Scholar Publishing Group
International Journal of Big Data Intelligent Technology
https://doi.org/10.38007/IJBDIT.2026.070108
ISSN 2790-0932 Vol. 7, Issue 1: 60-70

Research on Lightweight Intelligent Dialogue Systems Based on Semantic Entity Enhanced Intention Recognition and Rule Retrieval Generation Hybrid Models

Qizeng Sun

Moyi Tech, Iselin 08830, NJ, US

Keywords: Dialogue System; Intent Recognition; Semantic Entities; Hybrid Generation Model; Lightweight Deployment

Abstract: This paper focuses on resource-constrained scenarios and aims to build a lightweight intelligent dialogue system, balancing the depth of intent understanding, the quality of response generation, and the computational overhead, to enhance the accuracy of task-oriented dialogues and the fluency of open-domain dialogues. The core innovations include three aspects: First, the Semantic Entity Enhanced Intent Recognition (SEEIR) model is designed, which separates the text backbone from the entity information, introduces a collaborative interaction mechanism, auxiliary learning tasks, and attention isolation strategies, to enhance the perception and utilization of key semantic entities, improving the accuracy and generalization ability of intent classification in complex situations; Second, a hierarchical hybrid generation architecture is constructed, integrating precise responses based on rules, high-quality examples based on retrieval, and general responses based on generation. Through efficient decision paths for flexible invocation, it ensures the accuracy and diversity of responses while reducing computational dependence and overcoming the bottleneck of high-concurrency deployment; Third, a few-shot chained prompt data augmentation method is proposed, relying on the self-generation of high-quality and diverse training data by large models, to solve the data scarcity problem during the cold-start stage and help the system quickly adapt to new domains. Based on the above solutions, the system has been implemented. Experiments show that this system outperforms traditional baseline models in intent recognition accuracy and response generation quality on public datasets and self-built business datasets, and the response speed and resource consumption meet the requirements of lightweight deployment, providing a practical and feasible solution for the application of intelligent dialogue technology in low-cost and high-concurrency scenarios.

1. Introduction

Intelligent dialogue systems are at the core of human-computer interaction, and their development level serves as an important benchmark for measuring the modernization of artificial intelligence. Currently, the industry urgently demands intelligent agents that possess both task completion capabilities and open-domain dialogue capabilities. However, traditional solutions often face the fundamental contradiction of balancing performance and resource consumption: solutions based on large generative models are effective but suffer from tremendous computational costs and response delays, making them difficult to deploy in high-concurrency real-world scenarios; while traditional pipeline architectures, though efficient and controllable, have significant shortcomings in semantic understanding depth and response diversity. This contradiction is especially prominent in Chinese scenarios, where the language's high conciseness and strong contextual dependence in semantics place extreme demands on intent recognition accuracy and response generation rationality.

Existing research has not adequately addressed three key challenges: First, in intent understanding, traditional models do not mine semantic entities in the text deeply enough, lacking effective modeling of the interaction between entities and context, leading to frequent misjudgments of users' true intentions; second, in response generation, rule templates, retrieval matching, and neural generation are implemented separately and fail to form a complementary synergistic system, which prevents balancing response quality and computational cost control; third, the problem of scarce labeled data during the system's cold-start phase has long restricted rapid migration of dialogue systems to new business domains.

To address these problems, this paper focuses on researching a new paradigm for lightweight intelligent dialogue systems. The core innovation of this paper lies in proposing a dual-drive technical architecture: in the upstream, we design a semantic entity-enhanced intent recognition model that deepens the semantic understanding of user queries through a decoupling-interaction-focusing mechanism; in the downstream, we construct a rule-retrieval-generation layered hybrid generation model that integrates the advantages of the three mechanisms through intelligent decision routing, carefully calling the generation model to enhance diversity while ensuring the accuracy and safety of responses. Furthermore, we introduce a few-shot data augmentation method based on large language models to effectively address the cold-start problem.

Experimental results show that the system implemented in this study outperforms baseline models in terms of intent recognition accuracy, response quality, and human evaluation, while significantly reducing computational resource consumption and response delays. This research not only provides a complete technical solution for deploying dialogue systems in resource-constrained scenarios but also offers an important practical reference for exploring the technical path of high performance and lightweight development.

