manually reviewed all the LLMs' annotation results to identify and correct the erroneous results generated by the LLMS. To quantify the quality of annotations, we selected 30% of the samples for cross-review. The calculated Kappa coefficient between annotators was 0.87, which proved the high consistency and reliability of the dataset. The results of data category annotation are shown in Table 1.

Based on the collected data, a total of 5237 visual business cards of ICH inheritors in China were successfully constructed. At the same time, considering the image distortion caused by wear and tear in practical use, 20% of the images were subjected to wear and tear processing in the data. To recognize the information in visual business cards, PaddleOCR was employed to recognize the text content and obtain the spatial coordinates of each element. Finally, the recognized text was highlighted with a red border according to its coordinates. The resulting images, along with the label annotations, are presented in Fig. 2, which includes the vertex coordinates of OCR-recognized text, node numbers (0–9), business card text content, and annotated labels.

Semantic feature extraction

Through semantic extraction methods, feature representations of text content in visual business cards can be obtained, such as text embeddings of inheritors' names, inherited project names, etc., which can be used as input features for training the model. To extract text embeddings more accurately, this section constructs a lexicon of ICH inheritors and semantic extraction methods.

ICH usually contains rich professional terminology and specific cultural background knowledge. Some general corpora of pretrained models may not fully capture the deep meanings of ICH terms. Building a proprietary lexicon of ICH information can better adapt to the specific context of ICH texts, prevent semantic ambiguity, and improve the accuracy and relevance of word embeddings. Therefore, based on the collected information of ICH inheritors, Jieba word segmentation is used to segment all texts, covering proprietary terms such as ICH names, regions, and skill introductions, and avoiding segmentation errors through manual review. Subsequently, remove duplicate content and eliminate irrelevant information in the text, such as punctuation marks, numbers, and non-critical particles. Ultimately, by numbering and sorting all the words, a dedicated word bank containing 4875 unique terms in the field of ICH is constructed. Some of the results are shown in Table 2.

To more accurately identify information on the visual business cards of ICH inheritors, this study proposes an enhanced semantic extraction method that addresses the unique complexity and specialized vocabulary of the ICH domain. While pre-trained models such as RoBERTa perform well on general corpora, their generalization is limited when handling ICH texts containing domain-specific terms (e.g., inheritor names, project titles, locations), which can lead to semantic embedding biases. To overcome this, a specialized ICH vocabulary is constructed and integrated with RoBERTa's original vocabulary, enabling the model to fully capture domain-specific terminology and generate more representative word embeddings. Based on this fused vocabulary, the word embedding layer is further updated via backpropagation using the encoder of a 12-layer Transformer model, improving adaptation to the linguistic characteristics of ICH texts. Finally, a BiLSTM network is employed to capture sequential relationships in long text sequences, ensuring that subtle semantic differences in ICH information are effectively learned. The overall process is illustrated in Fig. 3.

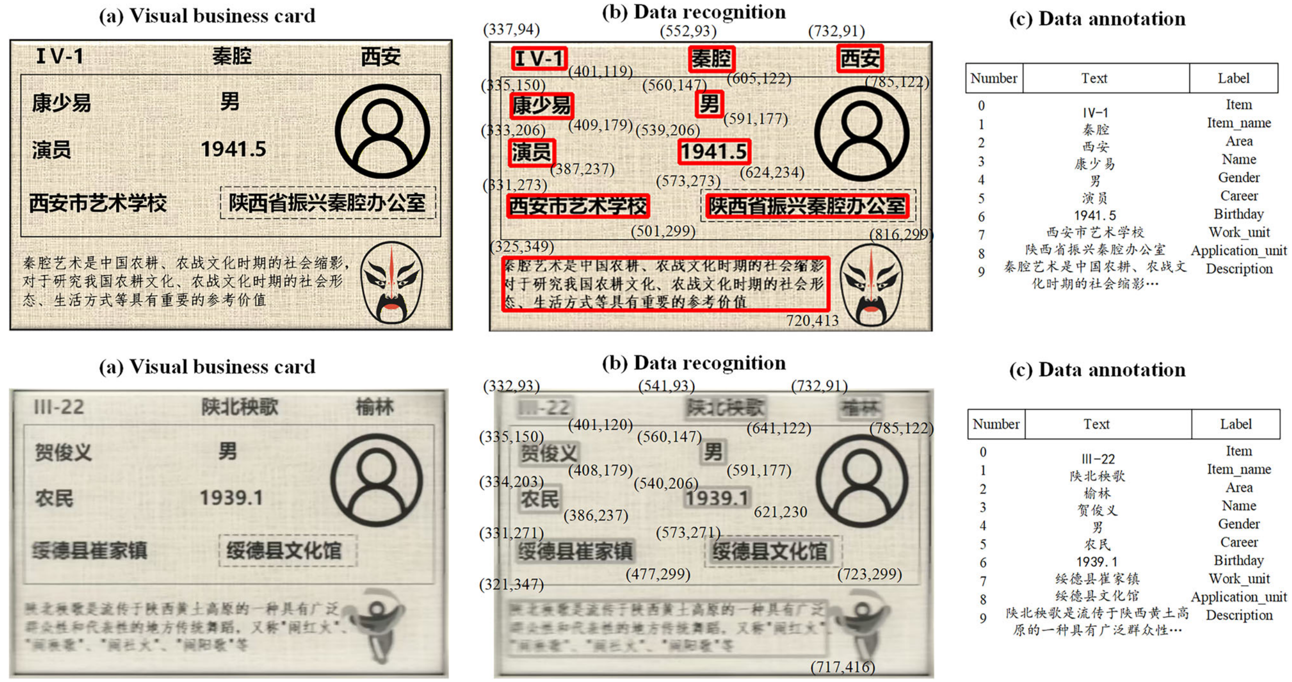
By re-standardizing the word numbers in the fusion vocabulary, corresponding searches are performed in the pre-trained word embeddings, and backpropagation is performed in the constructed model to continuously update the word embeddings. The specific process is shown in formula (1):

$$v_i = R(L(index(w_i), V)) \qquad (1)$$

Where v_i is the word embedding vector of the word w_i; index(w_i) is the index number of the word w_i in the summary lexicon; V is the constructed word vector matrix, with the horizontal dimension being the number of

Table 1 | Data annotation category

Describtion	Label	Description	Label
Project number	Item	Inheritor profession	Career
Project name	Item_name	Date of birth	Birthday
ICH region	Area	Inheritor's workplace	Work_unit
Inheritor name	Name	Inheritor application unit	Application_unit
Inheritor gender	Gender	Project description	Description



Number	Text	Label
0	IV-1	Item
1	秦腔	Item_name
2	西安	Area
3	康少易	Name
4	男	Gender
5	演员	Career
6	1941.5	Birthday
7	西安市艺术学校	Work_unit
8	陕西省振兴秦腔办公室	Application_unit
9	秦腔艺术是中国农耕、农战文化时期的社会缩影···	Description

Number	Text	Label
0	III-22	Item
1	陕北秧歌	Item_name
2	榆林	Area
3	贺俊义	Name
4	男	Gender
5	农民	Career
6	1939.1	Birthday
7	绥德县崔家镇	Work_unit
8	绥德县文化馆	Application_unit
9	陕北秧歌是流传于陕西黄土高原的一种具有广泛群众性···	Description

Fig. 2 | OCR identification results of visual business cards.

Table 2 | ICH inheritors lexicon

Term	Description	Category	Item	Region	Name	Lexicon
秦腔 (Qinqiang opera)	Traditional Chinese opera in Shaanxi Province is known for its high-pitched and passionate singing style and unique performance style	Traditional drama	Qinqiang opera	Shaanxi	Li Mei	秦腔/传统/戏曲/剧种/唱腔/表演 Qin opera opera/traditional opera/drama/singing/act
皮影戏 (Shadow play)	The traditional theatrical form of making character silhouettes from animal skin or cardboard and performing them on a screen through light projection	Traditional drama	Huaxian shadow play	Huaxian county	Wang Tianwen	皮影戏/兽皮/纸板/剪影/灯光/表演 Shadow play /animal hide/cardboard/silhouette/lighting/performance
剪纸(Paper-cut)	Traditional folk art of cutting and carving various patterns on paper with scissors or a carving knife	Traditional art	Shanbei paper-cut	North Shaanxi province	Gao Fenglian	剪纸/剪刀/刻刀/图案/民间/艺术 Paper cut/scissors/carving/pattern/folk/art
泥塑(Clay sculptures)	The traditional sculpture art of handcrafting various figures, animals, or objects using soil as raw material	Traditional art	Fengxiang clay sculptures	Fengxiang county	Hu Xinming	泥塑/泥土/手工/捏制/雕塑/艺术 Clay sculpture/clay/handicrafts/kneading/sculpture/art
社火 (Shehuo)	Collective dances and dramas performed in folk festivals, usually accompanied by elements such as gongs, drums, and stilts	Folk custom	Baoji shehuo	Baoji city	Wang Zhi	社火/民间/节庆/舞蹈/戏剧/锣鼓 Shehuo/folk/festivals/dance/drama/drums and drums
西府曲子(Xifu tune)	A traditional form of folk art popular in the Xifu area of Shaanxi Province, mainly performed through a combination of rap and singing	Tradition tune	Xifu tune	Baoji city	Li Aiqin	西府/曲子/曲艺/说唱/表演 Xifu/quzi/quyi/rap/performance

word lists and the vertical dimension being the embedding dimension of words; R is the RoBERTa pretrained model.

To fully capture long-range contextual dependencies in ICH texts, this study employs a BiLSTM network as the sequence modeling module. Traditional RNNs often suffer from gradient vanishing when processing long sequences and struggle to account for the influence of subsequent information on current words. In contrast, BiLSTM processes text sequences in both forward and backward directions, enabling a deeper understanding of contextual semantics in ICH visual business card information. For instance, an inheritor's skills may be closely related to the regional description that follows. By combining forward and backward LSTMs, entities such as names and projects can be efficiently modeled, allowing the model to fully capture contextual information and provide more accurate semantic representations for downstream entity recognition and information extraction. The operational process of the BiLSTM network is illustrated as follows:

$$h_{\mathrm{ft}} = LSTM(X_t, h_{t-1}) \tag{2}$$

$$h_{\mathrm{bt}} = LSTM(X_t, h'_{t-1}) \tag{3}$$

$$O_t = \mathrm{Concat}(\overrightarrow{\mathrm{h}_t}, \overleftarrow{\mathrm{h}'_t}) \tag{4}$$

Where h_{ft} represents forward LSTM operation, and h_{bt} represents backward LSTM operation. The result is obtained by connecting the output of h_{ft} and h_{bt}.

Random node masking

Visual business cards of ICH inheritors may contain incomplete information due to image wear, digitization errors, or other factors, posing challenges for accurate information recognition. To address this, the study introduces a random node masking strategy to enhance the model's robustness to noise. This method randomly masks certain node features in the graph relationship matrix, forcing the model to learn how to predict the features or categories of masked nodes using incomplete or noisy information. This approach not only reduces the risk of overfitting during training but also improves generalization in practical applications. The masking process involves randomly selecting a proportion of nodes from the graph relationship matrix in each iteration, hiding their features, and inputting the modified matrix into the GCN network for training, as detailed in formulas (5)–(7).

$$X_{mask} = X \times (1 - mask) \tag{5}$$

$$mask = P(x = k) = \binom{1}{k} \sigma^k (1 - \sigma)^{1-k} \tag{6}$$

$$H_{mask}^{(l+1)} = \sigma\left(\tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} X_{mask}{}^{(l)} W^{(l)}\right) \tag{7}$$

Where X_{mask} is the result after random node masking; X represents all node features of the input model; mask is the masking node, which follows a binomial distribution and has the same output dimension as the X dimension; k is 0 or 1, where 1 represents masking; σ is the mask probability.

