ISSN: 2664-9640

Application and Exploration of Large Model-Based Intelligent Service Ticket Generation Technology in Large-Scale Power Customer Service

Shuo Zhang, Qing Zhu, Jiaqi Shi, Liangfei Sun*

State Grid Customer Service Center, Tianjin, China

*sun_liangfei@163.com

Abstract

Addressing the complexities of conversational interactions, high information density, and stringent business logic in large-scale power customer service scenarios, this study investigates the application of large language model (LLM) agents in automated service ticket generation. We propose an intelligent ticket-generation architecture featuring "phased processing with multi-path validation," which decouples complex dialogue understanding into two sequential stages: critical information extraction and structured summary generation. This design enhances system accuracy and controllability. Based on real-world power customer service data, we systematically evaluate the performance of three LLMs-ERNIE 3.5-8K, Qwen2.5-32B, and Qwen3-32B-in ticket generation tasks. Experimental results demonstrate that the proposed method significantly improves the quality of information extraction and summary generation. Notably, Qwen3-32B achieves an overall accuracy of 96.33%, excelling particularly in complex semantic reasoning tasks such as customer request comprehension and emotion recognition. All evaluated models meet real-time operational efficiency requirements. This research validates the practical potential of LLM agents in industries with stringent requirements like power systems and provides actionable methodologies for constructing and optimizing intelligent customer service systems.

Keywords

Large Language Models; Power Customer Service; Intelligent Ticket Generation; Information Extraction; Semantic Understanding; Evaluation.

1. Introduction

The rapid advancement of artificial intelligence, particularly breakthroughs in large language models (LLMs) and agent technologies, offers novel pathways for intelligent transformation across industries. As cognitive entities capable of autonomous task planning, deep semantic understanding, and flexible tool invocation, LLM-based agents are evolving from conventional "communication tools" into "cognitive collaborative partners," demonstrating significant potential in handling complex tasks across diverse domains[1]. Research indicates that their capability framework integrates multifaceted theoretical foundations, including socio-cultural theory and knowledge construction theory, thereby establishing crucial theoretical support for next-generation intelligent systems[2].

In large-scale customer service systems, intelligent ticket generation scenarios are characterized by frequent conversational interactions, high information density, and intricate business logic, necessitating technical solutions capable of deep semantic comprehension and multi-turn reasoning. LLM agents address these challenges through a progressive workflow of "information extraction–summary synthesis," effectively integrating semantic parsing with business rules to significantly enhance service ticket accuracy and structural integrity. Based

ISSN: 2664-9640

on real-world operational scenarios from a major power customer service system, this study systematically explores the core technical pathways and implementation efficacy of LLM agents. By constructing a phased, multi-path ticket generation pipeline and conducting cross-model evaluations of varying scales, this research aims to provide methodological guidance and practical references for building high-accuracy, high-efficiency intelligent power service systems, thereby promoting the deepening application and paradigm innovation of this technology in critical industries.

2. Research Progress in Agent Technology

Agents-entities that perceive environments and autonomously act to achieve goals-have evolved through distinct technical paradigms: from early rule-based systems and perceptual intelligence to the current era of cognitive intelligence underpinned by big data and massive computing power[3–5]. LLMs serving as the "brain" of agents mark a new phase of "autonomous intelligence" and "collective intelligence." Their core capabilities encompass task planning, memory management, tool invocation, and action execution, transforming them from passive tools into collaborative partners with deep semantic understanding and autonomous decision-making abilities[6].

At the engineering level, development frameworks such as LangChain and AutoGen, alongside low-code platforms like Coze[7] and Dify[8], have substantially lowered barriers to agent development, enabling rapid deployment across diverse scenarios. Concurrently, multi-agent collaboration-through role specialization and task coordination-effectively overcomes the limitations of individual agents, emerging as a key developmental trend[9, 10]. Architectures such as the five-module framework (encompassing agent core and task planning) proposed by Yao et al., and the "Reason-Act-Observe" cyclic mechanism within the ReAct framework, provide critical theoretical and technical foundations for building agent systems in complex scenarios[11].

3. Overview of Large Model-Based Intelligent Ticket Generation

3.1. Common Paradigms and Technical Pathways

Power customer service operations dispatch tasks to departments, units, and personnel via service tickets, managing workflows through closed-loop business processes. Large model-based intelligent ticket generation refers to an intelligent task-processing system that leverages large language model (LLM) technology to perform deep semantic parsing of unstructured dialogues between customers and service systems, thereby automatically accomplishing intent recognition, critical information extraction, and structured summary generation. In power customer service scenarios, this system analyzes multi-turn dialogues to automatically extract multidimensional information-including service type, customer account number, contact details, core (service requests), emotional state, and latent risks-generating structured ticket summaries compliant with business specifications. This provides the technical foundation for precise ticket dispatching and closed-loop processing.