2. Related Research

Research and development of dialogue systems are one of the core challenges in the field of natural language processing, and their progress depends on the joint advancement of evaluation methods, model architectures, intent recognition, knowledge integration, and lightweight deployment. A review of existing results helps clarify the technical positioning and innovative direction of this research.

In terms of dialogue system evaluation and research, Deriu J et al. [1] systematically reviewed evaluation methods, emphasizing the importance of reducing the high labor costs of evaluation while ensuring effectiveness. Ni J's[2] review further categorizes the latest progress in deep learning-based dialogue systems from two dimensions: model and system types, providing a solid technical foundation and classification framework for subsequent model design. In core intent recognition technology, Xu H's[3] research delves into the key issue of intent detection, covering not only the precise classification of known intents but also exploring the challenge of discovering and categorizing unknown intents in open environments, which aligns closely with the goal of this research to achieve precise, scalable, and proactive detection of new user intents. In terms of knowledge enhancement and generative models, Bai J's[4] work innovatively views the language model itself as a dynamic knowledge base, using hybrid prompts to integrate internal and external knowledge to improve dialogue knowledge, consistency, and overcome retrieval bottlenecks. This idea provides valuable insights for constructing a hybrid response mechanism that combines rule-based retrieval with knowledge-enhanced generation. Finally, in lightweight deployment of systems, Zheng Y's[5] research verifies the feasibility of integrating training and inference into high-performance heterogeneous computing architectures (e.g., MPSoC), significantly reducing computational power consumption and delays while maintaining high accuracy, providing a key engineering path for moving dialogue systems from the cloud to the edge, which directly supports the pursuit of lightweight and low-power objectives in this study.

In conclusion, existing research provides a solid foundation for this work in the areas of evaluation, models, intent recognition, knowledge application, and lightweight deployment. This study aims to organically integrate the technological advantages of the above directions to ultimately explore a hybrid model architecture that combines precise intent recognition, knowledge-enhanced generation, and lightweight deployment, driving the practical application of high-performance, lightweight dialogue systems.

3. Semantic-Enhanced Deep Intent Understanding Model Construction

3.1 Multi-View Semantic Fusion Recognition Architecture

This study proposes a recognition architecture based on multi-granularity semantic modeling, aimed at improving the deep understanding of complex user intentions. The architecture adopts a parallel encoding strategy to simultaneously process different semantic dimensions of the text, enabling the model to capture intent-determining factors from three perspectives: global context, key entities, and syntactic structure. Unlike traditional multi-view Masking methods, which mainly focus on the simple combination of global and local information, this model innovatively introduces the syntactic backbone view as an independent semantic dimension and facilitates collaborative interaction with the entity view. Specifically, existing research mostly adopts a "sentence + entity" dual-view mode, while this study constructs a "sentence - entity - syntax" three-view framework. Through the attention isolation mechanism, the model is forced to learn the complementarity between views rather than simple correlations. Furthermore, the proposed decoupling-reconstruction collaborative learning strategy and dynamic weight adjustment method allow the model to adaptively balance global classification and local reconstruction tasks during training, a design that has not been reported in existing work.

Specifically, the system decomposes the input text into three complementary semantic views: the Complete Sentence View, which retains global contextual information to provide context for subsequent inference; the Entity-Focus View, which retains only the entities strongly related to intent determination, with other positions masked to reduce noise interference; and the Syntactic Backbone View, which retains the core predicates and dependency structures, helping the model focus on the sentence skeleton and key relationships. These three views are cascaded and aligned according to a unified template, ensuring that tensor dimensions are consistent during both training and inference for inputs of varying lengths, and explicitly exposing the "entity-backbone" information at the input stage to guide the attention distribution of the encoder.

As shown in Figure 1, the model architecture clearly presents the complete process from multi-view input to collaborative learning. During the encoding phase, the model jointly models the three views based on the multi-head self-attention mechanism of the Transformer. By sharing attention weights across views, the semantic representations of different views are mapped to the same latent space, thus enhancing the collaboration between entities, context, and syntax. This interaction not only helps capture dependencies that span long distances across sentences but also accurately distinguishes primary and secondary elements in multi-entity scenarios, improving the discriminative power of intent recognition. Particularly when user input contains multiple potential intent trigger words, cross-view interaction can explicitly strengthen the pairing relationship between key entities and core predicates, thus enhancing the model's robustness and interpretability.