Random edge deletion

Given the relatively fixed spatial layout of information categories in ICH inheritors' visual business cards, the graph adjacency matrix tends to be simple. To enable the model to learn more diverse graph representations, this study introduces a random edge deletion strategy. By randomly removing edges from the adjacency matrix during training, the structural uncertainty of the graph is increased, allowing the model to learn more varied graph patterns. This enhances the model's adaptability to different graph structures, improves its ability to handle previously unseen configurations, and strengthens overall robustness. Specifically, in each training

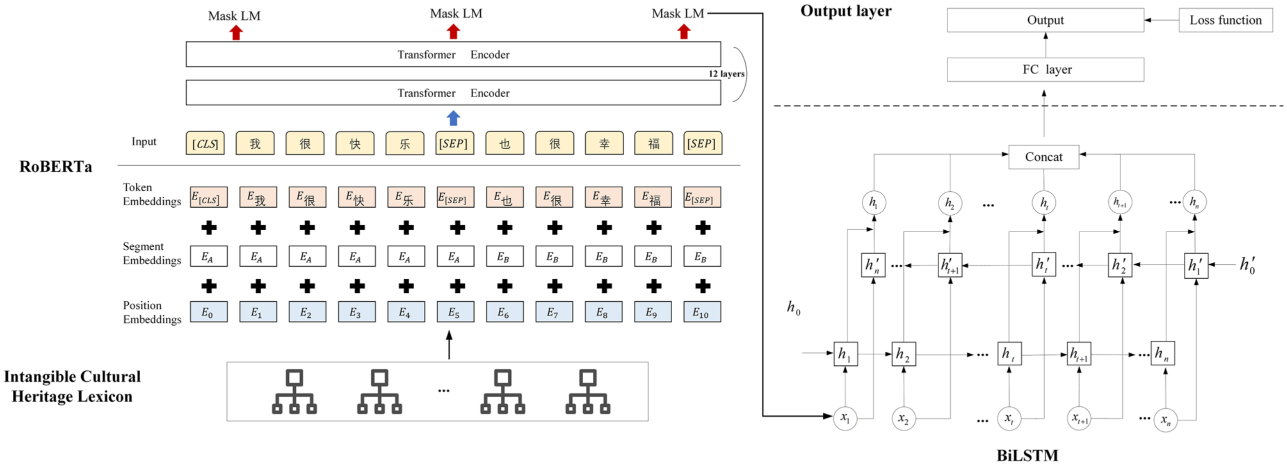

Mask LM
Mask LM
Mask LM
Transformer Encoder
12 layers
Transformer Encoder
Input
[CLS] 我 很 快 乐 [SEP] 也 很 幸 福 [SEP]
RoBERTa
Token Embeddings
E[CLS] E我 E很 E快 E乐 E[SEP] E也 E很 E幸 E福 E[SEP]
Segment Embeddings
EA EA EA EA EA EA EB EB EB EB EB
Position Embeddings
E0 E1 E2 E3 E4 E5 E6 E7 E8 E9 E10
Intangible Cultural Heritage Lexicon
Output layer
Output
Loss function
FC layer
Concat
h0
h'0
BiLSTM

Fig. 3 | Semantic extraction framework diagram.

iteration, a proportion of edges is randomly selected from the adjacency matrix and removed from the graph structure. The detailed calculation process is as follows:

Suppose there is a graph $G = (V, E)$, where V is the set of nodes and E is the set of edges $|E|$ is the integer number of edges in the graph.

Firstly, the parameter p is introduced to represent the random edge deletion probability, where $0 \le p \le 1$. The number of deleted edges is calculated and rounded down.

$$\text{num\_remove} = \lfloor p \cdot |E| \rfloor \tag{8}$$

Next, randomly select num_remove indices from all edge indices, which represent the set of indices for the edges that need to be deleted. Assuming that the randomly selected index set is $R = \{r_1, r_2, \cdots, r_{\text{num\_remove}}\}$, where r_i is the $i - th$ randomly selected index and $1 \le i \le \text{num\_remove}$.

Subsequently, for each edge $e = (u, v) \in E$, determine whether the edge needs to be deleted. If the index i of the edge is not in R, it means that the edge is not deleted, and a loop traversal is performed to construct the deleted edge set E'.

$$E' = \{e = (u, v) \in E | i \notin R\} \tag{9}$$

Finally, the set of deleted edges E' is transformed into an adjacency matrix $\tilde{D}_{\text{remove}}$ and input into the GCN network for learning.

$$H^{(l+1)}_{remove} = \sigma(\tilde{D}_{remove}^{-\frac{1}{2}} \tilde{A} \tilde{D}_{remove}^{-\frac{1}{2}} H^{(l)} W^{(l)}) \tag{10}$$

Positional attention mechanism

Different visual information data have different positional distributions. The positional attention mechanism incorporates the positional information of nodes into the self-attention framework, enhancing node feature representations. This allows the information extraction model to account for node positions and better learn the dependency relationships between different node orders. The positional encoding is computed using the steady-state distribution of random walks, which reflects the global importance of each node within the graph. By leveraging this distribution, long-term interactions between nodes and the overall graph structure can be captured, effectively encoding the global positional information of nodes.

Based on graph relationship matrix A, construct the transition matrix P of the graph. To control the global nature of the walk, add the restart probability α and define a random walk transition matrix with restart, I is the identity matrix:

$$P' = \alpha P + (1 - \alpha) I \tag{11}$$

By iteratively calculating the steady-state distribution π of the random walk, the position encoding pos can be obtained. Namely $\text{pos} = \pi \text{P}'$. Finally, by incorporating positional encoding into the attention mechanism, the model can ensure the learning of graph positional relationships.

$$\text{Attention}(Q, K, V) = \text{soft max}\left(\frac{((Q + Pos) \times (\text{K}^{\text{T}} + Pos))}{\sqrt{d_k}}\right) V \tag{12}$$

Where Q, K, V are matrix vectors obtained through linear transformation of input data, and d_k represents the dimension of K, which is used for numerical scaling; Pos represents the position embedding.

Graph convolutional neural network and graph relationship matrix construction

To effectively capture the network relationships among information in the visual business cards of ICH inheritors, this study employs graph convolutional neural networks (GCNs). Traditional approaches often treat this information as independent entities, ignoring inherent correlations and limiting recognition accuracy. In contrast, GCNs model ICH information as a graph, integrating entity nodes and their relationships to account for both semantic and relational features. Specifically, GCNs aggregate feature information from all first-order neighboring nodes using convolutional kernels to generate the central node's potential features, and employ pooling operations to remove redundant information, ultimately learning efficient node representations[47]. The core calculation formula of GCN is as follows:

$$H^{(l+1)} = \sigma(\tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} H^{(l)} W^{(l)}) \tag{13}$$

Where $H^{(l)}$ is the node feature matrix of the $l - th$ layer, $W^{(l)}$ is the weight matrix of the $l - th$ layer, σ is the activation function, $\tilde{A}$ is the normalized adjacent matrix considering node self-connection, and $\tilde{D}$ is the diagonal matrix.

The graph relation matrix is used to depict the spatial structure relationship of different information categories in ICH visual business cards and serves as the adjacency matrix of the GCN model. When constructing the connection edges, this study adopts the more k-nearest neighbor (k-NN) method to construct and calculate the connection relationship between nodes. Specifically, after calculating the Euclidean distance for any pair of nodes, select the top k nearest neighbors for each node to establish the initial connection. The k-NN method is widely used in visual structure modeling and spatial dependency learning. It can effectively improve the stability of graph construction and is applicable to more complex layout scenarios, such as multi-column and cross-block. Considering that the layout of visual business cards usually has a reading sequence from left to right and top to bottom. This study introduces directional filtering on the basis of

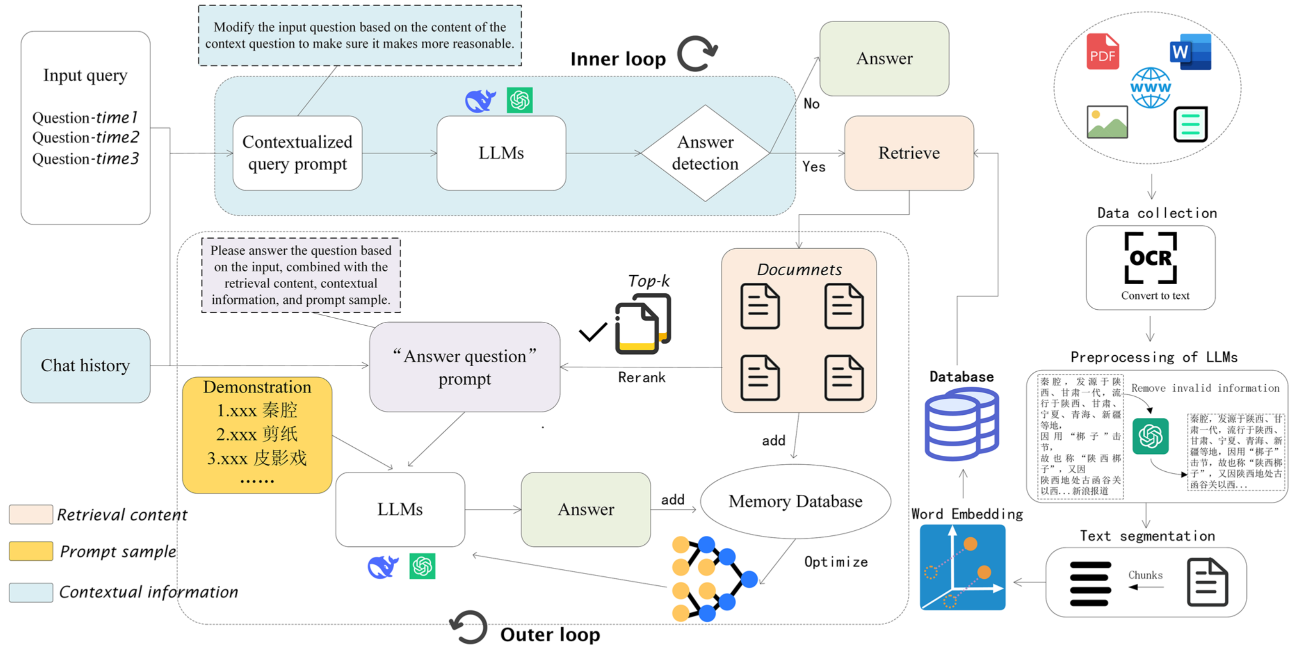

Modify the input question based on the content of the context question to make sure it makes more reasonable.
Inner loop
Input query
Question-time1
Question-time2
Question-time3
Contextualized query prompt
LLMs
Answer detection
No
Yes
Answer
Retrieve
PDF
W
WWW
Data collection
OCR
Convert to text
Preprocessing of LLMs
秦腔，发源于陕西、甘肃一代，流行于陕西、甘肃、宁夏、青海、新疆等地，因用“梆子”击节，故也称“陕西梆子”，又因陕西地处古函谷关以西...新浪报道
Remove invalid information
秦腔，发源于陕西、甘肃一代，流行于陕西、甘肃、宁夏、青海、新疆等地，因用“梆子”击节，故也称“陕西梆子”，又因陕西地处古函谷关以西...
Text segmentation
Chunks
Database
Word Embedding
Please answer the question based on the input, combined with the retrieval content, contextual information, and prompt sample.
Top-k
Documnets
Chat history
“Answer question” prompt
Rerank
Demonstration
1.xxx 秦腔
2.xxx 剪纸
3.xxx 皮影戏
……
add
LLMs
Answer
add
Memory Database
Optimize
Retrieval content
Prompt sample
Contextual information
Outer loop

Fig. 4 | Retrieval augmented generation framework.

composition, that is, prioritizing the retention of neighbors located to the right and below the current node, thereby ensuring layout consistency while reducing unnecessary long-distance connections. The specific construction process can be found in Algorithm 1.

Algorithm 1. Construction of Visual Business Card Graph with Directional k-NN

```
Input   List of nodes where each node contains bounding box
        diagonal corners (x1,y1), (x2,y2); integer k (number of
        nearest neighbors)
Output  Weighted adjacency dictionary graph_dict (convertible
        to adjacency matrix)
Initialize empty dictionary graph_dict ← {}
for each src_idx, src_box in nodes do
    src_cx ← (src_box.x1 + src_box.x2)/2, src_cy ← (src_box.y1 + src_box.y2)/2
    Initialize empty list candidates ← []
    for each dest_idx, dest_box in nodes do
        if dest_idx == src_idx then continue
        dest_cx ← (dest_box.x1 + dest_box.x2)/2, dest_cy ← (dest_box.y1 + dest_box.y2)/2
        distance ← √((src_cx − dest_cx)² + (src_cy − dest_cy)²)
        if dest_cx ≥ src_cx or dest_cy ≥ src_cy then
            candidates.append((distance, dest_idx))
    Sort candidates by distance ascending
    knn ← candidates[0:k] // top-k nearest neighbors
    Initialize empty list filtered_neighbors ← []
    for each (dist, idx) in knn do
        dest_box ← nodes[idx]
        if dest_box.x1 > src_box.x2 or dest_box.y1 > src_box.y2 then
            filtered_neighbors.append(idx)
    if filtered_neighbors is not empty then
        graph_dict[src_idx] ← {idx: 1/dist for (dist, idx) in knn if idx in filtered_neighbors}
Remove isolated nodes: graph_dict ← {u: v for u, v in graph_dict.items() if v ≠ {}}
return graph_dict
```

Loop-retrieval augmented generation strategy

RAG enhances generative models' ability to produce accurate and rich text by retrieving high-quality existing information. However, current RAG strategies are primarily static and lack a mechanism to optimize the overall retrieval process. This study extends traditional RAG by introducing inner and outer loop mechanisms, enabling dynamic decision-making within a single task while progressively accumulating optimal retrieval and prompt strategies over multiple task iterations. The Loop-RAG process consists of three key steps: knowledge base construction, inner-loop context fusion, and outer-loop memory optimization. By tightly integrating dynamic information retrieval with the generation process, the Loop-RAG strategy enables precise question answering and rich descriptions of ICH inheritor information. The detailed workflow is illustrated in Fig. 4.

Building the local RAG knowledge base: By collecting and organizing structured and unstructured data in the field of ICH, a multidimensional knowledge base covering the background, technical processes, representative figures, regional characteristics, and other aspects of ICH projects is formed. Using m3e-base for vectorization, knowledge is transformed into semantic vectors and stored in a vector database to support efficient similarity retrieval. m3e-base maintains a superior Chinese semantic representation effect while featuring lightweight characteristics.

