Coze + RAG Architecture-Driven Large Language Model Intelligent Question Answering in Libraries: Practice and Implications—A Case Study of Xiamen University' s Law Library Assistant

Authors: Xu Gonghan

Date: 2025-06-25T14:43:59+00:00

Abstract

Against the backdrop of rapid development of Large Language Models (LLMs), how to deeply integrate them with library local resources has become a key issue in the construction of smart libraries. Based on Retrieval-Augmented Generation (RAG) technology, this paper discusses in detail how to utilize the Coze platform to construct an intelligent question-answering service that integrates local knowledge without requiring high costs or complex technical investment. Through the practice of the "Fatu Assistant" project at the Law Branch Library of Xiamen University Library, it elaborates on core processes including knowledge base construction based on the Coze platform, agent configuration and debugging, multi-platform deployment, and post-maintenance. The results demonstrate that the visual operation platform provided by Coze lowers the threshold for system construction and maintenance, providing a feasible pathway for libraries to achieve efficient and reliable intelligent services amidst the wave of new-generation artificial intelligence.

Full Text

Large Language Model Intelligent Q&A Driven by Coze + RAG Architecture in Libraries: Practice and Insights—A Case Study of the "FaTu Assistant" at Xiamen University

Xu Gonghan
Xiamen University Library, Xiamen 361000, China

Abstract

Against the backdrop of rapid advancements in Large Language Models (LLMs), effectively integrating these models with local library resources has emerged as a critical challenge in smart library development. This paper explores how the Retrieval-Augmented Generation (RAG) approach can be leveraged through the Coze platform to build an intelligent question-answering service that incorporates domain-specific knowledge without requiring high costs or complex technical investments. Drawing on the implementation of the "Law Library Assistant" project at the Law Library of Xiamen University, we detail the core processes of knowledge base construction, AI agent configuration and testing, multi-platform deployment, and post-launch maintenance using the Coze platform. Our findings demonstrate that Coze' s visual, no-code interface significantly lowers the barriers to system setup and maintenance, offering libraries a viable pathway to deliver efficient and reliable AI-powered services amid the new wave of artificial intelligence.

Keywords: RAG; large language model; intelligent Q&A service; Coze; library

2 Methodological Foundation

2.1 Large Language Models and Library Intelligent Q&A

Large Language Models (LLMs) are deep learning models with billions of parameters trained on massive corpora, capable of performing language understanding and generation tasks such as question answering, translation, and content creation. General-purpose LLMs like DeepSeek-V3 and GPT-4o, with their hundreds of billions of parameters, exhibit human-like conversational abilities through extensive text learning. However, LLMs derive their knowledge solely from training data, lacking awareness of post-training developments and domain-specific details, which introduces uncertainty and may lead to incorrect or fabricated responses—a phenomenon known as "hallucination" [5]. In library intelligent Q&A scenarios, patrons frequently inquire about library-specific regulations, collection details, and other information not included in general LLM training corpora. Consequently, relying exclusively on LLMs cannot guarantee accurate, satisfactory answers. To address this limitation, LLMs must be granted access to local knowledge sources, enabling them to leverage authoritative domain information and enhance response reliability.

2.2 Retrieval-Augmented Generation (RAG) Technology

Retrieval-Augmented Generation (RAG) is a technical framework that combines information retrieval with LLM generation, designed to enable models to retrieve external knowledge sources before generating responses [9]. The fundamental principle involves: when a user poses a question, the system first retrieves relevant content fragments from a pre-constructed knowledge base based on the query, then provides both the retrieved information and the original question

to the LLM, which synthesizes the final answer. This approach allows LLMs to reference knowledge base information beyond their training corpora without requiring model retraining. RAG offers several proven advantages: (1) **Cost-effectiveness**—knowledge can be updated without training new models, reducing computational and maintenance costs; (2) **Information freshness**—models can access up-to-date data such as collection updates or real-time policy changes; (3) **Result interpretability**—responses can include source citations, enhancing user trust in answer accuracy; and (4) **Development controllability**—developers can correct errors or adapt to new requirements by simply updating knowledge sources. In library contexts, implementing RAG enables intelligent Q&A systems to query internal knowledge bases (e.g., library introductions, service guides, FAQs) and generate responses grounded in collection facts, significantly improving relevance and professionalism. This approach has become the mainstream solution for deploying LLMs in vertical domains, including libraries.