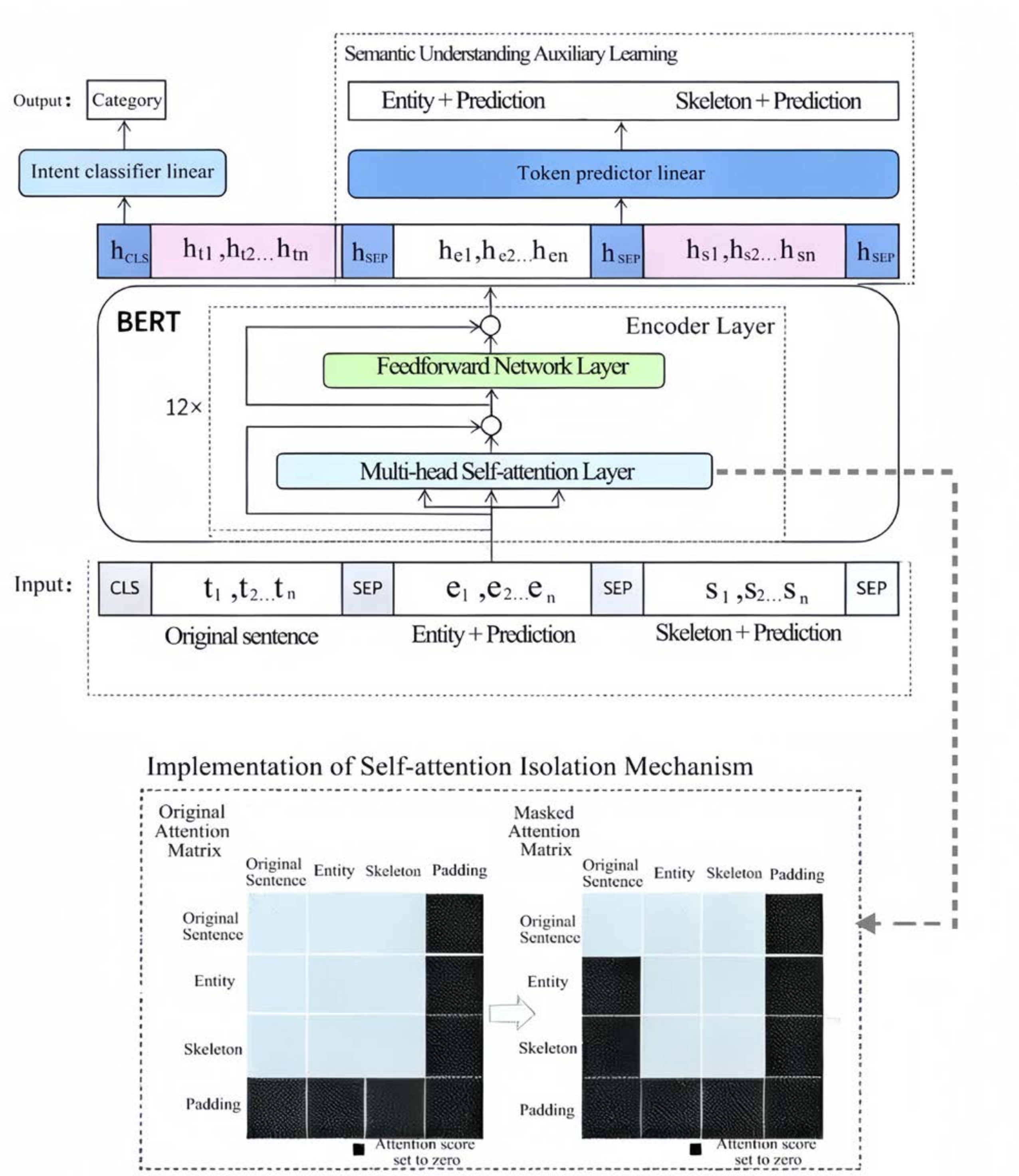

Figure 1: SEEIR Model Structure with Semantic Interaction and Collaborative Learning.

