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: *With the development of the digital information environment and the high demand for knowledge, the old model of library reference service can no longer handle open-ended, multi-turn and semantic-rich questions. Large Language Models are based on the Transformer architecture, have achieved good results in deep semantic understanding and generation, and are now being used in all sorts of reference services. First, we will present the theoretical basis for integrating large language models with library reference services; next, we will construct a technically layered and decoupled architecture of an intelligent reference system that includes domain adaptation and knowledge enhancement; finally, we will analyze the working mode of this system. Researchers have found all kinds of ways to add domain knowledge in Retrieval-Augmented Generation and fine-tuning of large models, as well as context-understanding mechanisms that support multi-turn dialogue and the combination of information from various sources. Finally, some directions for the development of intelligent reference systems are also put forward in this paper, such as personalised services, reconstruction of human-computer collaboration and trustworthy assurance systems. Present the theoretical basis and technical design ideas for building a new generation of intelligent knowledge service systems in libraries based on large language models in this paper.*

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

Introduction

Library reference services are changing now; they are moving from providing information to fostering deep knowledge collaboration. Given the vast amount of information and changing needs of the people, the keyword matching and fixed knowledge base model of the old digital reference model can no longer provide timely and personalised knowledge services. Large Language Models are one of the most outstanding achievements in artificial intelligence at present, and they have good abilities in natural language generation and context awareness, etc. Given the high demand for accurate and reliable information in the specialised service environment of libraries, this is not only a technical problem but also a problem of quality. There are many theoretical modifications, complicated reconstructions of technical architecture and reorganisations of service processes. Therefore, it is necessary to systematically study the integration logic of large language models and library reference services, design a reliable, scalable and sustainable intelligent system architecture, foresee its development path, promote the intelligent upgrading of library knowledge services, and optimise the academic information ecosystem.

1. Theoretical Basis for the Integration of Large Language Models and Library Reference Services

1.1 The Basic Technical Ideas of Large Language Models and Their Information Processing Features

First of all, the foundation of Large Language Models (LLMs) is the attention mechanism in the Transformer architecture. The above way can learn representations of a large amount of text in parallel and is also good at capturing long-range dependencies. The main attributes of information processing by these are probabilistic generation of natural language and semantic understanding. Two stages of

pre-training and fine-tuning are used to learn grammar rules, facts and logical relations from a large corpus. LLMs can consider the context better and provide answers to the questions than the previous retrieval model. Organise the unstructured user query and, according to the parameters, obtain several corresponding answers from the knowledge base. Based on the technical characteristics mentioned above, a new calculation model has been introduced to handle open-ended, multi-turn and semantically rich reference questions in libraries[1].

1.2 The Evolutionary Path and Current Problems of Library Reference Services

In recent years, library reference services have moved away from the old model of desk consultation and introduced many types of online and network services. The first service mode was based on the professional knowledge of reference librarians and lacked local collection resources. With the spread of information technology, many places have begun to introduce asynchronous consultation forms that use databases and email replies. With the development of the digital information era, the main problem in recent years has been how to handle the large amounts of data accurately and in a timely manner. The demands of the users have changed; now they want more personalisation, immediacy and detailed information. The old way of answering is to match keywords; it cannot handle vague questions or complex cross-disciplinary integration. Information provision has reached the stage of knowledge construction and interpretation now, but at the same time, a problem has arisen in this process that new reference service systems urgently need to address.

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 should be in close coordination with the existing library knowledge infrastructure, and it will serve as the main reasoning engine. Do not assume too readily that the construction of the "user-model-collection" tripartite collaborative interaction framework is in line with the theory of cognitive computing at present. At the theoretical level, the generative ability of large language models should be combined with the principles of information accuracy, authority and knowledge organisation theory (such as ontology and knowledge graphs), which have been well-supported by library science for a long time. Prompt engineering, Retrieval-Augmented Generation (RAG), domain-specific fine-tuning and other methods have been used to restrict the generation scope of a model and organise domain knowledge in an organised way. The above goals are to enable the system to have human-like conversational comprehension and fluency, as well as the ability to perform reliable and traceable reasoning and response based on the authoritative knowledge in the library. Therefore, a new type of consultation has been introduced; it is no longer a simple Q&A mode but rather a way to learn.