Currently, intelligent ticket generation systems predominantly adopt three technical paradigms:

End-to-end generation paradigm: Relies on structured prompts to output complete tickets in a single inference step. While streamlined, this approach is prone to information omissions or structural deviations.

Phased extraction-synthesis paradigm: Decomposes the task into critical information extraction followed by summary synthesis. This "extract-then-integrate" workflow enhances accuracy and controllability, establishing it as the mainstream approach in high-stakes domains like power systems.

ISSN: 2664-9640

Multi-agent collaboration paradigm: Employs specialized agents working collaboratively to generate tickets. This paradigm offers strong scalability and robustness, making it suitable for future scenarios with complex business logic and stringent fault tolerance requirements.

3.2. Core Challenges and Key Issues

Practical implementation of knowledge understanding and generation faces significant challenges:

Information extraction: Colloquial expressions, implicit references, and pronoun ambiguities compromise extraction accuracy and completeness.

Domain knowledge integration: Models must accurately interpret specialized terminology (e.g., "metering anomalies," "electricity bill refunds/rebates") and business workflows.

Risk-aware generation: Systems require capabilities to identify latent emotional cues and potential complaint risks while effectively suppressing hallucinated outputs.

Additionally, intelligent customer service systems demand stringent real-time responsiveness and operational stability, necessitating an optimal balance between model performance and inference efficiency.

3.3. Data Example for Intelligent Ticket Generation

The input for intelligent ticket generation systems consists of authentic customer service dialogue records, typically multi-turn text interactions between customer service representatives and clients. An example is illustrated in Figure 1.

[1] **[Agent]**: Happy New Year! This is Customer Service Agent 02012341. How may I assist you today?
(System Note: Binding your customer account number enables faster service. Please remain on this page during the chat, as switching applications may terminate the session. We appreciate your understanding.)
[2] **[Customer]**: I strongly suspect inaccurate metering on February 5th.
[3] **[Agent]**: May I ask what specific issue you encountered, sir/madam?
[4] **[Customer]**: Is a single-day consumption of 97.97 kWh reasonable?
[5] **[Agent]**: Were you checking this via the mobile application?
[6] **[Customer]**: Yes.
[7] **[Agent]**: If you suspect billing irregularities, please disconnect all indoor circuit breakers and observe the meter for 5–10 minutes. This helps identify potential issues such as leakage or post-meter power theft.
[8] **[Customer]**: No leakage or other anomalies detected.
[9] **[Agent]**: We recommend consulting a certified electrician to inspect your post-meter circuits.
[10] **[Customer]**: Customer account number: 4312345678959
[11] **[Customer]**: Please check this for me.
[12] **[Agent]**: Please hold while I retrieve your account details.
[13] **[Agent]**: To assist with troubleshooting, could you confirm whether high-power appliances (e.g., underfloor heating, space heaters, water heaters, or air conditioners) have been used frequently recently? Extended home occupancy and increased appliance usage may also contribute to higher consumption.
[14] **[System Alert]**: No message received for 90 seconds. Session will terminate in 60 seconds.
[15] **[Customer]**: Air conditioner is not in use.
[16] **[Agent]**: I will escalate this to our field technician team for resolution.
[17] **[Agent]**: Kindly provide your contact number for follow-up.
[18] **[Customer]**: 15212345678, Mr. Zhang
[19] **[Agent]**: Noted. Please keep your line available. We have logged your request, and a specialist will contact you shortly.
[20] **[System Alert]**: No message received for 90 seconds. Session will terminate in 60 seconds.
[21] **[System Prompt]**: Thank you for contacting us. Goodbye!

Figure 1. Dialogue Data Example

In this conversation, the system must accurately identify the following elements:

① Service type

② Customer account number

③ Contact information

ISSN: 2664-9640

④ Account holder name
⑤ Abnormal occurrence time
⑥ Customer service request
⑦ Prior incident reports
⑧ Customer sentiment
⑨ Risk of escalation
⑩ Customer expectations

A structured summary is then synthesized according to business logic and compilation guidelines, incorporating the following fields:

Service type, account number, contact details, account holder name, month of anomaly occurrence, address, customer request, agent handling actions, prior incident history, emotional state, risk assessment, and customer expectations. An example output is shown in Figure 2.

