MCP Service-Driven Intelligent Library Reference Consultation Pathways

Authors: Zhang Guangzhao, Wang Zhongyi, Jin Guanglong, Yang Fan, Zhongyi Wang

Date: 2025-07-09T09:44:20+00:00

Abstract

【Purpose/Significance】 Addressing the issues of insufficient intelligence, passive service models, and inefficient resource integration in library reference consulting services, this study explores an MCP (ModeIContextProtocol) service-driven intelligent transformation path, aiming to break down the barriers between large language models (LLM) and library heterogeneous resource systems through standardized protocols, construct a new paradigm of proactive, precise, and personalized intelligent services, and provide theoretical support and practical solutions for the intelligent development of libraries.

【Method/Process】 Based on the adaptability analysis of MCP technical characteristics and library business logic, this study constructs a four-layer interaction model of "Resources-MCP-LLM-Users" and designs a layered service system architecture that includes the MCP service layer, MCP protocol layer, foundation large model layer, and user layer. Through protocol parsing, tool encapsulation, and context management mechanisms, it enables LLM to achieve dynamic scheduling and knowledge integration of multi-source resources.

【Results/Conclusion】 The research finds that MCP services can dynamically extend the boundaries of intelligent reference consulting services according to library capabilities, providing an extensible theoretical framework and technical implementation guidelines for library intelligent transformation, and driving the evolution of public cultural services toward a proactive, knowledge-oriented direction.

Full Text

MCP-Driven Pathways for Intelligent Reference Services in Libraries

Zhang Guangzhao[12], Wang Zhongyi[1], Jin Guanglong[12], Yang Fan[2]

Abstract

[**Purpose/Significance**] To address the issues of insufficient intelligence, passive service models, and inefficient resource integration in library reference services, this study explores intelligent transformation pathways driven by the Model Context Protocol (MCP) service. The goal is to leverage standardized protocols to break down barriers between large language models (LLMs) and heterogeneous library resource systems, thereby constructing a new paradigm of proactive, precise, and personalized smart services, and providing both theoretical and practical solutions for the development of smart libraries.

[**Methods/Process**] Based on an analysis of the technical features of MCP and its compatibility with library business logic, a four-layer interaction model of "Resource-MCP-LLM-User" is constructed. A layered service system architecture is designed, comprising the MCP service layer, MCP protocol layer, foundational LLM layer, and user layer. Through protocol parsing, tool encapsulation, and context management mechanisms, dynamic scheduling and knowledge integration of multi-source resources by LLMs are achieved.

[**Results/Conclusions**] The study finds that MCP services can dynamically extend the boundaries of intelligent reference services according to library capabilities, providing a scalable theoretical framework and technical implementation guide for the intelligent transformation of libraries, and promoting the evolution of public cultural services towards proactivity and knowledge orientation.

Keywords: smart library; reference services; large language model; MCP

Classification Number: G252.61

1 Introduction

The 20th National Congress of the Communist Party of China emphasized optimizing service supply through deepening cultural institutional reforms to advance the building of a socialist cultural powerhouse. As a core component of the public cultural service system, libraries bear the important mission of providing knowledge and cultural services. Smart libraries have been formally included in the national 14th Five-Year Plan projects. As a core business function of libraries, the intelligent and smart transformation of reference consultation services represents a critical task in realizing smart libraries.

Recent research on large language models (LLMs) for library reference services has made significant progress, with main approaches falling into two categories: (1) LLM fine-tuning, where domain knowledge is used to fine-tune models for better understanding and more accurate responses; and (2) knowledge base retrieval, where domain knowledge bases are constructed, chunked into vector storage, and retrieved via similarity search to generate high-quality answers while mitigating hallucination issues. However, existing methods struggle to

achieve real-time integration with library business systems and third-party services, cannot provide one-stop problem-solving solutions deeply integrated with business operations, and cannot leverage third-party open tools to expand service capabilities.

MCP (Model Context Protocol), an open-source protocol introduced by Anthropic, aims to establish a standardized communication framework between LLMs and external systems. Unlike traditional API calls, MCP provides structured request-response patterns and context management mechanisms, enabling models to access and operate external resources in a consistent manner. Through unified interface specifications, this protocol achieves seamless integration between AI and databases, APIs, and file systems, offering new solutions to challenges facing library reference services.

This study explores MCP-driven pathways for the intelligent transformation of library reference services, constructs a "Resource-MCP-LLM-User"smart library reference service model, designs an MCP service architecture adapted to library scenarios, and analyzes its working mechanisms in intelligent reference applications. The research provides theoretical foundations and practical guidance for capability expansion and service model innovation in library intelligent reference services, promoting a shift from "capability-limited" to "capability-horizontally-extended" and from "traditional Q&A" to "one-stop problem solving."

This research employs a combination of literature review and system design, starting from MCP technical principles, analyzing the compatibility between library reference services and MCP integration, constructing theoretical models, designing service architectures, and exploring implementation pathways.

