Construction of a Localized Traditional Chinese Medicine Knowledge Base Question-Answering System Based on MaxKB + Ollama + Qwen3.6

Xu Li-qin^1^, Han Wen-zhe^1*^

2026-07-27

ChinaRxiv

Abstract

[Objective] To construct a fully localized traditional Chinese medicine knowledge-base question-answering system using the national higher education "14th Five-Year Plan" textbooks for the traditional Chinese medicine industry as the knowledge source, meeting the rigid requirements for medical data privacy protection and offline deployment. [Methods] Based on MaxKB, Ollama, and Qwen3.6-35B-A3B, systematic engineering exploration was conducted around textbook cleaning, 10-category intent recognition, hybrid vector and full-text retrieval (threshold 0.7 + Top-K convergence), prompt engineering, and parameter tuning. [Results] The system achieved a 100% execution success rate, an answer accuracy rate $\geq$ 95%, user waiting time $\leq$ 3 s, and satisfaction $\geq$ 95%. [Conclusion] A reproducible, low-threshold, and highly reliable local traditional Chinese medicine knowledge-base RAG solution was developed, providing a technical pathway and engineering reference for the digital inheritance and intelligent application of traditional Chinese medicine knowledge.

Full Text

Construction of a Localized Traditional Chinese Medicine Knowledge Base Question-Answering System Based on MaxKB + Ollama + Qwen3.6

Xu Li-qin[1], Han Wen-zhe^1*^

(1. Fengdu County Hospital of Traditional Chinese Medicine, Chongqing 408200)

Abstract:

[**Objective**] Using the national higher-education "14th Five-Year Plan" textbooks for the traditional Chinese medicine industry as the knowledge source, this study constructs a fully localized traditional Chinese medicine knowledge base question-answering system to meet the rigid requirements of medical data privacy protection and offline deployment.

[**Methods**] Based on MaxKB, Ollama, and Qwen3.6-35B-A3B, systematic engineering exploration was carried out around textbook cleaning, recognition of 10 categories of intent, vector + full-text hybrid retrieval (threshold 0.7 + Top-K convergence), prompt engineering, and parameter tuning.

[**Results**] The system execution success rate was 100%, the answer accuracy was $\geq$ 95%, user waiting time was $\leq$ 3 s, and satisfaction was $\geq$ 95%.

[**Conclusion**] A replicable, low-threshold, and highly reliable localized traditional Chinese medicine knowledge base RAG solution was formed, providing a technical pathway and engineering reference for the digital inheritance and intelligent application of traditional Chinese medicine knowledge.

Keywords: traditional Chinese medicine knowledge base; large language model; retrieval-augmented generation; localized question-answering system

Construction of a Localized Traditional Chinese Medicine Knowledge Base Question-Answering System Based on MaxKB + Ollama + Qwen3.6

XU Li-qin[1], HAN Wen-zhe[1]

(1. Fengdu County Hospital of Traditional Chinese Medicine, Chongqing, 408200, China)

Abstract:

[**Objective**] To build a fully localized TCM knowledge base QA system using the national "14th Five-Year Plan"textbooks as the knowledge source, addressing the needs of medical data privacy and offline deployment.

[**Methods**] Built on MaxKB, Ollama, and Qwen3.6-35B-A3B, systematic engineering

[1] **Funding project:** Medical Research Project of Fengdu County Health Commission (2024FDKW026)

Author profile: Xu Li-qin (1998-), female, Han ethnicity, from Fengdu County, Chongqing; traditional Chinese medicine physician; master' s student; research direction: integrated Chinese and Western medicine treatment of internal diseases.

Corresponding author: Han Wen-zhe (1996-), male, Han ethnicity, from Xiping County, Zhumadian City, Henan Province; attending traditional Chinese medicine physician; master' s student; research direction: traditional Chinese medicine and artificial intelligence. E-mail: wenzhe_han@126.com.

was conducted across textbook cleaning, 10-category intent classification, hybrid vector + full-text retrieval (threshold 0.7 + Top-K convergence), prompt engineering, and parameter tuning.

[**Results**] Workflow success rate 100%, answer accuracy $\geq$ 95%, user waiting time $\leq$ 3 s, satisfaction rate $\geq$ 95%.

[**Conclusion**] A reproducible, low-barrier, high-reliability localized TCM knowledge base RAG solution was developed, providing a technical pathway and engineering reference for the digital preservation and intelligent application of TCM knowledge.

