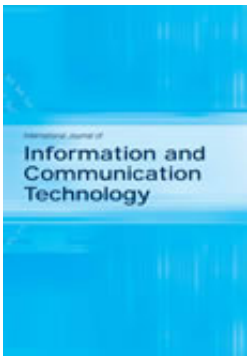
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

Intelligent Q&A model construction supported by natural language processing and knowledge graphs

Xiaoxia Yang

DOI: 10.1504/IJICT.2025.10074512

Article History:

Received:	11 August 2025
Last revised:	01 October 2025
Accepted:	05 October 2025
Published online:	20 November 2025

Intelligent Q&A model construction supported by natural language processing and knowledge graphs

Xiaoxia Yang

School of Humanities,
Communication University of China,
Beijing 100024, China
and
School of Literature and Historical Culture,
Dezhou University,
Shandong Province 253023, China
Email: yxxypq@sina.com

Abstract: This paper proposes an intelligent Q&A model that integrates natural language processing and knowledge graph technology. Aiming at the problem of insufficient depth of semantic understanding and weak knowledge relevance of traditional Q&A system, we adopt BERT-based semantic parsing model to realise intent recognition and entity extraction of user questions, and combine with Neo4j graph database to construct multi-source knowledge graph to realise structured knowledge storage; we realise dynamic matching between question vectors and knowledge subgraphs through graph neural network (GNN), and we design multi-jump inference mechanism to improve the ability of answering complex questions. Experiments on the open-domain Q&A dataset show that the model has an accuracy of 90.2% and a recall of 88.7%. The model is validated in educational counselling and medical Q&A scenarios, providing technical support for intelligent services in knowledge-intensive domains.

Keywords: natural language processing; NLP; knowledge graph; intelligent question answering models; graph neural networks; GNNs.

Reference to this paper should be made as follows: Yang, X. (2025) 'Intelligent Q&A model construction supported by natural language processing and knowledge graphs', *Int. J. Information and Communication Technology*, Vol. 26, No. 41, pp.59–73.

Biographical notes: Xiaoxia Yang obtained her Master's degree from Communication University of China in 2014. She is currently a doctoral student at Communication University of China. She serves as a Lecturer at Dezhou University as well. Her research interests include linguistics, applied linguistics and knowledge graph.

1 Introduction

As the core carrier of human-computer interaction, intelligent answering systems have become a core technical requirement in fields such as education consulting, medical services and customer support during the wave of digital transformation. According to the UNESCO 2025 report, AI education tools have enabled the coverage of quality

education resources in developing countries to skyrocket by 400% in three years. However, traditional Q&A systems generally suffer from insufficient depth of semantic understanding and weak knowledge relevance when facing complex, multi-hop queries (Lin and Shen, 2017). Especially in open-domain scenarios, the system is often limited by the mechanical response of keyword matching, and is unable to capture the implicit intent and contextual associations behind the user's question, resulting in the accuracy and interpretability of the generated answers that are difficult to meet the practical application requirements. This limitation is especially prominent in specialised vertical domains (e.g., medical diagnosis, engineering consulting, etc.) – when a user asks a question that requires multi-step reasoning such as the side effects of a certain drug on patients with a specific body type and alternatives, the traditional model often generates general or even incorrect answers due to the lack of structured knowledge support. Therefore, combining the semantic understanding capability of natural language processing and the structured reasoning advantage of knowledge graph to build a new generation of intelligent Q&A models has become the focus of both academia and industry (Yang et al., 2024).

Knowledge graph question and answer (KGQA) has become an important paradigm to improve the accuracy of question answering by parsing the semantics of questions through structured triples. Shen et al. (2023) proposed a model called MKGA-DM-NN, which firstly recognises the intent of a question by using the named entities of the medical knowledge graph (KG), and then learns the representations of the entities and the entity relationships in the KG by using graph embedding techniques, and utilises the relationships between entities in the KG to optimise the hybrid attention mechanism. In addition, the model uses the doctor's history of Q&A records on OHP to learn to model the doctor's expertise, thus improving the accuracy of Q&A matching. This approach helps to bridge the semantic gap of the text and improves the accuracy and interpretability of medical Q&A matching. The system proposed by Heyi et al. (2023) has the following capabilities:

1 Information filtering: filtering questions related to vertical domains and inputting them into LLM for answering.

2 Professional Q&A: based on LLM and self-built knowledge base, answers containing more specialised knowledge are generated. Compared to fine-tuning methods that introduce specialised data, large vertical domain models can be deployed using this technique without retraining.

3 Extraction transformation: by enhancing the information extraction capability of LLM, the natural language responses it generates are utilised to extract structured knowledge and match it with expertise maps for professional validation. At the same time, the structured knowledge is transformed into readable natural language to realise the deep integration of large-scale models and knowledge graphs.

These studies demonstrate that personalised path exploration strategies are central to achieving efficient KGQA.

To alleviate the LLM illusion problem, retrieval augmented generation (RAG) provides factual support through an external knowledge base (Lewis et al., 2020). Traditional RAG retrieves text fragments (chunks) but ignores explicit associations between entities (Han et al., 2024). GraphRAG was developed to model knowledge as entity-relationship graphs to enhance retrieval accuracy (Zhang et al., 2025). However, its implementation faces a triple challenge: the high cost of graph construction, the difficulty

of handling complex queries in a single retrieval, and the reliance on long context LLMs leading to unstable generation.