2.1 Research Status of Library Intelligent Reference Consultation

As a vital component of smart libraries, library intelligent reference consultation research has made considerable progress, which can be categorized into practical and theoretical studies.

Practical Studies primarily focus on answering user questions through corpus and knowledge base construction. Early research relied on rule engines, inverted index retrieval, and statistical models to build basic Q&A frameworks. For instance, Yao Fei et al. proposed using TF-IDF statistical models for question similarity matching combined with AIML rule engine technology. Hu Chaoming employed dictionary-based word segmentation and inverted index retrieval for question retrieval and ranking. Li Wenjiang et al. used AIML rule engines combined with dictionary-based word segmentation and inverted index retrieval. With the development of machine learning, researchers introduced classification algorithms like Support Vector Machines (SVM) to enhance intelligent processing, such as Li Xueting et al.' s use of SVM for user question classification combined with question paraphrasing methods. To overcome traditional limitations, word vector technologies (e.g., Word2Vec) were introduced to enhance semantic

understanding, as seen in Zhang Le et al.'s work using Word2Vec for synonym expansion and co-occurrence-based sentence similarity matching. Recently, breakthroughs in LLM fine-tuning technology have provided new paradigms for intelligent reference consultation. Wang Yihu et al. proposed LLM fine-tuning and knowledge base integration to reduce model hallucination. Guo Limin et al. suggested combining LLMs with ReAct patterns for dynamic reasoning to provide more intelligent and flexible information services. Zhang Guangzhao et al. proposed BERT-based classification and question similarity retrieval strictly based on knowledge base answers, combined with AIML multi-turn dialogue for user information collection and HTTP communication technology to achieve one-stop Q&A and business processing services.

Theoretical Studies have explored the role of ChatGPT-like AI in library services. Wang Yi et al. discussed automated classification, resource collection, and integrated resource consolidation. Zhang Peng examined the value and advantages of generative AI in library services. Wang Chao et al. explored ChatGPT' s powerful generation capabilities and immersive experiences combined with the metaverse. Pan Xuefeng discussed ChatGPT fine-tuning and integration with knowledge bases and databases to enhance intelligent consultation services. Wu Ruohang et al. examined how ChatGPT can empower library services from perspectives of personalized, precise, and diversified services. Guo Yajun et al. explored the logic, scenarios, and systems of LLM-empowered intelligent reference consultation. PANDA et al. discussed how ChatGPT can provide more accurate and personalized responses to user queries. RODRIGUEZ et al. explored how libraries can develop AI chatbots to meet growing demands for virtual reference services.

2.2 Overview of MCP Protocol Application Research

The Model Context Protocol (MCP), introduced by Anthropic in late 2024, is an open standard protocol designed to standardize interactions between AI models and external tools and data sources. MCP adopts a client-server architecture based on JSON-RPC 2.0, comprising three core components: MCP host, client, and server. It achieves unified interface standards through three primitives: Tools, Resources, and Prompts. The protocol' s core advantage lies in enabling AI agents to autonomously discover, select, and coordinate tools rather than relying on predefined tool mappings.

KRISHNAN' s research demonstrates that MCP can effectively solve context preservation and coordination disconnection issues in multi-agent systems, significantly improving performance in knowledge management and distributed decision-making tasks through standardized context sharing mechanisms. SZEIDER et al. achieved incremental verification and structured model editing through MCP' s standardized interfaces, addressing LLM limitations in formal logic reasoning and enabling handling of complex constraint optimization problems.

As an open protocol for standardized AI tool interaction, MCP has demonstrated significant potential in multi-agent systems, enterprise applications, and robot control. With continuous improvements in security mechanisms and ongoing technical architecture innovations, MCP is poised to become the foundational protocol for next-generation AI-native applications, driving AI toward greater autonomy and collaborative capabilities.

In summary, practice-oriented research has been largely confined to knowledge base Q&A models, unable to connect with business systems and third-party service capabilities. Methods using AIML rules integrated with business systems for one-stop user services face technical bottlenecks in text noise recognition and multi-system decoupling, limiting practical application effectiveness. In theoretical studies, some scholars overestimate LLM capabilities, neglecting their fundamental nature as generative models that lack proactive data collection, system integration, and external data source connection capabilities. These studies often lack feasible implementation pathways, remaining at the idealistic conceptual level of technology-library service integration, failing to recognize that LLMs alone cannot comprehensively address complex library service needs. However, with the release of the Model Context Protocol (MCP), new technical pathways and possibilities have emerged for LLM empowerment. Based on this context, this study systematically elaborates on how LLMs can dynamically expand intelligent reference service capabilities through MCP integration, and presents technical architectures and practical solutions for MCP-driven intelligent reference services.

3 The Theoretical Model of Library Reference Service Driven by MCP Service

3.1 Collaboration Principle Between MCP Service and Large Models

MCP, as an open standardized protocol, provides unified specifications for LLMs to acquire contextual information. Analogous to a "USB-C port for AI applications" —where USB-C provides standardized interfaces for connecting various peripherals—MCP offers standardized ways for AI models to connect with different data sources and tools. In library intelligent reference consultation scenarios, MCP enables efficient collaboration between LLMs and diverse library resource systems through a client-server architecture.

