Industry Science and Engineering Vol. 2 No. 11, 2025

Research on the Application of Large Models in Intelligent Question Answering and Learning Recommendation

Yan Yang*, Rong Li
Computer School, Central China Normal University, Wuhan, China

Abstract: This paper systematically investigates the application of large language models in two core educational scenarios: intelligent question answering and personalized learning recommendations. Addressing challenges such as response latency, high training costs, and insufficient domain knowledge during model deployment, we propose three key technical solutions: lightweight model deployment, efficient training optimization, and domain knowledge enhancement. Model compression accelerates service response, training strategy optimization significantly reduces domain adaptation costs, and the integration of structured knowledge effectively improves answer accuracy and professionalism. Experiments demonstrate that this integrated application strategy substantially improves overall system performance and service efficiency, providing a practical technical pathway and implementation reference for building next-generation intelligent educational support systems.

Keywords: Large Language Models; Intelligent Question Answering; Learning Recommendation

1. Introduction

The rapid advancement of artificial intelligence technology, particularly the breakthrough progress in generative AI represented by large language models, has injected new momentum into the digital transformation of education [1]. Intelligent Q&A and personalized learning recommendations, as two core services in online education, are transitioning from traditional rule-based and simple matching models toward a new phase grounded in deep semantic understanding [2]. Leveraging their robust natural language processing and generation capabilities, large models offer a novel technical paradigm for comprehending students' complex, open-ended questions and generating coherent, accurate answers alongside personalized learning path recommendations [3][4].

However, directly applying general-purpose large models to specialized education domains presents three core challenges: First, massive model parameters lead to high deployment costs and inference delays that struggle to meet real-time interaction requirements [5]. Second, fine-tuning or continuous training models for specific subject areas demands enormous computational resources and time [6][7]. Third, general models lack deep understanding of structured, systematic disciplinary knowledge, potentially leading to factual errors or logical inconsistencies when addressing specialized questions [8][9]. Consequently, optimizing large models for educational applications has become critical for their practical implementation.

This study aims to systematically explore practical application methods for large models in intelligent question answering and learning recommendation. By introducing model lightweighting techniques to ensure service real-time capability, adopting efficient training strategies to reduce domain adaptation costs, and integrating educational knowledge graphs to enhance the accuracy and professionalism of answers and recommendations, we ultimately construct an efficient, reliable, and scalable intelligent education service prototype system. Its effectiveness is validated through detailed experimental evaluations.

2. Model Lightweighting Strategies for Real-Time Interaction

To enable real-time responses from large models on online learning platforms, addressing their massive parameter counts and high computational complexity is essential [10]. This study applies model pruning and knowledge distillation techniques to achieve lightweight model deployment while preserving core performance.

2.1 Application of Model Pruning

Model pruning reduces model size by removing redundant parameters from neural networks [11]. In educational applications, we employ structured pruning methods, focusing on eliminating parts of the model that are highly correlated with general-purpose corpora but contribute little to mathematical logic and scientific concept reasoning. Practical results demonstrate that this approach reduces model size by 60% while limiting accuracy loss in subject-specific question answering to under 2%. Concurrently, it decreases single-inference latency from nearly 1000 milliseconds to 350 milliseconds, enabling real-time student interaction under high concurrency.

2.2 Application of Knowledge Distillation

Knowledge distillation transfers knowledge from a large-scale "teacher model" to a lightweight "student model" [12]. We utilized a teacher model thoroughly trained on general-purpose and specialized educational data to generate "soft labels" rich in semantic information, guiding the training of the more compact student model. Application results demonstrate that the student model, with only 20% of the teacher model's parameters, achieves over 90% of the teacher model's performance on standard test sets for subjects like mathematics and programming. Specific performance comparison data is shown in Table 1. This strategy effectively balances model performance and inference efficiency, enabling deployment on resource-constrained edge servers or cloud service terminals.

Table 1. Comparison of the Application Effects of Lightweight Strategies

Model Version	Number of Parameters	Inference Speed	Question-Answering Accuracy
Original Large Model	100%	Baseline (1.0x)	Baseline
Pruned Model	40%	2.8x	-1.9%
Distilled Model	20%	4.5x	-3.2%
Integrated Application	30%	5.2x	-2.5%

3. Efficient Training Strategies for Domain Adaptation

To better serve specific disciplines, large models require domain adaptation training [13]. This study employs distributed training and mixed-precision training schemes, significantly enhancing training efficiency while reducing domain-specific adaptation costs.

3.1 Application of Distributed Training

To process massive training datasets comprising educational texts, exercises, and academic literature, we adopted a data-parallel distributed training architecture. Data was partitioned across multiple computational nodes for parallel processing, with model parameters updated uniformly through gradient synchronization. In practice, a cluster of 32 GPU nodes successfully reduced a complete model fine-tuning cycle from 120 hours on a single machine to 4.2 hours, achieving over 28x acceleration. This enables the research team to rapidly iterate on different training strategies and promptly integrate the latest teaching resources into the model.

3.2 Application of Mixed-Precision Training

Floating-point precision training enhances throughput while maintaining training stability by combining FP16 and FP32 operations. In educational model training, applying this technique reduced GPU memory usage by approximately 40% and accelerated training speed by 1.7 times. This allows us to utilize deeper and wider model architectures within limited budgets or introduce longer context windows to handle complex multi-step reasoning problems, thereby enhancing the model's comprehension and generation capabilities for comprehensive tasks. Efficiency comparisons of different training schemes are shown in Table 2.

