WHIOCE PUBLISHING PTE. LTD.
PROVIDING
FIRST-CLASS SCIENTIFIC INFORMATION
FOR TOP SCHOLARS

2025 Volume 3, Issue 11
Educational Innovation Research
ISSN: 3029-1852 (Online)
ISSN: 3029-1844 (Print)

Research Review on Practical Course Design for Localization Deployment of Open-source Large Language Models — Practical Path of Industry-Education Collaboration Based on Ollama/vLLM Toolchain

Genyuan Wang, Wenshuang Li
Hainan Vocational University of Science and Technology, Haikou 571126, Hainan, China

Author to whom correspondence should be addressed.

Abstract: Addressing the widespread reliance on cloud-based Application Programming Interfaces (APIs) in AI courses at vocational colleges and students' limited practical experience in localized model deployment, this study systematically reviews the research status and implementation approaches of localized deployment of open-source large language models (LLMs) in vocational AI training programs. Through bibliometric analysis, case study analysis, and theoretical framework construction, we synthesize domestic and international research findings since 2020, focusing on the technological evolution of open-source LLMs, existing challenges in vocational AI education, and critical elements of localized course design. The study reveals a significant structural imbalance in vocational AI education characterized by "technological maturity versus pedagogical lag": while open-source models like LLaMA3 and Qwen2.5 can efficiently run on consumer-grade GPUs, and toolchains such as Ollama and vLLM have substantially reduced deployment barriers, over 90% of institutions still operate AI courses at the cloud API invocation level. Building on this, we propose a localized course design framework encompassing three dimensions: technology stack selection, curriculum module design, and industry-academia collaboration mechanisms. The Ollama platform integrated with the vLLM toolchain employs Low-Rank Adaptation (LoRA) lightweight fine-tuning technology, representing the most technically viable and pedagogically applicable localized AI training solution for vocational colleges. The school-enterprise "dual-track" curriculum collaboration mechanism forms the institutional foundation for ensuring high-quality implementation of such courses. The proposed "theory-instrument-practice" three-stage framework provides replicable reference models for peer institutions across the province.

Keywords: Open-source large language model; Localized deployment; Vocational AI education; Industry-education collaborative training; LoRA fine-tuning; Ollama

Online publication: Dcecmber 20, 2025

1. Introduction

Artificial intelligence technology is reshaping industrial ecosystems at an unprecedented pace, with generative AI

applications centered on large language models now fully entering industrial implementation phases. However, university AI talent development systems have shown significant lag in responding to this technological wave. Research indicates that over 90% of AI courses in Chinese undergraduate and vocational colleges still primarily utilize cloud-based APIs like ChatGPT and Wenxin Yiyang as teaching tools. Students demonstrate limited understanding of core competencies such as model architecture design, local inference environment setup, and parameter optimization, only approximately 12% of graduating students can independently build local inference environments. This cloud dependency phenomenon has driven continuous growth in industry demand for versatile AI professionals with localized deployment capabilities, exacerbating the widening supply-demand gap [1].

From a technical feasibility perspective, the current open-source LLM ecosystem has provided robust support for educational transformation in higher education institutions. Models such as Meta's LLaMA3 series, Mistral AI's Mistral-7B, Alibaba Cloud's Qwen 2.5, and Tsinghua University's KEG Lab ChatGLM can achieve inference speeds of approximately 150 tokens per second on standard computers equipped with consumer-grade GPUs like RTX 4090 or RTX 4060 through 4-bit quantization technology. Meanwhile, the Ollama tool compresses the entire model deployment process into three command-line instructions via containerization, enabling students without specialized backgrounds to complete the workflow from model download to local inference within two hours. vLLM, leveraging its PagedAttention inference acceleration mechanism, has significantly improved batch inference efficiency and has been validated in training platforms at Tsinghua University and other institutions.

Current research remains insufficient in systematically addressing the localized deployment of open-source LLM courses within the specific context of vocational colleges. Existing literature either focuses on performance evaluation of model technologies or remains confined to general discussions on AI education concepts, with few studies establishing comprehensive localized training curriculum frameworks tailored to vocational colleges' educational positioning, hardware infrastructure, and talent development objectives. Additionally, the institutional safeguards provided by school-enterprise collaborative education mechanisms in such courses have not been thoroughly investigated.