MCP employs a one-to-many client-server architecture, allowing a single library intelligent consultation system to simultaneously connect with multiple specialized servers. The architecture comprises five key components: (1) **MCP Host** carries the library' s intelligent consultation application (e.g., DeepSeek-based reference consultation system), serving as the AI tool that plans access to external data; (2) **MCP Client** maintains one-to-one connections with each server, managing protocol-level communications; (3) **MCP Server** acts as lightweight specialized programs, each exposing specific library service function modules through standardized protocols; (4) **Local Data Sources** encompass internal

resource systems, collection data, and business services; and (5) **Remote Services** connect to external academic resources and API services through the MCP service marketplace.

In library application scenarios, different MCP servers assume specialized responsibilities, forming a complementary service matrix. Literature retrieval servers connect to library OPAC systems and various document databases, providing precise resource discovery; user service servers manage user information, borrowing history, and personalized preference data; subject service servers integrate professional vocabularies, citation networks, and academic evaluation tools; external resource servers connect to open-access journals, academic search engines, and knowledge graph services via APIs. LLMs can dynamically select and combine functions from different servers according to consultation needs, achieving cross-resource intelligent integration.

When users submit reference consultations, the system initiates a standardized collaboration workflow. The LLM first analyzes consultation content, identifies required information types, and plans execution strategies. The MCP client routes requests to corresponding servers: sending retrieval instructions to literature retrieval servers, querying personal preferences from user service servers, and obtaining professional term mappings from subject service servers. Servers process requests in parallel, returning retrieval results, user profiles, and subject knowledge respectively. The LLM integrates multi-source information to generate personalized, professional consultation responses. The entire process ensures standardized and secure data transmission through the MCP protocol. The collaboration principle is illustrated in Figure 1 [Figure 1: see original paper].

3.2 Construction of Intelligent Reference Service Model

Based on MCP's standardized communication framework and synergistic advantages with LLMs, this study proposes a four-layer theoretical model for smart library reference services: "Resource-MCP-LLM-User."This model uses resource integration as its foundation, MCP as its hub, LLM as its intelligent core, and user needs as its driver. Dynamic interaction among the four layers enables the paradigm shift from static resource management to intelligent, personalized knowledge services.

As shown in Figure 2, the overall logical framework and collaboration mechanism consists of four components: (1) **Resource Layer** integrates multi-source heterogeneous library data (e.g., collection data, business systems, remote business systems, network resources, open knowledge bases) into unified semantic resource objects through standardized abstraction, laying the data foundation for intelligent services; (2) **MCP Layer** serves as the communication and mediation core for LLMs, managing bidirectional data flow. It encapsulates multi-source resources into standardized toolsets (e.g., collection retrieval, borrowing queries) while maintaining interaction states and user profiles for session context

management; (3) **LLM Layer** relies on domain-knowledge-enhanced language models to intelligently parse user natural language needs, dynamically invoke various tools under the MCP layer, achieve resource scheduling and knowledge integration, and generate personalized, context-aware consultation responses; and (4) **User Layer** interacts with the system through multi-modal interfaces (mobile terminals, intelligent assistants, etc.). Users input personalized needs, the system processes them intelligently, and outputs precise services, forming a closed-loop of "demand input-intelligent processing-service output"that connects resources, protocols, intelligence, and users comprehensively.

4 Architecture Design of Library Reference Consultation Service System Driven by MCP

4.1 Overall Architecture Design

Based on the "Resource-MCP-LLM-User" theoretical model, this study designs a "four-layer separation, vertical-horizontal integration" intelligent reference service system architecture (shown in Figure 2). The architecture adopts layered decoupling as its core design philosophy, comprising four layers: (1) **User Layer** includes various mobile devices and computers as interaction interfaces; (2) **LLM Layer** responsible for understanding user needs, planning execution strategies, and summarizing data output; (3) **MCP Protocol Layer** provides execution protocols between LLMs and MCP services, plus MCP service management, connection, and execution; and (4) **MCP Service Layer** provides resource responses for various professional knowledge, intelligent processing, and business transactions. This creates a complete chain from data resources to intelligent services with flexibility, security, and scalability.

4.2 User Layer

The user layer provides multi-modal interaction entry points for the intelligent reference consultation system, supporting interaction through voice, text input, and attachment uploads. User groups encompass general readers, researchers, educators, and other stakeholders who can express information needs through natural language. LLMs can accurately understand user intentions and respond in various forms including text, tables, charts, audio, or video to enhance interaction experience. Users can intuitively perceive intelligent outcomes during service usage, and evaluation and feedback mechanisms drive continuous system optimization. The user layer adopts unified API design, supporting flexible deployment across multiple channels including library websites, mobile applications, and social media, ensuring service consistency and broad accessibility.