Inner loop: context fusion and dynamic retrieval: In the generation process, the model first forms a query vector based on the input question q and the current context content C_t:

$$q_t = f_{\text{enc}}(q, C_t) \tag{14}$$

Where f_{enc} is the encoder and C_t represents the current task context. the model retrieves the candidate document set $D_t = \{d_1, d_2, \cdots, d_k\}$ from the knowledge base and calculates the similarity:

$$\text{sim}(q_i, d_i) = \frac{q_i \cdot d_i}{||q_i||\,||d_i||} \tag{15}$$

Subsequently, select Top-k documents as candidate knowledge and fuse them to form context enhanced representations:

$$C'_t = g(C_t, D_t) \tag{16}$$

Then the model generates the response y_t based on the enhanced context. If a decrease in information gain is detected during the generation process, the next round of adaptive re-retrieval will be triggered. The next round of query vectors is updated through the reconstruction function h. By iteratively updating q_t, the model gradually optimizes the retrieval and generation results in a single task:

$$q_{t+1} = h(q_t, y_t, D_t) \tag{17}$$

The inner loop terminates when the following conditions are met: the similarity is too low, sim < 0.2, and the information gain is lower than the threshold, $\triangle_{\text{sim}} < 0.1$.

Outer loop: optimization of long-term memory and retrieval strategies: In multiple rounds of tasks, the model stores the query q_t of each inner loop, the candidate document set D_t, the generated result y_t and its effect evaluation E_t in the memory database M:

In multi-round tasks, the outer loop continuously learns the optimal retrieval path by accumulating the retrieval and behavior of all tasks. For the $t - th$ task, the model stores the retrieval sequence $\left(q_t, D_t, y_t\right)$ generated by the inner loop and its effect score E_t in the memory M:

$$\mathcal{M} = \{(q_t, D_t, y_t, E_t)\}_{t=1}^{T} \tag{18}$$

The outer loop aims to maximize long-term performance and optimizes the retrieval strategy π based on the memory database:

$$\pi^* = \text{argmax}_\pi \mathbb{E}_{(\text{q,y})\sim \text{D},\tau\sim\pi}[\text{E}_\tau] \tag{19}$$

Among them, π^* represents the optimal strategy after multiple rounds of optimization, and τ is the retrieval path of the inner loop. The outer loop gradually leads the system towards a more stable and low-illusion retrieval path.

The outer loop terminates when any of the following conditions is met: the score gain of the most recent multiple strategy optimizations is less than the threshold ($\triangle_E < 0.05$); Reach the maximum number of outer loops $T_{\max} = 10$.

The generation phase: Finally, the model combines the optimized contextual representation C'_t with retrieval knowledge and inputs it into the generative model G to obtain the results of ICH Q&A and content generation:

$$y_t = G(q_t, {C'}_t, \pi^*) \tag{20}$$

In conclusion, the loop-RAG mechanism effectively reduces the risk of hallucinations and enhances the controllability of the generated content through real-time dynamic retrieval in the inner Loop and optimized learning in the outer loop.

Results

Experimental environment and evaluation metric

The experiments in this study were conducted on a high-performance computing platform. The hardware environment included an Intel processor (2.40 GHz, 16 cores) and two NVIDIA GeForce RTX 4090 GPUs with 24 GB of video memory each, supporting large-scale parallel computing and deep model training. The software environment consisted of Windows 11, Python 3.9 as the primary programming language, and PyTorch 1.13.1 as the deep learning framework, with GPU acceleration enabled via CUDA 11.7.

The cross-entropy loss is selected as the loss function to measure the difference between the true label and the probability distribution predicted by the model. When there is a significant difference between the prediction results of the model and the true labels, the cross-entropy loss will assign higher loss values to these errors, thereby effectively guiding the model to better approach the values of the true labels. The specific calculation process is shown in formula (21).

$$\text{Cross entropy} = \frac{1}{N}\sum_i L_i = -\frac{1}{N}\sum_i \sum_{c=1}^{M} y_{ic} \ln(p_{ic}) \tag{21}$$

Where M represents the number of classification categories, and y_{ic} represents the sign function. Taking the binary classification task as an example (with labels of 0 or 1), if the predicted sample label is equal to the true category c, $y_{ic} = 1$, Otherwise, $y_{ic} = 0$. p_{ic} represents the probability that the predicted sample i belongs to category c.

In addition, the confusion matrix, as an evaluation metric for measuring the performance of classification models, consists of four categories: true examples (TP); False positive cases (FP); False counterexample (FN); True Counterexample (TN). Based on the confusion matrix, the calculation of four basic evaluation index, namely precision (pre), recall, and F1 value, can be calculated as follows.

$$pre = \frac{TP}{TP + FP} \tag{22}$$

$$recall = \frac{TP}{TP + FN} \tag{23}$$

$$F_1 = \frac{2 \times pre \times recall}{pre + recall} \tag{24}$$

Considering the imbalance in the number of different entity categories in the experimental dataset, macro average and weighted average indicators are added[48]. Macro average indicators address imbalanced datasets by giving equal attention to each category. Weighted average measure considers the sample size of each category, resulting in a more balanced contribution of large and small categories to the overall results.

For text generation evaluation, we adopt the commonly used metrics ROUGE-L, BLEU, and METEOR. ROUGE-L primarily measures recall based on n-grams, and BLEU emphasizes precision on n-grams. These metrics are suited for the report generation task, which emphasizes the completeness and linguistic quality of the generated text. In contrast, the Q&A task aims to assess whether the model provides correct answers to domain-specific ICH questions rather than text similarity. Therefore, precision, recall, and F1 are more appropriate, as they directly measure factual accuracy and answer correctness.

After extracting specific information from the visual business cards of ICH inheritors, the RAG strategy is applied to downstream tasks such as ICH report generation and intelligent Q&A, enabling richer and more personalized dissemination of ICH. The specific model configurations are presented in Table 3.

Training of the visual information extraction model

This section focuses on training and validating the Graph-Retrieval model for visual information extraction. The preprocessed ICH inheritor business card data is divided into training, validation, and testing sets in a 6:2:2 ratio. The training set ensures sufficient data for feature learning, the validation set is used to tune hyperparameters and optimize recognition performance, and the test set independently evaluates the model's generalization ability.

To improve model performance, this study systematically tested multiple sets of hyperparameters and optimized them using grid search. The key parameters included learning rate, batch size, number of training epochs, optimizer type, number of network layers, word embedding dimension, chunk size and overlap rate, dropout rate, mask probability, edge deletion probability, and retrieval top-k value. The specific settings are listed in Table 3. Throughout the experimental process, the model followed a step-by-step workflow of semantic extraction, graph convolution, and feature enhancement, enabling it to effectively identify key information in the visual business cards of ICH inheritors. The detailed hyperparameter configurations used in the grid search are provided in Table 4.

Table 3 | Model selection

Model category	Model	Model introduction
LLMs	GPT-4	GPT-4 is the latest generation of generative pre-trained models released by OpenAI, which demonstrates outstanding performance in language modeling and generation tasks.
	Llama-3 (Llama-3-7b-chat)	Llama-3 is a large-scale language model released by Meta, known for its efficient model structure and training methods.
	ChatGLM-4	ChatGLM-4 is a generative dialogue model proposed by Zhipu AI company, which performs excellently in dialogue generation and language understanding.
RAG-LLm	RAG-Llama3 RAG-GPT4 RAG-ChatGLM4	Introducing the RAG strategy constructed in this article, which combines retrieval knowledge, contextual information, and prompt examples to guide LLMs to better adapt to downstream tasks.

Table 4 | Hyperparameter setting

Hyperparameter	Value
Learning rate	0.01, 0.001, 0.0001
Batch size	32, 64, 128
Epoch	30, 60, 100
Optimizer	Adam
RoBERTa layer	12-Layers transformer
BiLSTM layer	1, 2, 3
GCN layer	1, 2, 3
word embedding dimension	256, 512, 768
Random masking probability	0.1, 0.2, 0.3
Random edge deletion probability	0.01, 0.05, 0.1
Dropout	0.2
Topk	3
Chunk size	64
Chunk overlap	0.3

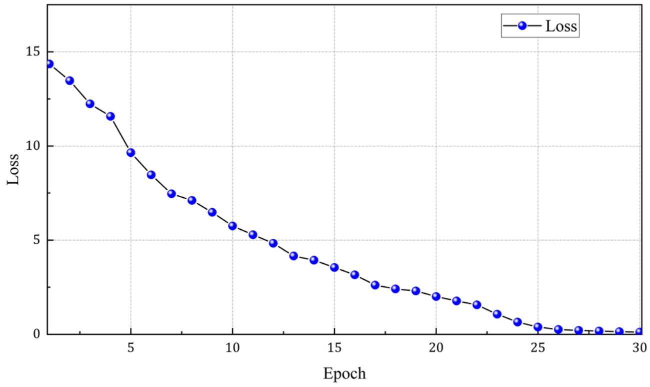


Fig. 5 | Model training loss changes.

Through extensive grid search, the optimal combination of hyperparameters was identified to minimize the loss of the Graph-Retrieval model on the 3142 training samples. The model achieved the best convergence under the following settings: a learning rate of 0.001, batch size of 64, 100 training epochs, a 12-layer Transformer architecture for RoBERTa, a 2-layer BiLSTM network, a 2-layer GCN, word embedding dimension of 512, random node masking probability of 0.05, and random edge deletion probability of 0.05. Under these conditions, the model's loss curve remained stable and consistently converged toward the minimum value, as shown in Fig. 5.

From the analysis of Fig. 5, it can be concluded that after 30 training epochs, the loss of the Graph-Retrieval model on the training set decreased steadily from 14.354 to 0.117, eventually stabilizing at this level. Throughout the entire training process, the model exhibited excellent stability, with a consistent decline in loss and no significant fluctuations. Notably, between the 20th and 30th epochs, the model showed clear signs of convergence, further demonstrating its robustness and efficiency during training.

Validation of the visual information extraction model

Although the Graph-Retrieval model performs well on the training set, its recognition performance must still be evaluated on the validation set, with potential adjustments to network parameters made based on the results. Accordingly, 1047 validation samples were input into the trained Graph-Retrieval model to assess its performance using precision (P), recall (R), and F1 score (F1). The recognition results are presented in the confusion matrix in Fig. 6.

As shown in Fig. 6, on the validation set, the Graph-Retrieval model achieved macro-averaged precision, recall, and F1 scores of 0.9313, 0.9304, and 0.9305, respectively. Overall, the model demonstrated strong recognition performance, with an average prediction accuracy exceeding 0.91. In particular, the highest accuracies were observed for category 4 (Gender) and category 6 (Birthday), both reaching 98%. This can be attributed to the relatively simple and short semantic content of gender information in category 4, and the distinct numeric patterns characterizing birthdate information in category 6. Consequently, the model achieves optimal predictive performance for these two categories. By contrast, accuracy was lower for category 3 (Name) and category 8 (Application_unit), with correct predictions of 93% and 89%, respectively. Category 3 involves personal names, which vary widely among ICH inheritors and thus increase prediction difficulty. Category 8 represents application unit information, which overlaps considerably with work unit information in category 7, leading to similar textual features and a reduction in prediction accuracy.

Testing of the visual information extraction model

To thoroughly assess the Graph-Retrieval model's performance in recognizing ICH inheritors' business card information on the test set, comparative experiments were conducted with several benchmark models. The results of these comparisons are summarized in Table 5.

BERTgrid[22]: Based on the grid method, the character encoding in Chargrid is replaced with word granularity BERT encoding, and information recognition is performed using a large-scale pre-trained language model.

VisualWordGrid[24]: Introduce visual information, overlay document images with grids, and extract text features.

BROS[30]: Combining the BERT model and embedding a relative position encoding into the self-attention mechanism to fully explore the deep connections between semantic and layout modalities.

GC-few shot[31]: A pretrained model based on RoBERTa and graph networks extracts semantic and layout information, while also incorporating a font encoding module.

Improved-GNN[35]: Visual image information is added during the initialization of document graph nodes, and the text embedding is combined with the corresponding visual image embedding to enhance the model's representation ability.

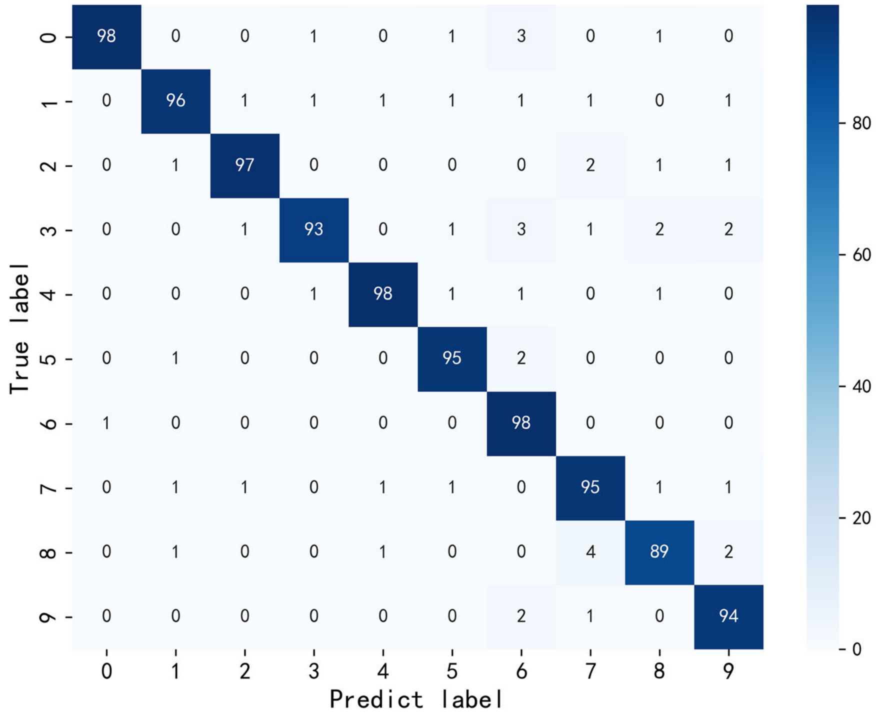


Fig. 6 | The confusion matrix of model in the validation set.