2.3 Knowledge Base Construction and Local Knowledge Integration

Implementing RAG requires building a high-quality local knowledge base—a corpus collection of domain-specific knowledge that may include Q&A pairs, policy documents, user guides, database introductions, and bibliographic lists. In this case, the knowledge base primarily covers information related to the Law Library of Xiamen University. The construction process typically involves three stages: data collection, content processing, and knowledge base establishment. First, we gather library documentation relevant to frequently asked patron questions, such as collection distribution, opening hours, borrowing policies, circulation procedures, database usage instructions, and discipline-specific reading lists. Next, we process the raw materials by condensing complex policy texts into key Q&A formats or converting web content into plain text. Finally, the processed documents are imported into a searchable knowledge base for subsequent semantic matching. Once established, the knowledge base enables LLMs to reference its content during answer generation, preventing unfounded speculation. It is important to note that knowledge base construction is an ongoing process: as collections and patron needs evolve, entries must be regularly updated to ensure content remains current and accurate. Only through the integration of local knowledge bases with LLMs can library intelligent Q&A systems achieve both general language intelligence and specialized domain expertise.

3.1 Overview of the Coze Platform

To cost-effectively and rapidly implement the aforementioned local knowledge-based intelligent Q&A solution, Xiamen University Library selected ByteDance' s Coze platform. Coze is a next-generation AI agent development platform offering zero-code visual tools that enable users to create and customize AI agents without programming and deploy them across multiple channels with a single click. Unlike general chatbots, Coze focuses on supporting vertical

domain assistants, employing RAG technology to enable models to access local knowledge bases and allowing developers to combine domain knowledge with plugin extensions for precise scenario-based services. Coze integrates powerful LLMs, including the latest DeepSeek series of domestic models, and continuously updates its model repository. By shielding developers from the complexities of underlying model fine-tuning and deployment, Coze empowers non-IT teams to conveniently build required applications using large models.

3.2 Advantages Over Self-Built Systems

The Coze platform offers numerous advantages over traditional self-built systems. **Low cost** is a primary benefit: using Coze' s AI agent services eliminates the need to purchase expensive servers or GPU computing resources and avoids high API call fees. The platform provides foundational models and runtime environments, allowing developers to simply prepare knowledge base content and configure basic agent settings to launch services free of charge or at low cost—particularly valuable for budget-constrained libraries. **Low technical barrier and high usability** represent another key advantage: Coze' s graphical interface and modular configuration eliminate coding requirements. Users complete most functional settings through simple parameter configuration and prompt writing. The platform' s design is clear, logical, and straightforward, enabling library staff without AI or programming backgrounds to build agent prototypes quickly and conveniently debug and optimize them later.

Superior model performance with timely updates: Coze' s integrated underlying LLMs are provided and maintained by the platform. For instance, its access to DeepSeek-V3 and DeepSeek-R1 models, which demonstrate outstanding performance in Chinese comprehension and Q&A, is regularly upgraded by the platform (e.g., the recent DeepSeek-V3-0324 update). This allows administrators to enhance agent performance with a single configuration change, avoiding the cumbersome redeployment required in self-built solutions and significantly reducing maintenance burden.

Built-in knowledge base support for convenient maintenance: Coze' s distinctive advantage lies in its simple yet powerful knowledge base module for managing and storing local information. Users can upload local files or online content to the knowledge base for model reference, thereby supplementing the LLM' s limited vertical domain knowledge. The knowledge base supports multiple document formats including txt, pdf, csv, Word, and Excel, and can even scrape and synchronize web text via URL links. Compared to self-built systems that require constructing knowledge bases and writing retrieval code from scratch, Coze' s knowledge base functionality is ready-to-use and provides a visual interface for easy content management. This enables libraries to rapidly build and iterate local knowledge bases: when regulations are updated or new FAQs emerge, staff can directly edit corresponding entries in the platform' s knowledge base, ensuring the agent' s knowledge remains current.