Keywords: Traditional Chinese Medicine knowledge base; Retrieval-Augmented Generation; Large Language Model; localized question-answering

system

Introduction

With the advancement of the "Healthy China 2030" strategy [1] and the "15th Five-Year Plan" for the revitalization and development of traditional Chinese medicine [2], the accessibility, standardization, and intelligentization of TCM knowledge have attracted increasing attention from the industry. Large Language Models (LLMs) have demonstrated strong capabilities in medical question answering and clinical decision support, such as Baichuan-M2 Plus, "the first evidence-enhanced medical large model" developed by Baichuan Intelligence [3]. However, most existing systems rely on the cloud, showing clear deficiencies in data privacy protection, offline deployment, and customized adaptation, and making it difficult to meet the stringent security requirements of the medical industry. Meanwhile, Retrieval-Augmented Generation (RAG) technology, by integrating external retrieval with large-model generation, can effectively reduce the risk of hallucination and make professional question answering more feasible.

Based on this, this project integrates MaxKB, Ollama, and the Qwen3.6 large model to build a localized question-answering system using "14th Five-Year Plan" textbooks as the knowledge source. In addition to system construction and joint debugging, engineering exploration and experimental validation were carried out around textbook selection, intent classification, retrieval-strategy optimization, prompt engineering, and model quantization tuning, forming a reproducible, low-barrier, and highly reliable RAG solution for a TCM knowledge base.

1 Key Technical Overview

1.1 Docker Containerization Technology

Docker is an open-source application container engine that enables rapid application deployment and environment isolation through containerization technology. This study requires the deployment of the MaxKB service in Docker.

1.2 MaxKB Knowledge Base Management Platform [4]

MaxKB is an LLM-based knowledge-base question-answering system that supports the RAG architecture and can connect with local LLMs through Ollama to achieve intelligent question-answering capabilities.

1.3 Ollama Large-Model Service Framework

Ollama is a high-performance service framework designed specifically for local LLM inference. It can quickly download and run various open-source large models through the command line. In this study, the Qwen3.6-35B-A3B model is run in Ollama.

1.4 Qwen3.6-35B-A3B Model

Qwen3.6-35B-A3B is the latest generation of hybrid Mixture-of-Experts (MoE) large model open-sourced by the Qwen team on April 15, 2026[5]. The model performs well on tasks such as Chinese semantic understanding, medical-knowledge question answering, and programming assistance, and its inference speed is significantly better than that of dense models with an equivalent total parameter count.

2 Knowledge-Base Construction and Preprocessing

2.1 Data Sources

The TCM knowledge base uses as its standardized data source the national higher-education textbooks for the Traditional Chinese Medicine industry published by China Press of Traditional Chinese Medicine under the "14th Five-Year Plan." A total of nine textbooks were selected, including four foundational TCM courses—*Basic Theory of Traditional Chinese Medicine*, *Diagnostics of Traditional Chinese Medicine*, *Chinese Materia Medica*, and *Formulas*—as well as five clinical disciplines: *Internal Medicine of Traditional Chinese Medicine*, *Surgery of Traditional Chinese Medicine*, *Gynecology of Traditional Chinese Medicine*, *Pediatrics of Traditional Chinese Medicine*, and *Acupuncture and Moxibustion*. Together, these form a complete closed loop from the foundations of "principle, method, formula, and medicinals" to the various clinical disciplines.

2.2 Text Cleaning and Segmentation

The above PDF electronic textbooks were uniformly converted into docx format, and systematic data cleaning was performed:

(1) Removal of irrelevant content: prefaces, conclusions, exercises, references, appendices, index pages, tables of contents, copyright pages, and other chapters unrelated to the main text were removed. If these purely list-based data (such as formula indexes and Chinese-medicine pinyin indexes) are chunked and vectorized, they will severely contaminate the vector space—for example, when a user searches for "Ephedra Decoction," the system may retrieve an isolated sentence from an index page with no context, such as "Ephedra Decoction······145," causing the large model to be unable to generate an effective answer.

(2) Processing of headers, footers, and figures/tables: header and footer information was removed. For tables in the textbooks, simple comparative tables retained their textual content; pure images and complex charts/tables were removed because their textual information could not be completely extracted during the PDF-to-docx conversion process.

(3) Title standardization and segmentation: segmentation was performed intelligently using textbook-outline headings as the basis for partitioning,

Figure 1 Workflow orchestration flowchart: Start → User input → Intent recognition → categories including TCM theory, TCM diagnosis, Chinese materia medica, formulas, acupuncture, internal medicine, surgery, gynecology, pediatrics, and other; retrieval of the corresponding knowledge base or retrieval of the knowledge bases for Chinese Internal Medicine, Basic Theory of TCM, TCM Diagnostics, Chinese Materia Medica, and Formulas; AI dialogue → End.