Multimodal weighted graph[37]: Abstracting the document into a fully connected graph, embedding the semantic and layout information of each text segment into the graph structure, and fully integrating node information based on attention mechanism.

DocExtractNet[49]: A module-based framework utilizing LayoutLMv3 fully leverages the features of image and text modalities to extract key information from receipts.

Donut[50]: It adopts an end-to-end Transformer-based architecture to directly analyze document images, making them more accurate in handling different languages and formats.

Graph-Retrieval: The visual information recognition model proposed in this study combines semantic information with enhanced graph features for accurate prediction.

The experimental results in Table 5 clearly demonstrate that the Graph-Retrieval model proposed in this study significantly outperforms the benchmark models in recognizing visual information on ICH inheritors' business cards. Specifically, the Graph-Retrieval model achieved excellent recognition performance, with macro-average F1 scores of 0.928. In comparison, the BERTgrid model showed poor performance in identifying the "Name,""Work unit," and "Application unit" categories, with average accuracy below 0.8. The VisualWordGrid method improves feature extraction by integrating grid layout and visual information, achieving a macro-average F1 score of 0.823, a 1.35% increase over BERTgrid. The BROS method enhances semantic recognition by incorporating position encoding embeddings with a pre-trained model. Although its macro-average F1 score reaches 0.826, it still does not fully capture the spatial layout relationships among node features, limiting its overall recognition performance.

Secondly, GC-FewShot introduces a search-word strategy and pre-training, resulting in a macro-average F1 score of 0.851, which is 3.03% higher than that of the BROS method. Improved-GNN further enhances performance by incorporating visual image information, combining text embeddings with the corresponding visual embeddings of each region. This approach not only accounts for spatial layout but also significantly strengthens the model's feature representation capability, achieving a macro-average F1 score of 0.858. Donut, as an end-to-end visual document model, has a macro avg F1 value of 0.873, indicating that the pure vision Transformer architecture has a strong modeling ability in image-level semantic alignment. However, Donut only relies on visual encoders, and its ability to model the structural relationships between local text regions is limited. The M-Weighted Graph method introduces an attention mechanism in feature extraction to construct a weighted graph, effectively integrating semantic information with node relationship features. Compared to Improved-GNN, it further improves predictive performance, with a macro-average F1 score of 0.893, a relative increase of 2.29%. Furthermore, DocExtractNet makes full use of the features of image and text modalities, and the F1 value corresponding to its macro avg can reach 0.907.

Finally, the Graph-Retrieval model proposed in this study considers both semantic information and graph relationships, while enhancing robustness through random node masking, random edge deletion, and positional attention mechanisms. As a result, it achieves a macro-average F1 score of 0.928, which is 2.32% higher than that of DocExtractNet. This demonstrates that the fusion of semantic information and graph features not only achieves the best recognition performance but also ensures stable results across different categories.

Ablation experiment of visual information extraction

To further evaluate the contribution of each module in the Graph-Retrieval model to the overall network performance, this study conducted ablation experiments. These experiments compared recognition performance when removing semantic information, node masking, random edge deletion, and positional attention mechanisms. Using the constructed ICH inheritor summary lexicon as the basis, each ablation model is described below. The specific results of these experiments are presented in Fig. 7.

W/O WB-EGraph: Semantic extraction uses RoBERTa's lexicon to obtain more accurate word embeddings, and the GCN network has not undergone any graph feature enhancement.

W/O W-EGraph: The semantic extraction uses RoBERTa's lexicon to obtain more accurate word embeddings. The BiLSTM network learns contextual information, while the GCN network does not undergo any graph feature enhancement.

W/O EGraph: The semantic extraction uses the summary lexicon of ICH inheritors, and the semantic extraction method is the same as above; However, no graph feature enhancement was performed during the process of graph structure learning.

Table 5 | Test set comparison experiment

Category		BERTgrid	VisualWordGrid	BROS	GC-few shot	Improved-GNN	Donut	M-weighted graph	DocExtractNet	Graph-Retrieval
Item	P	0.811	0.822	0.827	0.853	0.863	0.882	0.902	0.913	0.932
	R	0.816	0.806	0.814	0.853	0.872	0.885	0.902	0.911	0.931
	F	0.814	0.814	0.820	0.853	0.868	0.884	0.902	0.912	0.931
Item-name	P	0.789	0.801	0.817	0.847	0.855	0.874	0.894	0.909	0.929
	R	0.774	0.821	0.818	0.848	0.847	0.871	0.894	0.907	0.929
	F	0.781	0.811	0.817	0.847	0.851	0.873	0.894	0.908	0.929
Area	P	0.825	0.822	0.837	0.859	0.863	0.866	0.891	0.905	0.925
	R	0.833	0.823	0.828	0.859	0.866	0.867	0.891	0.906	0.926
	F	0.829	0.823	0.832	0.859	0.865	0.866	0.891	0.906	0.925
Name	P	0.799	0.817	0.803	0.851	0.844	0.852	0.876	0.903	0.929
	R	0.786	0.822	0.823	0.851	0.840	0.854	0.876	0.904	0.928
	F	0.792	0.819	0.813	0.851	0.842	0.853	0.876	0.903	0.929
Gender	P	0.838	0.837	0.848	0.857	0.874	0.891	0.906	0.913	0.933
	R	0.841	0.853	0.841	0.856	0.875	0.887	0.905	0.914	0.932
	F	0.839	0.845	0.844	0.857	0.874	0.889	0.905	0.913	0.932
Career	P	0.822	0.834	0.836	0.853	0.853	0.867	0.895	0.907	0.929
	R	0.815	0.830	0.827	0.854	0.861	0.869	0.896	0.906	0.929
	F	0.819	0.832	0.832	0.853	0.857	0.868	0.895	0.906	0.929
Birthday	P	0.844	0.854	0.862	0.857	0.885	0.898	0.908	0.909	0.932
	R	0.839	0.849	0.860	0.856	0.881	0.896	0.909	0.910	0.932
	F	0.842	0.852	0.861	0.857	0.883	0.897	0.909	0.910	0.932
Work unit	P	0.802	0.812	0.833	0.845	0.866	0.883	0.896	0.899	0.926
	R	0.793	0.825	0.825	0.845	0.873	0.885	0.895	0.900	0.925
	F	0.797	0.818	0.829	0.845	0.869	0.884	0.896	0.899	0.925
Application unit	P	0.774	0.786	0.792	0.842	0.833	0.852	0.884	0.899	0.921
	R	0.792	0.779	0.807	0.842	0.829	0.850	0.884	0.901	0.922
	F	0.783	0.783	0.799	0.842	0.831	0.851	0.884	0.900	0.922
Description	P	0.814	0.827	0.814	0.851	0.844	0.863	0.877	0.913	0.927
	R	0.824	0.837	0.823	0.849	0.841	0.866	0.877	0.911	0.927
	F	0.819	0.832	0.818	0.850	0.842	0.865	0.877	0.912	0.927
Macro avg	P	0.812	0.821	0.827	0.851	0.858	0.873	0.893	0.907	0.928
	R	0.811	0.824	0.826	0.851	0.858	0.873	0.893	0.907	0.928
	F	0.812	0.823	0.826	0.851	0.858	0.873	0.893	0.907	0.928
Weighted avg	P	0.803	0.815	0.826	0.857	0.863	0.879	0.899	0.916	0.931
	R	0.822	0.828	0.836	0.858	0.866	0.876	0.901	0.915	0.932
	F	0.812	0.821	0.831	0.858	0.864	0.876	0.900	0.916	0.932

W/O GCN1: The semantic extraction process is the same as above, and a random node masking is introduced in the graph feature enhancement process.

W/O GCN2: The semantic extraction process is the same as above, and the random edge deletion method is introduced in the graph feature enhancement process.

W/O GCN3: The semantic extraction process is the same as above, and the image feature enhancement process uses a positional attention mechanism.

Graph-Retrieval: The model proposed in this study combines semantic extraction with three graph enhancement methods: random node masking, random edge deletion, and a positional attention mechanism.

As shown in Fig. 7, the F1 value of W/O WB-EGraph is the lowest in the ablation experiment. Using W/O WB-EGraph as the benchmark, subsequent module additions progressively improved recognition performance, with the Graph-Retrieval model ultimately achieving the best predictive results. Specifically, W/O W-EGraph incorporates the BiLSTM network to enhance contextual semantic learning based on the pre-trained RoBERTa model, achieving an F1 value of 0.8566. Next, by introducing the proprietary vocabulary of ICH inheritors, W/O EGraph strengthened the training on specialized terminology and proper nouns, reaching an F1 value of 0.8767, a 2.35% improvement over W/O W-EGraph.

Furthermore, to capture spatial relationships in the visual business card data, graph feature methods were incorporated. With the addition of random node masking, the W/O GCN1 model showed a significant improvement, achieving an F1 value of 0.8915-1.68% higher than that of W/O EGraph. This demonstrates that random node masking helps the model learn from noisy features, enhancing its robustness to data imperfections such as wear and tear. The W/O GCN2 model employs random edge deletion, yielding an F1 value of 0.9061. Although its improvement is smaller than W/O GCN1, it enables the model to learn diverse graph structures and further strengthens robustness. By incorporating the sequential order of nodes, W/O GCN3 achieved the greatest performance gain, with an F1 value of 0.9167, 4.56% higher than W/O EGraph. Ultimately, the Graph-Retrieval model achieved optimal predictive performance by integrating semantic features with multiple graph feature enhancement strategies. Its ROC curve remained the highest, with an F1 value of 0.9299, representing an 11.6% improvement over the benchmark model W/O WB-EGraph.

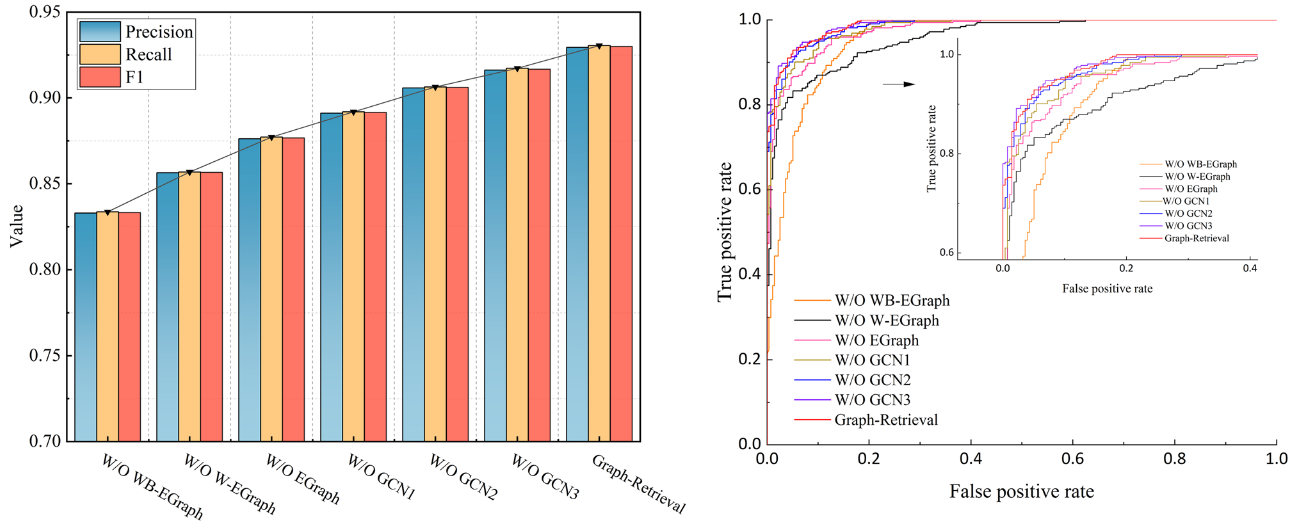
The influence of model parameters on visual information extraction experiments

This section investigates the impact of various parameters in the graph feature enhancement module on the information recognition performance of the Graph-Retrieval model. Specifically, random node masking, random edge deletion, and the number of GCN layers are examined as the primary factors. Since the positional attention mechanism is automatically learned during training rather than manually configured, it is excluded from this parameter analysis. The parameter settings and corresponding experimental results are presented in Fig. 8.