Team collaboration and permission management: Coze supports multi-person collaborative development of one or multiple agent projects through team workspaces. Project members can assume different responsibilities such as knowledge base maintenance, prompt design, and testing, overcoming the limitations of single-account operations. This collaborative mechanism proves highly practical for organizations like libraries that require multi-person service maintenance.

Rich plugin extension capabilities: Coze integrates an agent plugin store offering plugins covering news search, link reading, image understanding, table operations, and various other functions. Through plugins, agents can invoke external tools to perform diverse tasks, transcending the limitations of language-only responses and providing tool-enhanced service capabilities. Compared to self-built systems requiring manual API integration programming, Coze' s plugin mechanism significantly reduces the difficulty of extending functionality, granting agents greater customizability and usability.

Multi-channel publishing and integration: Coze supports one-click deployment of agents to numerous platforms including the Coze Store, Feishu, Douyin Mini Programs, WeChat Official Accounts, and others. Coze also provides Chat SDKs and APIs for embedding agents into library websites or OPAC systems. This multi-platform adaptability enables intelligent Q&A services to rapidly reach users through convenient channels. In contrast, developing WeChat Official Account interfaces or web widgets for self-built systems would be considerably more time-consuming.

In summary, Coze' s characteristics of low cost, low barrier, and high flexibility align well with library needs for building intelligent Q&A services. After evaluation, the Law Library of Xiamen University determined that Coze could fully satisfy the functional and performance requirements of the "FaTu Assistant" project and thus selected the platform for development. The following sections detail the implementation process and lessons learned.

4 Case Study: Construction Method and Process of "FaTu Assistant"

In early 2025, the Law Library of Xiamen University launched the "FaTu Assistant" intelligent Q&A service project, leveraging the Coze platform to build an AI assistant serving the university' s faculty and students. The implementation process included account and team setup, knowledge base construction, agent configuration (including model selection, persona and response logic design, knowledge base integration, and plugin integration), debugging, deployment, and post-launch maintenance (see Figure 1 [Figure 1: see original paper]). Each phase is detailed below.

4.1 Account Registration and Team Setup

Project team members first registered accounts on the Coze official website (coze.cn), a process requiring only a mobile phone number and SMS verification code. Since the project required multi-member collaboration, one team member created a team workspace and invited others to join. Through Coze' s team space functionality, all members could collaborate within the same platform environment, significantly improving development efficiency.

4.2 Knowledge Base Construction

The knowledge base construction formed the foundation of the "FaTu Assistant" project. We comprehensively collected and organized materials relevant to the Law Library of Xiamen University for Q&A services, then imported them into the system. The specific steps were as follows.

Information collection and organization: We first compiled a list of frequently asked patron questions covering library opening hours (including regular and holiday schedules), location, borrowing policies, circulation service usage methods, and contact information. Simultaneously, we collected authoritative guide content from the Xiamen University Library website—ideal material for knowledge base construction due to its standardized language and authoritative content. Additionally, we curated lists of popular books in the Law Library, such as high-circulation and high-reservation titles, to provide valuable recommendations when patrons sought reading suggestions. All collected content was verified to ensure accuracy and reliability.

Importing into Coze knowledge base: After processing, we entered the "Resource Library" module of the Coze platform workspace and created several knowledge bases according to file types, including text knowledge bases, popular book list databases, and image libraries. For text and table files, we imported them by synchronizing with Feishu shared documents. For image files, we directly uploaded local images to the corresponding image libraries.

4.3 Agent Configuration

After knowledge base construction, we entered the "Project Development" module of the Coze workspace, created the "FaTu Assistant" agent, and accessed its editing page for configuration.

