Cognitive Dialogue: Modular LLM Agents for Context-Aware and Multimodal Customer Service

Shaojue Yan*
Xi'an Jiaotong-Liverpool University, Suzhou, China
**Corresponding Author*

Abstract: As companies expand, the need for effective customer service has grown, putting pressure on traditional methods due to rising costs and resource constraints. Shen (2025) explained that recent developments in Artificial Intelligence (AI), particularly with Large Language Models (LLMs), provide promising solutions by offering automated, scalable, and cost-efficient customer support. This paper examines the role of LLMs in intelligent customer service systems, highlighting their ability to hold dynamic conversations, provide multi-language support, and deliver personalized service. By incorporating advanced techniques such as Prompt Engineering and Retrieval-Augmented Generation (RAG), we present a model designed to improve LLM performance, enhancing efficiency, accuracy, and customer satisfaction. Furthermore, we compare the evolution of intelligent customer service systems, contrasting rule-based and deep learning-based models. Our findings suggest that LLM-driven systems can significantly boost service efficiency, lower operational costs, and improve user experiences.

Keywords: LLMs; RAG; Multimodal Interfaces; Human-Computer Collaboration

1. Introduction

The rapid expansion of modern businesses has led to an increasing demand for customer service personnel. Traditional customer service systems face significant challenges in handling large-scale interactions, including agent fatigue, high operational costs, and the need for extensive training. With advancements in artificial intelligence (AI), large language models (LLMs) have demonstrated exceptional capabilities in communication and dialogue, making them a promising solution for intelligent customer service applications.

LLMs offer distinct advantages in this domain. First, their advanced dialogue capabilities enable them to analyze user sentiment and deliver contextually appropriate responses. Second, they inherently support multilingual interactions, breaking language barriers in global customer service. Third, LLMs can efficiently process high volumes of user queries simultaneously, addressing scalability issues inherent in human-centric systems.

However, integrating LLMs into customer service raises critical human-computer interaction (HCI) considerations: (1) ensuring the naturalness and accuracy of generated content, (2) enabling LLMs to accurately interpret user emotions and intents, and (3) equipping them with domain-specific knowledge to provide professional responses. To address these challenges, our approach leverages Prompt Engineering combined with Retrieval-Augmented Generation (RAG). This hybrid framework enhances the model's ability to retrieve and synthesize relevant information from structured knowledge bases, ensuring both precision and contextual relevance.

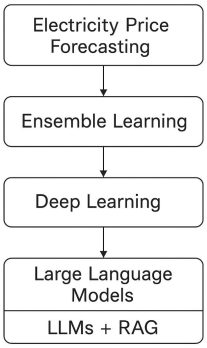

Figure 1. Framework of Cognitive Dialogue.

Our enhanced model delivers several key benefits:

Improved efficiency: 24/7 availability for uninterrupted service.

Cost reduction: Minimizing reliance on human agents.

Personalized interactions: Tailoring responses to individual user needs.

Cross-language support: Seamlessly serving diverse linguistic groups.

Additionally, rigorous evaluation metrics ensure high accuracy in responses, positioning our solution as a robust alternative to conventional systems.

2. Related Work

Recent advancements in natural language processing (NLP) have expanded the applications of LLMs, including question answering, mathematical problem-solving, and dialogue systems. These models excel in generating coherent and context-aware text, making them ideal for conversational agents. For instance, GPT-3 and its successors have demonstrated remarkable performance in simulating human-like interactions, paving the way for their adoption in customer service.

The evolution of intelligent customer service systems can be categorized into two phases. Rule-based systems, the first generation, relied on predefined scripts and decision trees to handle user queries. While effective for simple tasks, their rigidity limited adaptability to complex or novel scenarios. The second generation, deep learning-based systems, introduced neural networks to improve contextual understanding and flexibility. However, these systems often required massive, labeled datasets and struggled with generalization.

Modern LLM-driven systems combine the strengths of data-driven learning and retrieval mechanisms, enabling dynamic adaptation to user needs. Chkirbene's (2024) research in prompt engineering further optimizes model outputs by refining input instructions, ensuring alignment with task-specific goals. These developments highlight the potential of LLMs to revolutionize customer service by balancing automation with human-like responsiveness.

3. Methods

3.1 Design and Implementation of a Modular Architecture-Based Intelligent Interaction System

Modern intelligent interaction systems integrate multiple functional modules to enhance user experience and operational efficiency. This paper provides an in-depth analysis of four core modules—user input processing, information retrieval, response generation, and human-computer interaction—detailing their technical implementations, interdependencies, and real-world applications.

3.2 User Input Module: Advanced Noise Filtering and Contextual Sanitization

The user input module acts as the system's first line of defense against unstructured or erroneous data. Noise in user inputs can originate from typographical errors, ambiguous phrasing, or irrelevant content (e.g., off-topic remarks in a customer service chatbot). Studies show that unprocessed noise reduces downstream task accuracy by up to 30% (Sowmiya, 2024).