This study focuses on the application of open-source LLM localization deployment in AI training courses for vocational colleges. By integrating bibliometric analysis with case studies, it systematically examines technological advancements, addresses pedagogical challenges, and establishes curriculum design frameworks. The research also proposes concrete recommendations for building industry-education collaborative training mechanisms, aiming to provide theoretical insights and practical guidance for AI training curriculum reforms in vocational institutions [2].

2. Current technical development status of open-source LLM localization deployment

2.1. Current status of mainstream open-source model technologies

In recent years, the open-source large language model ecosystem has demonstrated a development trend characterized by diversified parameter scales, mature quantization technologies, and continuously decreasing inference thresholds. Meta AI's LLaMA 2 series released in 2023 established the technical benchmark for open-source LLMs. The subsequent LLaMA 3 series achieved further breakthroughs in instruction adherence, code generation, and multilingual understanding. Its 8 B-parameter version, quantized using the 4-bit GGUF format, achieved real-time inference speeds exceeding 150 tokens/s on RTX 4090 GPUs. Mistral AI's Mistral-7B demonstrated comprehensive performance comparable to GPT-3.5 with relatively compact parameter sizes, particularly through innovations in sliding window attention mechanisms that significantly reduced video memory consumption during long-text inference.

In the open-source model domain in China, Alibaba Cloud's Qwen 2.5 series is renowned for its exceptional Chinese semantic understanding and tool invocation capabilities, offering parameter scales ranging from 0.5 B to 72 B to flexibly adapt to varying computational resources [3]. Meanwhile, Tsinghua University's KEG Lab's ChatGLM4 series excels in dialogue coherence and knowledge reasoning, providing localized deployment solutions optimized for Chinese scenarios. Technologically, these models universally support 4-bit quantization in GGUF format, achieving approximately 75% lower

video memory usage compared to FP16 precision. This enables models previously requiring professional-grade GPUs (e.g., A100 80 GB) to run smoothly on consumer-grade GPUs equipped with 8 GB memory. The combination of "reduced parameter scales + enhanced quantization accuracy + lowered deployment barriers" has fundamentally transformed hardware requirements for AI training in academic institutions.

2.2. Reasoning and deployment toolchain maturity

The maturity of deployment toolchains is a critical enabling factor for open-source LLMs to enter university teaching scenarios. Currently, mainstream tools primarily include three categories: Ollama, vLLM, and llama.cpp, each with distinct technical characteristics and teaching applicability scenarios (see **Table 1**).

Table 1. Comparison of mainstream open-source LLM deployment tools

Tool name	Inference back-end	Minimum video memory requirement	Install and deploy Command count	Batch inference support	Teaching suitability (★)
Ollama	llama.cpp	6 GB (4-bit quantization)	3 items	Limited support	★★★★★
vLLM	PagedAttention	16 GB	More than 5 items	Full support	★★★★
llama.cpp	CPU/CUDA mixing	0 GB (CPU-only)	More than 8 items	Basic Support	★★★
LM Studio	llama.cpp	8 GB	graphical interfaces	nonsupport	★★★★

Ollama stands as the most versatile tool for contemporary educational applications. Through containerized encapsulation, it abstracts complex environment configurations into three core commands: Ollama pull (model retrieval), Ollama run (model execution), and Ollama list (model catalog). Coupled with automated dependency management, this framework compresses full deployment time to under 30 minutes, enabling quantum models with 7 B parameters to operate on just 6 GB of GPU memory. vLLM excels in high-throughput batch processing, leveraging its innovative Paged Attention mechanism that dynamically manages GPU memory paging to achieve inference efficiency several times higher than traditional methods, making it ideal for advanced training requiring multi-user concurrency simulations. llama.cpp, a lightweight inference engine built entirely in C++, supports GPU-free CPU environments for quantum model execution, providing cost-effective entry points for institutions with limited hardware resources. However, its deployment requires complex configuration of compilation environments. The Ollama + vLLM integrated toolkit strikes a balance between "user-friendly accessibility" and "professional sophistication", emerging as the optimal technical solution for vocational AI training programs.

3. Current status and pain points of ai practical training courses in higher vocational colleges

3.1. Analysis of existing teaching models

Current AI training courses in vocational colleges can be broadly categorized into three teaching models. The first type employs pure theoretical instruction, focusing on classroom lectures about neural network principles and deep learning algorithm frameworks, supplemented by simple Python code exercises in practical sessions (accounting for approximately 30%). While this approach helps establish foundational knowledge systems, it remains severely disconnected from real-world industrial applications. The second model utilizes cloud-based API implementation, leveraging open APIs from platforms like ChatGPT, Wenxin Yiyang, and iFlytek's Starfire to guide students through prompt engineering and application layer development (approximately 60%). Although this method offers intuitive accessibility, students remain in a "black-box" operational state without access to underlying model configuration, optimization, or deployment logic. The third model emphasizes localized environment practice, requiring students to build complete inference or fine-tuning environments on local GPU devices (accounting for less than 10% of cases), primarily implemented in select "Double

High" institutions with strong AI education and research capabilities [4].