Model selection and parameter settings: Coze supports deployment of high-performance LLMs such as the DeepSeek series and Tongyi Qianwen series. We selected "DeepSeek-R1-Tool-Use" as the foundation model for "FaTu Assistant." DeepSeek-R1 is currently one of the most popular reasoning models, capable of precisely understanding user questions and providing targeted answers while supporting Coze' s rich plugin tools. Additionally, "DeepSeek-V3-0324," the latest non-reasoning model from DeepSeek, represents another ideal option. Coze allows configuration of base model parameters such as "genera-

tion randomness," "context retention rounds," and "maximum response length." We set "context retention rounds" to 5 and retained default settings for other parameters.

Persona and response logic: In the "Persona and Response Logic" section of the editing page, team members could freely edit the agent' s prompts to align its response logic and style with requirements. We envisioned "FaTu Assistant" as professional yet approachable, describing its persona in the basic prompt settings as "a professional and caring AI assistant for the Law Library of Xiamen University." We explicitly instructed the agent to "answer questions about the Law Library of Xiamen University with enthusiasm, patience, precision, and comprehensiveness," to "never fabricate answers or make unfounded assumptions," and to "provide clear, logically coherent, and easily understandable responses." When users ask questions, the agent determines its response tone and scope based on these prompts.

Knowledge base configuration: In the "Knowledge" section of the "FaTu Assistant" editing page, we added the previously established knowledge bases from the workspace. To ensure the assistant fully utilized the knowledge bases, we enabled automatic knowledge base invocation, meaning the agent first performs semantic searches in the connected knowledge bases whenever users ask questions, extracting relevant content for model reference. This effectively automates the RAG workflow. When relevant knowledge points match with high confidence, the model' s responses directly integrate knowledge base content, ensuring accuracy.

Plugin integration and function extension: Recognizing that some patron needs might exceed knowledge base scope (e.g., querying latest legal news), we employed Coze' s plugin extension functionality to enable the agent to invoke external services. We integrated plugins such as "Bing Search," "Link Reader," and "Academic Search" to retrieve real-time internet information when required to answer user questions.

4.4 Debugging

After agent configuration, we entered the debugging and optimization phase. In the "Preview and Debug" window of the agent editing page, team members could evaluate the agent' s response performance through conversation, enabling targeted improvements. We tested the agent with representative questions based on knowledge base content to verify correct answers, and posed open-ended questions not directly answerable from the knowledge base to assess its ability to invoke plugins for internet retrieval and answer synthesis. Through iterative cycles of testing and modification, we continuously refined knowledge base content and prompt strategies until satisfactory responses were achieved for the vast majority of preset questions.

4.5 Deployment and Launch

Upon completing debugging, the "FaTu Assistant" agent was ready for official service launch. Coze supports deploying agents to multiple platforms including the Coze Store, Feishu, Douyin Mini Programs, and WeChat Official Accounts. We chose to publish "FaTu Assistant" to the Coze Store, which provided the assistant with a dedicated page link and QR code for convenient mobile or desktop access via browser login or QR code scanning. We also embedded the assistant' s dedicated page link into the menu options of the Law Library' s WeChat Official Account for easy patron access. The user interface of "FaTu Assistant" is shown in Figure 2 [Figure 2: see original paper].

4.6 Post-Launch Maintenance

Following the launch of "FaTu Assistant," the team entered routine maintenance, which encompasses several aspects: (1) **Dynamic knowledge base maintenance**: Knowledge base content must synchronize with the latest library information. Whenever new announcements or service changes occur (e.g., holiday opening arrangements), team members promptly update the knowledge base accordingly. (2) **Conversation log monitoring and quality control**: Coze provides access to user-agent conversation records from the past seven days. The project team reviews and analyzes these interaction logs to evaluate the assistant' s answer accuracy and user satisfaction, using insights to continuously improve the knowledge base and configuration. (3) **Platform upgrades and model iteration**: We monitor Coze' s official updates and evaluate new, more powerful model versions or functional plugins for potential adoption in our project.