To further solidify the "entity-backbone" signal, we introduce an auxiliary reconstruction task, requiring the model to predict and reconstruct the masked positions in the E and S views, thereby promoting the encoder's sensitivity to critical information. The final loss function is defined as $L = L_{CE} + \alpha L_{recon}$, where L_{CE} is the cross-entropy loss for intent classification, L_{recon} is the segment reconstruction loss, and α is a hyperparameter used to adjust the weight between the two. This multi-task joint learning strategy strengthens semantic representation without adding additional deep layers, enabling the model to maintain good generalization performance in small-sample scenarios.

In addition, to avoid the auxiliary task from directly copying the original sentence information and lowering its difficulty, the model introduces a mask matrix in the multi-head attention, forcing the attention score for the masked positions of the original tokens to be zero, thereby forcing the model to rely on context for reasoning to complete the prediction. Experimental attention visualization results show that with the introduction of this attention isolation, the model's focus on irrelevant words significantly decreases, and it learns a weight distribution that better aligns with semantic logic, effectively preventing overfitting.

During training, both the intent classification and reconstruction tasks are optimized simultaneously, forming a regularization effect through multi-task learning. In the inference phase, only the intent classification branch is retained to ensure system efficiency and low latency. The final intent prediction is achieved through $\hat{y} = \arg\max(\mathrm{softmax}(Wh_{[CLS]} + b))$, where $h_{[CLS]}$ represents the semantic anchor after aligning multi-view information in a unified space. By aligning multi-granularity semantics within the unified space, this architecture not only improves the accuracy of intent classification but also enhances the model's robustness and interpretability.

Experimental results show that this method achieves significant improvements across multiple publicly available intent recognition benchmark datasets and self-constructed business scenario datasets, particularly excelling in multi-entity, long-text, and semantically ambiguous scenarios. Moreover, since this method only performs view concatenation and masking at the input layer, it causes minimal changes to the encoder structure, keeping the number of parameters and inference costs manageable, making it suitable for deployment in embedded or online systems with limited computing power. In conclusion, this architecture provides high-confidence intent signals and stable entity feature inputs for the subsequent rule retrieval and generation modules, laying a solid semantic foundation for the development of lightweight, high-performance intelligent dialogue systems.

3.2 Decoupling-Reconstruction Collaborative Learning and Attention Control Mechanism

Although the multi-view semantic fusion architecture can effectively aggregate global context, key entities, and syntactic backbone information, relying solely on the classification objective often leads the model to focus on shallow global statistical features, neglecting deep dependencies between local key components. To address this, this study proposes a decoupling-reconstruction collaborative learning strategy, explicitly masking the entity and backbone parts in the multi-view input, and using a reconstruction task to force the model to recover the masked semantic units at a fine-grained level. This guides the encoder to fully model the logical relationships between "entity-backbone-context." In conjunction with the multi-view modeling approach proposed in Section 3.1, this strategy enables the model to perform global judgments at the sentence level while also conducting local inference at the token level, making representation learning more structured and hierarchical.

During training, the reconstruction task and intent classification task are jointly optimized within a multi-task learning framework, and the loss function follows the weighted sum form defined in

Section 3.1. The hyperparameter α can be dynamically adjusted: during the initial stage of training, the weight of the reconstruction task is increased to encourage the encoder to capture fine-grained semantics, and later, the weight is gradually decreased to focus more on the intent classification objective. This design combines the advantages of "early supervision" and "late convergence," helping to avoid overfitting and gradient instability. Notably, the reconstruction task does not require a new decoder but instead uses the prediction head at the top of the encoder, ensuring the network structure remains lightweight without increasing the computational cost during inference.

