Application Research on Intelligent Reference Services in Libraries Based on Large Language Models

Pingfeng Xie*

Library of Jiangxi University of Science and Technology. Jiangxi, 341000 China
**Corresponding author: 9120100038@jxust.edu.cn*

Abstract: *As the digital information environment grows increasingly complex and users' knowledge needs deepen, traditional library reference services face efficiency bottlenecks when addressing open-domain, multi-turn, and semantically complex inquiries. Large language models, built on the Transformer architecture, demonstrate profound capabilities in deep semantic understanding and generation, offering a new technological pathway for reshaping the paradigm of reference services. This study aims to systematically explore the theoretical foundation for integrating large language models with library reference services, construct a technically layered and decoupled architecture for an intelligent reference system equipped with domain adaptation and knowledge enhancement capabilities, and elucidate its core operational mechanisms. The research provides a detailed analysis of domain knowledge integration methods based on Retrieval-Augmented Generation and fine-tuning of large models, as well as context understanding mechanisms that support multi-turn dialogues and the integration of multi-source information. Furthermore, the paper discusses future directions for the evolution of intelligent reference systems from three dimensions: personalized services, the reconstruction of human-computer collaboration processes, and trustworthy assurance systems. This study offers a systematic theoretical framework and technical design concepts for libraries to leverage large language models in building a new generation of intelligent knowledge service infrastructure.*

Keywords: *Large Language Models; Library Reference Services; Intelligent Reference System; Retrieval-Augmented Generation; Human-Computer Collaboration*

Introduction

Library reference services are at a critical stage of transitioning from traditional information transmission to deep knowledge interaction. Confronted with information overload and the growing user demand for precise, personalized, and immediate knowledge services, the traditional digital reference model, based on keyword matching and static knowledge bases, reveals inherent limitations. As a significant advancement in the field of artificial intelligence, large language models, with their powerful natural language generation and contextual reasoning capabilities, offer revolutionary potential to address these challenges. However, effectively embedding general-purpose large language models into the highly specialized service environment of libraries, which emphasizes information accuracy and authority, is not a simple matter of technological application. It involves profound theoretical adaptation, complex technical architecture reconstruction, and service process reengineering. Therefore, systematically studying the integration logic of large language models and library reference services, designing a reliable, scalable, and sustainable intelligent system architecture, and anticipating its development path are of urgent theoretical significance and practical necessity for promoting the intelligent upgrade of library knowledge services and optimizing the academic information ecosystem.

1. The Theoretical Foundation for Integrating Large Language Models and Library Reference Services

1.1 The Core Technical Principles of Large Language Models and Their Information Processing Characteristics

The core technical principle of Large Language Models (LLMs) is based on the attention mechanism within the Transformer architecture. This mechanism enables deep representational

learning of massive text sequences through parallel computation and its ability to capture long-range dependencies. Their information processing characteristics are primarily manifested in the probabilistic generation of natural language and semantic understanding. Through a two-stage paradigm of pre-training and fine-tuning, the models acquire grammatical rules, factual knowledge, and logical associations from large-scale corpora. Compared to traditional retrieval models, LLMs demonstrate powerful contextual awareness and generative question-answering potential. They can transform unstructured user queries into internal representations and generate dynamic responses based on parametric knowledge. This technical characteristic provides a new computational paradigm for handling open-domain, multi-turn, and semantically complex reference inquiries in libraries[1].

1.2 The Evolutionary Path and Contemporary Challenges of Library Reference Services

The historical evolution of library reference services exhibits a clear trajectory of migration from physical desk inquiries to digital and networked services. Early service models relied on the individual professional knowledge of reference librarians and limited local collection resources. With the penetration of information technology, they gradually developed into asynchronous consultation formats based on database retrieval and email responses. Entering the digital information age, the core challenge facing these services has shifted from information scarcity to the contradiction between information overload and precise acquisition. User needs demonstrate a trend towards high personalization, immediacy, and deep knowledge integration. Traditional retrieval-based response mechanisms, which rely on keyword matching, exhibit significant limitations in addressing implicit needs, cross-disciplinary knowledge fusion, and creative knowledge services. How to transition from "information provision" to "knowledge construction and interpretation" has become a theoretical and practical bottleneck that contemporary reference service systems urgently need to overcome.