According to Fig. 8, when the random node masking probability is set to 0.05, the model achieves an F1 value of 0.9302 on the test set, demonstrating high recognition performance. In contrast, with a masking probability of 0.025, the F1 value drops slightly to 0.9287. This indicates that an appropriate masking probability enhances the model's generalization ability, allowing it to maintain high predictive performance even with incomplete or damaged data. However, as the masking probability increases, node features are increasingly disrupted, resulting in incomplete feature learning; for example, at a masking probability of 0.3, the F1 value falls sharply to 0.8624.

Similarly, random edge deletion probability significantly affects model performance. When the edge deletion probability exceeds 0.05, predictive performance declines markedly. Specifically, at a probability of 0.07, the F1 value drops to 0.8838, a 1.95% decrease compared to 0.05. Within the range of 0.01 to 0.05, the F1 value remains relatively stable around 0.91, reaching 0.9299 at a probability of 0.02. This suggests that moderate edge deletion prevents the model from overfitting to a single graph structure, while excessive deletion damages the graph structure and impairs recognition performance.

Finally, the number of layers in the GNN network significantly affects model performance. When the GNN has two layers, the model achieves its best performance, with an F1 value of 0.9304. With only one layer, the recognition performance is limited, yielding an F1 value of 0.9169. As the number of layers exceeds two, model performance begins to decline; at four layers, the F1 value drops to 0.9114. This suggests that a two-layer GNN effectively captures node features and integrates higher-level abstract information, enhancing the model's understanding of graph structure. In contrast, a single-layer GNN has limited capacity to learn complex



(a) Ablation experiment

(b) Roc curve

Fig. 7 | Results of ablation experiment. a The ablation experiment of the graph feature enhancement module. **b** The ROC curve.

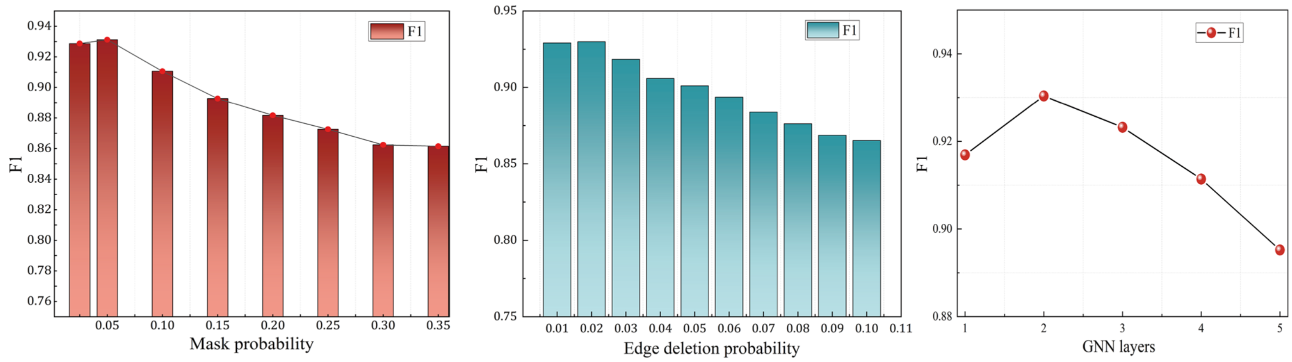


Fig. 8 | Experimental results of parameter influence.

Table 6 | Comparative experimental results of public datasets

Model	Dataset	$K=2$	$K=3$	$K=4$	$K=5$
BROS	SROIE	0.813	0.861	0.880	0.872
	FUNSD	0.732	0.781	0.790	0.771
	CORD	0.805	0.852	0.860	0.852
	RVL-CDIP	0.743	0.791	0.800	0.791
	Business cards	0.764	0.823	0.830	0.821
GC-few shot	SROIE	0.823	0.852	0.860	0.851
	FUNSD	0.745	0.791	0.801	0.783
	CORD	0.813	0.863	0.870	0.861
	RVL-CDIP	0.754	0.801	0.822	0.801
	Business cards	0.776	0.837	0.840	0.830
Multimodal weighted graph	SROIE	0.834	0.866	0.870	0.861
	FUNSD	0.753	0.791	0.810	0.791
	CORD	0.823	0.873	0.885	0.875
	RVL-CDIP	0.765	0.811	0.830	0.812
	Business cards	0.783	0.842	0.851	0.841
DocExtractNet	SROIE	0.843	0.871	0.880	0.870
	FUNSD	0.765	0.805	0.822	0.803
	CORD	0.833	0.887	0.890	0.882
	RVL-CDIP	0.771	0.821	0.847	0.821
	Business cards	0.793	0.851	0.860	0.851
Graph-Retrieval	SROIE	**0.893**	**0.911**	**0.919**	**0.915**
	FUNSD	**0.873**	**0.901**	**0.908**	**0.905**
	CORD	**0.903**	**0.921**	**0.931**	**0.927**
	RVL-CDIP	**0.883**	**0.911**	**0.925**	**0.922**
	Business cards	**0.893**	**0.911**	**0.924**	**0.920**

The values in bold are the maximum values obtained from the comparative experiments.

relationships, while deeper networks may suffer from over-smoothing, where node representations become overly similar, potentially causing information loss.

Comparative experiments on public datasets of visual information extraction

To evaluate the generality and scalability of our method, we selected five publicly available datasets-SROIE, FUNSD, CORD, RVL-CDIP, and Business_Cards_Images, and compared our approach with four representative visual information extraction methods: BROS, GC-few shot, Multimodal Weighted Graph, and DocExtractNet. Furthermore, we also verified the influence of the number of ks in the k-nearest neighbor method on the extraction performance when constructing the graph matrix. The experiment takes the F1 value as the evaluation index. The specific results are shown in Table 6.

From the overall experimental results, Graph-Retrieval achieves the best performance on all five cross-domain visual information extraction datasets, demonstrating significant generalization ability and cross-scenario robustness. In the SROIE dataset, Graph-Retrieval achieved the highest F1 value of 0.919 under the $k=4$ setting, significantly outperforming all benchmark methods. In the more structurally complex form dataset FUNSD, its F1 value reached 0.908, and the improvement rates compared with Multimodal Weighted Graph and DocExtractNet reached 9.8% and 8.6% respectively. This indicates that Graph-Retrieval also has certain feature modeling capabilities when dealing with semi-structured layouts and cross-text block relationships. In the complex ticket CORD dataset, the F1 value of Graph-Retrieval was further increased to 0.931. In the RVL-CDIP dataset, the performance of Graph-Retrieval reached 0.925 and achieved an improvement of over 10% on multiple benchmark models, demonstrating its robustness in a multi-layout document environment with significant layout differences. In the Business Cards Images dataset, the F1 value of Graph-Retrieval reaches 0.924. Overall, Graph-Retrieval demonstrates stable and consistent advantages in various structures, visual layouts, and cross-domain document scenarios.

In terms of graph structure construction strategies, different k values have a significant impact on model performance. When $k=4$, it represents the regular trend of the optimal graph structure configuration. When $k=2$, since each node is only connected to two neighbors, the graph structure is overly sparse, resulting in limited information dissemination capacity and relatively low overall performance of each model. As k increases to 3, the connectivity of the graph is enhanced, and performance generally improves. For instance, on the CORD dataset, it rises from 0.903 to 0.921, but it is still slightly below the optimal state. When $k=4$, the Graph structure achieves the optimal balance between connectivity and noise control, enabling all models to achieve the highest performance. Particularly, Graph-Retrieval achieves peak performance in all datasets, such as 0.908, 0.931, and 0.925 on FUNSD, CORD, and RVR-CDIP, respectively. When k increases to 5, due to too many adjacent edges, some noisy connections are introduced, resulting in a slight decline in performance. For example, in the Business Cards dataset, it drops from 0.924 to 0.920. Overall, the experimental results verify that an appropriate neighborhood size is crucial for the quality of the graph structure. When $k=4$, it can achieve a balance between the sparsity and expressive power of the graph, thereby maximizing the performance of the visual information extraction model.

Generation of ICH report

The above experiments verified the performance of the Graph-Retrieval model in visual information extraction. Based on the extraction results, this section uses LLMs to perform report generation and intelligent question answering tasks. Meanwhile, based on the powerful semantic capabilities of LLMs, the Loop-RAG enhancement strategy is adopted to introduce high-quality external knowledge. Therefore, different input templates have been set for each task. In this study, the test text is concatenated with Loop-RAG result as the input content, and the output results are obtained through multiple LLMs. Finally, the performance of the model is evaluated through the quantitative assessment of the output results by evaluation indicators. It is specifically shown in Fig. 9.

Where Prompt is the prompt word, Demonstration is the prompt example, Input is the input question, and the contextual information is the historical information in multiple rounds of Q&A. Ultimately, the LLMs generate more accurate and effective answers by combining the above content for retrieval.

LLMs can generate cultural background stories about ICH inheritors based on key information from visual business cards. For instance, they can integrate details such as inheritors' skills, cities, and affiliated units to produce coherent cultural introductions suitable for promotional materials or exhibition displays. For this purpose, this study adopted Llama3, ChatGLM-4, GPT-4, and multiple benchmark models based on RAG. A total of 5237 ICH inheritors' business cards were tested, with each model generating corresponding summary reports, including introductions of the inheritors, descriptions of their skills, and the status of the respective ICH.

For automated report generation tasks, performance was evaluated using BLEU, METEOR, and ROUGE-L metrics. It also includes the assessment of generation time, text length, and the number of inner and outer loop size. The report generation results of the benchmark models are presented in Table 7.

According to the results in Table 7, it can be seen that there are significant differences in the performance of the basic model in the task of generating intangible cultural heritage reports. Among them, GPT-4 achieved the highest scores in the three evaluation metrics of BLEU,

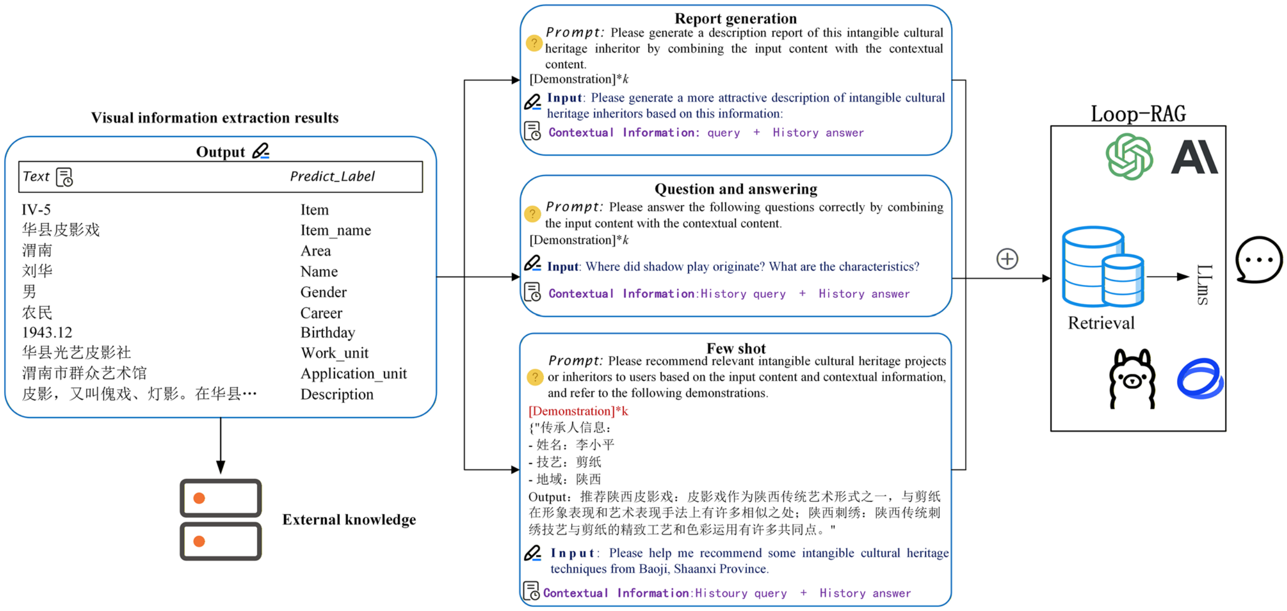

Visual information extraction results
Output
Text
Predict_Label
IV-5
Item
华县皮影戏
Item_name
渭南
Area
刘华
Name
男
Gender
农民
Career
1943.12
Birthday
华县光艺皮影社
Work_unit
渭南市群众艺术馆
Application_unit
皮影，又叫傀戏、灯影。在华县…
Description
External knowledge
Report generation
Prompt: Please generate a description report of this intangible cultural heritage inheritor by combining the input content with the contextual content.
[Demonstration]*k
Input: Please generate a more attractive description of intangible cultural heritage inheritors based on this information:
Contextual Information: query + History answer
Question and answering
Prompt: Please answer the following questions correctly by combining the input content with the contextual content.
[Demonstration]*k
Input: Where did shadow play originate? What are the characteristics?
Contextual Information:History query + History answer
Few shot
Prompt: Please recommend relevant intangible cultural heritage projects or inheritors to users based on the input content and contextual information, and refer to the following demonstrations.
[Demonstration]*k
{"传承人信息：
- 姓名：李小平
- 技艺：剪纸
- 地域：陕西
Output：推荐陕西皮影戏：皮影戏作为陕西传统艺术形式之一，与剪纸在形象表现和艺术表现手法上有许多相似之处；陕西刺绣：陕西传统刺绣技艺与剪纸的精致工艺和色彩运用有许多共同点。"
Input: Please help me recommend some intangible cultural heritage techniques from Baoji, Shaanxi Province.
Contextual Information:Histoury query + History answer
⊕
Loop-RAG
Retrieval
LLms

Fig. 9 | Test method.

