Journal of Artificial Intelligence Practice (2026)
Clausius Scientific Press, Canada

DOI: 10.23977/jaip.2026.090204
ISSN 2371-8412 Vol. 9 Num. 2

Research on the Construction of Intelligent Q&A System for University Scientific Research Management Based on Retrieval-Augmented Generation

Yang Lei

Yunnan Open University, Kunming, China
271092522@qq.com

Keywords: Retrieval-Augmented Generation; Large Language Model; Research Management; Intelligent Question-Answering System; Educational Digital Transformation

Abstract: With the deepening of digital transformation in university scientific research management, traditional manual consultation services can no longer meet the growing demand for research management inquiries. Taking the research qualification review scenario at Yunnan Open University as a case study, this paper designs and implements an intelligent question-answering system based on Retrieval-Augmented Generation (RAG) technology. The system innovatively adopts a four-layer hybrid architecture integrating a Drools rule engine, vector retrieval, large language models, and knowledge graphs. By deploying open-source large models locally and constructing a multi-source knowledge base with vector and graph databases, the system effectively addresses the hallucination problem and insufficient interpretability of LLMs in the compliance review scenario of research management. This study provides a replicable, low-cost technical path and practical reference for the intelligent transformation of university research management services.

1. Introduction

In recent years, artificial intelligence technology represented by Large Language Models (LLMs) has achieved groundbreaking progress, profoundly transforming the service models and management paradigms in the field of higher education[1]. The clearly proposes advancing the strategic initiative of educational digitalization and building intelligent campuses and integrated education management and service platforms[2]. Against this backdrop, the digital transformation of service models for university research management — a core link supporting teachers' academic development and innovative research — has become one of the key topics in the construction of educational informatization[3].

2. Practical Dilemmas of Research Management Services and the Practical Application Value of RAG in the Field of Research Management

Current university research management services still face numerous practical dilemmas. On the one hand, research management systems and project application rules are highly complex and

dynamically updated, covering extensive professional specifications including professional title evaluation criteria, categorized project requirements, fund utilization regulations, and academic achievement accreditation standards[4]. Taking Yunnan Open University as a typical case, the university has more than 1,000 faculty members and undertakes over 30 types of research projects, receiving more than 20 research management consultations on average every day. Research management staff are chronically under high operational pressure. On the other hand, traditional research consultation modes rely predominantly on manual replies and manual retrieval of institutional documents, resulting in prominent problems including delayed responses, inconsistent service standards, and isolated information silos. These issues not only reduce service efficiency but also trigger potential management risks such as misjudgment of application eligibility caused by information asymmetry[5].

As a new-type higher education institution undertaking core responsibilities for open education, vocational education and continuing education across Yunnan Province, Yunnan Open University employs over 1,000 faculty members engaged in teaching, scientific research, administrative management and other professional roles. Each year, the university manages the application and review of more than 30 categories of research projects, including the National Social Science Fund of China, National Natural Science Foundation of China, Yunnan Provincial Philosophy and Social Science Projects, Yunnan Provincial Department of Education Projects, and university-level research projects. Field investigations summarize four major pain points in the current research management and consultation system. First, policy documents are fragmented and decentralized. Research management regulations are scattered across multiple departmental websites, internal circulars and paper manuals, making it difficult for faculty to acquire the latest and complete application information in a timely manner. Second, qualification verification procedures are extremely complicated. The application criteria for different project categories involve multi-dimensional constraints related to professional title, educational background, age, preliminary academic achievements, and the number of ongoing research projects. Manual review is prone to omissions of key clauses, resulting in a low project application success rate. Third, consultation response is inefficient. Research management staff are tasked with project review, fund administration, achievement statistics and other routine work, making it difficult to cope with concentrated consultation demands, with an average response time exceeding 24 hours. Fourth, historical research data is underutilized. Years of application experience and frequently asked questions (FAQs) have not been standardized into institutional knowledge assets, forcing newly recruited research administrators to repeat trial-and-error exploration.

In this context, the deployment of intelligent question-answering systems has become an effective approach to improve the overall quality of research management services. However, relying solely on general LLMs for research management consultation poses two core challenges. First, LLM hallucinations may generate erroneous information that deviates from official research management regulations, leading to substantial compliance risks in rigorous scenarios such as qualification verification. Second, general LLMs lack in-depth domain knowledge of the internal research management systems and standardized workflows of specific universities, failing to provide accurate, traceable and compliant judgment evidence[6].