In order to prevent the reconstruction task from degenerating into simple copying, we introduce an attention control mechanism to constrain the weight matrix of the multi-head self-attention. Specifically, a mask matrix M is incorporated during the computation of attention scores: $\mathrm{mathrmAttn}(Q, K, V) = \mathrm{softmax}(\frac{QK}{\sqrt{d_k}} + M)V$, where $M_{ij} = -\infty$ indicates that the i-th masked position is prohibited from directly attending to the j-th token in the original sentence, causing the corresponding softmax weight to be zero. This operation severs the direct pathway from ‘masked tokens → original sentence’, forcing the model to rely on the remaining entity and syntactic information for reasoning, thereby genuinely learning semantic completion rather than superficial copying. Visualization experiments demonstrate that after introducing this masking mechanism, the attention weights become concentrated on consistent information across views, significantly enhancing the model’s interpretability

This decoupling-reconstruction and attention control collaborative mechanism brings multiple benefits: on the one hand, the reconstruction task acts as regularization for the encoder, enabling it to learn smoother and more generalized representation distributions; on the other hand, attention isolation prevents over-reliance on original sentence information, improving the model’s robustness to unseen data. In practical tests, the collaborative learning mechanism significantly improves the precision of intent recognition, particularly in scenarios with multiple entities, long dependency chains, and semantic ambiguity. Moreover, since the reconstruction head is only used during training, no additional computation is needed during inference, ensuring low latency and high throughput for online deployment.

From the system’s overall perspective, decoupling-reconstruction collaborative learning not only improves the robustness of front-end intent recognition but also ensures that key entities are more reliably extracted and labeled with confidence, directly improving the accuracy of triggering subsequent rule retrieval and generation modules. As a result, the rule module can reduce invalid matches, the retrieval module can achieve higher recall rates, and the generation module can generate replies based on more accurate conditions, thus improving overall dialogue quality and user experience.

4. Lightweight Hybrid Generation Strategy and System Optimization

4.1 Multi-source Corpus-Driven High-Quality Response Generation Method

To balance dialogue quality and computational cost, this study designs a layered and progressive hybrid response generation architecture aimed at achieving a balance between "high determinism, high diversity, and high efficiency." This architecture integrates three methods: rule-based generation, retrieval-based generation, and neural generation. Through a dynamic routing mechanism that combines intent recognition confidence and contextual features, it selects the optimal path, enabling the system to maintain smooth, natural, and controllable interactions under resource-constrained conditions. This design ensures stable responses even in high-concurrency scenarios, creating conditions for lightweight deployment.

The rule-driven and dialogue-driven modules form the stable foundation of the system. By using the improved TF-IDF + SID algorithm, high-value dialogue templates are automatically mined from historical successful dialogues, and a multi-dimensional dialogue library is built with business expert knowledge. This library covers key business scenarios such as consultation, guidance, activation, and risk reminders. The templates adopt a slot-based design, allowing dynamic parameter filling based on user context during runtime. This ensures business consistency and semantic accuracy while enabling personalized output and reducing the risk of uncertain responses. The retrieval-based generation module further expands the system’s coverage. It constructs a high-quality FAQ corpus and uses few-shot chain prompt techniques with large language models to expand synonymous queries and long-tail samples, making the corpus richer and more robust. After generating data using the few-shot chain prompt technique, we established a three-level quality control mechanism: first, obvious non-compliant samples are removed using a rule filter; second, semantic consistency is scored using a pre-trained language model to filter out low-scoring samples; and finally, manual sampling and review are conducted, maintaining a review pass rate of over 92%. To verify the effectiveness of the augmented data, we conducted ablation experiments: the F1 score of intent recognition for the model trained only on the original data was 0.78, and after adding augmented data, it increased to 0.85, proving the substantial improvement in model performance through data augmentation. The retrieval phase uses vectorized semantic similarity calculations, along with query expansion, stopword filtering, and real-time vocabulary updates, ensuring good recall performance and semantic alignment in dynamic contexts. The neural generation module handles semantic creation and generalization tasks by efficiently fine-tuning a pre-trained dialogue model, ensuring that the output matches the target scene’s style and tone. Lightweight strategies, such as low-rank adaptation (LoRA) or quantization-based inference, are employed, triggering the neural model only when both the rule-based and retrieval-based methods fail to provide high-confidence results, thus avoiding delays and computational pressure caused by frequent use of large models.