4.3 LLM Layer

The LLM layer serves as the intelligent core of the entire architecture. Through domain-adaptive training, it possesses deep understanding of library professional knowledge and reference consultation service capabilities. Relying on advanced

natural language understanding technology, it accurately captures and parses user intentions, automatically decomposes complex consultation needs into executable operation steps, and intelligently generates tool invocation plans containing parallel and serial tasks. The LLM then initiates standardized requests through MCP to schedule required resources and tools. For example, when addressing academic literature research needs, the model can sequentially invoke tools for subject term mapping, advanced retrieval, literature analysis, and citation network construction to form a complete task chain. MCP service objects receive and process these requests, returning execution results to the LLM. The LLM achieves multi-round dynamic interaction during tool invocation, continuously adjusting subsequent calls based on each tool's feedback to ensure high flexibility and targeting of the consultation process. Based on tool-returned data, the LLM performs knowledge integration and logical reasoning to ultimately generate structured, content-rich, and highly professional responses, providing users with intelligent and precise reference consultation services. The entire interaction process supports multi-round progression, with the LLM automatically loading historical session records to continuously deepen understanding of user needs, gradually achieving knowledge refinement and service personalization. MCP service' s context management mechanism ensures information consistency and coherence across multi-round dialogues, making the interaction experience as natural as conversing with professional librarians.

4.4 MCP Protocol Layer

The MCP protocol layer serves as the central hub of the intelligent reference consultation service system, undertaking multiple core functions including protocol parsing, resource scheduling, security control, and session management. It is the key bridge enabling efficient collaboration between LLMs and various library services. Designed with high cohesion and strong decoupling principles, this layer primarily comprises three functional modules:

(1) **Protocol Manager** responsible for dynamic pluggable management of MCP services, centrally maintaining and publishing capability descriptions and data protocol specifications for all MCP services. It synchronizes enabled service function information to LLMs in real-time, ensuring accurate perception and invocation of available tools and resources. Upon receiving standardized requests initiated by LLMs based on JSON-RPC 2.0 specifications, the protocol manager handles parsing, validation, and authentication, routing requests to corresponding resource or tool modules. It supports both synchronous and asynchronous invocation modes, ensuring efficient and secure request processing.

(2) **Tool Execution Encapsulator** responsible for further protocol encapsulation and parsing of requests routed by the protocol manager. It transforms LLM execution instructions into standard interface call formats understandable by MCP servers according to JSON-RPC 2.0 specifications, and normalizes execution results into unified data structures to feed back

to LLMs. This module supports concurrent invocation and result integration of multiple tools and resources, enhancing system execution efficiency and scalability.

(3) **Context Processor** focuses on maintaining user session states and profiles, supporting cross-round and multi-session request management. It records and tracks user historical interaction information, current dialogue states, and relevant parameters, ensuring each tool invocation aligns closely with user actual needs and historical behaviors, thereby improving service coherence and personalized experience. The context processor also collaborates with security policies to prevent unauthorized access and sensitive data leakage, providing robust contextual support for multi-round intelligent interactions and complex business processes.

Through collaborative operation of these three core modules, the MCP protocol layer achieves efficient, secure, and flexible connection between LLMs and MCP services and library resources. It simplifies integration complexity between LLMs and heterogeneous services, laying a solid foundation for subsequent service expansion, permission management, and data protection, while fully ensuring the stability, scalability, and security compliance of the entire intelligent reference consultation service system.

4.5 MCP Service Layer

The MCP service layer serves as the solid foundation of the overall architecture, responsible for integrating and managing multi-source heterogeneous data resources and service capabilities inside and outside the library. Through modular and pluggable approaches, it flexibly expands the service boundaries of intelligent reference consultation systems. This layer embraces service decoupling and standardization as its core philosophy, encapsulating various resources and functions into independent service modules that seamlessly connect with the protocol layer through MCP protocol specifications, enabling dynamic service access and expansion.

This layer encompasses various core service accesses including: (1) **Collection Resource Database** integrating OPAC systems, institutional repositories, and electronic resource platforms to uniformly manage various print and electronic collection data, supporting efficient literature retrieval and access; (2) **User Behavior Database** aggregating search logs, borrowing records, consultation history, and other behavioral data to support user profile construction and personalized service recommendations; (3) **Knowledge Rule Database** covering subject ontologies, citation networks, professional vocabularies, and other structured knowledge systems, providing support for intelligent applications like semantic retrieval, automatic classification, and knowledge reasoning; (4) **Domain Knowledge Base** incorporating library regulations, service scopes, and characteristic literature Q&A to strengthen intelligent response capabilities for local library operations and characteristic service scenarios; (5) **Library**

Business Systems interfacing with seat reservation, book renewal, research integrity detection service information collection, and status query systems to achieve intelligent processing and progress tracking of users' daily service needs, enhancing business linkage and automation levels of consultation systems; and (6) **Open MCP Services** connecting to external intelligent services provided by the MCP service marketplace such as academic resource search engines and public data query services, further enriching system service capabilities and application scenarios.