5 Functionality Implementation and Service Scenarios

Since its launch, "FaTu Assistant" has implemented various intelligent service functions for patrons. Primarily, it serves as an automated FAQ system for Law Library basic information, providing 24/7 online service that compensates for limited traditional human service hours. Patrons can obtain instant, accurate answers through the chat interface without consulting printed guides or waiting for human responses. In practical applications, "FaTu Assistant" handles diverse information inquiries: it explains borrowing policies and details specific service procedures such as "proxy borrowing" and "book recommendation" ; it recommends relevant professional titles from the built-in popular book list based on patrons' disciplinary backgrounds and interests. Furthermore, by continuously collecting user questions, "FaTu Assistant" provides data support for understanding evolving patron needs—identifying frequently asked questions and under-recognized services—to help improve human services. Overall, "FaTu Assistant" has demonstrated the value of the "AI + library services" model: reducing pressure on human consultation while innovatively bridging the gap between patrons and the library.

6 Conclusion and Outlook

The construction of "FaTu Assistant" at the Law Library of Xiamen University demonstrates that libraries can rapidly develop intelligent Q&A services tailored to their needs by leveraging emerging low-barrier platforms amid the flourishing development of LLM technology. This paper, using "FaTu Assistant" as an example, illustrates how the Coze platform combines general LLM capabilities with local collection knowledge to deploy library intelligent Q&A services with relatively small investment—solutions previously requiring complex technical approaches. Such intelligent Q&A services complement traditional consultation methods, offering new possibilities for expanding library service models. Naturally, current attempts are not yet perfect, and the AI assistant' s role requires further clarification—it should serve as an assistant and tool for human librarians rather than a complete replacement. We will continue monitoring patron acceptance of and preferences for intelligent services to find the optimal balance in human-machine collaboration.

Acknowledgments

The development of "FaTu Assistant" would not have been possible without library leadership support and the collective efforts of the librarian team. Dai Xiancong, Director of the Law and Arts Branch Library of Xiamen University Library, conceived the idea of creating an AI Q&A assistant for the Law Library and led the completion and launch of the "FaTu Assistant" project. Teachers Su Qi, Chen Mengyi, and Zhu Qiaoqing from the Law and Arts Branch Library also contributed to building and improving "FaTu Assistant." We express our sincere gratitude to them.

References

[1] Chu Jiewang, Du Xiuxiu, Li Jiaxuan. Impact and Application Prospects of Artificial Intelligence Generated Content on Smart Library Services[J]. Information Studies: Theory & Application, 2023, 46(05): 6-13.

[2] Ji Ting, Zhou Gang, Xu Lei. Research on Innovative Application Requirements and Scenarios of Large Models in Libraries[J]. Journal of Information and Management Research, 2024, 9(05): 1-13.

[3] Harbin Institute of Technology Library. Bokan AI Librarian[EB/OL]. [2025-04-07]. http://www.lib.hit.edu.cn/bkAIgy/list.htm.

[4] Zhongnan University of Economics and Law Library. The AI Librarian of Zhongnan University of Economics and Law Library is Launched[EB/OL]. [2025-04-07]. https://library.zuel.edu.cn/2025/0403/c5996a388202/page.htm.

[5] Liu Zeyuan, Wang Pengjiang, Song Xiaobin, Zhang Xin, Jiang Benben. A Survey on Hallucination Problems in Large Language Models[J]. Journal of Software, 2025, 36(03): 1152-1185.

[6] Wang Yihu, Bai Haiyan, Meng Xuyang. Intelligent Practice Exploration of Large Language Models in Library Reference Services[J]. Information Studies: Theory & Application, 2023, 46(08): 96-103.

[7] Zhao Yang, Zhang Bei, Dou Tianfang. Exploration and Practice of Generative AI Application in Tsinghua University Library[C]// Proceedings of the 17th Library Management and Service Innovation Forum 2024. Tsinghua University Library, 2024: 85.

[8] Beijing Chuntian Zhiyun Technology Co., Ltd. Coze[EB/OL]. [2025-04-07]. https://www.coze.cn/.

[9] Liu Xueying, Yun Jing, Li Bo, et al. A Survey on Retrieval-Augmented Generation Based on Large Language Models[J/OL]. Computer Engineering and Applications, 1-31[2025-04-07].

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