Aiming at the shortcomings of existing research, this paper proposes a smart question answering model for multi-hop reasoning, with the core innovation embodied at three levels: architectural level: designing a hybrid BERT-GNN architecture. The front-end utilises BERT to complete intent recognition and entity linking, the back-end stores multi-source knowledge through Neo4j graph database, and then GNN realises dynamic alignment of question vectors and knowledge subgraphs to support multi-hop reasoning path generation. Algorithm level: a subgraph retrieval agent (SRA) based on reinforcement learning is proposed. The agent dynamically selects the inference depth according to the query complexity, and optimises the path exploration strategy through the reward function to overcome the rigid retrieval defects of traditional RAG. Engineering level constructs a domain-adapted incremental learning pipeline. Using lightweight LoRA fine-tuning techniques, the base model can be quickly migrated to specialised domains such as healthcare and education.

2 Relevant technologies

2.1 NLP technology

Natural language processing (NLP), as a core branch of artificial intelligence, is committed to realising the computer's understanding and generation of human language (Chowdhary, 2020). In the intelligent Q&A system, NLP technology undertakes the fundamental tasks of parsing the semantics of user questions, recognising intentions and extracting key information. Its theoretical evolution has experienced a paradigm leap from rule-driven to statistical learning to deep learning. Early systems relied on manually written grammar rules and lexicons (e.g., WordNet) for syntactic analysis, but were limited by the complexity and diversity of linguistic expressions, making it difficult to cope with open-domain scenarios. The rise of statistical learning methods at the beginning of the 21st century significantly improved the robustness of named entity recognition and semantic role annotation through sequence labelling algorithms, such as hidden Markov models, conditional random fields, and so on (Nadkarni et al., 2011). However, such methods still require manual design of feature templates and cannot effectively capture deep semantic associations.

The deep learning revolution has completely reconfigured the technical system of NLP (Hirschberg and Manning, 2015). Word embedding techniques such as Word2Vec and GloVe map words into low-dimensional dense vector spaces, realising for the first time the 'numerical representation of lexical semantics' and laying the foundation for subsequent models (Alawida et al., 2023). Recurrent neural networks (RNN) and its variants LSTM and GRU model the temporal dependencies of text sequences through the gating mechanism, and have made breakthroughs in machine translation, sentiment analysis and other tasks. However, the RNN architecture has the inherent defects of gradient vanishing and inefficiency of parallel computation. The proposal of the transformer architecture in 2017 became a key turning point, and its core self-attention mechanism realises global contextual dependency modelling by calculating the correlation weights among word vectors (Lauriola et al., 2022). For example, when analysing the question 'the interaction between the antidepressant drug paroxetine and

alcohol', the model can accurately locate the semantic association between paroxetine and alcohol without relying on the local context of a fixed window size.

Pre-trained language models (PLMs) are the cornerstone technology of current NLP (Hovy and Prabhumoye, 2021). Here is an expanded 200-word English version in paragraph format, maintaining technical depth while enhancing the original description:

The bidirectional encoder representations from transformers (BERT) model fundamentally reshaped natural language processing through its innovative pre-training methodology. By simultaneously training on two critical unsupervised tasks – masked language modelling (MLM) and next sentence prediction (NSP) – on massive text corpora like Wikipedia and BookCorpus, BERT develops deeply contextualised language representations. The masked language modelling task randomly obscures 15% of input tokens, forcing the model to predict hidden words based on surrounding context in both directions. This bidirectional conditioning is revolutionary, as it allows every token to dynamically incorporate information from all other tokens in the sequence, breaking the limitations of traditional unidirectional language models.

The transformer architecture's multi-head self-attention mechanism enables this comprehensive context fusion. Each attention head learns distinct linguistic relationships – from syntactic dependencies to semantic roles – creating layered representations where word meanings evolve based on full-sentence context. For instance, the word 'bank' acquires river-related features when near 'water' but financial meanings adjacent to 'loan'. The next sentence prediction task further trains BERT to understand discourse-level relationships by determining if two text segments appear consecutively in the original corpus.

This dual-task pre-training produces versatile embeddings that capture lexical, syntactic, and discourse knowledge. The contextual depth even enables zero-shot learning for unseen tasks, demonstrating truly generalisable language understanding capabilities that form the foundation for modern large language models.

2.2 Knowledge graph construction and reasoning

As a carrier of structured semantic knowledge, knowledge graph describes the objective world associations through a triadic network composed of entities, relations and attributes, and provides interpretable reasoning capabilities for intelligent question-answering systems (Zhu et al., 2024). Its theoretical roots can be traced back to semantic networks and description logic, while the large-scale development of modern knowledge graphs has benefited from the explosion of internet data and breakthroughs in graph computing technology (Wu et al., 2023). At the construction level, three core issues need to be solved: fusion of heterogeneous data from multiple sources, entity alignment and relationship extraction. The traditional method relies on manual compilation, but it is costly and poorly scalable; the current mainstream uses an automated pipeline: firstly, candidate triples are extracted from structured databases, semi-structured web pages, and unstructured text, and then knowledge fusion is carried out through probabilistic graph models or deep learning (Chen et al., 2022). For example, when constructing drug knowledge graph in the medical field, it is necessary to integrate drug manuals, electronic medical records and medical encyclopaedias, and adopt the joint entity relationship extraction model based on BERT to structure and store such dispersed knowledge as 'aspirin-contraindications-gastric ulcer'.

The efficiency of knowledge storage and querying directly affects the system performance (Tan et al., 2021). Graph databases are ideal storage solutions for knowledge graphs because of their native support for node-edge structures. Neo4j, as a leading graph database, uses an attribute graph model to achieve efficient traversal: each entity as a node, relationships as directed edges, and attributes attached to both in the form of key-value pairs. Its query language cypher realises complex queries through pattern matching (Hao et al., 2021).