Figure 1: Figure 1 Workflow orchestration flowchart: Start → User input → Intent recognition → categories including TCM theory, TCM diagnosis, Chinese materia medica, formulas, acupuncture, internal medicine, surgery, gynecology, pediatrics, and other; retrieval of the corresponding knowledge base or retrieval of the knowledge bases for Chinese Internal Medicine, Basic Theory of TCM, TCM Diagnostics, Chinese Materia Medica, and Formulas; AI dialogue → End.

with a maximum of 4,096 characters per segment. Title-based segmentation is the core strategy for ensuring the contextual integrity of TCM "principle, method, formula, and medicinals." Each segment corresponds as much as possible to a complete teaching unit, ensuring that the segment contains a relatively complete logical chain of principle, method, formula, and medicinals.

After cleaning was completed, the docx-format textbooks were again converted into markdown format for import into the MaxKB knowledge base.

2.3 Knowledge-Base Construction

The textbooks were imported into the MaxKB knowledge base, intelligently segmented, and vectorized using the vector model qwen3-embedding:4b; questions were generated based on the segment contents.

3 Construction of a Local TCM Knowledge-Base Question-Answering System

3.1 Workflow Orchestration Process

An advanced orchestration application— "Qihuang Assistant" —was created in MaxKB to serve as the front-end interactive interface for the local TCM knowledge-base question-answering system.

Workflow orchestration process: user input → intent recognition → knowledge-base retrieval → AI dialogue. For example,

As shown in Figure 1.

Figure 1 Workflow orchestration flowchart

3.2 Core Nodes in Workflow Orchestration

(1) User input node: Receives the user's text question and passes it to subsequent processing modules. For security reasons, the system currently supports only text input for the time being; it does not support voice input, image input, or file upload.

(2) Intent recognition node: The Qwen3.6-35B-A3B model classifies the user' s question into 10 intent categories: Chinese materia medica, formulas, pediatrics, gynecology, surgery, acupuncture, internal medicine, TCM theory, TCM diagnostics, and other. According to the classification result, the question is routed to the corresponding knowledge-base subset. The "other" category serves as a fallback option and is triggered only when the question cannot be assigned to the preceding nine categories. Its retrieval scope covers five core foundational textbooks: Chinese Internal Medicine, Basic Theory of TCM, TCM Diagnostics, Chinese Materia Medica, and Formulas.

(3) Knowledge-base retrieval node: Retrieval is performed within the limited scope of the knowledge base. The retrieval mode is hybrid retrieval (vector + full-text retrieval). The similarity threshold is set to ≥ 0.7, and the Top-3 cited segments are selected (5 for the "other" category, to ensure that sufficient context can still be recalled for broad questions).

(4) AI dialogue node: The Qwen3.6-35B-A3B model in Ollama is called for inference and generation. The system prompt and user prompt are adjusted, the number of historical chat records is set to 5, and the AI response content is output.

3.3 Optimization of Retrieval Strategy

(1) Retrieval mode: Hybrid retrieval is adopted (vector + full-text retrieval). Pure vector retrieval performs poorly when encountering specialized TCM terms, such as "Guizhi plus Gegen Decoction" and "Eighteen Incompatibilities" ; in searches of medical textbooks, literal exact matching is often more reliable than fuzzy semantic matching.

(2) Similarity threshold: It is set to 0.7 (based on cosine similarity, with a range of 0–1). Only segments with a similarity score ≥ 0.7 are allowed to enter the context of the large model. If the highest scores of all segments in the retrieval results are below 0.7, it is determined that there is no directly matching content in the knowledge base. In this case, the system does not send low-scoring segments to the large model; instead, the model answers based on its own knowledge. At the same time, the front-end interface displays the number of retrieved segments as 0, for the u-

provide users with complete transparency.

(3) Number of cited segments (Top-K): uniformly set to 3 except for the "Other" category. Under a simple coarse-ranking architecture without a

Reranker, feeding the large model 3 absolutely precise segments yields answer quality far higher than feeding it 10 segments mixed with noise. Each segment contains about 200–1500 words; the total length of 3 highly similar segments is about 3000–4500 words, which is sufficient for the Qwen3.6 35B-A3B model to extract the core conclusions and professional analysis.