The three methods work together through a unified routing mechanism. When the intent recognition output has high confidence, it directly matches the dialogue library to generate a response. If the match fails or the confidence is insufficient, it proceeds to the retrieval step, recalling relevant questions from the FAQ corpus and outputting answers. If the retrieval result still does not cover the user’s needs, the optimal retrieval result is passed to the generation module for completion and rewriting. This "evidence-first — generation fallback" strategy balances interpretability, coverage, and diversity, effectively reducing hallucinations and uncertainties that may arise from generative methods. To ensure the system continues to perform efficiently as the business evolves, a dynamic data maintenance mechanism is also designed: the dialogue and rule libraries are incrementally updated, the FAQ corpus is automatically mined and audited for parallel expansion, the retrieval dictionary and index support hot updates, and the generation model undergoes continuous fine-tuning and distillation as needed.

Experiments show that the hybrid generation architecture significantly improves response accuracy and user satisfaction, reduces average response latency by more than 15%, and decreases GPU usage by about 30%, fully verifying the advantages of the lightweight design. To quantitatively assess the response generation quality, we compared the proposed hybrid model with several baseline models on two datasets: Initial FAQ and Qingyun. The results are shown in Table 1.

Analysis of Table 1 shows that the hybrid generation model proposed in this study achieves the best performance in both domain-specific (Initial FAQ) and open-domain (Qingyun) scenarios. Its ROUGE-L and human evaluation scores significantly outperform all baseline models. This not only validates the effectiveness of the hybrid architecture in improving response quality but also demonstrates its excellent generalization capability. Importantly, the hybrid model received the highest and most stable human evaluation scores, which proves the success of its "evidence-first" routing strategy in ensuring response accuracy and controllability. This provides crucial support for the practical deployment of lightweight dialogue systems. This method forms a closed loop with the previously discussed semantic-enhanced intent recognition module, ensuring that the system maintains both accuracy and controllability while preserving natural fluency and efficient utilization of computational resources.

Table 1: Comparison Experiment of Hybrid Model and Multiple Baseline Models

Model	Dataset	ROUGE-1	ROUGE-2	ROUGE-L	Human Evaluation
TF-IDF	InitialFAQ	0.065	0.004	0.057	0.14
	Qingyun	0.078	0.010	0.067	0.16
ElasticSearch	InitialFAQ	0.085	0.013	0.082	0.27
	Qingyun	0.088	0.013	0.076	0.33
seq2seq	InitialFAQ	0.107	0.045	0.091	0.39
	Qingyun	0.198	0.107	0.181	0.54
GPT2	InitialFAQ	0.430	0.171	0.143	1.26
	Qingyun	0.463	0.184	0.464	1.35
PLATO	InitialFAQ	0.462	0.192	0.353	1.51
	Qingyun	0.510	0.203	0.461	1.45
Hybrid Model	InitialFAQ	0.522	0.218	0.484	1.64
	Qingyun	0.588	0.216	0.570	1.61

4.2. Cascaded Decision and Dynamic Routing Mechanism

Based on the multi-source corpus-driven three-channel response framework, the system constructs a cascaded routing mechanism focused on cost and risk, using three signals— “intent, entity, and dialogue state” —as decision-making criteria. The core idea is to always prioritize the channel with higher certainty and lower computational cost in each round of dialogue, and only extend to downstream channels when evidence or coverage is insufficient. Specifically, the router first reads the intent label and confidence, entity slots and their confidence from the SEEIR output, and combines it with the current stage of the dialogue state machine (such as guidance, confirmation, clarification, etc.). It then determines whether the conditions for triggering each channel in the “Rule → Retrieval → Generation” sequence are met. Once a match is found at any upstream module, it directly returns without invoking downstream modules. The routing thresholds are dynamically set using a statistical learning method based on historical dialogue data: the initial thresholds are set based on the validation set performance (e.g., rule module confidence > 0.9, retrieval module similarity > 0.85), and are adjusted every 24 hours through an online learning mechanism. The adjustment strategy is based on a feedback loop: if any module continuously misroutes beyond a preset number of times (e.g., 10 times), threshold recalibration is automatically triggered. In practical deployment, the error routing rate decreased from an initial 5.2% to 1.8% after two weeks of optimization, significantly improving system stability. This strategy complements the system’s hybrid generation design: on one hand, it significantly reduces the average computational cost of the natural language generation pipeline and improves concurrency performance, while on the other hand, it stabilizes the interpretability and consistency of response selection by relying on high-confidence intent and entity anchors at the front end.