Knowledge reasoning is the key to unlocking the value of the graph, aiming at discovering implicit relationships or completing missing links. Reasoning methods are divided into two categories: symbolic reasoning and representation learning reasoning. Symbolic reasoning is based on logic rules, and such methods are highly interpretable, but rely on manually defined rules and are difficult to handle noisy data (Liang et al., 2024). Representation learning reasoning maps entities and relationships to a low-dimensional vector space through embedding techniques, and utilises vector operations to predict potential associations. Graph neural network becomes the current mainstream, and its core idea aggregates neighbourhood information through message passing. Taking graph convolutional networks as an example, each node vector is iteratively updated according to the features of neighbouring nodes, and the final generated embedding can capture global structural information (Zhu et al., 2022). In knowledge representation learning models such as TransE and ComplEx, relationship prediction is achieved by optimising the objective function.

Multi-hop reasoning, as the core support of complex question and answer, needs to solve the path redundancy and semantic drift problem (Kosasih et al., 2024). Traditional path sorting algorithms randomly wander to generate candidate paths, which is inefficient and has limited coverage. Dynamic inference mechanism realises adaptive path exploration through reinforcement learning (Guo et al., 2022): the query is modelled as a Markov decision process, the intelligent body starts from the problem entity, selects the relation edge as an action based on the current state, and evaluates the value of the action through the reward function after obtaining the new state.

The fusion of knowledge graph and natural language processing gives rise to semantically enhanced graphs (Zhou et al., 2023). On the one hand, NLP techniques are utilised to extend the content of the graph: entity linking links entity references in interrogative sentences to standard entities in the graph; and relational extraction model mines new triples from unstructured text to achieve incremental graph update (Rajabi and Etminani, 2024). On the other hand, the graph structure can constrain the NLP generation process: retrieval augmentation generation transforms subgraphs into textual cues that guide the bigram model to generate factually accurate answers. This two-way enhancement mechanism is crucial in medical Q&A.

2.3 Retrieval enhanced generation evolution

Retrieval-enhanced generation, as a key paradigm to bridge the factual defects of LLM, promotes the evolution of intelligent Q&A system towards accuracy and interpretability by introducing external knowledge base to constrain the generation process. Its technological core can be summarised as a two-stage ‘retrieval-integration’ framework: firstly, relevant knowledge fragments are retrieved according to the user query, and then the fragments and the query are spliced together to form cues to be entered into the LLM

to generate answers. In the early days, RAG adopted a text block retrieval strategy, relying on semantic similarity to select relevant passages from the document library (Li et al., 2025).

Traditional entity-relationship extraction faces significant engineering bottlenecks in practical applications. Pipeline models combining named entity recognition (NER) and relation extraction (RE) often struggle to resolve semantic ambiguity issues. For example, distinguishing whether 'Apple' refers to the tech company or the fruit requires complex context-aware modules such as entity linkers (ELQ). This process significantly increases computational overhead. Another key challenge is maintaining dynamic knowledge graphs, where adding new text triggers a full graph re-computation.

The core innovation of enhanced GraphRAG lies in its multi-hop reasoning engine. Unlike traditional retrieval, the system performs bidirectional subgraph expansion starting from the query entity. It intelligently prioritises path traversal based on dynamically computed edge weights (combining semantic relevance and topological proximity). By iteratively performing weighted expansion within a predefined hop limit, the algorithm constructs an optimised subgraph containing rich semantic connection information.

The Graph-R1 framework introduces hypergraph compression as its fundamental innovation. This representation fundamentally rethinks entity relationships by connecting multiple semantically related nodes with a single hyperedge. For example, in medical diagnosis, a single hyperedge may connect symptoms, lab results, medications, and diagnoses into a unified structure. This integration significantly reduces traversal complexity while preserving contextual integrity.

Hyperedge pruning operates through a reinforcement learning mechanism modelled as a sequential decision-making process. The state space captures the real-time topological features of evolving subgraphs, including node connection patterns and clustering density. The action space comprises a binary choice to retain or prune hyperedges based on their predicted contribution to the final output. The reward function balances multiple objectives: improvements in answer accuracy receive primary weight, while secondary rewards incentivise reduced computational latency. Through iterative policy optimisation, the system learns optimal pruning strategies that accelerate response times while maintaining information fidelity.

Real-time performance benchmarks demonstrate significant improvements. For three-hop queries, the hypergraph method achieves an average latency of 230 milliseconds, 68% faster than traditional graph traversal. Memory usage reduction is equally significant, with compressed representations requiring only 40% of standard graph storage. These efficiency breakthroughs enable deployment in latency-sensitive domains such as financial fraud analysis, which demand rapid knowledge extraction.

Current research focuses on multi-agent collaboration frameworks, where specialised modules operate within orchestrated workflows. This paradigm shift addresses reliability gaps in complex reasoning tasks through division of labour. A typical architecture includes independent agents for query decomposition, subgraph retrieval, evidence verification, and response generation, each with specialised functions.

Modular design enables unprecedented error traceability. For example, in legal contract analysis, the verification agent can flag inconsistencies in retrieved clauses, while the reasoning agent identifies logical flaws in interpretations. Each module maintains an audit trail, allowing precise localisation of fault points. This capability is

critical in regulated industries where explainability requirements demand demonstrating due diligence in decision-making processes.