1.3 The Embedding Logic and Theoretical Adaptation of Large Language Models in Intelligent Reference Systems

The embedding logic of large language models in intelligent reference systems lies crucially in their deep integration, as the core reasoning engine, with the existing library knowledge infrastructure. This adaptation process is not a simple substitution; rather, it involves constructing a "user-model-collection" tripartite collaborative interaction framework based on cognitive computing theory. At the theoretical level, the generative capabilities of large language models must be combined with the principles of information accuracy, authority, and knowledge organization theories (such as ontology and knowledge graphs) long upheld by library science. Through technical pathways such as prompt engineering, Retrieval-Augmented Generation (RAG), and domain-specific fine-tuning, the model's generation scope is constrained, and structured domain knowledge is injected. This adaptation aims to endow the system not only with human-like conversational understanding and fluency but also with the ability to conduct trustworthy and traceable reasoning and responses based on the library's authoritative knowledge sources. Consequently, it realizes a paradigm upgrade in the consultation process, moving from simple question-answering to knowledge discovery within dynamic interactions.

2. The Technical Architecture and Core Mechanisms of the Intelligent Reference System

2.1 The Architecture Design of the Consultation Question-Answering Engine Based on Large Language Models

2.1.1 Core Module Decoupling and Collaborative Workflow

This architecture typically consists of a user interface layer, a dialogue management engine, a core large language model, a knowledge retrieval and augmentation module, and a security and ethics filtering layer. The user interface layer is responsible for multi-channel access and the standardized preprocessing of queries. The dialogue management engine, in turn, serves as the "command center" for session state, maintaining the contextual history of multi-turn dialogues and executing dialogue act planning, such as clarifying questions, confirming needs, or guiding the topic. After receiving the planned query and the retrieval-augmented context, the core large language model generates an initial response. The entire workflow operates asynchronously in a pipeline manner, with each module communicating through well-defined interface protocols, thereby ensuring the system's scalability and fault tolerance.

2.1.2 Technical Pathways for Domain Adaptation and Knowledge Enhancement

To overcome the inherent hallucination issues and knowledge obsolescence of large language models, domain adaptation is a critical step. This is primarily achieved through two complementary technical pathways: Retrieval-Augmented Generation (RAG) and targeted fine-tuning. In the RAG pathway, before generating a response, the system retrieves relevant information in real-time from the library's authoritative knowledge sources (such as the integrated library system OPAC, academic databases, self-built special collections, and standardized knowledge graphs) and provides this information as supplementary context to the model. This transforms the model's role from a "knowledge repository" into a "knowledge interpreter and synthesizer." The targeted fine-tuning pathway, on the other hand, uses high-quality library consultation dialogue corpora and specialized professional texts to perform supervised fine-tuning on the base model, or employs parameter-efficient fine-tuning techniques. This enables the model to deeply internalize the professional terminology, service logic, and response style of library science.

2.1.3 Interface and Process Design for Human-Computer Collaboration

Considering the uncertainty inherent in complex consultation requests, the architecture must reserve an interface for seamless human-computer collaboration. The system should possess metacognitive abilities, enabling it to evaluate the confidence level of its own generated answers. When the confidence level falls below a threshold, or when the question involves significant value judgments or in-depth academic analysis, the dialogue management engine should be capable of triggering a collaboration mechanism. It can then transfer the current session state, retrieved information, and the model's preliminary analysis to a human reference librarian's processing platform in the form of a structured work order, and subsequently reintegrate the response after the librarian's intervention. This design is not a simple handoff of tasks; rather, it achieves a streamlined collaboration between the intelligent system and the expertise of the librarian.