The above distribution pattern indicates that most vocational colleges still teach AI courses from the perspective of "application consumers" rather than "technology builders", resulting in students lacking in-depth understanding of the underlying operational logic of AI systems and practical hands-on skills. Relevant survey data also corroborates this judgment, the "2024 China Higher Education AI Course Survey" reveals that 90% of undergraduate institutions rely on cloud API teaching for AI courses, while only 12% of graduates possess the ability to independently build local inference environments. The situation in vocational colleges is largely similar to or even more severe than that in undergraduate institutions.

3.2. Core pain points summary

This reveals a profound structural contradiction in the current pedagogical approach. On one hand, the technological development trajectory is clear: the maturation of open-source model ecosystems has led to reduced barriers to quantitative deployment, which in turn creates emerging opportunities for technology adoption in universities. On the other hand, current teaching practices remain largely unchanged: vocational institutions still rely heavily on cloud APIs for instructional models, resulting in significant delays in curriculum iteration. This divergence leads to two core pain points: first, a lack of customized private knowledge base capabilities, and second, severe deficiencies in localized deployment practices. This stark "technology-teaching" structural disparity forms the fundamental research question of this study.

Currently, the open-source large language model ecosystem continues to mature, with models such as LLaMA3, Qwen 2.5, and Mistral-7 B emerging steadily. Coupled with the gradual adoption of 4-bit quantization technology and consumer-grade GPUs, the barrier to local model deployment has significantly decreased, creating favorable conditions for universities to integrate large model technologies into teaching practices. However, AI course instruction in vocational colleges remains heavily reliant on cloud APIs like ChatGPT, Wenxin Yiyang, and iFlytek Spark, with curriculum updates lagging markedly, over 90% of institutions still operate at the API invocation level. The gap between technological advancements and current teaching practices has become increasingly pronounced, manifesting in two core challenges: first, students lack the ability to customize private knowledge bases, as cloud APIs cannot integrate domain-specific knowledge, leaving them generally deficient in RAG (Research and Generation) practical skills; second, localized deployment practices are severely inadequate, with industry demand for relevant talent growing by 38% annually while graduates demonstrate insufficient capability to independently establish local inference environments, a widening supply-demand gap. This structural contradiction between technological progress and teaching lag serves as the fundamental motivation for this study to develop a localized training curriculum system for open-source large language model deployment in vocational colleges.

(1) Key challenge 1

Lack of customized private knowledge base capabilities. Cloud APIs typically provide standardized universal services, making it impossible for educators and students to integrate domain-specific subject knowledge at the API level. Taking the "Machine Learning" course as an example, instructors aim to develop a customized Q&A system capable of accurately answering course-specific terminology queries, exercise analyses, and experimental guidance. However, cloud APIs fail to meet these customized requirements due to data privacy constraints, content limitations, and inability to incorporate proprietary corpora. Retrieval-Augmented Generation (RAG) frameworks represent the mainstream technical solution for this issue, but their full implementation requires locally deployed large language models (LLMs), which cannot achieve end-to-end control in cloud API environments.

(2) Key challenge 2

Substantial Shortage in Localized Deployment Practices. According to the "2024 Hainan Free Trade Port AI Talent White Paper", local enterprises in Hainan face an annual demand growth rate of 38% for professionals skilled in localized AI model deployment, yet only 12% of current college graduates meet these requirements, highlighting a critical talent gap. Pilot studies conducted by institutions like Beijing University of Posts and Telecommunications through vLLM training workshops demonstrate that systematic localized deployment training enables 85% of participants to master basic optimization techniques, proving the effectiveness of structured training programs in

addressing competency gaps. However, constrained by hardware procurement costs, faculty capacity limitations, and institutional inertia in curriculum development, most vocational colleges have yet to establish effective localized deployment training channels [4].