Each service module adheres to JSON-RPC 2.0 specifications, enabling flexible and dynamic plugging into the MCP protocol layer to achieve standardized integration with the upper-layer intelligent reference consultation system. Through this mechanism, libraries can continuously expand service content and scope according to business development and user needs, dynamically adjust and optimize service boundaries of intelligent reference consultation systems, and enhance system flexibility, maintainability, and sustainable innovation capabilities.

5 Implementation Paths and Challenges of MCP-Driven Library Reference Services

5.1 Technical Support Challenges and Solutions

(1) Technical Support Challenges

MCP service-driven intelligent reference consultation faces multiple technical support challenges: (a) **Technical Complexity** - As an emerging technical specification, MCP requires complex middleware design and adaptation work to achieve seamless integration with traditional library management systems, institutional repositories, and digital resource management systems. Interface compatibility issues are particularly prominent when dealing with unstructured data and legacy systems. (b) **Infrastructure Requirements** - MCP service and LLM collaboration demands substantial computing resources, including high-performance servers, adequate storage space, and stable network environments. Many small and medium-sized libraries cannot afford these infrastructure investments, resulting in high technical barriers. Additionally, ensuring system stability and response speed requires establishing comprehensive load balancing, disaster recovery, and elastic scaling mechanisms, further increasing deployment complexity. (c) **Professional Talent Shortage** - MCP service deployment, maintenance, and optimization require interdisciplinary talent proficient in AI technology, distributed systems, and library operations. However, library IT staff currently lack sufficient expertise in LLMs and MCP, making it difficult to meet continuous system maintenance needs. Moreover, the contradiction between rapid technology iteration and long-term system evolution poses challenges. LLM technology and MCP are still in rapid development, and frequent version updates may cause system compatibility issues. Libraries need to maintain system stability while continuously adapting to technological evo-

lution, imposing higher requirements on forward-looking design and flexibility of technical architecture.

(2) Solutions

To address technical support challenges in MCP protocol-driven AI intelligent applications for library reference services, we recommend a three-pronged approach focusing on industry-academia-research collaboration, infrastructure optimization, and talent cultivation to advance MCP service deployment and sustainable development in phases, effectively reducing technical implementation risks:

(a) **Deepen Industry-Academia-Research Collaboration** to address technical complexity challenges. Libraries can partner with academic institutions and enterprises to conduct in-depth research on MCP protocol integration with existing library management systems, institutional repositories, and digital resource management systems. Through multi-party collaboration, develop customized middleware to enhance data compatibility and interface adaptation capabilities between heterogeneous systems, particularly achieving technical breakthroughs in unstructured data processing and legacy system transformation. Industry-academia-research collaboration can introduce cutting-edge technologies and innovative solutions while providing continuous technical support and upgrade guarantees for libraries, facilitating successful MCP service deployment.

(b) **Leverage Campus LLM Platforms** to optimize infrastructure investment. To address computing resource and operational pressures faced by small and medium-sized university libraries, we recommend fully utilizing locally deployed campus LLM platforms (such as DeepSeek) as the underlying support for MCP services. By connecting with campus LLM resources, libraries can significantly reduce costs and technical barriers of building high-performance computing platforms, achieving resource sharing and collaborative innovation. Additionally, load balancing, elastic scaling, and disaster backup mechanisms should be implemented to ensure high availability and system stability of MCP services, enhancing user experience.

(c) **Systematically Advance Professional Talent Cultivation and Capacity Building**. To alleviate talent shortages, libraries should develop specialized training programs focused on AI and distributed systems to enhance existing IT staff' s understanding and application capabilities of LLM technology and MCP. By introducing open-source software and low-code/zero-code tools, MCP service deployment and maintenance difficulty can be reduced, gradually building the library' s own AI technical team. Simultaneously, continuous learning and knowledge updating mechanisms should be established to enhance team adaptability to new technologies, achieving long-term sustainable development of MCP services.

Through industry-academia-research collaborative innovation, infrastructure optimization, and talent pipeline development, libraries can effectively overcome multiple technical challenges in MCP service application and steadily advance the intelligent and smart transformation of reference consultation services.

5.2 Implementation Plan Design

The technical implementation plan for MCP-empowered library reference consultation services must balance feasibility and practicality. Overall implementation is divided into three steps:

(1) **Containerized Deployment of MCP Server Core Components**. We recommend containerized deployment strategies, encapsulating core MCP service components (e.g., protocol parsers, tool managers) as Docker containers. Mainstream open-source MCP service core components such as ChatMCP[1], AIaW[2], and FLUJO[3] can be selected for system construction. This approach reduces dependence on library IT technical capabilities and infrastructure while effectively achieving application environment isolation and enhancing system security and portability. Libraries can flexibly choose between single-machine or cluster deployment schemes based on actual scale and needs, achieving reasonable balance between performance and cost.