Routing decisions are not simple threshold judgments but instead work in tandem with a state-rule joint mechanism for dialogue management. The system continuously receives abstracted dialogue states as inputs and, based on predefined state transitions and decision logic, identifies the current stage and user intent to choose the actions to execute and determine the channel. For example, in transformation tasks, if the initial guidance receives positive feedback, the system directly proceeds with the rule-based dialogue and completes the process; if feedback is unclear, a few rounds of clarification/chat are allowed to gather the missing points; if valid evidence is still not formed, the retrieval channel is triggered to obtain traceable answers; when retrieval still cannot meet the needs or the contextual constraints are strong, top-k evidence is used to invoke the generation module; if repeated attempts fail, the system reverts to human intervention to ensure service timeliness and user experience. This "State-Action-Channel" integrated design ensures the orderly progression of processes and flexible responses in abnormal situations.

To maintain robust and lightweight routing under real traffic conditions, the system introduces a set of adjustable control and optimization strategies. First, there is an activation and circuit breaker mechanism: dynamic quotas and trigger cooling times are set for the generation channel, adjusting the trigger thresholds based on business risk levels and current load to prevent frequent recalls of large models during peak periods. Second, there is state-aware online updates: dialogue states and routing strategies can be hot-updated during runtime, working with multiple chat/business subsystems to cover diverse needs. Third, there is evidence-first caching and reuse: short-term caching is maintained for high-frequency questions and stable phrases, with retrieval results and structured slots reused at the session level, reducing redundant recall and inference. Fourth, there is a feedback loop: routing decisions, channel usage, failure reasons, and human intervention triggers are all recorded for offline threshold calibration and strategy optimization. These measures bring quantifiable benefits in practice: without altering the core model structure, system response latency and computational load significantly decrease, while the continuity of user interactions and the success conversion rate improve.

Under this mechanism, cascaded routing and hybrid generation form a clear division of responsibilities: rule and state logic govern the "deterministic first" fast path, retrieval provides "evidence-verifiable" knowledge completion, and the generation module serves as the "semantic creation" fallback. The three modules share SEEIR’s high-confidence intent and entity outputs as common semantic anchors, ensuring that response quality and safety are guaranteed while controlling overall computational consumption within acceptable limits for lightweight deployment.

5. Conclusion and Outlook

This paper presents and implements a complete technical path from "high-confidence intent understanding" to "evidence-first hybrid generation" for constructing intelligent dialogue systems in resource-constrained environments. The upstream semantic entity-enhanced intent recognition model (SEEIR) strengthens the role of key entities and syntactic structures through multi-view inputs, decoupled-reconstruction collaborative learning, and attention isolation, maintaining higher discriminative power and robustness in complex contexts and multi-entity interference scenarios. The downstream lightweight hybrid generation combines rule-based and retrieval-based methods with neural generation as a fallback, achieving a balance between quality and cost with the aid of cascaded decision-making and dynamic routing; combined with dynamic corpus construction and online updating mechanisms, the system ensures explainability, coverage, and response efficiency in real-world scenarios. Experiments show that this paradigm outperforms traditional baselines on both public and self-constructed datasets and meets lightweight deployment requirements in terms of latency and computational cost, providing a replicable engineering model for low-cost, high-concurrency human-computer interactions.

From a methodological perspective, the research proposes an "architecture prior + controlled learning" framework for intent recognition, explicitly injecting semantic skeletons and entity clues via multi-view construction, reinforced with reconstruction tasks for fine-grained modeling, and controlled by attention mechanisms to avoid information leakage, significantly reducing overfitting and dependency on annotated data. It also presents an "evidence-first — fallback generation" hybrid generation solution: rules ensure determinism and compliance, retrieval provides verifiable evidence, and generation handles semantic creation and style consistency. The dynamic routing and threshold adjustment effectively suppress large-model hallucinations and reduce overall computational costs. The system also forms a closed-loop of training, inference, and operation with data, including low-shot chain-prompt corpus augmentation, session-level cache reuse, and online feedback-driven strategy iterations to support business migration and scenario expansion.