4. Design framework for open-source LLM localization deployment training course

4.1. Overall design concept

This study proposes a curriculum design framework structured around a three-tier progression of "theory-instrument-practice", with industry-education integration serving as institutional safeguards and quantitative competency assessment forming a quality control loop. The three-tier architecture comprises: Theoretical Level (understanding AI fundamentals and open-source model technologies), Tool Level (operating Ollama/vLLM toolchains and environment adaptation), and Practical Level (building proprietary knowledge bases and LoRA fine-tuning practices). The framework adheres to three core design principles: Technical Implementability Principle, all training tasks must run on consumer-grade GPUs (e.g., RTX 4060 with 8 GB VRAM) to ensure hardware accessibility; Transferability Principle – ≥ 80% of tasks should be executable on devices with ≤ 8 GB VRAM to ensure replicability across institutions with limited hardware resources; Curriculum Standardization Principle – developing scalable course packages and assessment tools to enable implementation at over two peer institutions within the province, thereby achieving inclusive AI education.

4.2. Design of six major course modules

Based on the aforementioned design philosophy, this study categorizes practical training courses into six functional modules (see **Table 2**), with a total instructional duration of no less than 36 class hours, thereby establishing a comprehensive competency development pathway that progresses from foundational environmental training to integrated practical application.

Table 2. Competency objectives and assessment indicators for six major practical training modules

Module number	Module name	Core skill objectives	Propose class hour	Evaluation mode	Competency achievement indicator
Module 1	Environmental adaptation	Basic environment setup for linux/ python/cuda and installation testing of Ollama toolchain	4 class hours	Operational assessment	Environmental setup success rate ≥ 95%
Module 2	Model selection and download	Selection principles for mainstream open-source models (e.g., LLaMA3, Qwen2.5) and analysis of GGUF quantitative format	4 class hours	Answer by word of mouth	can explain the selection basis
Module 3	Localized deployment training	Master the three commands: pull, run, and list to perform local model inference	6 class hours	Practical test	Deployment success rate ≥ 95%
Module 4	vLLM inference acceleration	Practical training on paged attention principle understanding and batch processing inference	6 class hours	Practical training + Report	Throughput increase ≥ 50%
Module 5	Private knowledge base construction	Vectorization of PDF/markdown courseware and construction of RAG framework	8 class hours	Project works	Question-answer accuracy ≥ 80%
Module 6	LoRa lightweight fine-tuning	Complete domain fine-tuning with ≤ 50 lines of code and enhance professional term recognition capabilities	8 class hours	Task acceptance	Improvement in professional term recognition accuracy by ≥ 15%

(1) Module 1

Linux/Python/CUDA Environment Adaptation (4 hours) introduces students to GPU driver installation, CUDA environment configuration, and Ollama one-click deployment toolchain testing, establishing essential software/hardware foundations for subsequent modules.

(2) Module 2

Mainstream Open-Source Model Selection & Download (4 hours) guides learners in understanding application scenarios and parameter scale differences of popular models like LLaMA3 and Qwen2.5, while mastering GGUF quantization format download and validation methods.

(3) Module 3

"Three Commands for Local Deployment" Standardized Training (6 hours) requires independent execution of three core commands: Ollama pull (model retrieval), Ollama run (model execution), and Ollama list (model catalog), with deployment success validated through standardized evaluation tasks.

(4) Module 4

vLLM Inference Acceleration Integration (6 hours) explains Paged Attention architecture principles, demonstrates throughput optimization through batch inference training, and covers basic vLLM service-oriented deployment configurations.

(5) Module 5

Private Knowledge Base Construction (8 hours) teaches vectorization processing of PDF/Markdown course materials into local vector databases, followed by domain-specific Q&A system development using RAG framework.

(6) Module 6

Covers practical LoRA lightweight fine-tuning with 8 class hours, employing a design constraint of ≤ 50 lines of code. It guides students through domain-specific fine-tuning tasks for professional term recognition, demonstrating quantifiable improvements in professional vocabulary response accuracy before and after fine-tuning.

4.3. Quantitative competency assessment criteria

This study proposes three quantifiable objectives for student competency development to establish clear benchmarks for teaching effectiveness evaluation.

(1) Objective 1

Through Module 3 practical training, 90% of students should acquire localized deployment capabilities for single models with a deployment success rate of no less than 95%.

(2) Objective 2

After completing Module 6 training, 85% of students should independently complete domain-specific fine-tuning tasks, achieving at least a 15% improvement in professional term recognition accuracy compared to baseline models.

(3) Objective 3

Following Module 5 training, 80% of students should independently construct discipline-specific private knowledge bases and accomplish autonomous deployment and testing of question-answering systems.