Customer Account Number: 4312345678959
Contact Number: 15212345678
Account Holder: Mr. Zhang
Anomaly Occurrence Date: February 5
Issue: Customer disputes electricity billing for February 5 despite standard troubleshooting procedures.
Prior Reports: None
Customer Sentiment: Calm
Risk Assessment: No escalation risk identified
Customer Expectation: "Prompt resolution"

Figure 2. Summary Example

3.4. Non-Functional Requirements

Beyond functional specifications, the system must satisfy the following critical non-functional requirements:

Accuracy: Field extraction accuracy must meet operational standards (typically >95% overall accuracy), with zero tolerance for errors in core identity fields (e.g., account number, contact information).

Real-time performance: End-to-end system response latency must remain under 8 seconds to ensure seamless real-time interaction in online customer service.

Explainability: Model decision processes-particularly for critical information extraction and risk assessment-must be interpretable to facilitate operational troubleshooting and iterative optimization.

Scalability: System architecture must support low-cost, flexible expansion of new service types and information fields to accommodate evolving business rules.

Security: Strict confidentiality of user privacy data must be maintained, with all data processing workflows compliant with State Grid's Data Security Management Specifications.

4. Application Development of Large Model-Based Intelligent Ticket Generation

4.1. Model Selection

Model selection for power sector ticket generation requires balancing semantic comprehension depth, domain knowledge alignment, complex reasoning capability, and deployment costs. Three representative models were comparatively evaluated:

ISSN: 2664-9640

ERNIE 3.5-8K: Demonstrates stable performance in Chinese linguistic understanding and instruction following. Benefits from domain adaptation via State Grid's "Guangming Large Model," providing robust prior knowledge of power systems.

Qwen2.5-32B: Leverages its parameter scale advantage for superior long-text comprehension and complex logical reasoning, particularly effective for causal analysis and multi-step inference in power business scenarios.

Qwen3-32B: As a next-generation model, integrates dual-mode reasoning (deliberative and reactive mechanisms), offering flexible and efficient support for scenarios demanding both deep analysis and real-time responsiveness in intelligent ticket generation.

4.2. Scenario Implementation Strategy

To address the inherent tension between accuracy and performance, a hierarchical system architecture was developed based on the core design principle of "phased processing with multi-path validation" (Figure 3). This architecture strategically decomposes workflows and implements cross-verification mechanisms to ensure high accuracy in critical information extraction while maintaining system-wide response efficiency.

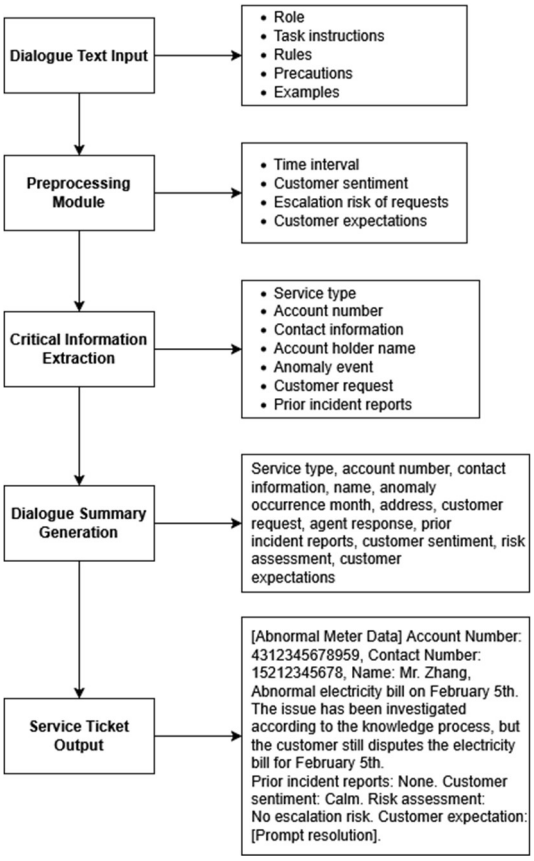

Figure 3. Intelligent Ticket Generation Logic

The core of the system architecture decouples ticket generation into two primary phases:

(1) Critical Information Extraction Phase

A meticulously designed structured prompt template guides the model to precisely extract four error-prone critical fields from preprocessed dialogue text. Subsequently, the large language model (LLM) extracts remaining conventional fields, ensuring comprehensive and accurate identification of all required data points.