2. Technical Architecture and Core Mechanism of the Intelligent Reference System

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

2.1.1 Core Module Decoupling and Collaborative Workflow

The sections of the general structure are the user interface, the dialogue management module, the main large language model, the knowledge retrieval and augmentation unit, and the security and ethics filter. The UI layer has many ways to get the query and normalises them. The dialogue management engine will also remember the state of the session and, in cooperation with dialogue act planning, handle all kinds of situations in multi-turn conversation, such as asking for clarification, requesting confirmation or changing the topic. Based on the planned query and the retrieved-augmented context, the main large language model produces the first answer. The whole process is asynchronous and pipeline-based, all modules have good interface protocols, so the system will have good scalability and fault tolerance.

2.1.2 Technical Paths for Domain Adaptation and Knowledge Augmentation

Hallucination and outdated knowledge are problems of large language models, so domain adaptation has been introduced to solve them. The two main ways to achieve this are Retrieval-Augmented Generation (RAG) and some fine-tuning. In the RAG path, before generating a response, relevant data from all kinds of reliable sources is obtained in real time, such as the OPAC of

the integrated library system, academic databases, custom special collections, standardised knowledge graphs, etc., and then this information is added as extra context for the model. Thus, the function of the model will be that of a "knowledge interpreter and synthesiser", not a "knowledge base". Therefore, for the fine-tuning path, a high-quality library of consultation dialogues and specialised professional texts can be used to perform supervised fine-tuning of the base model, or parameter-efficient fine-tuning methods can be employed. It will help the model learn the professional terms, service logic and other response modes of library science better.

2.1.3 Interface and Process Design for Human-Computer Collaboration

As most of the consultation problems are not known in advance, there should be a place for human-computer interaction in the system. The system should have a meta-cognitive function to ascertain how certain the answer it has just produced is. If the confidence level is too low, or if the question is very subjective and needs to be studied in detail academically, then the dialogue management engine should be able to switch to collaboration mode. Then it will pass the data of the current session, the retrieved information and the model's initial analysis to the human reference librarian's handling platform as an organised work order, and finally combine the response after the librarian has made adjustments. It is not only the division of labour but also the cooperation of the intelligent system and the professional skills of librarians.

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

2.2.1 Unified Representation of Multi-Source Heterogeneous Information via Vectorization

The first step in good fusion is to build a common-sense framework for all kinds of data. Key information from text, tables and even pictures are all in the same high-dimensional vector space after being processed by pre-trained models. It is not the simple conversion performed by a Bag-of-Words model; rather, it is an encoding based on deep semantics so that concepts with different expressions but similar meanings, such as "cardiovascular disease prevention" and "primary prevention of coronary heart disease", are close to each other in this vector space. At the same time, the library has also built its own knowledge graph. Embedding vectors can show the real relationship and give a reason for the integration process.

2.2.2 Context Modeling Based on Dynamic Attention Mechanisms

Understanding multi-turn conversation is not to add previous comments. Dynamically change the attention paid to the previous turns in the conversation history and assign different weights to various parts of the history for the current question. For instance, the user's next clarification (e.g., "I mean research in the last three years") will be given more weight and thus change the semantic vector of the query dynamically. At the same time, there is an "information state" tracker that continuously logs the entities, attributes and intentions of confirmed or pending clarification information in the dialogue. Therefore, we can build a detailed and abstract model of the consultation for the next stage of information retrieval and generation without redundancy or modification of the subject matter.