(2) **Intelligent Model Selection and Integration**. For daily reference consultation services, we recommend prioritizing lightweight open-source models (e.g., deepseek-r1-7B, ChatGLM-6B) for efficient and low-cost intelligent Q&A. For complex scenarios like academic research assistance, higher-performance professional models (e.g., deepseek-70B or adaptively fine-tuned versions) can be deployed. Libraries can flexibly configure model resources according to their technical capabilities and funding conditions. For university libraries with weak technical strength and limited funding, priority should be given to accessing existing basic LLMs deployed on campus, with the university responsible for LLM deployment and maintenance, further reducing technical barriers and operational burden.

(3) **MCP Service System Integration and Expansion**. When integrating with existing library systems, we recommend loose coupling architecture, achieving efficient communication between heterogeneous systems through a middleware layer. The middleware can connect to library management systems via REST or GraphQL APIs while providing required JSON-RPC 2.0 interfaces for the MCP protocol layer. To meet high-concurrency service demands, the system can adopt Redis-based distributed session storage mechanisms for service state management and support multi-channel real-time interaction through WebSocket interfaces. For libraries with limited technical capabilities and funding, implementation can start from simple application scenarios, such as using open-

source frameworks like haystack , RAGFlow , and Diffy to implement basic knowledge base Q&A functions, and building literature resource knowledge graphs and knowledge graph-based precise Q&A services through LightRAG and GraphRAG . Meanwhile, open academic search services from the MCP service marketplace can be utilized. Libraries can gradually promote software service providers to offer local data MCP service interfaces or independently develop local data access services based on actual conditions, achieving continuous expansion and localized deployment of MCP service capabilities. The detailed implementation plan is shown in Figure 4 [Figure 4: see original paper].

[1] https://github.com/daodao97/chatmcp
[2] https://github.com/NitroRCr/AIaW
[3] https://github.com/mario-andreschak/FLUJO
https://github.com/deepset-ai/haystack
https://github.com/infiniflow/ragflow
https://github.com/opendiffy/diffy
https://github.com/HKUDS/LightRAG
https://github.com/microsoft/graphrag

Libraries can flexibly advance implementation based on the above plan and actual funding conditions. For university libraries with limited funding, they can apply to campus network centers for virtualized Linux servers, access the DeepSeek LLM deployed by the university, and utilize open-source software to build the overall service architecture, achieving low-cost or even zero-cost service deployment. When conditions permit, they can gradually conduct independent model deployment and integrate local business system interfaces to continuously expand and improve service capabilities.

5.3 Data Security Challenges and Protection Strategies

(1) Core Security Challenges

(a) **Privacy Leakage Risk**. During LLM-MCP service interactions, user consultation content (e.g., borrowing records, research topics) faces leakage risks. If intermediate data is not encrypted during storage and transmission, it may be intercepted by malicious parties or reconstructed through reverse engineering, leading to exposure of user privacy information.

(b) **User Permission Grading and Control Difficulties**. Different user types (e.g., ordinary users, faculty, library administrators) have varying access permissions to sensitive data. For example, ordinary users can only query their own borrowing information, while administrators can access all user data for statistics and management. If permission control mechanisms are not robust, sensitive data may be illegally obtained or misused by unauthorized users, affecting library data asset security.

(c) **Model Jailbreak and Permission Escalation Risks**. Malicious users may craft carefully designed prompts to induce LLMs to bypass permission controls (e.g., executing administrator operations for ordinary users). Additionally, when models autonomously invoke related services, unreasonable permission boundary settings may cause them to exceed expected permission scopes, resulting in unauthorized access and sensitive operations that pose serious security threats.

(2) Multi-Level Protection Strategies

(a) **Enhanced Data Encryption and Privacy Protection**. All user data in LLM-MCP service interactions should be encrypted using industry-standard algorithms (e.g., TLS, AES) during transmission and storage to prevent data theft during transmission or unauthorized access during storage. Simultaneously, access permissions to intermediate data should be strictly restricted, and sensitive information should be de-identified and desensitized to reduce risks of data reconstruction and privacy leakage.

(b) **Fine-Grained User Permission Management**. Establish robust hierarchical permission control mechanisms that assign access permissions based on user identity (ordinary user, faculty, administrator, etc.), ensuring sensitive data can only be accessed by authorized users. Design and implement fine-grained access control strategies such as Role-Based Access Control (RBAC), and regularly review and update permission configurations to prevent permission abuse or unauthorized access. Operations involving whole-library data aggregation and asset statistics require enhanced approval and operation logging.

(c) **Prevention of Model Jailbreak and Permission Escalation**. Strictly filter and review LLM input and output content to prevent malicious prompts from inducing models to execute operations beyond permission scopes. Introduce prompt detection and response mechanisms to timely identify and intercept suspicious requests. Simultaneously, restrict permission boundaries for model-calling backend service interfaces to ensure models can only operate within authorized scopes. Regularly assess model security through red team testing that simulates jailbreak attacks to discover and patch potential vulnerabilities.

(d) **Security Auditing and Monitoring**. Establish comprehensive security auditing mechanisms for real-time monitoring and logging of critical system operations, user behaviors, and model invocations. Regularly analyze audit logs to promptly detect anomalous behaviors and security incidents for rapid response and handling. Strengthen security awareness training to enhance recognition and prevention capabilities of operational and development personnel regarding relevant risks.