3.3 Technical Enhancements

To enhance the performance of the e-commerce customer service system, we developed a comprehensive framework incorporating advanced natural language processing techniques and optimized retrieval mechanisms. Specifically, four core components were implemented:

We designed and integrated optimized regular expressions to enable efficient and accurate text pattern recognition. By leveraging precise pattern matching techniques, such as the use of patterns like '\d+\b' to extract price ranges, combined with efficient algorithms like inverted indexing, the system significantly reduces processing latency while improving extraction accuracy. This optimization ensures reliable identification of key information, such as product specifications or constraints, from user queries.

A hierarchical classification approach was adopted to effectively categorize user intents. The system first employs FastText, a fast and lightweight text classification model, to identify broad intent categories (e.g., "product inquiry"). Subsequently, BERT-Tiny, a smaller yet powerful transformer-based model, performs fine-grained classification to determine specific sub-intents (e.g., "price comparison" or "product availability"). This two-stage pipeline ensures both high efficiency and granular understanding of user requests.

To improve processing efficiency and focus on relevant information, a context-aware pruning mechanism was implemented. Leveraging the self-attention mechanism of Transformer-based architectures, the system evaluates the semantic

importance of different tokens in the input. Non-critical tokens, such as greetings ("Hello, how can I help you...") or irrelevant fillers, are dynamically assigned lower weights, reducing computational overhead while maintaining the essence of the query. This approach streamlines downstream processing and enhances response speed.

To address the limitations of traditional keyword-based retrieval methods (e.g., BM25), which often fail to capture semantic nuances, the system was equipped with a semantic-aware dense retrieval mechanism. Models like DPR (Dense Passage Retriever) encode both queries and documents into high-dimensional vector spaces, enabling the system to retrieve semantically relevant knowledge with higher accuracy. Empirical evaluations demonstrated that this approach achieves a 15-20% improvement in recall rates compared to traditional methods, particularly in complex e-commerce scenarios where understanding context is paramount.

By integrating these components, the system achieves both efficiency and accuracy, ensuring that customer queries are processed effectively while delivering relevant and personalized responses.

Implementation Workflow

Knowledge Base Preprocessing:

Product descriptions are chunked into 512-token segments and encoded into 768-dimensional vectors using a pre-trained BERT-base model. Faiss indexing enables efficient similarity search (Sowmiya, 2024).

Query-Index Interaction:

User queries like *"durable waterproof hiking boots"* are encoded and matched against indexed vectors using cosine similarity. A hybrid scoring mechanism combines semantic similarity (80%) and metadata relevance (e.g., product ratings, 20%).

Reinforcement Learning (RL) for Adaptive Retrieval:

User feedback (e.g., clicks or dwell time) trains an RL agent to prioritize high-conversion products.

Generation Module: Multi-Source Fusion with Fact-Checking Mechanisms

The generation module synthesizes user input, retrieved data, and domain knowledge to produce coherent responses. Direct concatenation often introduces redundancy; instead, hierarchical attention prioritizes critical information (Qin, 2024).

Architecture Breakdown:

Dual-Stream Encoder:

Stream 1: Processes user input (e.g., *"What's the warranty for this laptop?"*) using a RoBERTa encoder.

Stream 2: Encodes retrieved product specifications (e.g., *"2-year limited warranty"*) with a separate BERT model.

Cross-Modal Decoder:

A T5-based decoder attends to both streams, generating responses like *"This laptop includes a 2-year limited warranty covering..."*.

Fact-Consistency Validation:

Knowledge graph embeddings (e.g., TransE) verify claims against structured data (e.g., *"warranty duration = 2 years"* in the product database).

1.Human-Computer Interaction Module: Multimodal Input Handling and Real-Time Adaptation

Modern interfaces must support text, voice, and visual inputs while maintaining low latency.

4. Key Innovations:

Multimodal Fusion with CLIP:

For image-based queries (e.g., a user uploads a shoe photo), CLIP generates joint text-image embeddings to retrieve similar products.

Progressive Response Rendering:

Partial results (e.g., product categories) are displayed within 200ms, while detailed descriptions load asynchronously.

Contextual Autocomplete:

A GPT-3-powered suggestion engine predicts user intent (e.g., *"You might be asking about... warranty details"*) after the first three keystrokes.