2.2.3 Knowledge Retrieval and Semantic Fusion Generation Strategy

Obtain the contextually augmented query vector, and at the same time, query the vector database and the knowledge graph[2]. The first kind of retrieval is carried out by a vector database for semantic matching, and a knowledge graph can be used to retrieve specific relations and facts. Re-rank the recalled fragments of information from all sources according to relevance, then deduplicate and merge them to form an "augmented context package". When the large language model produces the final answer, the attention mechanism will give more weight to the original question and the dialogue history, and this extended context package will also be taken into account at the same time. It is a kind of deep semantic fusion that can produce answers in line with the context and is based on official data sources.

2.3 Technical Path for Quality Assessment and Feedback Iteration of Consultation Services

2.3.1 A Multi-dimensional Automatic Evaluation Index System

The two kinds of information for evaluation are the scores of people and some other observable data. These are:

a: Factual Indicators: Automatically compare the main content of the answer with reliable knowledge bases or use model self-correction to determine how factually correct it is.

b. Faithfulness Metrics: Determine how well the generated content is in line with the retrieved reference materials, and do not allow the model to rely too much on parametric memory and ignore the provided evidence.

c. Completeness Metrics: Determine whether all the implicit sub-questions in the original consultation query have been covered by breaking down the question and matching it with an answer.

d: Fluency and Usefulness Metrics: These are relatively subjective, but a special quality evaluation model can be constructed, or one can observe how users behave at a particular time, such as whether a user immediately asks a follow-up question.

2.3.2 Collection and Structured Processing of Mixed Feedback Data

Based on the good feedback, make improvements and repeat. The two kinds of collections in the system will be explicit feedback collections and implicit feedback collections; that is to say, users can give explicit feedback (e.g., a five-star rating and an error report button) or provide implicit feedback (e.g., clicking on some citations, leaving early a session, modifying and resubmitting queries). Organize the raw feedback data in an orderly way and then train the model with this data. For instance, a certain negative rating and its corresponding session context can be considered a negative sample for the accuracy of a particular fact. The length of time a user has been viewing a particular retrieved document can also be considered a positive reinforcement signal that this document is relevant to the query.

2.3.3 Iterative Optimisation based on Reinforcement Learning and Continuous Learning

Therefore, Reinforcement Learning will be employed to optimise the processed feedback data according to the results mentioned above. The process of dialogue can be modelled as a sequence of decisions, and the model's action of generating a response is regarded as an action; user satisfaction and further interaction are considered environmental rewards. Proximal Policy Optimisation is a method that can be used to learn how to produce better answers incrementally. At the same time, new training data have been added regularly by adding high-quality human-computer conversations and user-corrected, high-quality answers, and thus the model has been continually improved. Therefore, the system will be able to handle new developments in academic thought and changes in people's demands, and will continue to be improved and developed.

3. Future Development Directions of Intelligent Library Reference Services

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

3.1.1 The basis of personalisation is to build a model of users' interests and knowledge

Based on the analysis of the above consultation records, literature search behaviour, digital resource access paths, learning process data, etc., an updated user interest vector and knowledge-state profile in line with the law have been constructed. This profile can show what people are interested in, give us some ideas about how knowledgeable they are, what research methods they prefer, and any shortcomings in their information needs.

3.1.2 Context-Aware Dynamic Resource Adaptation and Generation is an application based on the above model

As the user types the beginning of a question, based on the historical data it has gathered about the user's interests and thinking process, relevant suggestions will be displayed. Based on how much knowledge a user has, either a pre-written answer will be returned or one will be generated; for example, beginners will be given conceptual explanations and background information, and experienced researchers will be given new developments and methodological discussions[3]. At the same time, according to the long-term study plan of the user, some suitable materials that have not been searched for yet will also be proactively suggested by the system.

3.1.3 Adaptive Learning and Strategy Optimisation based on Interaction for Continuous Improvement of the System

Every time a user is used, it will help us improve our model of users. Based on the feedback users have given about the recommended materials (such as whether they have been used, ignored, or read completely), modify the following questions and adjust the recommendations and the level of explanation accordingly. Therefore, the consultation dialogue is in fact a joint construction of

knowledge through learning, and both the service provided and the development of user progress go hand in hand.