(2) Summary Synthesis Phase

ISSN: 2664-9640

This phase employs a hybrid strategy combining "template instantiation with semantic refinement." First, structured information extracted in Phase 1 is accurately populated into a business-compliant ticket template, guaranteeing data completeness and format standardization. Then, leveraging the LLM's semantic summarization and coherence generation capabilities, predefined content blocks within the template undergo natural language refinement and logical optimization. This yields high-quality ticket summaries that are both informationally precise and human-readable.

4.3. Workflow Implementation

Building upon the aforementioned architecture, we further developed a four-stage refined processing workflow (Figure 4) to ensure orderly and dependable data propagation throughout the system.

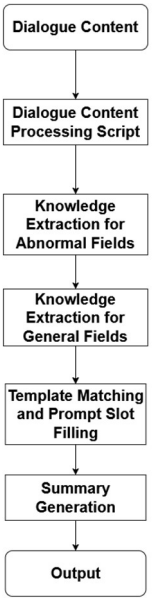

Figure 4. Intelligent Ticket Generation Workflow

Stage 1: Dialogue Preprocessing

Raw dialogue texts undergo cleansing and standardization to eliminate noise such as agent scripted responses and system prompts irrelevant to customer requests. Concurrently, colloquial and abbreviated user expressions are normalized into purified, structured text, providing high-quality input for subsequent deep analysis.

Stage 2: High-Risk Field Identification

Specialized processing targets format-sensitive yet error-prone "high-risk" fields (e.g., account numbers, dates). A dual safeguards mechanism is implemented: initial identification and localization by the LLM, followed by rigorous format validation and correction via a high-precision regex engine. This approach prevents critical information extraction errors at the source.

Stage 3: Key Field Extraction

Leveraging purified text, iteratively optimized prompt engineering systematically guides the LLM to extract all 10 critical information dimensions. For complex scenarios requiring cross-turn reasoning or involving semantic ambiguity, the workflow supports multi-turn interactive extraction. This mimics agent clarification logic by appending targeted instructions to the model, ensuring extraction completeness.

Stage 4: Dialogue Summary Generation

ISSN: 2664-9640

In the final stage, the system synthesizes all extracted fields and invokes the standardized ticket template corresponding to the identified service type. Structured information is intelligently populated into designated template slots, followed by coherent organization and natural language rewriting based on business logic. This yields standardized ticket summaries that are format-compliant, semantically precise, and meet downstream dispatch requirements—completing the intelligent transformation from unstructured dialogues to operational service tickets.

5. Agent Evaluation

To objectively assess LLM agent efficacy in power sector ticket generation, a systematic evaluation framework was designed around the core tasks of critical information extraction and ticket synthesis. This provides quantitative decision support for model selection and system optimization.

5.1. Evaluation Objectives

This evaluation systematically measures the performance of diverse LLMs implementing the "phased processing with multi-path validation" workflow in power ticket generation scenarios. It specifically validates the approach's effectiveness in:

Information extraction accuracy

Intent recognition reliability

Sensitivity in sentiment and risk assessment

Ticket generation compliance

Additionally, it analyzes model performance variations across fields of differing complexity to inform model selection that balances high accuracy and efficiency.

5.2. Evaluation Methodology

Dataset: 500 authentic dialogue records covering major power customer service scenarios, ensuring diversity and representativeness in business contexts and linguistic expressions.

Metrics:

Per-field accuracy: Proportion of fields where extraction matches manual annotations exactly.

Holistic accuracy: Proportion of dialogues where all fields are correctly extracted (a strict all-or-nothing criterion reflecting real-world operational requirements).

Procedure:

Automated testing via a self-developed batch tool invoking workflow APIs deployed across platforms.

Manual verification of all model outputs by domain experts to ensure evaluation reliability and impartiality.

Test results imported into a dedicated verification tool for human review, with final metrics computed automatically.

5.3. Evaluation Conclusions

Systematic analysis of evaluation data yielded the following conclusions:

(1) Comparative Results

Models with varying parameter scales demonstrated differentiated performance on power ticket generation tasks (detailed comparisons in Table 1). Qwen3-32B achieved leading results across most fields, exhibiting significant advantages in complex reasoning tasks requiring deep semantic analysis-particularly for anomaly occurrence month, customer request, and risk assessment. Qwen2.5-32B delivered balanced performance, achieving superior precision in

ISSN: 2664-9640

intricate fields such as address recognition. ERNIE 3.5-8K maintained high stability in fundamental fields, demonstrating robust domain adaptation capabilities within the power sector context.