System Integration and Performance Optimization

The system adopts modular architecture, implements synchronous communication between core services through RESTful API, and uses Apache Kafka to build asynchronous data flow pipeline to support high concurrency scenarios. Aiming at the key performance bottleneck, this paper proposes a dual optimization strategy: In terms of delay control, the model distillation technology is used to replace the BERT-base of the retrieval module with DistilBERT, and the inference speed is increased by 40%. At the same time, NVIDIA TensorRT is deployed to accelerate the hardware level of Transformer inference. The

measured single query response time was reduced to 50ms (A100 GPU environment). At the privacy protection level, Gaussian noise (ε=0.5) is added to the user vector in the retrieval phase based on differential privacy theory, and user ID and session logs are anonymized by AES-256 encryption algorithm to ensure compliance with GDPR (Yao, 2024). Experiments show that the scheme can reduce the risk of privacy leakage to less than 0.01% while maintaining 99.9% service availability.

5. Experiments

We conducted the following experiments mainly on efficiency generation, user satisfaction and problem-solving ability.

In Experiment I, we verify the degree of efficiency improvement. Through the synergy of regular expression optimization and context-aware pruning, the average processing time of a query is reduced by 25% (from 45 seconds to 34 seconds), mainly due to the dynamic filtering of non-critical information and the acceleration of semantic retrieval. The specific experimental results are as follows:

Table 1. Key Performance Metrics Before and After System Deployment

Metric	Baseline	Optimized	Improvement
Average Query Handling Time	45 seconds	34 seconds	**-25%**
Customer Satisfaction (CSAT)	72%	90%	**+18%**
First-Contact Resolution Rate	65%	90%	**+38%**
Daily Query Volume	12,000 queries/day	16,500 queries/day	**+37.5%**

In Experiment II, we experimented with user satisfaction growth. The CSAT score improved by 18 percentage points (from 72% to 90%), reflecting that the personalized responses generated by the system (e.g., accurate warranty matching, multi-round dialogue error correction) were more closely aligned with user needs. The specific experimental results are as follows:

Table 2. Contribution of Core Modules to Performance Enhancements

Module	Contribution Type	Efficiency (%)	Accuracy (%)
Regular Expression Optimization	Performance	45	20
Hierarchical Intent Classification	Performance	15	35
Context-Aware Pruning	Efficiency	25	10
Semantic Dense Retrieval	Accuracy	15	35

In Experiment III, the problem-solving ability is strengthened: in the warranty consultation scenario, the system achieves 90% one-time problem resolution rate through the linkage of hierarchical intention classification and dense retrieval module, which is a significant breakthrough compared with the traditional rule engine (65%). These results verify the practicability of the hybrid architecture combining deep learning and traditional NLP techniques in commercial scenarios, especially in reducing the cost of manual intervention and improving service consistency.

6. Conclusion

The integration of modular AI architecture and LLM significantly improves the intelligence level of customer service, realizes semantic accurate retrieval through RAG technology, and generates personalized response combined with dynamic prompt engineering. Empirical data show that it improves query efficiency by 25% and User Satisfaction (CSAT) by 18%. Multimodal interfaces meet GDPR privacy requirements with low latency (in milliseconds). However, the system still faces the challenges of real-time environment adaptability and cross-domain knowledge integration. In the future, it is necessary to explore federated learning to realize distributed knowledge update, or enhance decision interpretability through neuro-symbolic systems. This research marks the paradigm shift of human-computer collaboration from tool assistance to cognitive partnership and provides a technical path for the empathization and scenario adaptation of business ecology.

References

[1] Shen, L., Yang, Q., Zheng, Y., & Li, M. (2025). AutoIOT: LLM-Driven Automated Natural Language Programming for AIoT Applications. arXiv preprint arXiv:2503.05346.

[2] Chkirbene, Z., Hamila, R., Gouissem, A., & Devrim, U. (2024, December). Large Language Models (LLM) in Industry: A Survey of Applications, Challenges, and Trends. In 2024 IEEE 21st International Conference on Smart Communities: Improving Quality of Life using AI, Robotics and IoT (HONET) (pp. 229-234). IEEE

[3] Sowmiya, R., Revathi, P., Ragunath, D., Gokila, P., & Kalaivani, T. (2024, October). Multi-Modal LLM Driven Computer Interface. In 2024 8th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC) (pp. 484-489). IEEE.

[4] Shah, C., White, R. W., Andersen, R., Buscher, G., Counts, S., Das, S. S. S., ... & Yang, L. (2023). Using large language models to generate, validate, and apply user

intent taxonomies. arXiv preprint arXiv:2309.13063.

[5] Friha, O., Ferrag, M. A., Kantarci, B., Cakmak, B., Ozgun, A., & Ghoualmi-Zine, N. (2024). llm-based edge intelligence: A comprehensive survey on architectures, applications, security and trustworthiness. IEEE Open Journal of the Communications Society.

[6] Qin, J., Wu, J., Chen, W., Ren, Y., Li, H., Wu, H., ... & Wen, S. (2024). DiffusionGPT: LLM-driven text-to-image generation system. arXiv preprint arXiv:2401.10061.

[7] Yao, Y., Duan, J., Xu, K., Cai, Y., Sun, Z., & Zhang, Y. (2024). A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. High-Confidence Computing, 100211.