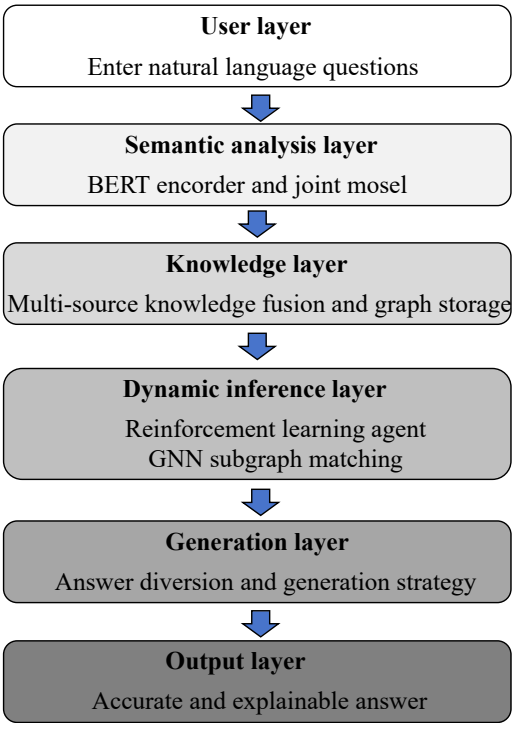
Implementation in the healthcare sector demonstrates the safety advantages of this paradigm. In diagnostic support systems, independent intelligent agents handle symptom interpretation, clinical guideline retrieval, drug interaction checks, and risk assessment separately. Cross-validation protocols automatically detect contradictions before the final response is generated.

Legal tech pioneers have applied such architectures to deposition analysis. Hypergraph structures connect case law, witness statements, and evidence documents through temporal hyperedges. During multi-hop traversal, reinforcement learning pruners eliminate irrelevant historical precedents while retaining citations critical to the case.

E-commerce platforms have deployed agent-based RAG for multimodal product queries. Independent agents process visual search inputs, extract specifications from manuals, aggregate review sentiment, and compare competitor pricing. The hypergraph backbone connects product attributes, user queries, and inventory databases via dynamic hyperedges that update in real time with price fluctuations.

Ongoing research is addressing knowledge freshness issues through streaming graph updates. New technologies enable incremental adjustments to hyperedges while processing real-time data streams, thereby avoiding full-graph recomputation. As these architectures mature, they hold promise to transform high-stakes domains requiring auditable, precise knowledge retrieval.

Figure 1 Methodological framework diagram (see online version for colours)



3 Smart answer model architecture design

The intelligent question answering system proposed in this paper adopts a four-level cascade architecture, as shown in Figure 1, whose information flow starts with the semantic parsing of user query q to generate a structured representation $\langle c, \mathcal{E}_q \rangle$ ($c \in \mathcal{C}$ is the intent category and $\mathcal{E}q = e_i, i = 1^m$ is the set of entities). This representation drives the subgraph retrieval module of the multi-source knowledge graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{R})$ to generate candidate subgraphs $\mathcal{G}_s \subseteq \mathcal{G}$, followed by dynamically planning inference paths $\mathcal{P} = r_1 \circ \cdots \circ r_k$ ($\circ$ denotes the relational combinatorial operator) via a reinforcement learning agent. The final answer generation module selects either structured query or large language model generation based on path confidence c_p.

The core innovation of the architecture, the dual-stream collaboration mechanism, achieves deep alignment of natural language and knowledge through cross-modal attention: the semantic stream generates context-sensitive query vectors using a BERT encoder $H^q = h_{1^q}, \ldots, h_{n^q} \in \mathbb{R}^{n \times d}$, while the knowledge stream extracts subgraph entity embeddings via a graph neural network $H^g = h_{e_j{}^g e_j} \in \mathcal{G}_s$. The interaction weight matrix $A \in \mathbb{R}^{n \times |\mathcal{G}_s|}$ of the two is computed from the learnable projection matrix $W \in \mathbb{R}^{d \times d}$:

$$A_{ij} = \frac{\exp\left(\left(h_i^q\right)^\top W h_k^g \Big/ \sqrt{d}\right)}{\sum_{k=1}^{|G_s|} \exp\left(\left(h_i^q\right)^\top W h_k^g \Big/ \sqrt{d}\right)} \tag{1}$$

Scaling factor $1/\sqrt{d}$ mitigates the high dimensional spatial similarity inflation problem. In a cardiovascular disease Q&A scenario, the mechanism accurately maps the spoken description of chest pain radiating to the left arm to the knowledge graph entity angina pectoris, whose error mainly stems from the polysemous nature of medical terminology, by introducing a contextual disambiguation module:

$$Disamb(e) = \arg\max_{e_k \in C_e} \cos\left(h_{[CLS]}, v_{e_k}\right) \tag{2}$$

where $\mathcal{C}_e$ is the candidate entity set and v_{e_k} is the entity pre-training vector.