Despite achieving good results, there is still room for improvement. SEEIR’s adaptation to proprietary entities and heterogeneous expressions still relies on some manual calibration, and the cost of rapid cross-domain migration can still be reduced. The quality and coverage of rule and retrieval resources determine the boundaries of explainability, and long-tail and timeliness knowledge updates still require more automated governance. The fallback generation may suffer from style drift and safety alignment risks in open-domain dialogues and multi-turn noisy interactions. Routing thresholds and strategies mainly rely on offline tuning and have not fully introduced uncertainty estimation and long-term return optimization for global resource scheduling. Future work may explore cross-domain continual learning and parameter-efficient fine-tuning, build pluggable entity ontologies and task descriptions to support rapid cold start, introduce knowledge graphs and evidence tracing mechanisms for dynamic governance and automated evaluation of retrieval and rule resources, incorporate adversarial training, refusal learning, and uncertainty calibration to further reduce generation hallucination risks, optimize inference scheduling under edge-cloud collaboration using quantization, batching, and reinforcement learning scheduling to improve peak throughput, and introduce multi-dimensional evaluation metrics like user experience and takeover rates to drive long-term strategy iteration, allowing the system to continuously approach the expected goals during business evolution.

References

[1] Deriu J, Rodrigo A, Otegi A, et al. Survey on evaluation methods for dialogue systems. Artificial Intelligence Review, 2021, 54(1): 755-810.

[2] Ni J, Young T, Pandelea V, et al. Recent advances in deep learning based dialogue systems: A systematic survey. Artificial intelligence review, 2023, 56(4): 3055-3155.

[3] Huang, J. (2025). Research on Cloud Computing Resource Scheduling Strategy Based on Big Data and Machine Learning. European Journal of Business, Economics & Management, 1(3), 104-110.

[4] Xu H, Zhang H, Lin T E. Intent Recognition[M]//Intent Recognition for Human-Machine Interactions. Springer Nature Singapore, 2023: 7-29.

[5] Liu X. The Role of Generative AI in the Evolution of Digital Advertising Products. Journal of Media, Journalism & Communication Studies, 2025, 1(1): 48-55.

[6] Bai J, Yan Z, Zhang S, et al. Infusing internalized knowledge of language models into hybrid prompts for knowledgeable dialogue generation. Knowledge-Based Systems, 2024, 296: 111874.

[7] Zhou, Y. (2025). Improvement of Advertising Data Processing Efficiency Through Anomaly Detection and Recovery Mechanism. Journal of Media, Journalism & Communication Studies, 1(1), 80-86.

[8] Huijie Pan. Design of Data-Driven Social Network Platforms and Optimization of Big Data Analysis. Machine Learning Theory and Practice (2025), Vol. 5, Issue 1: 133-140.

[9] Yixian Jiang. Research on Integration and Optimization Strategies of Cross-platform Machine Learning Services. Machine Learning Theory and Practice (2025), Vol. 5, Issue 1: 141-148.

[10] Jiangnan Huang. Application of AI-driven Personalized Recommendation Technology in E-commerce. International Journal of Neural Network (2025), Vol. 4, Issue 1: 40-47.

[11] Huijie Pan. Discussion on Low-Latency Computing Strategies in Real-Time Hardware Generation. International Journal of Neural Network (2025), Vol. 4, Issue 1: 48-56.

[12] Huijie Pan. Discussion on Low-Latency Computing Strategies in Real-Time Hardware Generation. International Journal of Neural Network (2025), Vol. 4, Issue 1: 57-64.

[13] Chuying Lu. Object Detection and Image Segmentation Algorithm Optimization in High-Resolution Remote Sensing Images. International Journal of Multimedia Computing (2025), Vol. 6, Issue 1: 144-151.

[14] Buqin Wang. Research on Load Balancing Technology in Distributed System Architecture. International Journal of Multimedia Computing (2025), Vol. 6, Issue 1: 152-159.

[15] Yajing Cai. Distributed Architecture and Performance Optimization for Smart Device Management. International Journal of Big Data Intelligent Technology (2025), Vol. 6, Issue 2: 130-138.