2.2 Construction of Multi-Source Information Fusion and Context Understanding Mechanisms

2.2.1 Unified Representation of Multi-Source Heterogeneous Information through Vectorization

To achieve efficient fusion, the system must first create a unified representation of heterogeneous information. Key information from text, tables, and even images is embedded into the same high-dimensional vector space using pre-trained models. This process is not a simple transformation based on a bag-of-words model; rather, it is an encoding process grounded in deep semantics, such that concepts with different expressions but similar meanings, such as "cardiovascular disease prevention" and "primary prevention of coronary heart disease," are positioned closely within this vector space. The library's self-built knowledge graph, meanwhile, provides a structured network of entities and relationships. Its embedding vectors can explicitly encode factual associations, thereby furnishing a logical framework for the integration process.

2.2.2 Context Modeling Based on Dynamic Attention Mechanisms

Understanding multi-turn dialogues goes beyond simply concatenating historical utterances. The system employs a dynamic attention mechanism to model the conversation history, distinguishing the contribution weight of different turns to the current query. For instance, a user's subsequent clarifying statement (e.g., "I mean research from the last three years") will receive higher attention, thereby dynamically adjusting the semantic vector of the query. Concurrently, the system maintains an "information state" tracker, which continuously records the entities, attributes, and intentions that have been confirmed or require clarification within the dialogue. This process forms a deep, abstract representation of the consultation session, guiding subsequent information retrieval and generation directions and preventing redundancy or topic deviation.

2.2.3 Knowledge Retrieval and Semantic Fusion Generation Strategy

Once the contextually enhanced query vector is obtained, the system simultaneously queries the vector database and the knowledge graph[2]. The vector database is responsible for coarse-grained recall based on semantic similarity, while the knowledge graph enables precise relationship inference and factual chain tracing. After the recalled multi-source information fragments undergo re-ranking for

relevance, deduplication, and conflict resolution, they are integrated into a coherent "augmented context package." When the large language model generates the final answer, its attention mechanism simultaneously focuses on the original question, the dialogue history, and this augmented context package. It performs a form of deep semantic fusion to generate a response that is both contextually coherent and strictly anchored to authoritative information sources.

2.3 Technical Pathways for Consultation Service Quality Assessment and Feedback Iteration

2.3.1 A Multi-Dimensional Automated Evaluation Indicator System

Evaluation must transcend reliance solely on human scoring and instead construct a set of quantifiable computational metrics. This includes:

a: Factual Metrics: Automatically comparing key claims in the response against trusted knowledge sources, or employing model self-checking techniques, to calculate a factual consistency score.

b: Faithfulness Metrics: Assessing the degree to which the generated content adheres to the retrieved reference sources, thereby preventing the model from over-relying on parametric memory and neglecting the provided evidence.

c: Completeness Metrics: Analyzing the response's coverage of all implicit sub-questions contained within the original consultation query, which can be calculated through question decomposition and answer matching.

d: Fluency and Usefulness Metrics: Although relatively subjective, these can be predicted by training specialized quality evaluation models or inferred by analyzing subsequent user interaction behaviors, such as whether the user poses an immediate follow-up question, as a proxy indicator.

2.3.2 Collection and Structured Processing of Mixed Feedback Data

Optimization and iteration rely on a high-quality stream of feedback data. The system must be designed with a mixed feedback collection mechanism that integrates both explicit feedback (such as five-star ratings and error flagging buttons) and implicit feedback (such as user clicks on specific citations, whether a session is terminated prematurely, or queries that are modified and resubmitted). This raw feedback data must be transformed into structured signals usable for model training. For instance, a specific negative rating, combined with the corresponding session context, can be labeled as a negative sample regarding the accuracy of a particular factual statement. Similarly, a user's sustained browsing behavior on a specific retrieved document can be quantified as a positive reinforcement signal for the relevance of that document to the query.