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

3.2.1 Role Reconstruction

The new direction of change for librarians' work is now to move from direct response to process management and knowledge organisation. The intelligent system will be able to answer most of the people's daily questions and simple inquiries promptly, reducing the time librarians need to spend on such matters. The tasks above are to provide all kinds of difficult academic consultation, organise and coordinate cross-disciplinary knowledge networks, establish and supervise rules for intelligent consultation strategies and knowledge bases, and carry out quality review and ethical calibration of the system's output.

3.2.2 Process Optimisation

Based on what has been achieved so far, a good cooperative pipeline for intelligent agents and librarians needs to be built. We will arrange a time for the cooperation trigger and handoff. When the intelligent system finds that the consultation request falls under the category of value judgement, is an ambiguous domain, requires creative synthesis, or the user has expressed a wish for transfer, it should be able to automatically create an organised work order that includes a session summary, analysis results and outstanding questions, and then hand over this work order at the librarian's station. Librarians' interventions and decisions can be used as good training data to improve the judgment and handoff logic of the system.

3.2.3 Interface Design

In order to make people feel a sense of belonging and have a place to share their thoughts, we need to cooperate. The consultation interface should show the original evidence of the answer, the degree of certainty of the system, and any corrections or additions made by librarians, so that the service is more transparent and reliable. Together, they form a common "cognitive space" for people and machines, and based on this foundation of unified information, users, intelligent systems, librarians, and so on can communicate and cooperate effectively.

3.3 Trustworthy Assurance and Sustainable Development Mechanisms for Intelligent Reference Systems

3.3.1 The first basis of technological trust is Trusted Computing and output verifiability assurance

There should be several ways to check the structure of the system; for example, automated fact-checking of the generated content can be done based on knowledge graphs, sources of information should be clearly marked and tracked, and modules for quantifying uncertainty need to be added so that the system can proactively indicate how certain it is about answering some questions. At the level of algorithms, we need to continuously address bias and hallucination to ensure that the service is fair and unbiased by carrying out all kinds of audits on the training data and debiasing methods.

3.3.2 Cost-Effectiveness and Evolutionary Model of Sustainable Operation refer to the economic and technical feasibility of the system

A good model and a hybrid cloud will be employed to reduce the amount of computing power and achieve a good balance of performance and cost. At the same time, a continuous update mechanism for knowledge should be established to automatically or semi-automatically add new library resources, academic results and other reliable online information to the knowledge base of the system and avoid the problem of outdated knowledge[4].

The organisation will improve the governance system, regularise it, provide institutional support for integration with the library ecosystem, and continue to add value.

Therefore, a cross-departmental working group will be established in the library to be responsible for the daily operation and maintenance, impact assessment, system planning, etc. Establish clear policies and ethical norms for the use of data, protection of privacy, scope of services, division of responsibilities, etc. Periodically carry out technical audits and impact assessments to ensure that the development direction of the intelligent reference system is still in line with its original purpose and the

actual needs of library users; thus, it will continue to provide good academic services.

Conclusion

Organise systematically the theoretical basis and technical implementation methods for using large language models in the intelligent reference service system of libraries in this paper. Based on the above research, the technical architecture of Retrieval-Augmented Generation and domain-specific fine-tuning can be built; thus, the intelligent system will inherit the strong conversational ability of large language models and be grounded in the reliable knowledge sources of the library to effectively answer all kinds of consultation questions.

In the future, the following three problems will be solved to achieve this goal: First, build personalised consultation and adaptive learning systems based on detailed user data to keep up with changes in people's cognitive abilities; Second, improve the design of roles and processes in human-computer collaboration models to create an efficient workflow that utilises the good points of intelligent agents and reference librarians; Finally, establish a trustworthy guarantee and sustainable development plan for technical verification, cost control and organisational adjustment to ensure that the intelligent reference system can continue to meet the basic needs of the library stably for a long time. The development of the above areas will raise the level of library reference services to that of knowledge services and make them 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.