3.4 Prompt Design

(1) Role setting

The system' s role is set as "Qihuang Assistant," a senior traditional Chinese medicine expert with rich clinical experience and deep academic accomplishments. All answers must be strictly based on the nationally planned TCM textbooks of the "14th Five-Year Plan."

(2) Opening statement

Hello! I am "Qihuang Assistant," a TCM knowledge retrieval expert based on the "14th Five-Year Plan" textbooks.

I can look up content in fundamentals, formulas, and various clinical specialties. For example:

"Etiology and pathogenesis of yin deficiency with exuberant fire?"

"Drug composition of Sijunzi Decoction?"

"Clinical manifestations of cough due to wind-cold invading the lung?"

This system is for learning reference only and cannot replace an in-person consultation with a licensed physician. For emergencies or severe illness, please seek medical care promptly.

(3) AI dialogue prompt

Role

You are "Qihuang Assistant," a senior traditional Chinese medicine expert with rich clinical experience and deep academic accomplishments. All your answers must be strictly based on the [**reference materials**] provided below (nationally planned TCM textbooks of the "14th Five-Year Plan").

Core Principles

1. **Absolute fidelity**: Answers must be 100% derived from the [**reference materials**]. When the [**reference materials**] are empty or not provided, the model may answer based on its own knowledge, but it must state at the beginning of the answer: "The following content is based on the model' s knowledge and was not retrieved in this query." Fabrication is strictly prohibited.

2. **Safety and pattern identification first**: Always place medication safety (such as contraindications in pregnancy, compatibility contraindications, and toxic or side effects) and the logic of pattern identification and treatment first. If high-risk operations or acute/severe conditions are involved, a clear safety warning must be given.
3. **Objectivity and citation of classics**: Maintain an objective, rigorous academic tone. When citing textbook content, mark the source in parentheses in the main text, in the format: (*Textbook Title • Chapter/Section*). For example: "the representative ···for wind-cold common cold

the exterior-releasing formula is Jingfang Baidu San (*Formulary Science · Exterior-Releasing Formulas*)."

Language Style (dual-track language style – strictly enforce)

- [**General Analysis**] (for etiology and pathogenesis, clinical manifestations, daily nursing, health preservation and prevention, etc.): Use **professional yet plain and easy-to-understand** language. Transform profound TCM theory into logic that is easy to understand, ensuring that non-professionals can also grasp it accurately.
- [**Formula Analysis**] (for formula interpretation, the roles of sovereign-minister-assistant-envoy, the significance of drug compatibility, therapeutic method, formula and medicinals composition, etc.): **Highly standardized and rigorous medical written language** must be used. Wording should be concise and precise, conforming to the standards for writing TCM clinical medical records, so as to ensure that its degree of professionalism can be directly recorded by physicians in the medical record.

Output Structure

Core Conclusion

(Use 1-3 sentences to give a direct answer in a targeted manner, such as the disease name, pattern type, first-choice formula, or safety conclusion; do not add any empty words.)

Professional Analysis

(Use an unordered list and boldface, and expand in detail. Please automatically switch according to the above "dual-track language style" based on the content.)

- **Point 1**: ···
- **Point 2**: ···

Safety/Differential-Diagnosis Reminder

(Output this module only when the materials explicitly mention medication contraindications, cautious use in special constitutions, or the need for differential diagnosis. If not, do not output it.)

3.5 Debugging Evaluation and Optimization

(1) Optimization objectives:

Execution success rate (all workflow nodes execute normally without interruption due to errors) = 100%;

Answer accuracy (the core conclusion of the answer is consistent with the textbook content, without hallucination) $\geq$ 95%;

User waiting time (total time from the user' s question to the answer response) $\leq$ 3 s;

User satisfaction (based on user feedback) $\geq$ 95%.

(2) Optimization measures and parameter settings

Improve execution success rate: conduct multiple rounds of testing by multiple people to identify possible abnormal errors.

Improve answer accuracy: perform knowledge-base retrieval by segment according to intent recognition; adjust retrieval threshold and number of segments, threshold 0.7 + Top-3; adjust model parameter settings to temperature 0.3, maximum output Token count 1024, thinking mode false, Top P 0.9, Top K 40, Min P 0.03, repetition penalty 1.05.

Reduce user waiting time: Optimize the retrieval process; retrieve from categorized knowledge bases based on intent recognition; remove question optimization and multi-route recall; and disable model reasoning.

Improve user satisfaction: Use a gentle tone in the opening remarks; add emojis to increase interactivity and salience; use a concise and orderly format so that answers are easy to read and understand; avoid refusing to answer, and provide suggestions.

