Research on Optimization of Large Model Recommendation Algorithm for Intelligent Customer Service

Yan Yang, Sai Wang
Computer School, Central China Normal University, Wuhan, China

Abstract: This paper addresses the bottlenecks of large language models in intelligent customer service scenarios, such as high response latency, significant resource consumption, and lack of domain knowledge, by conducting systematic optimization research on recommendation algorithms. Model lightweighting is achieved through the integration of model pruning and knowledge distillation techniques. Training efficiency is enhanced by combining distributed and mixed-precision training. Furthermore, domain knowledge graphs are innovatively introduced to improve the accuracy and reliability of generated responses. Experiments demonstrate that the optimized system significantly accelerates response times while effectively improving accuracy in handling complex specialized queries, providing a viable pathway for efficient and precise application of large models in vertical business scenarios.

Keywords: Recommendation Algorithm; Large Language Model; Intelligent Customer Service

1. Introduction

As enterprise digital transformation deepens and customer service standards rise, intelligent customer service systems have become critical hubs connecting businesses with clients, handling massive volumes of inquiries, troubleshooting, and transaction processing [1]. To enhance service efficiency, quality, and reduce operational costs, this paper proposes a large-model-based intelligent customer service ticket knowledge recommendation system with systematic algorithmic optimizations. The paper first outlines the system's overall architecture, then focuses on algorithm optimization strategies for large model deployment. These encompass three key directions: model refinement, training acceleration, and domain knowledge integration. Finally, comprehensive performance evaluations are conducted on the optimized system, covering critical metrics such as accuracy, recall rate, and response time. Experimental results demonstrate significant improvements in both recommendation precision and processing efficiency, proving the system's broad application potential in enterprise-level intelligent customer service scenarios.

Traditional customer service heavily relies on human agents' experience and memory, suffering from slow response times, inconsistent service standards, and difficulty handling complex multi-turn conversations and personalized requests. While rule-based or traditional machine learning-based chatbots have seen limited adoption, their comprehension and generalization capabilities remain limited [2][3]. Large language models (LLMs), exemplified by Transformers, demonstrate exceptional performance in natural language understanding and generation, offering a new technical pathway for intelligent customer service upgrades [4][5]. However, directly applying large models to real-time customer service scenarios still faces challenges such as massive model parameters, high computational resource consumption, significant real-time response latency, and insufficient domain expertise [6][7]. Therefore, this paper aims to address these challenges by implementing targeted algorithmic optimizations for large models, thereby enhancing system usability and cost-effectiveness in real business environments while ensuring service quality.

2. Model Pruning

Despite their formidable capabilities, large models' massive parameter counts (often reaching hundreds of billions) result in prohibitively high deployment costs and failure to meet real-time interaction latency requirements [8]. To effectively apply large models in intelligent customer service, a scenario demanding extreme real-time responsiveness, this study employs two mainstream model reduction techniques: model pruning and knowledge distillation, aiming to achieve an optimal balance between performance and efficiency [9].

Model pruning reduces model size and computational load by identifying and removing redundant weights or neurons with minimal contribution to outputs [10]. Research indicates that many large models contain a significant number of parameters with extremely low activation rates during inference [11]. By applying structured pruning techniques, we successfully reduced the model's parameter count by approximately 60%, while only incurring a 1.5% accuracy loss on the customer service intent classification task, dropping from the baseline 94.2% to 92.7%. This demonstrates significant parameter redundancy within the model, confirming pruning as an effective lightweighting approach.

Knowledge distillation employs a "teacher-student" framework to transfer rich knowledge, including "soft labels" (output probability distributions) and intermediate layer features, from a large teacher model to a more compact student model [12]. In this system, we employed a 10-billion-parameter Teacher Model (95.1% accuracy) to guide training of a 1-billion-parameter Student Model. Experimental results show the distilled Student Model achieved 93.8% accuracy, significantly outperforming a similarly scaled model trained from scratch (91.2% accuracy), while accelerating inference speed by nearly fivefold. Table 1 illustrates the comparative optimization effects of model pruning versus knowledge distillation. In practical AI customer service scenarios, we adopted a synergistic strategy of pruning followed by distillation. The final optimized model reduced average response time during peak concurrent user inquiries from 3.5 seconds to 1.2 seconds while maintaining over 93% intent recognition accuracy, laying a solid foundation for seamless customer service interactions.

Table 1. Comparison of Model Pruning and Knowledge Distillation Effects

Optimization Method	Reduction in Parameters	Accuracy Change	Inference Speed Improvement
Model Pruning	~60%	-1.5% (94.2% → 92.7%)	Approx. 2.3 times
Knowledge Distillation	~90%	-1.3% (95.1% → 93.8%)	Approx. 5.0 times
Pruning + Distillation	~85%	-2.1% (95.1% → 93.0%)	Approx. 7.5 times

3. Accelerated Training

Building efficient large models requires massive domain-specific data for training, while traditional single-machine training is time-consuming and costly [13]. To shorten model iteration cycles and rapidly adapt to customer service business changes, this study introduces two key technologies: distributed training and mixed-precision training, significantly enhancing training efficiency.

3.1 Distributed Training

We adopted a distributed training strategy based on data parallelism. Large-scale customer service dialogue datasets were partitioned and distributed across multiple GPU computing nodes (e.g., 64 nodes) for parallel processing. Each node maintained an identical model replica, independently computed gradients, and then synchronized gradients and updated model parameters via All-Reduce operations. This approach nearly linearly increased training throughput. Experiments demonstrate that compared to single-machine training, 64-node distributed training reduces the time required to complete one model training cycle from 120 hours to approximately 2.5 hours, achieving a 48x acceleration ratio. This significantly enhances the agility of algorithm development and model updates.