2.3.3 Iterative Optimization Based on Reinforcement Learning and Continuous Learning

Leveraging the processed feedback data, the system can employ a reinforcement learning framework for optimization. By modeling the dialogue process as a sequential decision-making problem, the model's action of generating a response is considered an action, while user satisfaction and subsequent interactions constitute the environmental reward. Through algorithms such as Proximal Policy Optimization, the model progressively learns strategies for generating higher-quality responses. Concurrently, the system establishes a continuous learning pipeline, periodically incorporating high-quality human-computer dialogues and user-corrected, high-quality responses as new training data for incremental fine-tuning of the model. This enables the system to adapt to new developments in academic discourse and emerging trends in user needs, thereby achieving self-evolution and capability enhancement.

3. Future Development Dimensions of Intelligent Library Reference Services

3.1 Directions for the Technical Realization of Personalized Consultation and Adaptive Learning

3.1.1 Modeling fine-grained user interests and cognitive states forms the basis for personalization.

The system needs to construct a continuously updated user interest vector and knowledge state profile by analyzing the user's historical consultation records, literature search behaviors, digital

resource access paths, and even learning process data within permissible boundaries. This profile not only describes the user's focused subject areas but can also infer their knowledge level, research paradigm preferences, and potential blind spots in their information needs.

3.1.2 Context-Aware Dynamic Resource Adaptation and Generation represents an application built upon this model.

When a user initiates a consultation, the system performs a contextualized interpretation of the current question by placing it within the user's long-term interest and cognitive profile. It then retrieves or generates answers that align with the user's knowledge level — for example, providing more conceptual explanations and background information for beginners, while focusing on cutting-edge developments and methodological discussions for experienced researchers[3]. Simultaneously, the system can proactively recommend prospective resources that are highly relevant to the user's long-term research interests but have not been explicitly searched for.

3.1.3 Adaptive Learning and Strategy Optimization during Interaction ensure the system's continuous improvement.

Each interaction is regarded as an opportunity to test and update the user model. By analyzing user feedback on recommended resources (such as adoption, ignoring, or in-depth reading) and the evolution of subsequent consultation questions, the system adaptively adjusts its recommendation strategy and depth of explanation. This approach makes the consultation dialogue itself a co-constructive learning process, achieving a synchronous evolution of the service and the user's development.

3.2 Role Reconstruction and Process Optimization in the Human-Computer Collaborative Consultation Model

3.2.1 Role Reconstruction

The shift from direct responding to process management and knowledge curation becomes the key to the transformation of librarian responsibilities. The intelligent system will undertake the immediate response to routine, factual, and basic analytical inquiries, thereby freeing up librarian resources and enabling them to focus on higher-value duties. These duties include in-depth involvement in complex academic consultations, organizing and curating interdisciplinary knowledge networks, establishing and supervising rules for intelligent consultation strategies and knowledge bases, and conducting quality reviews and ethical calibration of the system's output.

3.2.2 Process Optimization

Constructing an efficient collaborative pipeline between the intelligent agent and librarians is essential for realizing this role reconstruction. It is necessary to design standardized protocols for collaboration triggers and handovers. When the intelligent system identifies that a consultation request involves value judgments, ambiguous domains, creative synthesis, or when the user expresses a desire to be transferred, it should be capable of automatically generating a structured work order containing a session summary, analyses performed, and pending questions, and smoothly transferring it to the librarian's workstation. The librarian's intervention and decisions can then serve as high-quality feedback data, reciprocally training and optimizing the system's judgment and handover logic.

3.2.3 Interface Design

Supporting transparent interaction and the construction of a shared cognitive space is key to the collaborative experience. The consultation interface should clearly display the source evidence for answers, the system's confidence level, and instances where a librarian's corrections or additions have been incorporated, thereby enhancing the service's explainability and trustworthiness. This collectively builds a shared "cognitive space" for humans and machines, within which users, the intelligent system, and librarians can communicate and collaborate efficiently based on a consistent foundation of information.