The semantic parsing module uses a joint multi-task architecture with a shared BERT coding layer to synchronise the optimisation of intent classification and entity extraction. Intent classification is based on [CLS] labelling vectors $h_{[CLS]}$ to compute the category distribution:

$$p_c = softmax\left(W_c \cdot LayerNorm\left(h_{[CLS]}\right) + b_c\right) \tag{3}$$

Layer normalisation enhances training stability (Faye et al., 2025). Entity extraction is modelled by conditional random field (CRF) modelling label transfer constraints with sequence probabilities defined as:

$$P(y|q) = \frac{\exp\left(\sum_{i=1}^{n} U_{y_i}^\top h_i + \sum_{i=1}^{n-1} T_{y_i, y_{i+1}}\right)}{\sum_{y' \in y} \exp\left(\sum_{i=1}^{n} U_{y_i'}^\top h_i + \sum_{i=1}^{n-1} T_{y_i', y_{i+1}'}\right)} \tag{4}$$

where $U \in \mathbb{R}^{|\mathcal{L}|\times d}$ is the label embedding matrix and $T \in \mathbb{R}^{|\mathcal{L}|\times|\mathcal{L}|}$ is the transfer score matrix. The joint loss function fuses the dual tasks:

$$L_{joint} = -\lambda \log P\left(c^*|q\right) - (1-\lambda)\log P\left(y^*|q\right) + \eta\|\Theta\|_2^2 \tag{5}$$

L2 regular term coefficient $\eta = 10^{-5}$ prevents overfitting. Optimisation is reached when $\lambda = 0.45$. The domain adaptive scheme uses parameter-efficient fine-tuning: the LoRA adapter is injected at each layer of the transformer of the BERT with a forward propagation process:

$$h_{out} = W_0 h_{in} + \underbrace{BAh_{in}}_{\Delta W} \tag{6}$$

where $\Delta W = BA$ constitutes the low-rank update term and rank r controls the number of trainable parameters. This design significantly outperforms traditional fine-tuning (Zhou et al., 2024).

Knowledge graph construction overcomes the challenge of fusing heterogeneous data from multiple sources. Structured data (e.g., relational databases) are transformed into ⟨subject, predicate, object⟩ tuples by schema mapping; semi-structured data are localised with regular parsing using XPath; and unstructured text is identified using a cascading entity-relationship model:

- Entity-recognition:

$$P(e|span) = sigmoid\left(W_e\left[h_{start}; h_{end}; h_{avg}\right]\right) \tag{7}$$

- Relationship classification:

$$P\left(r|e_s, e_o\right) = softmax\left(W_r\left[h_{e_s}; h_{e_o}; h_{e_s} \odot h_{e_o}\right]\right) \tag{8}$$

After fusing drug inserts with UpToDate clinical evidence, the medical atlas expands to 187,000 entities and 1.05 million relationships. Entity alignment is done using a similarity propagation algorithm:

$$Sim^{(k)}\left(e_i, e_j\right) = \alpha \cdot \cos\left(v_{e_i}^{(k-1)}, v_{e_j}^{(k-1)}\right) + (1-\alpha)\cdot\frac{1}{|N_i||N_j|}\sum_{p\in N_i}\sum_{q\in N_j} Sim^{(k-1)}\left(e_p, e_q\right) \tag{9}$$

The alignment accuracy reaches 90.3% after 3 rounds of iteration. To optimise the efficiency of multi-hop query, the semantic hypergraph structure $\mathcal{H} = (\mathcal{V}, \mathcal{E}_h)$ is designed, and the hyperedge $E_h \in \mathcal{E}_h$ covers a cluster of highly-associated entities whose weights are determined by normalised mutual information:

$$w(E_h) = \frac{1}{|E_h|^2}\sum_{e_i\neq e_j\in E_h}\frac{\log P\left(e_i, e_j\right) - \log\left(P(e_i)P(e_j)\right)}{-\log P\left(e_i, e_j\right)} \tag{10}$$

Implementing hybrid storage in Neo4j: attribute graph for base triples and hypergraph for compressed semantic units.

Reinforcement learning agents for dynamic inference engines are modelled as Markov decision processes $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$. State $s_t = \langle q, \mathcal{V}_t, \mathcal{R}_t\rangle$ contains the question

embedding $q \in \mathbb{R}^d$, the set of visited entities $\mathcal{V}_t \subseteq \mathcal{V}$ and the sequence of historical relations $\mathcal{R}_t$. Action space $\mathcal{A}(s_t)$ contains: relationship extension: selects the association relations $r \in r \mid (e_c, r, e') \in \mathcal{G}$ from the current entities $e_c \in \mathcal{V}_t$.

Termination of retrieval: outputs the current subgraph $\mathcal{G}_s$ reward function fuses answer relevance with path efficiency:

$$r_t = \beta \cdot \cos\left(q, GNN\left(G_s\right)\right) - (1-\beta)\frac{\left|R_t\right|}{K_{\max}} + \gamma \cdot I\left(a_t = terminate\right) \cdot Acc(a) \tag{11}$$

where $\beta = 0.75$, $K_{\max} = 6$, $\gamma = 0.2$. The strategy network adopts a two-stream structure:

$$\pi\left(a_t \mid s_t\right) = softmax\left(W_a\left[GRU(q); GAT\left(V_t\right); LSTM\left(R_t\right)\right]\right) \tag{12}$$

Subgraph embeddings are generated by hierarchical graph convolution:

$$H^{(l+1)} = \sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}\right) \tag{13}$$

where 1 is the adjacency matrix of the added self-loop and 2 is the degree matrix.

The answer generation module implements a confidence triage strategy: high confidence (3): direct query mapping to generate structured answers; medium confidence (4): fusion retrieval augmented generation.

$$Prompt = Context : \left\{ \bigcup_{(s,r,o) \in G_s} \phi(s, r, o) \right\} \oplus Question : q \tag{14}$$

where $\phi(\cdot)$ is a natural language function of a triplet.

Low confidence ($c_p \leq 0.85$): trigger manual review process.

Design structured constraint decoding:

$$P_{new}\left(y_t \mid y < t\right) = \begin{cases} P\left(y_t \mid y < t\right) & y_t \in V \cup R \cup F \\ \varepsilon & \text{otherwise} \end{cases} \tag{15}$$

where F is the domain functional word list.