(3) **Test results:**

Execution success rate: 100%. The workflow executed stably across all types of question scenarios, with no node exceptions or timeout interruptions.

Answer accuracy: Approximately 96%. Among more than 200 test cases, deviations occurred only in a very small number of questions involving high ambiguity or conflicts across textbooks. Analysis showed that these were all caused by differences in wording between textbooks, rather than model hallucinations.

User waiting time: Within 3 seconds, averaging 2–2.5 seconds. Intent classification took about 1–1.5 seconds, and retrieval took about 0.5–1.5 seconds. AI dialogue time depended on the length of the answer, at about 5–15 seconds.

User satisfaction: Approximately 96%. Invited users expressed approval of the accuracy of the answers, the clarity of the structure, and the response speed.

(4) **Key experience**

The accuracy of intent classification is the key bottleneck in system performance. Classification errors cause deviations in the retrieval scope and directly affect recall quality.

The combination of a threshold of 0.7 + Top-3 performs robustly in scenarios without a Reranker. Raising the threshold to 0.75 slightly improves accuracy, but causes some valid content to be truncated; lowering it to 0.65 leads to a noticeable increase in noise. 0.7 is the optimal balance point determined through testing.

Compared with the default 0.7, a Temperature of 0.3 significantly improves faithfulness, especially in scenarios requiring precise citation, such as prescription analysis. The variation in wording is reduced, while the accuracy of the core content is significantly improved.

Disabling the reasoning function reduces the average response time from 30 seconds to within 20 seconds, with no perceptible decline in answer quality, verifying the applicability of non-reasoning models in RAG scenarios.

4 Limitations and Prospects

Because research on AI large models is advancing rapidly, models that were the highest-configuration options one year ago have now been eliminated. Therefore, the deployment plan for the local traditional Chinese medicine question-answering system that was considered the best solution at the time now has many limitations. After multiple rounds of testing and research, these limitations are listed below, together with corresponding improvement measures and research prospects.

4.1 Better and Faster Open-Source Large Language Models

If more extreme inference speed is desired, the GPT-oss series models open-sourced by OpenAI in August 2025 can be used [6], including two scales, 120b and 20b. Actual tests show that their inference speed is faster, and the 120B version can also run locally after quantization. Alternatively, smaller 4B or 2B models can be used, but their model capability is not as high as Qwen3.6-35B-A3B. If on-device operation on mobile phones is required, evaluation and migration can be carried out in subsequent research.

If better inference quality is desired, the latest model released by DeepSeek can be used.

DeepSeek-V4-Flash model [7], but the model has a relatively large number of parameters. Although it can still run locally after quantization, its inference speed is relatively slow, making it suitable only for non-real-time knowledge-base retrieval and dialogue.

4.2 A More Professional RAG Platform

As a simple and easy-to-use RAG platform, MaxKB is suitable for deployment and use by users with no prior foundation. However, with the rapid development of large language models, more specialized and more flexible tools have emerged, such as RAGFlow and Dify. Model selection and migration can be carried out according to actual needs, in order to obtain more flexible retrieval-strategy configuration and more fine-grained workflow orchestration capabilities.

4.3 More Diverse Types of Question Answering

The development direction of AI large models has evolved from pure text toward multimodal models. Future research can use multimodal models that not only support text-based question answering, but can also incorporate voice input (streaming ASR) and image input (tongue images, complexion images, imaging data, etc.). Voice input is especially in demand in traditional Chinese medicine scenarios—there are many rare characters and variant characters in TCM terminology (such as 癥瘕, 瘢痕, 瞤动, etc.). Voice input can perfectly bypass the barrier of "not knowing how to type," and streaming recognition technology can smooth out the sense of latency. For image input, the system may first support OCR-type images (photographed book pages, photographed prescriptions), while temporarily postponing support for consultation-type images (photographed tongues, photographed complexions) to avoid medical compliance risks.

4.4 A More Targeted Question-Answering System

The question-answering system in this study is designed for the broad category of traditional Chinese medicine. At the same time, because orthopedics and traumatology more often adopts surgery and manipulation therapies, and uses relatively few internal treatment methods, it has not yet been incorporated into the knowledge base. In the future, when constructing local models, more targeted question-answering systems can be built separately for corresponding specialties, such as TCM internal medicine, TCM surgery, TCM gynecology, and TCM pediatrics, thereby achieving more fine-grained knowledge-base partitioning and more precise intent routing. In addition, *TCM Orthopedics and Traumatology* and *Tuinaology* may be incorporated to supplement the TCM "external treatment methods" module, so as to meet retrieval needs for high-frequency conditions such as neck, shoulder, waist, and leg pain.