Retrieval-Augmented Generation (RAG) technology provides a viable solution to the aforementioned problems. RAG retrieves policy document fragments relevant to user queries from external knowledge bases and feeds them as contextual prompts into LLMs. This framework enables the model to generate responses based on verified and authoritative knowledge sources, thereby substantially mitigating hallucination risks and improving the accuracy and traceability of output results[7]. Existing studies have verified the outstanding application value of RAG technology in various vertical domains such as medical teaching assistance[8]. In the field of higher education

management, RAG-based intelligent consultation systems have been preliminarily applied to library services and course Q&A scenarios[9][10]. Nevertheless, systematic research targeting the highly specialized and compliance-sensitive scenarios of university research management remains in its infancy, leaving considerable room for technological exploration and practical application.

3. System Demand Analysis and Architecture Design

3.1. Principles of Retrieval-Augmented Generation (RAG) Technology

Retrieval-Augmented Generation (RAG) technology has emerged as a pivotal research direction in the practical application of large language models in recent years. Traditional generative LLMs produce responses entirely based on internalized knowledge learned from pre-training parameters, which leads to inherent limitations including outdated knowledge boundaries, insufficient domain-specific expertise, and pervasive hallucination problems[11]. By embedding an external knowledge retrieval module into the model generation pipeline, RAG constructs a three-stage operational framework: retrieval, augmentation, and generation.

This study develops a RAG-powered intelligent question-answering system to realize precise push of research management policies, automatic pre-review of project application qualifications, real-time response to daily consultation questions, and structured accumulation of research management knowledge. Taking the qualification review workflow of Yunnan Open University as the practical application scenario, this study designs and implements a customized RAG-based intelligent consultation system. The core innovation lies in the construction of a novel four-layer hybrid architecture integrating the Drools rule engine, Milvus vector database, ChatGLM-2-6B large language model, and Neo4j knowledge graph. Specifically, the rule engine formalizes and precisely matches standardized research application constraints; vector retrieval enables semantic recall of unstructured policy documents; the knowledge graph supports associative reasoning of multi-source research data; and the large language model realizes natural language interaction and interpretable result generation. Adopting a fully open-source technology stack, the system features high replicability and popularization value. This paper addresses three core research questions: (1) How to construct a RAG system architecture adaptive to the complex rule system of university research management? (2) How to effectively suppress LLM hallucinations in research qualification review and guarantee the interpretability of system outputs? (3) What is the practical operational performance of the proposed system? This research aims to provide theoretical references and practical paradigms for the digital transformation of research management in peer higher education institutions.

3.2. Overall System Architecture Design

Following the software engineering principle of "high cohesion and low coupling", the system adopts a four-layer hybrid architecture, consisting of a data resource layer, a knowledge processing layer, a core service layer, and an application interaction layer in a bottom-up hierarchy, which is showed in figure 1.

(1) Data Resource Layer. This layer undertakes the collection and aggregation of multi-source heterogeneous research data. The data repository covers the MySQL-based research management system of Yunnan Open University, annual project application guidelines in PDF and Word formats, fund administration regulations issued by the Finance Department, and professional title evaluation documents released by the Personnel Department. The system synchronizes structured personnel and project data at regular intervals via API interfaces and captures and parses unstructured policy documents through customized crawlers and file parsing tools.

(2) Knowledge Processing Layer. Serving as the core knowledge hub of the entire system, this

layer comprises three functional processing engines. First, the text vectorization engine employs the bge-large-zh model to segment policy documents into 512-token semantic chunks and generate 768-dimensional feature vectors, which are stored in the Milvus vector database. Second, the knowledge graph construction engine utilizes Neo4j to store core entities (teachers, projects, academic achievements, management rules) and their interactive relationships, supporting complex multi-dimensional graph queries. Third, the rule formalization engine converts rigid quantitative and qualitative constraints in research management policies (e.g., "age ≤ 45 years old", "holding associate senior or above professional titles") into executable DRL rule files for the Drools engine.

(3) Core Service Layer. This layer deploys a locally deployed large language model (ChatGLM-2-6B) and a dedicated RAG scheduler. Upon receiving user queries, the RAG scheduler dynamically optimizes retrieval strategies based on intent recognition results. For explicit condition-matching queries (e.g., "Am I eligible to apply for the provincial youth project?"), the system preferentially invokes the Drools rule engine and knowledge graph for deterministic and accurate judgment. For open-ended consultative queries (e.g., "What are the key precautions for provincial social science fund application?"), the system adopts Milvus vector retrieval to obtain relevant document fragments and generates standardized, professional answers with the assistance of the large language model.