3.3 Trustworthy Assurance and Sustainable Development Mechanisms for Intelligent Reference Systems

3.3.1 Trusted Computing and Output Verifiability Assurance is the cornerstone of technological trustworthiness.

This requires the system architecture to embed multiple verification mechanisms, including an automated process for fact-checking generated content against knowledge graphs, explicit labeling and traceability of information sources, and the design of uncertainty quantification modules that enable the system to proactively express its confidence level in answers to specific questions. At the algorithmic level, it is necessary to continuously combat bias and hallucinations, ensuring the neutrality and fairness of the service through diversified training data audits and debiasing techniques.

3.3.2 The cost-effectiveness and evolutionary model for sustainable operation focuses on the economic and technical feasibility of the system.

This involves optimizing the consumption of computational resources, for example, by exploring more efficient model architectures and adopting hybrid cloud deployment strategies to balance performance and cost. Simultaneously, it is necessary to establish a sustainable closed loop for knowledge updating, continuously integrating newly acquired library resources, academic outputs, and authoritative online information into the system's knowledge base through automated or semi-automated processes, thereby preventing knowledge obsolescence[4].

3.3.3 Organizational Adaptation and a Long-term Governance Framework serve as the institutional guarantee for integrating the system into the library ecosystem and continuously creating value.

This necessitates establishing a cross-departmental collaboration team within the library, responsible for the system's daily operation and maintenance, effect evaluation, and strategic adjustment. It requires formulating clear policies and ethical guidelines concerning data usage, privacy protection, service boundaries, and the attribution of responsibility. Furthermore, through regular technical audits and impact assessments, it is essential to ensure that the development direction of the intelligent reference system aligns with the library's core mission and societal expectations, thereby realizing its long-term, stable value in academic services.

Conclusion

This study systematically demonstrates the theoretical feasibility and technical implementation pathways for integrating large language models into the intelligent reference service system of libraries. The research indicates that by constructing a technical architecture that integrates Retrieval-Augmented Generation and domain-specific fine-tuning, the intelligent system can inherit the fluent conversational capabilities of large language models while effectively anchoring itself to the library's authoritative knowledge sources, thereby significantly improving its effectiveness in handling complex consultation requests.

Realizing this vision requires future work to focus on three interrelated dimensions: first, developing personalized consultation and adaptive learning mechanisms based on fine-grained user modeling, enabling the service to evolve in tandem with users' cognitive growth; second, deepening the design of roles and processes within human-computer collaboration models to construct efficient workflows that leverage the complementary strengths of intelligent agents and reference librarians; and third, establishing a trustworthy assurance and sustainable development framework encompassing technical verification, cost governance, and organizational adaptation, thereby ensuring the intelligent reference system serves the library's core mission in a long-term and stable manner. The continued pursuit of advancements in these directions will lead library reference services into a new phase of knowledge services that are more interactive, insightful, and inclusive.

References

[1] Fu Guorui, et al. "Research on the Construction and Application of Intelligent Reference Services in University Libraries Based on Large Language Models: A Case Study of Shandong University

Library." Library Journal, pp. 1-14.
[2] Gao Jiaqi. Research on the Construction of Service Models in University Libraries Empowered by Artificial Intelligence. 2025. Qufu Normal University, MA thesis.
[3] Wang Gefei. "Development Strategies for Intelligent Reference Services in University Libraries Based on Large Language Models." Library Work and Study, no. 7, 2025, pp. 80-87.
[4] Liang Xiang. "Application and Exploration of Intelligent Consultation in Public Libraries from the Perspective of Smart Libraries: A Case Study of Sun Yat-sen Library of Guangdong Province." Journal of Library Science, vol. 46, no. 6, 2024, pp. 93-97.