Table 7 | Report generation evaluation results

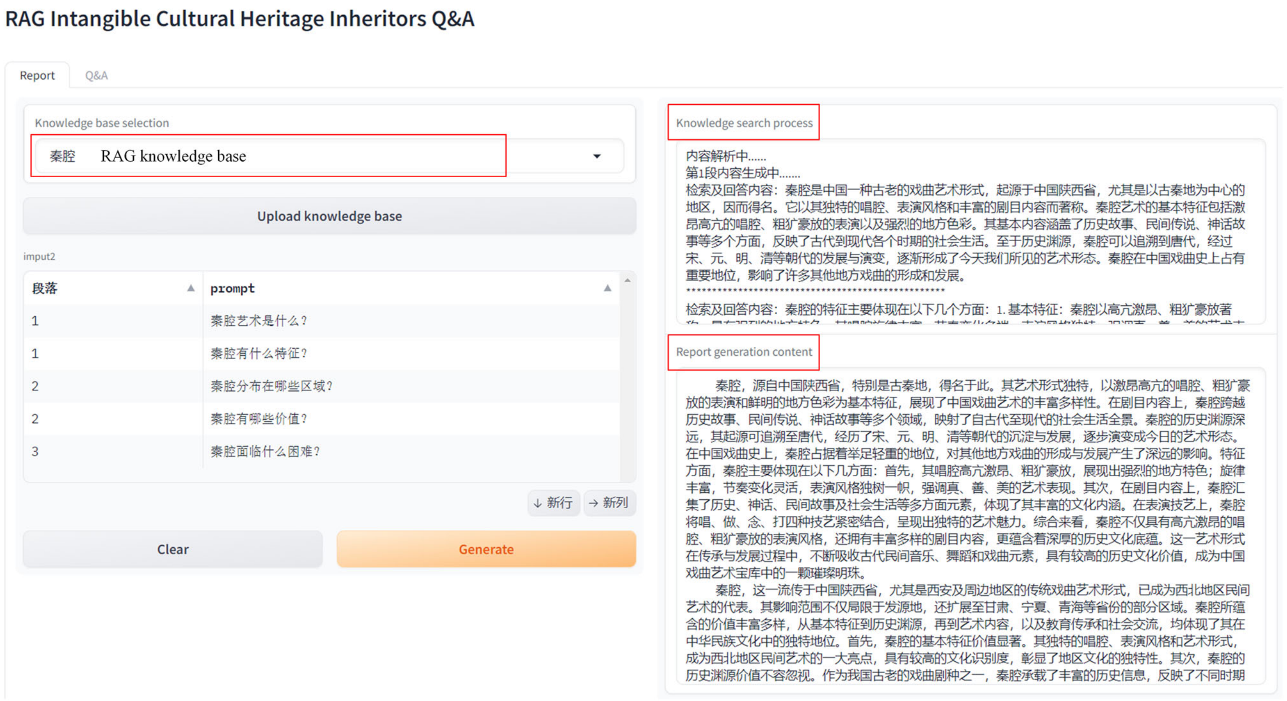
Model	BLEU	METEOR	ROUGE-L	Average generation time(s)	Average generation length	Average Inner-loop size	Average outer-loop size
Llama3	21.50	22.10	23.80	15.3	532	-	-
ChatGLM-4	23.20	23.75	24.60	16.9	426	-	-
GPT-4	26.40	26.90	27.50	14.8	584	-	-
RAG-Llama3	22.84	22.91	24.12	17.8	577	-	-
RAG-ChatGLM4	24.33	24.35	24.74	17.9	453	-	-
RAG-GPT4	25.12	27.24	27.53	16.5	623	-	-
Loop-RAG-Llama3	24.80	25.30	26.20	22.5	1038	3	4
Loop-RAG-ChatGLM4	25.60	26.20	26.90	18.2	1242	3	5
Loop-RAG-GPT4	27.10	27.60	28.30	21.4	896	4	6

METEOR and ROUGE-L, which were 26.40%, 26.90% and 27.50% respectively, demonstrating the best semantic expression ability and text organization ability. ChatGLM-4 came second. Although its BLEU decreased by 3.2% compared to GPT-4, it was still significantly better than Llama-3. The overall indicators of Llama-3 are relatively low, especially ROUGE-L, which is only 23.80%, indicating that there are still deficiencies in terms of content coverage and factual consistency.

On this basis, this study further introduces two types of retrieval enhancement strategies, namely the common RAG and Loop-RAG, to evaluate the impact of external knowledge injection on the generation quality. Ordinary RAG can already bring about certain performance improvements. For instance, the BLEU, METEOR, and ROUGE-L of RAG-Llama3, RAG-ChatGLM4, and RAG-GPT4 have all been enhanced compared to their corresponding base models. Among them, the improvement of RAG-GPT4 is the most stable. Its ROUGE-L has increased from 27.50% to 27.53%. However, such improvements mainly come from one-time retrieval enhancements, and the overall extent is limited.

In contrast, Loop-RAG further introduces the dynamic optimization mechanism of the inner and outer loops, making the gain of the generated quality more significant. Loop-rag-gpt-4 achieved the highest performance in this study, with its BLEU, METEOR, and ROUGE-L increasing by 27.10%, 27.60%, and 28.30% respectively, verifying the effectiveness of multi-round retrieval decisions and cross-task strategy optimization in factual reinforcement and contextual consistency. The enhancement for weaker models is more obvious. For example, the BLEU of Loop-RAG-Llama3 has increased from 21.50% of the basic model to 24.80%, with an increase of 3.3%, indicating that Loop-RAG is particularly suitable for compensating for the insufficient knowledge coverage of the basic model. Meanwhile, the performance improvement of Loop-RAG comes at the cost of computing. Both its generation time and text length have increased significantly. For example, the generation delay of Loop-RAG-Llama3 has risen from 15.3 s to 22.5 s, and the text length is approximately twice that of the base model. Although loop-Rag-GPT-4 performs 4 policy updates in the outer Loop and triggers an average of 4 inner loop searches, its generation time is 21.4 s, achieving a better balance between quality and efficiency.

Finally, the generated report for “Qinqiang Opera (秦腔)” is shown in Fig. 10. The left side displays the selected Loop-RAG knowledge base and the user requirements for report writing. The upper right corner illustrates the search process within the Loop-RAG database for each query. The report generation relies on the high-quality knowledge retrieved from this process. The bottom right corner presents the final generated report content. Overall, by leveraging the Loop-RAG strategy, the generated Q&A content is more diverse, and the answers are more accurate.



RAG Intangible Cultural Heritage Inheritors Q&A

Knowledge base selection

秦腔 RAG knowledge base

imput2

段落	prompt
1	秦腔艺术是什么？
1	秦腔有什么特征？
2	秦腔分布在哪些区域？
2	秦腔有哪些价值？
3	秦腔面临什么困难？

Knowledge search process

内容解析中......

第1段内容生成中.......

检索及回答内容：秦腔是中国一种古老的戏曲艺术形式，起源于中国陕西省，尤其是以古秦地为中心的地区，因而得名。它以其独特的唱腔、表演风格和丰富的剧目内容而著称。秦腔艺术的基本特征包括激昂高亢的唱腔、粗犷豪放的表演以及强烈的地方色彩。其基本内容涵盖了历史故事、民间传说、神话故事等多个方面，反映了古代到现代各个时期的社会生活。至于历史渊源，秦腔可以追溯到唐代，经过宋、元、明、清等朝代的发展与演变，逐渐形成了今天我们所见的艺术形态。秦腔在中国戏曲史上占有重要地位，影响了许多其他地方戏曲的形成和发展。

****

检索及回答内容：秦腔的特征主要体现在以下几个方面：1. 基本特征：秦腔以高亢激昂、粗犷豪放著

Report generation content

秦腔，源自中国陕西省，特别是古秦地，得名于此。其艺术形式独特，以激昂高亢的唱腔、粗犷豪放的表演和鲜明的地方色彩为基本特征，展现了中国戏曲艺术的丰富多样性。在剧目内容上，秦腔跨越历史故事、民间传说、神话故事等多个领域，映射了自古代至现代的社会生活全景。秦腔的历史渊源深远，其起源可追溯至唐代，经历了宋、元、明、清等朝代的沉淀与发展，逐步演变成今日的艺术形态。在中国戏曲史上，秦腔占据着举足轻重的地位，对其他地方戏曲的形成与发展产生了深远的影响。特征方面，秦腔主要体现在以下几方面：首先，其唱腔高亢激昂、粗犷豪放，展现出强烈的地方特色；旋律丰富，节奏变化灵活，表演风格独树一帜，强调真、善、美的艺术表现。其次，在剧目内容上，秦腔汇集了历史、神话、民间故事及社会生活等多方面元素，体现了其丰富的文化内涵。在表演技艺上，秦腔将唱、做、念、打四种技艺紧密结合，呈现出独特的艺术魅力。综合来看，秦腔不仅具有高亢激昂的唱腔、粗犷豪放的表演风格，还拥有丰富多样的剧目内容，更蕴含着深厚的历史文化底蕴。这一艺术形式在传承与发展过程中，不断吸收古代民间音乐、舞蹈和戏曲元素，具有较高的历史文化价值，成为中国戏曲艺术宝库中的一颗璀璨明珠。

秦腔，这一流传于中国陕西省，尤其是西安及周边地区的传统戏曲艺术形式，已成为西北地区民间艺术的代表。其影响范围不仅局限于发源地，还扩展至甘肃、宁夏、青海等省份的部分区域。秦腔所蕴含的价值丰富多样，从基本特征到历史渊源，再到艺术内容，以及教育传承和社会交流，均体现了其在中华民族文化中的独特地位。首先，秦腔的基本特征价值显著。其独特的唱腔、表演风格和艺术形式，成为西北地区民间艺术的一大亮点，具有较高的文化识别度，彰显了地区文化的独特性。其次，秦腔的历史渊源价值不容忽视。作为我国古老的戏曲剧种之一，秦腔承载了丰富的历史信息，反映了不同时期

Fig. 10 | Report generation content. This is the report generation page. On the left is the constructed knowledge base, in the upper right corner is the knoeledge retrieval process, and in the lower right corner is the generated result.

Table 8 | Q&A test results

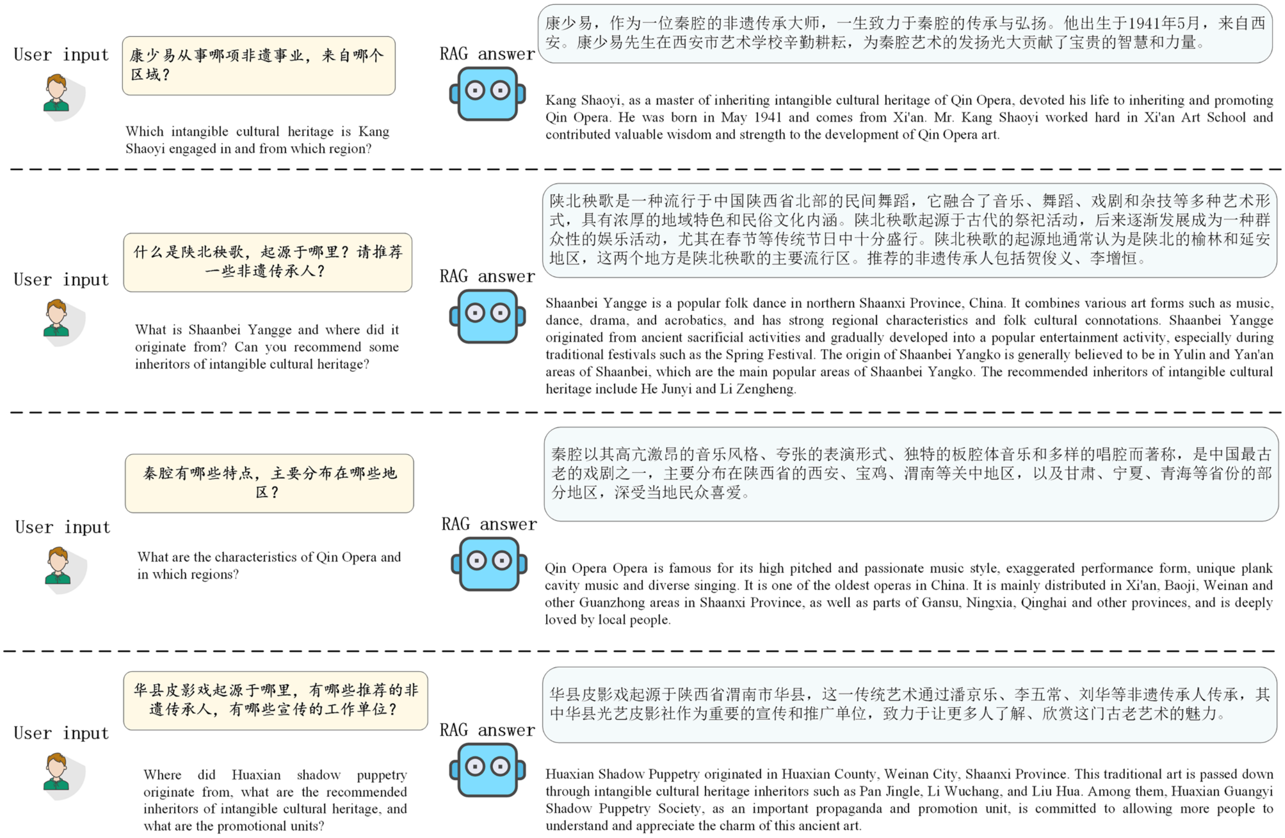
Model	Precision	Recall	F1	Average Inner-loop size	Average outer-loop size	Retrieval steps (avg)	Convergence (avg)
Llama3	0.859	0.865	0.862	-	-	-	-
ChatGLM-4	0.841	0.847	0.844	-		-	-
GPT-4	0.877	0.881	0.879	-	-	-	-
RAG-Llama3	0.873	0.877	0.875				
RAG-ChatGLM4	0.861	0.865	0.863				
RAG-GPT4	0.886	0.891	0.889				
Loop-RAG-Llama3	0.922	0.926	0.924	2	3	5	3
Loop-RAG-ChatGLM4	0.906	0.910	0.908	2	3	6	3
Loop-RAG-GPT4	0.939	0.943	0.941	2	4	4	2