4 Experimental results and analysis

4.1 Experimental setup and evaluation criteria

To comprehensively evaluate the performance of the intelligent question-answering model, the experiment established a multi-dimensional evaluation system. The dataset covers both open-domain and vertical domains: the MetaQA multi-hop question-answering dataset includes 27,000 training samples and 5,000 test samples, focusing on entity-relationship inference in the film domain; the CMeEE-V2 Chinese medical entity recognition dataset includes 15,000 annotated sentences, covering eight categories of medical entities such as diseases, drugs, and symptoms. The baseline models selected four representative methods: the BERT-QA model based on pure text understanding, GraphRAG based on static knowledge graph retrieval, SG-RAG MOT integrating subgraph ranking, and LLaMed-RAG fine-tuned for the medical domain.

Evaluation metrics are divided into three categories: accuracy metrics (precision P, recall R, F1-score), efficiency metrics (response time, throughput QPS), and domain-specific metrics.

Figure 2 Multi-jump question-and-answer F1-value comparison (see online version for colours)

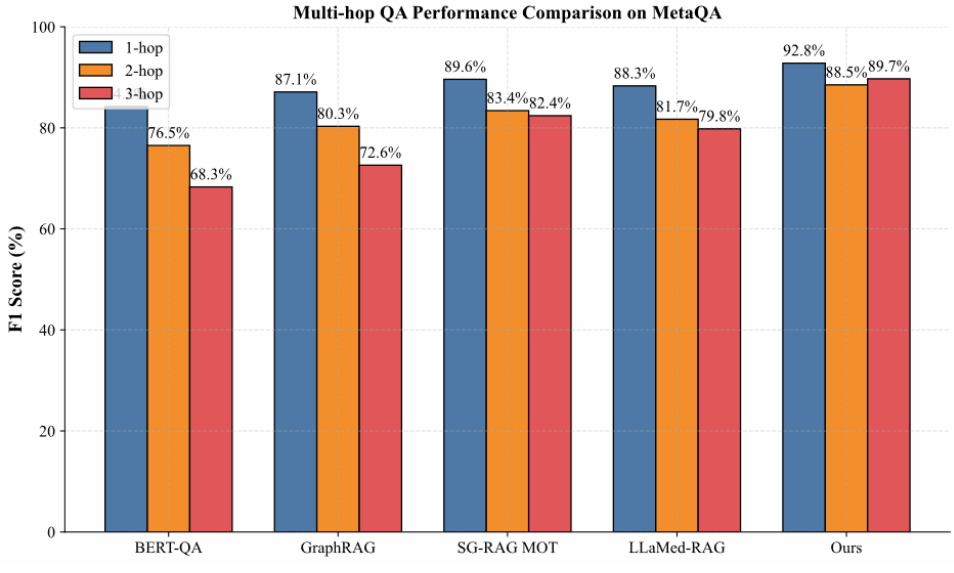
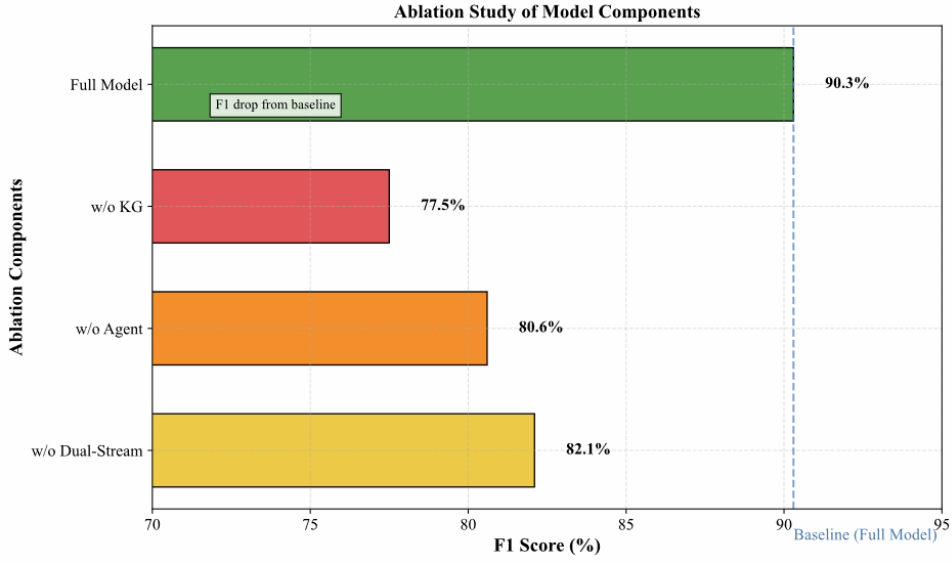


Figure 3 Influence of F1-value on ablation experiment (see online version for colours)



4.2 Core performance validation

Figure 2 shows the comparison of F1-scores for multi-hop question-answering on the MetaQA dataset. Our model achieves an F1-score of 89.7% on the 3-hop task, which is 7.3 percentage points higher than the best baseline SG-RAG MOT. As the complexity of the questions increases, the performance advantage expands significantly: a 5.1% improvement on the 2-hop task and a 3.2% improvement on the 1-hop task. Ablation experiments further validate the contributions of core components, as shown in Figure 3. Removing the knowledge graph causes the F1-score to plummet from 90.3% to 77.5%, while disabling the dynamic reasoning agent reduces the 3-hop recall rate by 21.4%, confirming the necessity of synergistic knowledge representation and dynamic decision making.