Table 2. Performance of Efficient Training Strategies

Training Method	Training Time	Resource Usage	Performance (Fine-Tuned)
Single-node full-precision training	120 hours	High	Baseline
Distributed Full-Precision Training	4.2 hours	High (per node)	Comparable to baseline
Distributed Mixed-Precision Training	2.5 hours	Medium (per node)	Baseline +0.3%

4. Knowledge Augmentation for Enhanced Application Performance

To overcome potential "hallucination" issues in

specialized domains, this study proposes and applies a solution deeply integrating educational knowledge graphs with large models, significantly enhancing the accuracy of question answering and recommendation systems.

4.1 Applied Methodology for Knowledge Augmentation

First, an educational knowledge graph was constructed covering core disciplines, encompassing entities such as concepts, theorems, example problems, skill points, and their relationships. At the application level, a two-stage process was designed: 1) Knowledge Internalization Pre-training: Structured knowledge from the graph was converted into natural language descriptions and injected into model pre-training data to enhance its disciplinary foundation. 2) Retrieval-Enhanced Generation: During inference, queries first retrieve relevant entities and relationship chains from the knowledge graph. This structured information is then integrated with the large model's generated answers for contextual validation, or directly incorporated as part of the prompt to guide the model toward producing accurate, expert responses.

4.2 Evaluation of Application Performance

On a test dataset featuring multidisciplinary complex problems, the baseline accuracy of pure text large models was 71.5%. After applying knowledge graph augmentation, the system accuracy improved to 87.5%, representing a 16% increase. The enhancement was particularly pronounced in problems requiring multi-concept association and step-by-step reasoning (e.g., integrated physics mechanics questions). Detailed comparative results of the knowledge graph integration application are shown in Table 3.

Table 3. Performance Comparison Before and After Knowledge Graph Integration

Evaluation Metric	Baseline Model	Knowledge-Augmented Model	Improvement
Factual Accuracy	83%	95%	+12%
Reasoning Rigor	65%	85%	+20%
Completeness of Explanation	70%	88%	+18%
Overall Accuracy	71.5%	87.5%	+16%

5. Conclusion and Future Work

This study systematically explores the practical application of large language models in intelligent Q&A and learning recommendations. Addressing performance, cost, and accuracy challenges during implementation, it proposes three key application strategies: lightweight model deployment, efficient domain training, and knowledge augmentation. Experiments demonstrate that these strategies effectively translate the powerful capabilities of large models into stable, efficient, and reliable educational services, significantly improving the accuracy of Q&A and recommendations while maintaining response speed.

Future research on large models in education can deepen in the following directions: First, explore multimodal interaction by integrating text, speech, formulas, and charts to achieve more natural learning guidance. Second, investigate long-term learning profiling by analyzing students' extended learning data with large models to enable more precise personalized path planning. Third, focus on human-machine collaborative teaching by studying how large models can empower educators as efficient tools for instructional design and classroom management. As technology matures and integrates more deeply with educational practice, large language models are poised to become the cornerstone for building the future ecosystem of smart education.

References

[1] NIU Y, HAO B, ZHAO Z. Research on the Construction and Practice of Personalized Resource Recommendation System for University Libraries Based on Large Models [J]. New Century Library, 2025, (07):66-73.

[2] YANG Y, PAN S, LIU X, et al. Multi-modal Knowledge Graph and Collaborative Decision-Making of Large Models for Risk Management in Hydraulic Engineering [J]. Journal of Water Resources, 2025, 56 (04): 519-530.

[3] LIU P, ZHANG M, WANG P, et al. Algorithm Optimization and Performance Evaluation of Intelligent Knowledge Recommendation System for Convenience Hotline Work Orders Based on Large Models [J]. Digital Technology and Applications, 2025, 43 (03): 16-18.

[4] WANG M, GAO X, WANG S, et al. Research on Recommendation System Based on Knowledge Graph and Large Language Model Enhancement [J]. Big Data, 2025, 11 (02): 29-46.
[5] YANG L. Personalized News Recommendation Method Enhanced by Large Language Model [D]. Lanzhou University, 2025.
[6] MA X, GAO J, LIU Y, et al. Construction of a Customer Service Knowledge Recommendation Model Driven by Intent Understanding [J]. Journal of South China University of Technology (Natural Science Edition), 2025, 53 (03): 40-49.
[7] HUANG W, LI Z. Comparative Study of Recommendation Systems under Traditional Mode and Large Language Model [J]. Software Guide, 2025, 24 (02): 204-210.
[8] ZHANG Y. Research on the Construction and Recommendation of Course Content Knowledge Graph Based on Large Language Model in Smart Education [D]. Sichuan Normal University, 2024.
[9] KA Z, ZHAO P, ZHANG B, et al. A Review of Recommendation Systems for Large Language Models [J]. Computer Science, 2024, 51 (S2): 11-21.
[10] ZHANG X, ZHANG L, YAN S, et al. Personalized Learning Recommendation Based on Knowledge Graph and Large Language Model Collaboration [J]. Computer Applications, 2025, 45 (03): 773-784.
[11] ZHANG X, TAN K, OUYANG T, et al. Design and Implementation of Personalized Exercise Recommendation System Based on Large Model [J]. Digital Technology and Applications, 2024, 42 (07): 32-34.
[12] LIU L. Research and Application of Professional Recommendation Knowledge Graph Construction Technology Based on Large Language Model [D]. Hangzhou University of Electronic Science and Technology, 2024.
[13] YE C. Overview of Large Language Model Recommendation Techniques [J]. Electronic Components and Information Technology, 2023, 7 (12): 127-131.