(4) Application Interaction Layer. Developed based on Vue3 and Element Plus, this layer provides Web and mobile H5 interactive interfaces, integrating natural language dialogue windows, project qualification self-assessment tools, and intelligent policy push dashboards. It enables faculty to acquire professional research management consultation services anytime and anywhere.

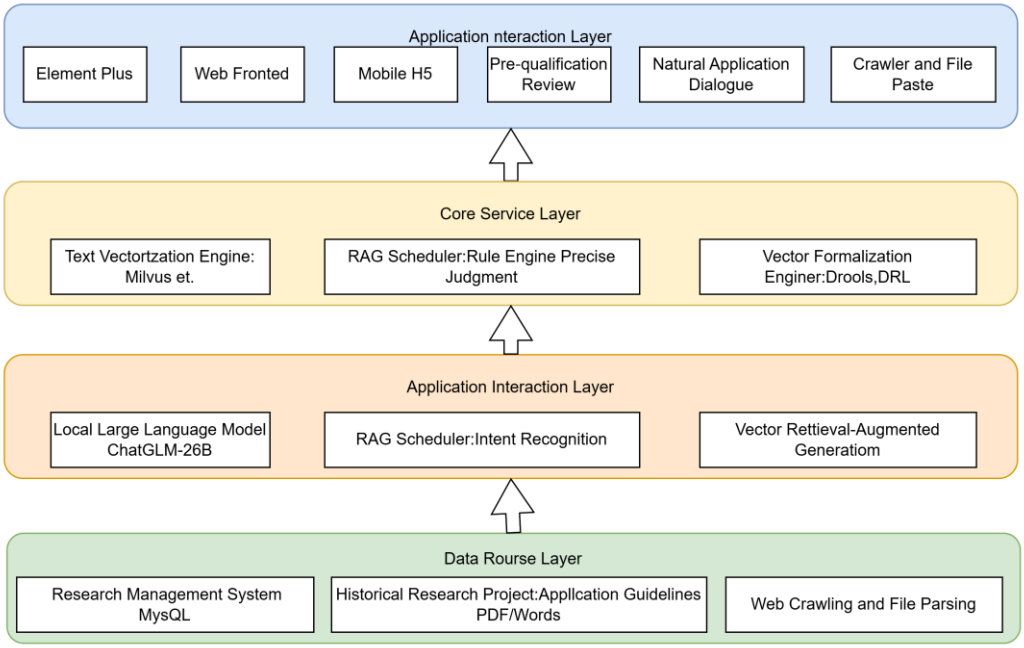


Figure 1 Four-layer hybrid architecture for Intelligent Q&A System for University Scientific Research Management Based on Retrieval-Augmented Generation

3.3. Functional Requirements and Non-functional Constraints

In terms of functional requirements, the system supports three core modules. First, intelligent interactive Q&A, which enables multi-turn dialogue, context memory, and voice input functions. Second, qualification self-test and evaluation, which automatically matches eligible research projects and calculates matching scores based on users' personal profile information. Third, intelligent policy interpretation, which extracts core key points and visualizes the content of lengthy project application

guidelines.

In terms of non-functional constraints, the system complies with the following technical specifications. First, the system response latency is limited within 2 seconds through vLLM inference acceleration and vector retrieval caching mechanisms. Second, full data security is ensured: personal faculty information and undisclosed project data are encrypted for storage, and all model inference processes are completed locally to eliminate data leakage risks. Third, the system supports scalable expansion, allowing seamless access to upgraded large language models (e.g., Qwen-2.5, Llama-3) and multi-modal data sources (e.g., research images and experimental data). Fourth, cost optimization is achieved by leveraging open-source ecological resources to reduce hardware procurement and cloud service operation costs.

4. Design and Implementation of Core System Modules

4.1. Construction of Multi-source Heterogeneous Scientific Research Knowledge Base

High-quality retrieval augmentation fundamentally depends on the standardization and completeness of the underlying knowledge base. This study constructs a full-process knowledge base development pipeline covering data cleaning, semantic chunking, vectorized storage, and knowledge graph association.