Intelligent Q&A on ICH

This study tested Llama3, ChatGLM-4, GPT-4, RAG-based models and Loop-RAG-enhanced models, for question answering. During the experiment, questions about ICH inheritors were input into each model, and the accuracy of their responses was evaluated. Each model was tasked with answering 2000 common questions about ICH information, with a response considered correct if it matched the real information. Detailed accuracy results and sample Q&A content are presented in Table 8 and Fig. 11. Among them, the retrieval step represents the total number of complete and repeated steps in the entire retrieval chain, that is, the complete number of times the model undergoes retrieval, filtering, recombination, reflecting the efficiency of the model in locating answers in the external knowledge base.

According to the data in Table 8, the overall question-answering performance of the benchmark models Llama3, ChatGLM4 and GPT-4 is relatively stable, with their F1 values being 0.862, 0.844, and 0.879, respectively. However, due to the fact that these models failed to cover the latest knowledge of intangible cultural heritage during the pre-training stage, they are prone to factual bias when dealing with problems involving specific inheritors' information, technical details or regional cultural practices, that is, they tend to give "illusory" answers. To alleviate the above problems, this study constructed a structured local knowledge base for intangible cultural heritage inheritors and introduced three types of enhanced retrieval mechanisms on this basis, including the common RAG and the Loop-RAG that further incorporates a loop optimization strategy. The experimental results of the common RAG show that the F1 values of all models have been improved compared with the benchmark model. Among them, the F1 values of RAG-Llama3, RAG-ChatGLM4 and RAG-GPT4 have increased to 0.875, 0.863, and 0.889, respectively. It is indicated that external knowledge retrieval can effectively supplement the domain knowledge blind spots of the model.

On this basis, the experimental results show that the Loop-RAG strategy has achieved significant improvements in precision, recall and F1 metrics. For example, the F1 value of Loop-RAG-Llama3 reaches 0.924, which is 7.19% higher than that of the benchmark model. The F1 value of Loop-RAG-ChatGLM4 has been raised to 0.908, an increase of 7.58%. The optimal model is Loop-RAG-GPT4, whose F1 reaches 0.941, which is 7.05% higher than the original GPT-4 model. Furthermore, the optimization process indicators presented in the table further verify the convergence behavior and knowledge fusion efficiency of Loop-RAG: Although Loop-RAG- GPT-4 goes through 2 inner loops and 4 outer loops, its overall



康少易从事哪项非遗事业，来自哪个区域？

Which intangible cultural heritage is Kang Shaoyi engaged in and from which region?

康少易，作为一位秦腔的非遗传承大师，一生致力于秦腔的传承与弘扬。他出生于1941年5月，来自西安。康少易先生在西安市艺术学校辛勤耕耘，为秦腔艺术的发扬光大贡献了宝贵的智慧和力量。

Kang Shaoyi, as a master of inheriting intangible cultural heritage of Qin Opera, devoted his life to inheriting and promoting Qin Opera. He was born in May 1941 and comes from Xi'an. Mr. Kang Shaoyi worked hard in Xi'an Art School and contributed valuable wisdom and strength to the development of Qin Opera art.

什么是陕北秧歌，起源于哪里？请推荐一些非遗传承人？

What is Shaanbei Yangge and where did it originate from? Can you recommend some inheritors of intangible cultural heritage?

陕北秧歌是一种流行于中国陕西省北部的民间舞蹈，它融合了音乐、舞蹈、戏剧和杂技等多种艺术形式，具有浓厚的地域特色和民俗文化内涵。陕北秧歌起源于古代的祭祀活动，后来逐渐发展成为一种群众性的娱乐活动，尤其在春节等传统节日中十分盛行。陕北秧歌的起源地通常认为是陕北的榆林和延安地区，这两个地方是陕北秧歌的主要流行区。推荐的非遗传承人包括贺俊义、李增恒。

Shaanbei Yangge is a popular folk dance in northern Shaanxi Province, China. It combines various art forms such as music, dance, drama, and acrobatics, and has strong regional characteristics and folk cultural connotations. Shaanbei Yangge originated from ancient sacrificial activities and gradually developed into a popular entertainment activity, especially during traditional festivals such as the Spring Festival. The origin of Shaanbei Yangko is generally believed to be in Yulin and Yan'an areas of Shaanbei, which are the main popular areas of Shaanbei Yangko. The recommended inheritors of intangible cultural heritage include He Junyi and Li Zengheng.

秦腔有哪些特点，主要分布在哪些地区？

What are the characteristics of Qin Opera and in which regions?

秦腔以其高亢激昂的音乐风格、夸张的表演形式、独特的板腔体音乐和多样的唱腔而著称，是中国最古老的戏剧之一，主要分布在陕西省的西安、宝鸡、渭南等关中地区，以及甘肃、宁夏、青海等省份的部分地区，深受当地民众喜爱。

Qin Opera Opera is famous for its high pitched and passionate music style, exaggerated performance form, unique plank cavity music and diverse singing. It is one of the oldest operas in China. It is mainly distributed in Xi'an, Baoji, Weinan and other Guanzhong areas in Shaanxi Province, as well as parts of Gansu, Ningxia, Qinghai and other provinces, and is deeply loved by local people.

华县皮影戏起源于哪里，有哪些推荐的非遗传承人，有哪些宣传的工作单位？

Where did Huaxian shadow puppetry originate from, what are the recommended inheritors of intangible cultural heritage, and what are the promotional units?

华县皮影戏起源于陕西省渭南市华县，这一传统艺术通过潘京乐、李五常、刘华等非遗传承人传承，其中华县光艺皮影社作为重要的宣传和推广单位，致力于让更多人了解、欣赏这门古老艺术的魅力。

Huaxian Shadow Puppetry originated in Huaxian County, Weinan City, Shaanxi Province. This traditional art is passed down through intangible cultural heritage inheritors such as Pan Jingle, Li Wuchang, and Liu Hua. Among them, Huaxian Guangyi Shadow Puppetry Society, as an important propaganda and promotion unit, is committed to allowing more people to understand and appreciate the charm of this ancient art.

Fig. 11 | Intelligent Q&A content.

retrieval path is the shortest, and the number of convergence rounds is 2, indicating that it can achieve the optimal effect at a lower optimization cost. In contrast, Loop-RAG-ChatGLM4 requires six retrieval steps and a longer convergence process, reflecting the differences in knowledge absorption efficiency among various models. Overall, the Loop-RAG- strategy has significantly improved the accuracy of the model in answering questions about intangible cultural heritage, especially when dealing with complex problems in intangible cultural heritage scenarios, the advantages are more obvious.

The question-answering case shown in Fig. 11 further demonstrates that with the support of hybrid retrieval and Loop optimization, loop-RAG can generate more complete and closely related answers to the knowledge system of ICH, providing a high-quality solution for knowledge question-answering in specific fields.

Few-shot test

Additionally, due to the limited information about ICH inheritors stored in the local knowledge base, LLMs often exhibit lower accuracy when answering unfamiliar questions. To evaluate the few-shot learning capability of the Graph-Retrieval model, 0-shot, 1-shot, 2-shot, and 3-shot experiments were conducted. The experimental setup involved providing the same type of prompt examples before inputting queries into the LLMs, enabling the model to reason effectively with fewer samples and handle previously unseen data. Following these steps, the performance of different benchmark models was tested under few-shot conditions. The results demonstrate that incorporating appropriate prompt examples can significantly improve question answering accuracy. The detailed experimental outcomes are presented in Fig. 12.

As shown in Fig. 12, increasing the number of prompt examples leads to improved performance. Compared with the 0-shot condition, under 3-shot conditions, the F1 values of Llama3, ChatGLM4, GPT-4, Loop-RAG-Llama3, Loop-RAG-ChatGLM4, and Loop-RAG-GPT4 increased by 6.21%, 5.48%, 5.65%, 5.17%, 4.94%, and 3.92%, respectively. Notably, the Loop-RAG-GPT4 model achieved an F1 value of 0.954 under the 3-shot condition, significantly outperforming all other benchmark models. This indicates that Loop-RAG can also make full use of the retrieval process in the few-shot environment and improve the accuracy value of the answers.

Discussion

The visual business card of ICH inheritors represents a novel digital form for disseminating ICH. This study utilizes visual recognition and question answering methods to not only effectively extract the profile of inheritors of intangible cultural heritage, but also achieve accurate and personalized question answering in the vertical field of ICH. Specifically, based on the visual business cards of national ICH inheritors, this study combines graph feature enhancement with Loop-RAG technologies to construct a Graph-Retrieval framework for visual information recognition and intelligent question answering, achieving excellent experimental performance.

By collecting and preprocessing information of ICH inheritors in China, a large number of visual business cards for ICH inheritors have been constructed. The business cards contain 10 types of information, including ICH project number, ICH name, inheritor information, and ICH content, which is helpful for the digital storage and sharing of ICH inheritor information.

At the performance level of visual information extraction, the graph feature enhancement method significantly improves the model's ability to extract information and model semantics. Comparative experiments demonstrate that the proposed model outperforms multiple benchmark models, achieving a macro-average F1 score of 0.928. Ablation experiments confirm the positive contributions of semantic feature enhancement, random node masking, random edge deletion, and the positional attention mechanism to model optimization. Additionally, parameter sensitivity

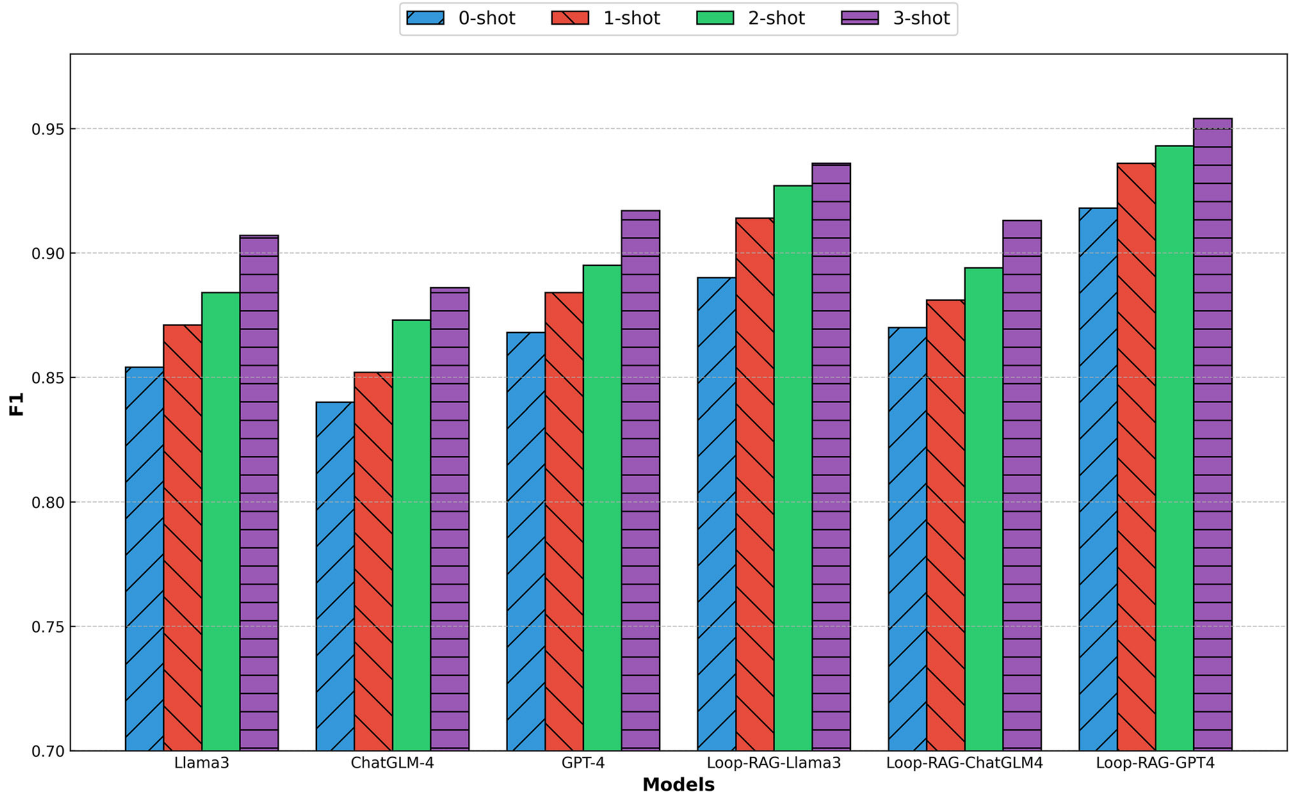


Fig. 12 | Few shot test result.

analysis clarifies how masking and edge deletion probabilities enhance robustness and balance performance. The model also achieves superior recognition results across multiple public datasets.