3.2 Mixed-Precision Training

Mixed-precision training simultaneously utilizes single-precision (FP32) and half-precision (FP16) floating-point numbers during model training. This approach substantially reduces memory consumption and accelerates computation while maintaining training accuracy. Specifically, FP16 is employed for forward and backward propagation computations, while weights are backed up in FP32 for precision-sensitive gradient update stages. After applying this technique, memory consumption during model training decreased by approximately 40% on identical hardware, while training speed increased by an additional 30%-50%. This enables training larger models or utilizing longer dialogue histories within limited resources.

Table 2 comprehensively presents performance metrics across different training configurations, clearly demonstrating the substantial benefits of distributed and mixed-precision training.

Table 2. Performance Comparison under Different Training Configurations

Training Configuration	Training Duration	GPU Memory Usage	Final Model Accuracy
Single-machine Training (FP32)	120 hours	48 GB	94.5%
Distributed training (64 nodes, FP32)	2.5 hours	48 GB per node	94.5%
Distributed + Mixed-Precision Training	1.7 hours	29 GB per node	94.3%

4. Integration of Knowledge Graphs

While large language models possess robust general language comprehension capabilities, they still exhibit "hallucinations" or knowledge lag when processing highly specialized, structured, and dynamically updated domain knowledge (e.g., product information, policy terms, error codes) in enterprise customer service scenarios [14][15]. To address this challenge, this paper deeply integrates domain knowledge graphs with large language models, constructing an intelligent recommendation system that combines powerful generalization capabilities with precise domain knowledge retention.

4.1 Integration Methodology

First, we constructed a customer service domain knowledge graph covering core entities including enterprise products, users, orders, fault symptoms, and solutions. The graph is organized as "entity-relationship-attribute" triples, e.g., (Product A, Supported Feature, Remote Upgrade), (Error Code E1001, Corresponding Solution, S_005).

Second, we designed a graph retrieval-enhanced generation framework. When a user inputs a query, the system first leverages the large model to parse the query intent and extract key entities. It then retrieves relevant subgraphs (entities and associated relationships) from the knowledge graph. Finally, the original query and retrieved structured knowledge are fed into the large model to guide it in generating accurate, reliable recommended answers or solutions. This paradigm ensures the large model's responses are strictly constrained by the official corporate knowledge base, significantly enhancing information accuracy and reliability.

4.2 Data Support and Case Analysis

We conducted testing on a dataset comprising 5,000 real customer service dialogues. The baseline pure large model system achieved an average response accuracy of 72%, with notably higher error rates when addressing specific product parameters and the latest promotional policies.

After integrating the knowledge graph, system performance improved significantly. For example, when a user asked, "My X-model phone hasn't been fast charging recently. What should I do?", the system accurately identified the entity "X-model phone" and linked it to graph nodes such as "charging protocol," "compatible charger list," and "common troubleshooting steps," ultimately providing standardized, step-by-step troubleshooting advice. Testing revealed that the system's average accuracy rose to 88% after integrating the knowledge graph. Its advantages became even more pronounced when handling complex, multi-constraint queries like "rules for using combined coupons."

Table 3 quantitatively demonstrates the system's performance comparison across different types of customer service inquiries before and after knowledge graph integration.

Table 3. System Performance Comparison Before and After Knowledge Graph Integration

Problem Type	Baseline Model Accuracy	Graph-Augmented Model Accuracy	Improvement Rate
Simple FAQ	85%	92%	+7%
Complex Troubleshooting	65%	87%	+22%
Policy and rule interpretation	68%	90%	+22%
Multi-entity Correlation Query	58%	82%	+24%
Overall Average	72%	88%	+16%

5. Conclusions

This paper designs and implements a ticket knowledge recommendation system tailored for intelligent customer service scenarios. Addressing efficiency and accuracy challenges in large model deployment, it proposes a series of algorithmic optimizations encompassing model refinement, accelerated training, and knowledge fusion. Experimental evaluations conclusively demonstrate that the optimized system achieves significant improvements in recommendation accuracy, response speed, and domain adaptability. This enables effective empowerment of customer service centers, simultaneously achieving cost reduction, efficiency gains, and enhanced user experience. In future, we will continue exploring directions such as multimodal information fusion (e.g., voice, screenshots), intelligent customer sentiment perception, and personalized recommendations. We are committed to building more efficient and versatile lightweight application frameworks for vertical domain large models, driving deeper advancements in intelligent customer service technology.

References

[1] YUAN P, ZHANG F. Application of Adaptive Large Model Architecture in Intelligent Recommendation [J]. Information Recording Materials, 2025, 26 (09): 95-97.

[2] WU Y, LU J. Multi-modal Information Generation and Recommendation Driven by Large Models [J]. Journal of Henan Normal University (Natural Science Edition), 2025, 53 (05): 145-151+181.

[3] 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.

[4] YOU X, LI S, SHAO H. Construction of an Intelligent Question and Answer System for Clothing Recommendation Based on Large Models [J]. Woolen Textile Technology, 2025, 53 (05): 87-94.

[5] 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.

[6] ZHOU X. Research on Efficient Recommendation Algorithm Based on Large Language Model [D]. University of Electronic Science and Technology of China, 2025.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] ZHU M. Research on Personalized Resource Recommendation Method Based on Large Language Model [J]. Journal of Lanzhou University of Arts and Sciences (Natural Science Edition), 2024, 38 (05): 59-64.

[13] ZHOU X, DENG X, HUANG W. Large Models and Recommendation Systems Open a New Chapter in Personalized Recommendation [J]. Shanghai Informatization, 2024, (09):35-38.

[14] WU G, QIN H, HU Q, et al. Research on Large Language Models and Personalized Recommendations [J]. Journal of Intelligent Systems, 2024, 19 (06): 1351-1365.

[15] 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.