To address common defects of policy documents, including irregular layout, nested tables, and missing formula content, the system adopts a LayoutParser-based layout analysis algorithm to accurately identify and restore document hierarchies, paragraph contents and table structures. A dynamic sliding window strategy is further applied for text segmentation. Compared with traditional fixed-length chunking methods, the dynamic segmentation mechanism identifies natural paragraph boundaries and avoids splitting a complete application rule into discrete chunks. The segmented semantic chunks are imported into the bge-large-zh model to generate 768-dimensional vectors, which are stored in the Milvus database with Hierarchical Navigable Small World (HNSW) indexes built to support efficient approximate nearest neighbor search for million-scale vector datasets.

Meanwhile, a standardized research knowledge graph is constructed based on Neo4j. The graph schema defines four core entity types: Teacher, Project, Achievement, and Rule, along with multiple relational predicates including HasQualification, Hosts, Participates in, and BasedOn. For example, the entity of "Teacher Zhang San" is associated with "Associate Senior Professional Title" via the HasQualification relationship, and the "Provincial Youth Project" entity is linked to the constraint rule "age $\leq$ 35 years old" through the BasedOn relationship. By executing Cypher graph queries, the system realizes efficient multi-hop logical reasoning (e.g., retrieving teachers with doctoral degrees who have presided over university-level projects in the past three years), providing structured prior knowledge for LLM generation.

4.2. Application of Drools Rule Engine in Qualification Review

Research qualification review requires rigorous and deterministic logical judgment, and vague or uncertain outputs will bring substantial compliance risks. General LLMs excel at open-ended semantic interaction but suffer from inherent weaknesses in Boolean logic operation and numerical value comparison. Therefore, this system integrates the Drools rule engine as the deterministic hard-constraint module for qualification review. The rule engine workflow consists of three core steps: rule extraction, DRL formal transformation, and logical reasoning execution. First, all types of research management regulations are systematically sorted out, and natural language-based application rules are abstracted into standardized formal logical expressions.

During system operation, user-submitted personal information is encapsulated into Fact objects

and injected into the working memory of the Drools engine. If all pre-set application conditions are satisfied, the system outputs a "application-qualified" result with detailed rule basis; if any constraint is violated, the system specifies the unmet requirements explicitly (e.g., "You meet the professional title and age criteria, but the number of ongoing participated projects exceeds the limit, so you are not eligible for application"). This mechanism deeply integrates the generative capability of LLMs with the deterministic logic of the rule engine, fundamentally eliminating hallucinations and misjudgments in research qualification review.

4.3. Hybrid Retrieval Strategy Based on Vector and Knowledge Graph Fusion

Single retrieval methods are incapable of adapting to the diversified and complex scenarios of research management consultation. Vector retrieval performs well in semantic similarity matching but lacks advantages in accurate conditional query and multi-hop relational reasoning. Knowledge graph retrieval excels at structured logical inference but has limited generalization ability for unstructured natural language queries. BM25 keyword retrieval shows superiority in proper noun matching but cannot capture deep semantic information. To solve these limitations, this study proposes a three-way hybrid retrieval strategy integrating vector retrieval, BM25 keyword retrieval, and graph retrieval.

The system firstly conducts user intent recognition and query rewriting based on prompt engineering. The large language model converts colloquial user queries into standardized structured query statements. For instance, the informal query "I want to apply for a well-funded project" is rewritten into "projects with the highest funding quota in official application guidelines". Subsequently, three retrieval branches run in parallel. The vector retrieval module recalls Top-K semantically relevant text chunks from Milvus based on cosine similarity. The keyword retrieval module matches high-correlation documents from Elasticsearch via the BM25 algorithm. The graph retrieval module acquires structured entity relationships by executing customized Cypher queries on Neo4j.

All candidate results retrieved by the three branches are imported into the re-ranking module. The system adopts the bge-reranker-large cross-encoder model to conduct fine-grained scoring. The model takes query-document pairs as input, captures subtle semantic interactions via attention mechanisms, and outputs correlation scores ranging from 0 to 1. After re-ranking, only high-quality fragments with a correlation score greater than or equal to 0.8 are retained and concatenated as augmented context for LLM input. Experimental results show that the proposed hybrid retrieval strategy improves Recall@5 by 18.6% compared with single vector retrieval, effectively solving the problem of missing retrieval of long-tail policy documents.

4.4. Prompt Engineering and Interpretable Answer Generation for Large Language Models

After acquiring high-quality augmented contextual information, the system adopts meticulously designed prompt templates to guide the large language model to generate accurate, compliant and interpretable responses. The prompt template follows a Chain-of-Thought (CoT) logical structure, which includes role definition, task specification, contextual input, output constraints, and demonstration examples.