Table 1. Model Performance Comparison

FIELE TYPE	ERNIE3.5-8K	QWEN2.5-32B	QWEN3-32B	DIFFICULTY LEVEL
Service Type	96%	98%	99%	Low
Account Number	95%	97%	95%	Medium
Contact Information	96%	98%	98%	Low
Account Holder Name	98%	98%	98%	Low
Anomaly Occurrence Month	76%	87%	95%	High
Address	82%	95%	98%	High
Customer Request	78%	85%	96%	High
Agent Handling Actions	80%	90%	94%	Medium
Prior Incident Reports	95%	97%	97%	Medium
Customer Sentiment	85%	90%	95%	Medium
Risk Assessment	78%	86%	95%	High
Customer Expectations	89%	93%	95%	Medium
Workflow Response Time (s)	6.71	5.36	6.12	—
Overall Accuracy	**87.33%**	**92.83%**	**96.33%**	—

(2) Comparative Conclusions

All three models met the operational requirement of <8-second response latency in power ticket generation scenarios, while exhibiting distinct capability profiles. Qwen3-32B demonstrated superior reasoning capabilities, achieving optimal performance in high-difficulty tasks such as anomaly occurrence month identification and risk assessment. Qwen2.5-32B delivered balanced performance across multiple tasks with robust comprehensive capabilities. Although ERNIE 3.5-8K has a smaller parameter scale, its domain adaptation strengths ensured stable performance in fundamental field extraction.

The evaluation results validate the effectiveness of the "phased processing with multi-path validation" architecture in enhancing system accuracy. They also reveal that model performance ceilings in complex business scenarios are jointly determined by parameter scale, domain knowledge integration, and reasoning capabilities. For practical deployment, we recommend differentiated model selection based on business complexity and performance requirements:

Qwen3-32B for mission-critical scenarios demanding high precision

ERNIE 3.5-8K for cost-sensitive, medium-to-low frequency scenarios

Qwen2.5-32B for optimal balance between effectiveness and resource consumption

6. Conclusion

This study applies large language model agent technology to intelligent ticket generation in power customer service through a phased processing workflow, enabling automated service ticket generation from customer dialogues. Evaluation results demonstrate that Qwen3-32B achieves optimal performance with 96.33% overall accuracy, exhibiting significant advantages in complex reasoning tasks such as anomaly detection and risk assessment. This technology substantially enhances power customer service efficiency, providing a feasible pathway for

ISSN: 2664-9640

industry digital transformation. Future work will focus on multi-modal data fusion and real-time performance optimization to further improve system practicality and generalization capabilities.

References

[1] RAPP A, LODOVICO C D, CARO L D. How do people react to ChatGPT's unpredictable behavior? Anthropomorphism, uncanniness, and fear of AI: A qualitative study on individuals' perceptions of LLM hallucinations[J]. International Journal of Human-Computer Studies, 2025, 198: 103125.

[2] LIAO J, LIU M, YANG M, et al. Development, application status and future prospects of educational large model agents[J]. Modern Educational Technology, 2024(11): 1–15. [in Chinese]

[3] SIHAI A N, QIU J, LIN J, et al. Multi-agent mobile robot-based adaptive charging network for enhancing power system resilience under extreme conditions[J]. Applied Energy, 2025, 381S: 125189.

[4] LI H, JIA X, DUAN S, et al. Accurate prescribed-time output consensus of heterogeneous multi-agent systems: A bounded time-varying gain approach[J]. Automatica, 2025, 182: 110387.

[5] ANWAR G A, AKBER M Z. Multi-agent deep reinforcement learning for resilience optimization of building structures considering utility interactions[J]. Computers & Structures, 2025, 310: 107703.

[6] TRUJILLO F, POZO M, SUNTAXI G. Artificial intelligence in education: A systematic literature review of machine learning approaches in student career prediction[J]. Journal of Technology and Science Education, 2025, 15(1): 145–168.

[7] COZE. Setup Guide[EB/OL]. (2025-11-26). https://www.coze.cn/docs/guide/welcome

[8] DIFY. Dify Official Documentation[EB/OL]. (2025-11-26). https://docs.dify.ai

[9] SONG J, ZHENG W, CHEN M, et al. Analysis of collusive bidding characteristics in the power spot market based on multi-agent reinforcement learning simulation[J]. Electrical Engineering, 2025, 107(8): 4211–4225.

[10] MA Q, YE Y, LIU Z, et al. Carbon cap based multi-energy sharing among heterogeneous microgrids using multi-agent safe reinforcement learning[J]. Applied Energy, 2025, 393: 124087.

[11] YAO S, ZHAO J, YU D, et al. ReAct: Synergizing reasoning and acting in language models[J]. arXiv preprint arXiv:2210.03629, 2022.