5 Conclusion

Based on Docker containerization technology, this study deeply integrates MaxKB, Ollama, and Qwen3.6-35B-A3B, and successfully constructs a fully localized question-answering system for a TCM knowledge base based on Fourteenth Five-Year Plan textbooks– "Qihuang Assistant." Without a reranking model, the system realizes high-quality RAG question-answering performance through a combined strategy of 10-class intent recognition and classification, hybrid retrieval, hard truncation with a similarity threshold of 0.7, Top-K convergence, and prompt engineering. The test results show that the system execution success rate is 100%, the answer accuracy rate is no less than 95%, the user waiting time does not exceed 3 seconds, and user satisfaction is no less than 95%.

The core contributions of this study are as follows: (1) it establishes an end-to-end methodology from textbook screening, data cleaning, and intent classification to prompt engineering; (2) it systematically summarizes practical experience in compensating for retrieval accuracy through engineering methods in scenarios without a Reranker; (3) it explores the optimal configuration of quantized model selection, sampling-parameter tuning, and reasoning-model decision-making for the open-source large model (Qwen3.6-35B-A3B) in vertical-domain RAG systems; and (4) it provides a replicable local TCM knowledge-base RAG solution, satisfying the rigid requirements of medical data privacy protection and offline deployment.

This study provides a solid technical pathway and engineering reference for the digital inheritance and intelligent application of TCM knowledge, and also lays a complete foundation for the subsequent introduction of multimodal interaction, specialized subsystems, and more efficient models.

6 Statement on AI Use

AI-assisted tools were used in the writing process of this paper, mainly in two respects: first, to polish the sentence structure and academic wording of some paragraphs in order to improve readability; second, the English title, abstract, and keywords were initially translated from the Chinese manuscript with the assistance of an AI translation tool, after which the authors checked and revised them.

The core research content of the paper—including system architecture design, experimental design, parameter tuning, test-data analysis, and the organization of professional knowledge in traditional Chinese medicine—was completed independently by the authors. AI did not participate in the generation of the core academic conclusions. All experimental data, technical schemes, and references are authentic and reliable.

The authors are responsible for the truthfulness of the above statement and are willing to assume corresponding responsibility for any inaccuracies.

7 Appendix

System access link: http://wenzhe.tech:8080/chat/69f093bcd4464bbd

Demonstration video link: https://v.douyin.com/du1Qh5wJStM/

8 Author Contribution Statement:

Han Wenzhe: proposed the research ideas, established the project topic, implemented all steps, optimized the system, and revised the final version of the paper.

Xu Liqin: collected and cleaned the knowledge-base documents, performed system debugging, and drafted the paper.

References

[1] CPC Central Committee and State Council. *Outline of the Healthy China 2030 Plan* [EB/OL]. (2016-10-25) [2026-07-26]. https://www.gov.cn/gongbao/content/2016/content_5133024.htm.

[2] National Administration of Traditional Chinese Medicine; National Development and Reform Commission. Notice on Issuing the *"15th Five-Year Plan" for the Revitalization and Development of Traditional Chinese Medicine* [EB/OL]. (2026-07-23) [2026-07-26]. http://www.natcm.gov.cn/guicaisi/zhengcewenjian/2026-07-22/38398.html.

[3] Baichuan Intelligence. Baichuan Intelligence Official Website [EB/OL]. [2025-12-12]. https://www.baichuan-ai.com.

[4] Feizhiyun. MaxKB Official Website [EB/OL]. [2025-12-12]. https://maxkb.cn.

[5] Qwen Team. Qwen3.6-35B-A3B: Agentic Coding Power, Now Open to All [EB/OL]. (2026-04-15) [2026-07-26]. https://qwen.ai/blog?id=qwen3.6-35b-a3b.

[6] OpenAI. gpt-oss-120b & gpt-oss-20b Model Card [DB/OL]. (2025-08-08) [2026-07-26]. https://arxiv.org/abs/2508.10925. doi:10.48550/arXiv.2508.10925.

[7] DeepSeek-AI. DeepSeek-V4: Towards Highly Efficient Million-Token Context Intelligence [DB/OL]. (2026-04-26) [2026-07-26]. https://arxiv.org/abs/2606.19348. doi:10.48550/arXiv.2606.19348.

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