To enhance the interpretability and credibility of system outputs, the system extracts the metadata of cited documents (file name, page number, source link) during response generation and presents them as standardized references at the end of each reply. Furthermore, a self-reflection mechanism is embedded in the generation workflow. Before outputting the final result, the model verifies the factual consistency and logical accuracy of preliminary outputs, which further reduces factual errors and operational risks[12].

5. Conclusion

The practical verification of this study demonstrates that the application of RAG technology in university research management is not a simple technical superposition, but a systematic integration based on in-depth comprehension of business logic and targeted technical architecture adaptation. The collaborative mechanism of rule engine and large language model is the core approach to resolve compliance bottlenecks, while the hybrid retrieval and re-ranking strategy ensures high-quality knowledge acquisition. In addition, the low-cost open-source technology stack provides a realizable path for the inclusive digital transformation of university research management.

Nevertheless, this study has several limitations. On the one hand, the current system is mainly optimized for textual and tabular data, with limited analytical capability for complex graphical research materials such as technical roadmaps. On the other hand, restricted by the fixed context window of large language models, the system cannot fully understand ultra-long policy documents. Future research will introduce multimodal large language models and long-text sliding summarization technology to further expand the knowledge processing boundary and improve the comprehensive service capability of the system.

Acknowledgement

Key Project of Yunnan Open University University-level Fund: Research and Application of RAG-based Intelligent Question-Answering System — Taking Application Qualification Verification in Scientific Research Management as an Example (25YNOU05).

References

[1] Liu Bangqi, Yu Yankun, Wang Tao, et al. Large Educational Models of Artificial Intelligence: System Architecture and Key Technical Strategies[J]. Open Education Research, 2024, 30(5): 76-86. DOI:10.13966/j.cnki.kfjyyj.2024.05.008.

[3] Xie Panke, Guo Weixiu, Wu Shaojing, et al. Research on AI-Driven Smart Campus Service System — A Case Study of Central China Normal University[J]. Modern Educational Technology, 2025, 35(10): 118-126.

[4] Chen Jianlong, Wang Xincai, Shao Yan, et al. Integrated Resource Service Guidance Report of University Libraries[J]. Journal of Academic Libraries, 2025, 43(1): 38-48. DOI:10.16603/j.issn1002-1027.2025.01.005.

[5] Zhang Tian. Intelligent Question-Answering System for Libraries Based on Large Language Models[J]. Computer Programming Skills & Maintenance, 2026, (4): 109-111. DOI:10.16184/j.cnki.comprg.2026.04.018.

[6] Zhang Qiang, Qu Pengyu, Zuo Lina. Intelligent Q&A System Based on Large Language Model and Improved RAG[J/OL]. Computer Systems & Applications: 1-10[2026-08-16]. https://doi.org/10.15888/j.cnki.csa.010297.

[7] Liu Xueying, Yun Jing, Li Bo, et al. A Survey of Retrieval-Augmented Generation Based on Large Language Models[J]. Computer Engineering and Applications, 2025, 61(13): 1-25.

[8] Ding Ning, Song Yuxin, Shan Zetian, et al. Exploration on the Construction of Medical Teaching-Assisted Intelligent Q&A System Based on Retrieval-Augmented Generation (RAG) Technology[J]. China Medical Education Technology, 2025, 39(1): 1-5. DOI:10.13566/j.cnki.cmet.cn61-1317/g4.202501001.

[9] Pan Xuefeng. Construction and Application of University Library Intelligent Reference Consultation System Based on RAG Architecture[J]. Journal of Academic Library & Information Science, 2026. https://openurl.ebsco.com/contentitem/gcd:195323047?sid=ebsco:plink:crawler-gcd&id=ebsco:gcd:195323047&crl=c&jrnl=10061525.

[10] Wang Huazhen, Lin Haibin, Lin Xinyu, et al. Research on Construction of University AI Project Tutors under the Collaboration of Large Language Model and Knowledge Graph — Taking "AI Innovative Project Design" as an Example[J]. Digital Education, 2024, 10(5): 70-76.

[11] Sun Haoran, Wang Zhihao, Wu Yifan, et al. Educational Large Model Q&A Enhancement Method Based on Re-ranking and Post-Retrieval Reflection[J]. Big Data, 2025, 11(5): 4-17.

[12] Li Shengfei, Lu Yujie, Chen Xiaoying, et al. Implementation of Construction Engineering Management Knowledge Q&A Model Based on Retrieval-Augmented Generation[J]. Science Technology and Engineering, 2025, 25(25): 10840-10849.