5.4 Analysis of Core Application Scenario Working Mechanisms

Focusing on several core application scenarios of library services, MCP-driven intelligent reference consultation systems can achieve the following working mechanisms:

(1) **Library Regulations and Service Scope Consultation**. The system integrates library regulations, service processes, opening hours, borrowing policies, etc., through the open-source platform Diffy to establish domain knowledge bases and set system prompts to create an intelligent agent. Users can pose questions in natural language, and the system leverages LLM semantic understanding capabilities to accurately identify user intentions, retrieve relevant regulation content, provide real-time standard answers, or guide users to obtain detailed service information as needed.

(2) **Academic Misconduct Research Output Detection Information Collection and Direct Business System Push**, referencing Zhang Guangzhao' s proposal for MCP servitization of academic misconduct detection services based on HTTP communication technology. For users' academic misconduct detection needs, the system provides not only intelligent Q&A regarding detection processes, specifications, and service entry points but also seamlessly integrates with business systems to achieve full-process information connection for detection applications, progress queries, and result feedback. When users submit detection requirements and upload attachments, the system automatically collects required information, assists in filling business forms, and directly pushes data to the academic misconduct detection system, significantly improving service efficiency and user experience. Simultaneously, the system can track detection progress in real-time and provide timely feedback, achieving closed-loop management of service processes.

(3) **Intelligent Q&A for Collection Literature Resources**. The system deeply integrates diverse resources including collection catalogs, electronic literature, and subject databases to provide efficient literature retrieval, access, and utilization support for users. Users can initiate natural language queries through keywords, subjects, authors, and other methods. Based on LLM semantic parsing capabilities, the system accurately understands user needs, links collection and external academic resources, and quickly generates optimal literature recommendations and access paths. For complex knowledge questions, knowledge graphs built with LightRAG can assist in reasoning, retrieving and recalling relevant content from book chapters to directly provide literature-based professional knowledge Q&A.

Through these mechanisms, library intelligent reference consultation systems can efficiently and accurately meet users' diverse information needs in core scenarios including regulation consultation, academic misconduct detection, and literature resource Q&A, powerfully driving library services toward intelligence, standardization, and refinement.

6 Conclusion and Prospect

Addressing the needs of library intelligent transformation, this study focuses on MCP service-driven intelligent reference consultation pathways, systematically exploring a new paradigm for deep integration between LLMs and library resource systems through MCP. Starting from the collaboration principle between MCP services and LLMs, the study constructs an integrated "Resource-MCP-LLM-User" smart library reference service model, designs a layered service architecture, analyzes working mechanisms in core application scenarios, and thoroughly discusses implementation challenges and coping strategies. By introducing MCP into the library reference consultation domain, the study effectively breaks information silos in traditional reference services, solves technical difficulties in LLM integration with heterogeneous library systems, and proposes a closed-loop optimization service system to transform reference services from passive response to proactive service models.

However, this study lacks in-depth research on specific service implementations and analysis of librarians' role transformation and professional development paths in the MCP service ecosystem. Future research will expand in the following directions: (1) Conduct empirical case studies of MCP services in different types of libraries to verify model adaptability and practical effectiveness; (2) Explore deep integration of MCP with specific professional domains, such as academic research support and specialized knowledge Q&A applications for characteristic literature resources; (3) Construct a comprehensive evaluation system for MCP services including multi-dimensional indicators such as service quality, user experience, and resource efficiency; (4) Explore MCP-based cross-library collaboration models to promote the formation of a broader intelligent consultation service ecosystem.

This study provides innovative solutions for library reference consultation services by introducing MCP, effectively addressing issues of insufficient intelligence, passive service, and lack of proactive service. As MCP technology continues to mature and library digital transformation accelerates, this service paradigm is expected to become an important component of smart library construction, providing solid support for improving public cultural service quality, promoting efficient utilization of knowledge resources, and meeting diverse user needs, ultimately maximizing library resource value and continuously optimizing user experience.

References

[1] XIAO P. Theoretical Framework, Institutional Logic, and Improvement Pathways of China' s Public Cultural Service System [J]. Journal of Library Science in China,2024,50(05):4-28.DOI:10.13530/j.cnki.jlis.2024035.

[2] LI L Z, WANG T. Research on the spatial construction of urban libraries from the perspective of spatial trichotomy dialectics [J]. Publishing Research,2025,(04):39-45.DOI:10.19393/j.cnki.cn11-1537/g2.2025.04.003.

[3] KE P. Smart Library Construction for the 15th Five-Year Plan [J]. Library Theory and Practice,2025,(02):1-13.DOI:10.14064/j.cnki.issn1005-8214.20250207.001.

[4] WANG Y H, BAI H Y, MENG X Y. An Intelligent Practical Exploration of Large Language Model in Library Reference Consulting Service [J]. Information Studies:Theory & Application,2023,46(08):96-103.DOI:10.16353/j.cnki.1000-7490.2023.08.012.