At the retrieval and generation level, the Loop-RAG strategy effectively reduces resource consumption during LLM fine-tuning, addressing challenges related to domain differences in ICH. It demonstrates strong performance in both report generation and intelligent question-answering tasks, with content accuracy, richness, and generation time surpassing the benchmark models. In addition, it also achieved relatively excellent performance in few-shot tests.

Compared with traditional static information channels, such as official ICH websites, the intelligent question-answering system developed in this study offers several distinct advantages: (1) Semantic retrieval and reasoning ability: It can achieve accurate matching based on context and semantic relations in the case of users' fuzzy queries or non-standard expressions, rather than being limited to keyword retrieval; (2) Interactivity and intelligence: Supports multi-round Q&A and context tracking, capable of generating more targeted responses based on users' dynamic needs, rather than merely returning to fixed pages or entries; (3) Information value-added and content expansion: Through graph feature enhancement and LLMs reasoning, it is possible to complete the context information not explicitly stated in the database, generating more abundant and interpretable content. In future research, knowledge graphs will be introduced to build richer multi-dimensional association relationships among ICH inheritors, and multi-modal features such as images and videos will be integrated to enhance the cross-modal understanding ability of the model. Meanwhile, the key factors that can suppress hallucinations during the retrieval process will also be focused on in the future.

In conclusion, this research provides a novel technical approach for the intelligent and interactive dissemination of ICH, offering practical significance for the digital preservation of ICH and the broader popularization of cultural resources.

Data availability

The datasets used during the current study are available from the corresponding author on reasonable request.

Code availability

The underlying code for this study is not publicly available but may be made available to qualified researchers on reasonable requests from the corresponding author.

Received: 30 June 2025; Accepted: 11 February 2026;
Published online: 20 February 2026

References

1. Li, X. et al. Research on the construction of intangible cultural heritage corridors in the Yellow River Basin based on geographic information system (GIS) technology and the minimum cumulative resistance (MCR) model. *Herit. Sci.* **12**, 271 (2024).
2. Cen, C. et al. Enhancing the dissemination of Cantonese Opera among youth via Bilibili: a study on intangible cultural heritage transmission. *Humanit. Soc. Sci. Commun.* **11**, 1–13 (2024).
3. Nebot-Gomez de Salazar, N., Chamizo-Nieto, F. J., Conejo-Arrabal, F. & Rosa-Jiménez, C. Intangible cultural heritage as a tool for urban and social regeneration in neighbourhoods. Participatory process to identify and safeguard ICH in the city of Malaga, Spain. *Int. J. Herit. Stud.* **29**, 524–546 (2023).
4. Gao, Y., Li, M., Li, Q., Huang, K. & Shen, S. Inheritors' happiness and its relevant factors in intangible cultural heritage. *Sustainability* **14**, 14084 (2022).
5. Xue, K., Li, Y. & Meng, X. An evaluation model to assess the communication effects of intangible cultural heritage. *J. Cult. Herit.* **40**, 124–132 (2019).

6. Kim, Y. & Byun, Y. C. Enhancing quality control in web-based participatory augmented reality business card information system design. *Sensors* **23**, 4068 (2023).
7. Su, X., Li, X., Wu, Y. & Yao, L. How is intangible cultural heritage valued in the eyes of inheritors? Scale development and validation. *J. Hosp. Tour. Res.* **44**, 806–834 (2020).
8. Yao, M., Liu, Z., Zhuang, L., Wang, L. & Li, H. A robust framework for one-shot key information extraction via deep partial graph matching. *IEEE Trans. Image Process.* **33**, 1070–1079 (2024).
9. Aumann, Y., Feldman, R., Liberzon, Y., Rosenfeld, B. & Schler, J. Visual information extraction. *Knowl. Inf. Syst.* **10**, 1–15 (2006).
10. Jung, K., Kim, K. I. & Jain, A. K. Text information extraction in images and video: a survey. *Pattern Recognit.* **37**, 977–997 (2004).
11. Guo, P. et al. DCMAI: A dynamical cross-modal alignment interaction framework for document key information extraction. *IEEE Trans. Circuits Syst. Video Technol.* **34**, 504–517 (2023).
12. Liu, J., Lin, L., Cai, Z., Wang, J. & Kim, H. J. Deep web data extraction based on visual information processing. *J. Ambient Intell. Humaniz. Comput.* **15**, 1481–1491 (2024).
13. Peanho, C. A., Stagni, H. & da Silva, F. S. C. Semantic information extraction from images of complex documents. *Appl. Intell.* **37**, 543–557 (2012).
14. Wu, T. et al. A brief overview of ChatGPT: the history, status quo and potential future development. *IEEE/CAA J. Autom. Sin.* **10**, 1122–1136 (2023).
15. Liu, C. et al. CPMI-ChatGLM: parameter-efficient fine-tuning ChatGLM with Chinese patent medicine instructions. *Sci. Rep.* **14**, 6403 (2024).
16. Xu, J., Zhang, H., Zhang, H., Lu, J. & Xiao, G. ChatTf: a knowledge graph-enhanced intelligent Q&A system for mitigating factuality hallucinations in traditional folklore. *IEEE Access* **12**, 162638–162650 (2024).
17. Kasneci, E. et al. ChatGPT for good? On opportunities and challenges of large language models for education. *Learn. Individ. Differ.* **103**, 102274 (2023).
18. Schuster, D. et al. Intellix--End-User trained information extraction for document archiving. In *Proc. 12th International Conference on Document Analysis and Recognition*, 101–105 (IEEE, 2013).
19. Dengel, A. R. & Klein, B. smartfix: a requirements-driven system for document analysis and understanding. In *Proc. International Workshop on Document Analysis Systems*, 433–444 (Springer, 2002).
20. Klink, S. & Kieninger, T. Rule-based document structure understanding with a fuzzy combination of layout and textual features. *Int. J. Doc. Anal. Recognit.* **4**, 18–26 (2001).
21. Katti, A. R. et al. Chargrid: towards understanding 2d documents. In *Proc. 2018 Conference on Empirical Methods in Natural Language Processing*, 4459–4469 (Association for Computational Linguistics, 2018).
22. Denk, T. I. & Reisswig, C. Bertgrid: contextualized embedding for 2d document representation and understanding. Preprint at arXiv https://doi.org/10.48550/arXiv.1909.04948 (2019).
23. Ha, H. T. & Horák, A. Information extraction from scanned invoice images using text analysis and layout features. *Signal Process. Image Commun.* **102**, 116601 (2022).
24. Kerroumi, M., Sayem, O. & Shabou, A. VisualWordGrid: information extraction from scanned documents using a multimodal approach. In *Proc. International Conference on Document Analysis and Recognition*, 389–402 (Springer International Publishing, 2021).
25. Liu, X. et al. Emotion classification for short texts: an improved multi-label method. *Humanit. Soc. Sci. Commun.* **10**, 1–9 (2023).
26. Jeong, S. et al. Real-time CNN training and compression for neural-enhanced adaptive live streaming. *IEEE Trans. Pattern Anal. Mach. Intell.* **46**, 6023–6039 (2024).
27. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
28. Zia, A. et al. Topological deep learning: a review of an emerging paradigm. *Artif. Intell. Rev.* **57**, 77 (2024).
29. Liu, Q. et al. A contrastive pretrain model with prompt tuning for multi-center medication recommendation. *ACM Trans. Inf. Syst.* **43**, 1–29 (2025).
30. Hong, T. et al. Bros: a pre-trained language model focusing on text and layout for better key information extraction from documents. *Proc. AAAI Conf. Artif. Intell.* **36**, 10767–10775 (2022). June.
31. Yang, Z. et al. Generative compositor for few-shot visual information extraction. *Pattern Recognit.* **165**, 111624 (2025).
32. Wang, J., Lin, Z., Huang, D., Xiong, L. & Jin, L. LiLTv2: language-substitutable layout-image transformer for visual information extraction. *ACM Trans. Multimed. Comput., Commun. Appl.* **21**, 1–27 (2025).
33. Zhu, P. et al. CCP-GNN: competitive covariance pooling for improving graph neural networks. *IEEE Trans. Neural Netw. Learn. Syst.* **36**, 6395–6406 (2024).
34. Prieto, J. R., Andrés, J., Granell, E., Sánchez, J. A. & Vidal, E. Information extraction in handwritten historical logbooks. *Pattern Recognit. Lett.* **172**, 128–136 (2023).
35. Cao, P. & Wu, J. GraphRevisedIE: multimodal information extraction with graph-revised network. *Pattern Recognit.* **140**, 109542 (2023).
36. Belhadj, D., Belaïd, A. & Belaïd, Y. Improving information extraction from semi-structured documents using attention based semi-variational graph auto-encoder. In *Proc. International Conference on Document Analysis and Recognition*, 113–129 (Springer Nature, 2023).
37. Gbada, H., Kalti, K. & Mahjoub, M. A. Multimodal weighted graph representation for information extraction from visually rich documents. *Neurocomputing* **573**, 127223 (2024).
38. Mizrahi, M. et al. State of what art? A call for multi-prompt LLM evaluation. *Trans. Assoc. Comput.Linguist.* **12**, 933–949 (2024).
39. Dai, M., Feng, Y., Wang, R. & Jung, J. Enhancing the digital inheritance and development of Chinese intangible cultural heritage paper-cutting through stable diffusion LoRA models. *Appl. Sci.* **14**, 11032 (2024).
40. Liu, J., Ma, X., Wang, L. & Pei, L. How can generative artificial intelligence techniques facilitate intelligent research into ancient books? *ACM J. Comput. Cult. Herit.* **17**, 1–20 (2024).
41. Zhang, J., Xiang, R., Kuang, Z., Wang, B. & Li, Y. ArchGPT: harnessing large language models for supporting renovation and conservation of traditional architectural heritage. *Herit. Sci.* **12**, 220 (2024).
42. Xu, L., Lu, L., Liu, M., Song, C. & Wu, L. Nanjing Yunjin intelligent question-answering system based on knowledge graphs and retrieval augmented generation technology. *Herit. Sci.* **12**, 118 (2024).
43. Stepchenkova, S., Kirilenko, A. & Yang, J. Capturing differences between culturally dissimilar audiences in the authentication of SMIs who organically promote destinations: the large language model approach. *J. Destin. Mark. Manag.* **35**, 100957 (2025).
44. Li, X., Li, W., Lu, L. & Fan, X. An artificial intelligence (AI) agent-based question-and-answer (Q&A) system for Tibetan Jiu chess. *ICGA J.* **47**, 13896911251343450 (2025).
45. El Idrissi, B. Utilizing chat generative pre-trained transformer (ChatGPT) to support the development of a domain ontology for world heritage. *Appl. Ontol.* **20**, 15705838251336686 (2025).
46. Ayash, L., Alhuzali, H., Alasmari, A. & Aloufi, S. Saudiculture: a benchmark for evaluating large language models' cultural competence within Saudi Arabia. *J. King Saud. Univ. Comput. Inf. Sci.* **37**, 123 (2025).
47. Zhou, J., Huang, J. X., Hu, Q. V. & He, L. Sk-gcn: modeling syntax and knowledge via graph convolutional network for aspect-level sentiment classification. *Knowl. Based Syst.* **205**, 106292 (2020).
48. Yan, C., Liu, J., Liu, W. & Liu, X. Research on public opinion sentiment classification based on attention parallel dual-channel deep learning hybrid model. *Eng. Appl. Artif. Intell.* **116**, 105448 (2022).

49. Yan, Z. et al. DocExtractNet: a novel framework for enhanced information extraction from business documents. *Inf. Process. Manag.* **62**, 104046 (2025).
50. Kim, G. et al. Donut: Ddocument understanding transformer without OCR. Preprint at arXiv https://doi.org/10.48550/arXiv.2111.15664 (2021).

Acknowledgements

This work was supported by the Humanities and Social Sciences Research Planning Fund Project of the Ministry of Education (No.24YJA630129); Shaanxi Provincial Social Science Fund (No.2024R055). Science Research Project of Shaanxi Provincial Education Department (No.25JK023).

Author contributions

All the authors contributed to the current work. R.Z.W.: conceptualization, data collection, methodology, writing-original draft preparation. X.S.Z.: investigation, funding acquisition, formal analysis, writing-review & editing. Q.L.L.: investigation, formal analysis. J.Q.S.: investigation. Y.Z.Z.: data curation, supervised. Y.L.M.: data collection.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Runzhou Wang.