This study recommends employing radar chart evaluation methods to visually assess students' comprehensive AI practical competencies Dashed lines indicate students' initial proficiency levels before course enrollment (approximately 20–30% across all dimensions), while solid lines represent expected proficiency levels after completing all six module training modules (90% for model deployment capability, 80% for knowledge base construction capability, 85% for model fine-tuning capability, 75% for problem-solving ability, and 70% for innovation capability). These evaluation metrics are implemented through a "teaching-learning-assessment" closed-loop mechanism: instructors design standardized assessment tasks based on module objectives, students validate their competencies through hands-on practice and report submissions, and educators continuously optimize course content and difficulty gradients according to quantitative results, thereby establishing a dynamic iterative course quality assurance system.

The course employs radar chart evaluation methodology to visually assess students' comprehensive AI practical competencies across five dimensions: model deployment capability, knowledge base construction capability, model fine-tuning capability, problem-solving ability, and innovation capability. Prior to course enrollment, students demonstrated generally low baseline proficiency, with all dimensions ranking between 20% and 30%, reflecting relatively weak foundational skills before systematic hands-on training. Upon completing all six practical modules, expected improvements include: model deployment capability reaching 90%, knowledge base construction capability achieving 80%, model fine-tuning capability reaching 85%, while problem-solving and innovation capabilities rise to 75% and 70% respectively, marking significant advancements across all dimensions. These evaluation metrics are implemented through a closed-loop "teaching-learning-assessment" mechanism. Instructors design standardized assessment tasks aligned with module objectives, while students demonstrate their competencies through practical operations and report submissions. Based on quantitative feedback, instructors continuously refine course content and difficulty levels, establishing a dynamic, iterative quality assurance system that ensures measurable, traceable, and improvable training outcomes.

5. Pathways for establishing industry-education collaborative talent cultivation mechanisms

5.1. Dual-track curriculum development model for schools and enterprises

The industry-academia collaborative education mechanism serves as the institutional foundation for ensuring high-quality implementation of localized training courses for open-source LLM systems. Taking the cooperative project between Hainan Vocational University of Science and Technology and Zhejiang Hangda Technology Development Co., Ltd. as a case study, this paper proposes a "dual-track" curriculum development model. The "dual-track" approach refers to the simultaneous operation of two parallel development pathways: the "technical track" and the "teaching track". The technical track is enterprise-led, with Zhejiang Hangda Technology Development Co., Ltd. providing industrial-grade toolchain licensing and compatibility support, delivering real-world AI system integration project case studies and datasets, and arranging on-campus training sessions led by enterprise engineers with frontline development experience to ensure curriculum alignment with industry advancements. The teaching track is university-driven, where academic teams systematically design course frameworks and pedagogically validate learning objectives. They develop standardized teaching resources using enterprise-provided technical materials while conducting real-time teaching process monitoring and academic research output [5].

The dual-track system achieves deep integration through a collaborative operational framework of "resource interdependence and joint talent development". Corporate engineers delivering on-campus lectures synergize with tiered practical training programs designed by faculty teams, balancing hands-on skill acquisition with adherence to pedagogical principles. The Ministry of Education's "Management Measures for Industry-Academia Collaborative Talent Cultivation Projects" provides clear institutional foundations and policy support for this mechanism, ensuring standardized division of responsibilities, outcome ownership, and fund utilization between universities and enterprises. This dual-track collaboration model, centered on the principle of "enterprises providing resources while universities establishing standards", effectively addresses the dual challenges of "low corporate engagement" and "lagging technological innovation in academia" prevalent in traditional industry-academia partnerships.

5.2. Replicable course promotion model

To ensure the scalability of this course framework, this paper proposes a "1 core curriculum package + N adaptation pathways" implementation strategy. The core curriculum package includes: standardized deployment command templates (comprehensive Ollama/vLLM workflow manuals), LoRA fine-tuning task libraries (task sets with ≤ 50 lines of code containing 5 domain-specific case studies), courseware digitization tools (supporting one-click vectorization from PDF/Markdown materials), and integrated quantitative evaluation tools. The N adaptation pathways are tailored to universities

with varying computing resources through differentiated hardware threshold solutions: Using RTX 4060 (8 GB VRAM) as benchmark configurations, the system supports over 80% of practical training tasks, enabling institutions with limited computing resources to rapidly replicate and implement this curriculum framework.