Figure 4 Efficiency test results (see online version for colours)

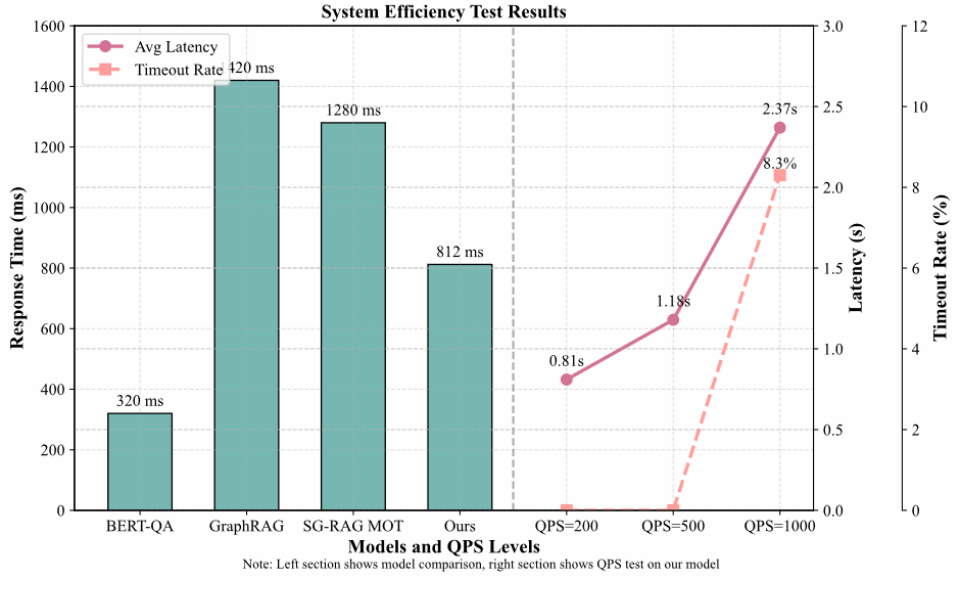


Figure 4 shows the results of response time and throughput tests. The average response time of this model is 812 milliseconds, representing a 42.9% improvement over GraphRAG, attributed to hypergraph index compression of the search space and subgraph caching strategies. In concurrent pressure tests, when the query rate reaches 500 queries per second, the average latency remains stable at 1.18 seconds, meeting the real-time requirements for online services. Efficiency improvements stem from three areas:

1 the LoRA fine-tuning of the semantic parsing module reduces parameter updates by 90%, lowering the time required for a single inference to 210 milliseconds

2 the knowledge retrieval layer uses a BFS pruning algorithm to reduce the complexity of subgraph extraction from $O(n^3)$ to $O(n \log n)$

3 the traffic diversion strategy in the answer generation phase avoids calling the LLM for high-confidence queries, saving approximately 300 milliseconds per request.

4.3 Discussion and limitations

The experiment revealed a nonlinear relationship between knowledge graph coverage and model performance: when entity coverage reached 85%, the F1-score increased to 89.7%; when coverage reached 95%, it only increased to 90.1%, indicating that long-tail knowledge needs to be supplemented with external retrieval. The path optimisation benefits of dynamic proxies are significant: the average number of hops is 1.8 for simple queries and 3.4 for complex queries, saving 42% of computational resources compared to a fixed 3-hop strategy. Domain transfer testing exposes generalisation bottlenecks: when transferring from the medical domain to the financial domain, the F1-score of the untuned model decreases by 15.7%, necessitating the activation of an incremental learning pipeline.

The current system has two limitations: first, knowledge updates are delayed by approximately 15 minutes, affecting time-sensitive tasks such as emergency pandemic medication guidelines; second, it lacks multimodal question-answering capabilities.

5 Conclusions

This paper proposes an intelligent question-answering model that deeply integrates natural language processing with knowledge graphs. By constructing a dual-stream semantic-knowledge collaborative architecture, a reinforcement learning-driven dynamic reasoning mechanism, and a lightweight domain-adaptive pipeline, the model significantly improves the accuracy, interpretability, and real-time response capabilities of complex question-answering tasks. The model innovatively achieves deep alignment between user intent and structured knowledge, overcoming key bottlenecks in traditional question-answering systems related to multi-hop reasoning and domain transfer. It provides a technical framework for intelligent service systems that combines theoretical foundations with practical feasibility, holding significant value for advancing the practical application of cognitive intelligence.

Declarations

The author declares that she has no conflicts of interest.

References

Alawida, M., Mejri, S., Mehmood, A., Chikhaoui, B. and Abiodun, O.I. (2023) ‘A comprehensive study of ChatGPT: advancements, limitations, and ethical considerations in natural language processing and cybersecurity’, *Information*, Vol. 14, No. 8, p.462.

Chen, Y., Li, H., Li, H., Liu, W., Wu, Y., Huang, Q. and Wan, S. (2022) ‘An overview of knowledge graph reasoning: key technologies and applications’, *Journal of Sensor and Actuator Networks*, Vol. 11, No. 4, p.78.

Chowdhary, K. (2020) ‘Natural language processing’, *Fundamentals of Artificial Intelligence*, Vol. 1, No. 4, pp.603–649.

Faye, B., Azzag, H., Lebbah, M. and Feng, F. (2025) 'Context normalization: a new approach for the stability and improvement of neural network performance', *Data & Knowledge Engineering*, Vol. 155, No. 4, p.102371.

Guo, L., Yan, F., Li, T., Yang, T. and Lu, Y. (2022) 'An automatic method for constructing machining process knowledge base from knowledge graph', *Robotics and Computer-Integrated Manufacturing*, Vol. 73, No. 1, p.102222.