[5] LUO Y, LI G T, HE J. The Path for DeepSeek to Empower Libary Reference Services [J/OL]. Library Tribune,1-9[2025-05-21].http://kns.cnki.net/kcms/detail/44.1306.G2.20250226.1248.00

[6] Anthropic. Model context protocol: A standard for AI system integration, oct 2024. URL: https://modelcontextprotocol.io.

[7] YAO F, JI L, ZHANG C Y, et al. New attempt on real - time virtual reference service - he smart chat robot of tsinghua university library [J]. Data analysis and knowledge discovery,2011(4):77-81.

[8] YAO F, ZHANG C, CHEN W. Smart talking robot Xiaotu: participatory library service based on artificial intelligence[J]. Library hi tech, 2015.33(2):245-260.

[9] HU C M. Improving Intelligent Service of CVRS Based on Word Segmentation [J]. Library and Information Service,2012,56(09):110-113.

[10] LI W J, CHEN S Q. Application of AIMLBot intelligent robot in real-time virtual reference service [J]. Data analysis and knowledge discovery,2012(Z1):127-132.

[11] LI X T, LI X. Research on language system of library automatic question-answering system based on wechat [J]. Journal of modern information, 2016, 36 (10): 99-101,122.

[12] ZHANG L. Application and implementation of word vector semantic extension technology in library intelligent consulting system [J]. Library and information service, 2020, 64 (18): 126-136.

[13] GUO L M, FU Y M. An LLM-based Knowledge Service System for Libraries Integrating the ReAct Model[J]. Library Tribune,2024,44(06):61-70.

[14] ZHANG G Z, WANG Z Y, YANG F, et al. Research on the Deep Integration of Intelligent Reference Consulting and Business in Libraries[J]. Library and Information Service,2025,69(06):33-45.DOI:10.13266/j.issn.0252-3116.2025.06.003.

[15] WANG Y, DONG Y T. Application and reflection of chatgpt-like artificial intelligence in library intelligence service [J]. Library theory and practice, 2023, (6): 129-136.

[16] ZHANG P. enerative Artificial Intelligence Chatbots: Redefining

User Service in Library [J]. Library Theory and Practice,2024,(02):123-129.DOI:10.14064/j.cnki.issn1005-8214.2024.02.006.

[17] WANG C, PAN X F. Empowering intelligent consultation in university libraries with ChatGPT [J]. Library science research & Work, 2023, (11): 49-53.

[18] PAN X F, WANG C. Research on the impact of chatgpt on intelligent consultation of university libraries from the functional perspective [J]. Journal of library and information science, 2023, 8 (5): 15-20.

[19] WU R H, MAO Y H. Library services under the chatgpt boom: concepts, opportunities, and breakthroughs [J]. Library & information, 2023, (2): 34-41.

[20] GUO Y J, KOU X Y, FENG S Q, et al. Large Language Model Empowers Library Reference Services :Logic,Scenarios,and Framework [J]. Library Tribune,2025,45(01):118-127.

[21] PANDA S, KAUR N. Exploring the viability of ChatGPT as an alternative to traditional chatbot systems in library and information centers[J]. Library hi tech news, 2023, 40(3): 22-25.

[22] RODRIGUEZ S, MUNE C. Uncoding library chatbots: Deploying a new virtual reference tool at the San Jose State University library[J]. Reference services review, 2022, 50(3/4): 392-405.

[23] HOU X Y, ZHAO Y J, WANG S N, et al. Model Context Protocol (MCP): Landscape, Security Threats, and Future Research Directions. ArXiv, 2025, abs/2503.23278:n.pag.

[24] NARAJALA V S, IDAN H. Enterprise-Grade Security for the Model Context Protocol (MCP): Frameworks and Mitigation Strategies. ArXiv, 2025, abs/2504.0862:n.pag.

[25] KRISHNAN N. Advancing Multi-Agent Systems Through Model Context Protocol: Architecture, Implementation, and Applications. ArXiv,2025, abs/2504.21030:n.pag.

[26] SZEIDER S. MCP-Solver: Integrating Language Models with Constraint Programming Systems. ArXiv, 2024, abs/2501.00539:n.pag.

[27] WANG L B. Towards Humanoid Robot Autonomy: A Dynamic Architecture Integrating Continuous thought Machines (CTM) and Model Context Protocol (MCP). ArXiv, 2025, abs/2505.19339:n.pag.

[28] ZHANG G Z. Research on the Optimization Practice of Scientific Research Integrity Services in Universities Based on HTTP Communication Technology [J]. Information Recording Materials,2022,23(05):110-112.DOI:10.16009/j.cnki.cn13-1295/tq.2022.05.025.

Author Contributions

Zhang Guangzhao: Topic selection, model framework construction, manuscript writing and revision;
Wang Zhongyi: Manuscript guidance, model framework guidance, manuscript revision;
Jin Guanglong: Literature investigation, manuscript revision;
Yang Fan: Manuscript guidance, model framework guidance.

Note: Figure translations are in progress. See original paper for figures.