The course rollout follows a four-phase implementation roadmap: Phase 1 (Months 1–3) focuses on curriculum framework design and experimental environment adaptation, culminating in standardized framework design reports, experimental environment adaptation guidelines, and initial syllabus drafts. Phase 2 (Months 4–6) involves deploying practical training modules and developing instructional resources, producing six standardized operational manuals for training units and interdisciplinary knowledge transformation toolkits. Phase 3 (Months 7–9) entails fine-tuning instructional unit designs and conducting pilot programs, including the development of LoRA fine-tuning task libraries (with five case studies) and pilot program reports demonstrating ≥ 15% improvement in professional term recognition accuracy. Phase 4 (Months 10–12) evaluates teaching effectiveness and promotes nationwide adoption, generating curriculum standardization reports applicable to at least two provincial institutions. Final deliverables include a comprehensive teaching resource package, three academic papers covering localized deployment methodologies, low-code fine-tuning practices, and industry-academia collaboration mechanisms, along with a curriculum standardization report scalable to similar institutions across the province.

6. Conclusion

This study focuses on the application of localized deployment of open-source large language models (LLMs) in AI training courses at vocational colleges, conducting a systematic review from four dimensions: current technological advancements, analysis of teaching challenges, curriculum design frameworks, and industry-academia collaboration mechanisms. Research findings indicate that open-source LLM technology ecosystems represented by LLaMA3 and Qwen2.5 have achieved high maturity. The widespread adoption of 4-bit quantization technology and consumer-grade GPUs has fundamentally eliminated hardware barriers for localized AI training in higher education institutions. The Ollama + vLLM toolchain demonstrates comprehensive advantages in deployment convenience, inference efficiency, and instructional applicability, establishing it as the optimal technical solution for AI training programs in vocational colleges.

The theoretical contributions of this study are reflected in three aspects. First, we established a three-tier progressive framework for open-source LLM localization deployment training courses encompassing “theory-instrument-practice”, providing actionable curriculum design paradigms for peer institutions. Second, we proposed a quantifiable student competency evaluation standard comprising three core metrics: model deployment success rate, LoRA fine-tuning effectiveness, and private knowledge base completion rate, facilitating the transition from subjective qualitative assessments to objective quantification in AI training program evaluation. Third, we explored and validated the feasibility of a “dual-track” curriculum collaboration mechanism between academia and industry, offering institutional model references for deep integration of industry-education partnerships in AI specialized courses [6].

This study also has several limitations requiring further exploration. Firstly, the current curriculum framework design and validation are primarily based on project practices from a single institution, with limited sample size, necessitating broader empirical validation to confirm the generalizability of research findings. Secondly, while this paper focuses on localized deployment of text-based LLMs, the design of localized deployment courses for multimodal models (such as visual-language models) remains unaddressed, representing an important area for future investigation. Additionally, the implementation effects of replicating the curriculum across multiple institutions require longitudinal tracking studies for evaluation. As open-source AI ecosystems continue to mature and hardware costs decline, localized AI training is expected to become a standard component in vocational colleges’ AI education programs. Concurrently, research on curriculum design theories and practical mechanisms will advance toward greater depth and systematicness.

Funding

Ministry of Education Industry-Academia Cooperation Collaborative Talent Cultivation Project: "Training Course Design for Localized Deployment of Open-Source AI Models" (Project No.: 250700409014913)

Disclosure statement

The authors declare no conflict of interest.

References

[1] Wang H, Lu W, Wang C, 2025, Design and Implementation of Distributed Local Large Model Services Based on OpenWebUI and LiteLLM. Yangtze River Information and Communications, 38(12): 20–24.

[2] Deng Y, 2025, Research on Information Security Risk Control Strategies for Localized Deployment of Large Models in Archival Management. Heilongjiang Archives, 2025(5): 141–143.

[3] Jin H, Cui C, 2025, Research on Customized AI Large Models Based on Localization and Privatized Databases. Architectural Technology, 56(20): 2480–2482.

[4] Song H, 2025, IT Operations Knowledge Base Based on Localized Deployment of Artificial Intelligence Models. Computer Programming Techniques and Maintenance, 2025(9): 122–124 + 132.

[5] Li R, Yang J, 2025, Research on Localization Deployment and Security Protection System Based on DeepSeek Large Model. Wireless Interconnection Technology, 22(17): 1–6.

[6] Tao X, 2024, Research on Large Language Model-Based Intelligent Q&A System Based on Hybrid Architecture. Post and Telecommunications Design Technology, 2024(5): 48–55.