Han, B., Susnjak, T. and Mathrani, A. (2024) 'Automating systematic literature reviews with retrieval-augmented generation: a comprehensive overview', *Applied Sciences*, Vol. 14, No. 19, p.9103.

Hao, X., Ji, Z., Li, X., Yin, L., Liu, L., Sun, M., Liu, Q. and Yang, R. (2021) 'Construction and application of a knowledge graph', *Remote Sensing*, Vol. 13, No. 13, p.2511.

Heyi, Z., Xin, W., Lifan, H., Zhao, L., Zirui, C. and Zhe, C. (2023) 'Research on question answering system on joint of knowledge graph and large language models', *Journal of Frontiers of Computer Science & Technology*, Vol. 17, No. 10, pp.1–12.

Hirschberg, J. and Manning, C.D. (2015) 'Advances in natural language processing', *Science*, Vol. 349, No. 6245, pp.261–266.

Hovy, D. and Prabhumoye, S. (2021) 'Five sources of bias in natural language processing', *Language and Linguistics Compass*, Vol. 15, No. 8, p.e12432.

Kosasih, E.E., Margaroli, F., Gelli, S., Aziz, A., Wildgoose, N. and Brintrup, A. (2024) 'Towards knowledge graph reasoning for supply chain risk management using graph neural networks', *International Journal of Production Research*, Vol. 62, No. 15, pp.5596–5612.

Lauriola, I., Lavelli, A. and Aiolli, F. (2022) 'An introduction to deep learning in natural language processing: models, techniques, and tools', *Neurocomputing*, Vol. 470, No. 1, pp.443–456.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W-t. and Rocktäschel, T. (2020) 'Retrieval-augmented generation for knowledge-intensive nlp tasks', *Advances in Neural Information Processing Systems*, Vol. 33, No. 1, pp.9459–9474.

Li, X., Jin, J., Zhou, Y., Zhang, Y., Zhang, P., Zhu, Y. and Dou, Z. (2025) 'From matching to generation: a survey on generative information retrieval', *ACM Transactions on Information Systems*, Vol. 43, No. 3, pp.1–62.

Liang, K., Meng, L., Liu, M., Liu, Y., Tu, W., Wang, S., Zhou, S., Liu, X., Sun, F. and He, K. (2024) 'A survey of knowledge graph reasoning on graph types: static, dynamic, and multi-modal', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 46, No. 12, pp.9456–9478.

Lin, Y. and Shen, H. (2017) 'SmartQ: a question and answer system for supplying high-quality and trustworthy answers', *IEEE Transactions on Big Data*, Vol. 4, No. 4, pp.600–613.

Nadkarni, P.M., Ohno-Machado, L. and Chapman, W.W. (2011) 'Natural language processing: an introduction', *Journal of the American Medical Informatics Association*, Vol. 18, No. 5, pp.544–551.

Rajabi, E. and Etminani, K. (2024) 'Knowledge-graph-based explainable AI: a systematic review', *Journal of Information Science*, Vol. 50, No. 4, pp.1019–1029.

Shen, J., Pan, T., Xu, M., Gan, D. and An, B. (2023) 'A novel DL-based algorithm integrating medical knowledge graph and doctor modeling for Q&A pair matching in OHP', *Information Processing & Management*, Vol. 60, No. 3, p.103322.

Tan, J., Qiu, Q., Guo, W. and Li, T. (2021) 'Research on the construction of a knowledge graph and knowledge reasoning model in the field of urban traffic', *Sustainability*, Vol. 13, No. 6, p.3191.

Wu, X., Duan, J., Pan, Y. and Li, M. (2023) 'Medical knowledge graph: data sources, construction, reasoning, and applications', *Big Data Mining and Analytics*, Vol. 6, No. 2, pp.201–217.

Yang, T., Mei, Y., Xu, L., Yu, H. and Chen, Y. (2024) 'Application of question answering systems for intelligent agriculture production and sustainable management: a review', *Resources, Conservation and Recycling*, Vol. 204, No. 1, p.107497.

Zhang, Q., Fang, C., Zheng, Y. et al. (2025) 'Improving deep assertion generation via fine-tuning retrieval-augmented pre-trained language models', *ACM Transactions on Software Engineering and Methodology*, Vol. 15, No. 1, pp.145–166.

Zhou, B., Shen, X., Lu, Y., Li, X., Hua, B., Liu, T. and Bao, J. (2023) 'Semantic-aware event link reasoning over industrial knowledge graph embedding time series data', *International Journal of Production Research*, Vol. 61, No. 12, pp.4117–4134.

Zhou, C., Li, Q., Li, C., Yu, J., Liu, Y., Wang, G., Zhang, K., Ji, C., Yan, Q. and He, L. (2024) 'A comprehensive survey on pretrained foundation models: a history from BERT to ChatGPT', *International Journal of Machine Learning and Cybernetics*, Vol. 2, No. 1, pp.1–65.

Zhu, X., Li, Z., Wang, X., Jiang, X., Sun, P., Wang, X., Xiao, Y. and Yuan, N.J. (2022) 'Multi-modal knowledge graph construction and application: a survey', *IEEE Transactions on Knowledge and Data Engineering*, Vol. 36, No. 2, pp.715–735.

Zhu, Y., Wang, X., Chen, J., Qiao, S., Ou, Y., Yao, Y., Deng, S., Chen, H. and Zhang, N. (2024) 'LLMs for knowledge graph construction and reasoning: recent capabilities and future opportunities', *World Wide Web*, Vol. 27, No. 5, p.58.