Development of Knowledge Graph Construction and Intelligent Question Answering System in Education Based on Transformer Model

Zhuxue Bai[1], Qi Mu[2]
[1]Shool of Marxism, Changchun Humanities and Sciences College, Changchun 130000, Jilin, China
[2]Human Resources Division, Jilin University, Changchun 130000, Jilin, China

Corresponding author: Zhuxue Bai (e-mail: baizhuxue@126.com)

ABSTRACT Efficient semantic information processing and multi-hop knowledge reasoning have become essential technologies for intelligent information services and next-generation networked systems. To address inaccurate semantic understanding caused by short or ambiguous queries and insufficient reasoning capability under fragmented knowledge structures, this study proposes an intelligent question answering framework that tightly integrates Transformer-based semantic encoding with graph attention reasoning. The proposed architecture employs DeBERTa-v3-base and conditional random fields for entity recognition, combines dual-tower vector retrieval with cross-encoder reranking for semantic disambiguation, and constructs k-hop knowledge subgraphs enhanced by multi-layer Graph Attention Networks to achieve relation-aware information propagation and evidence aggregation. A multi-task joint optimization strategy incorporating adaptive gradient normalization, adjacency reconstruction regularization, and negative sampling is further introduced to improve long-tail generalization and reasoning robustness. Neo4j graph storage and FAISS vector indexing enable near real-time retrieval and scalable deployment. Experimental evaluation demonstrates high semantic understanding accuracy, interpretable multi-hop reasoning capability, and stable performance under short and colloquial queries, with superior Exact Match and path reasoning accuracy compared with baseline models. Beyond educational applications, the proposed framework provides an effective methodology for semantic information fusion, distributed knowledge reasoning, intelligent query processing, and adaptive decision support, offering valuable references for communication-enabled information systems, networked knowledge services, and intelligent information infrastructures related to Electromagnetic Waves, Antennas and Propagation engineering applications.

INDEX TERMS semantic information fusion, graph attention networks, distributed knowledge reasoning, intelligent query processing

I. INTRODUCTION

With the popularization of digital teaching resources and online learning platforms, natural language-based question-and-answer services in educational scenarios have become an important tool for improving learning efficiency and personalized teaching [1]–[3]. Compared with general question answering, educational question answering places higher demands on the accuracy, interpretability, and consistency with the curriculum system of the answers [4], [5]. However, students' inquiries are often expressed in short phrases, colloquial language, or omissions of information, and the knowledge in textbooks and question banks exists in discrete and distributed fragments, causing question answering based on single text retrieval or pure generative models to frequently fail in multi-hop reasoning, evidence tracing, and teaching adaptation [6], [7]. Therefore, how to use structured teaching knowledge for reliable, multi-step reasoning and interpretable output while ensuring semantic understanding has become a key problem that needs to be solved in current educational intelligent question answering.

In recent years, significant progress has been made in the research of intelligent question answering systems in the field of education. Aurpa TT et al. constructed a question answering system based on reading comprehension in the Bengali language environment. By applying Transformer architectures including BERT (Bidirectional Encoder Representation from Transformers) and ELECTRA (Efficiently Learning an Encoder that Classifies Token Replacements Accurately), the accuracy of question answering was significantly improved. However, the system is still

limited by the limited size of the language dataset and has difficulty in dealing with the semantic capture problem of user phrases or fuzzy queries [8]. Nassiri K reviewed the Transformer-based text question answering system and analyzed the implementation and performance of encoders, decoders and encoder-decoder architectures. However, his method has limited ability to handle complex multi-hop reasoning and knowledge fragmentation [9]. Dartiko F et al. compared RoBERTa (Robustly Optimized BERT Pretraining Approach) and IndoBERT (Indonesian BERT) in a campus orientation question-and-answer system. Although IndoBERT performed well in EM and F1, it still had shortcomings in handling cross-knowledge point reasoning and information missing problems [10]. Ba Y enhanced multimodal question-and-answer capabilities by integrating Mask R-CNN (Mask Region-based Convolutional Neural Network) and Word2Vec+LSTM (Long Short-Term Memory), but the model still had limitations in text semantic understanding and structured knowledge fusion [11].

Zhu Y achieved intelligent adaptive learning driven by learning objectives based on the KRO (knowledge-resource-objective) model of knowledge graph and BiLSTM-CRF (Bidirectional Long Short-Term Memory-Conditional Random Field), but the fragmentation and automatic construction capabilities of the knowledge graph were limited, resulting in limited multi-hop reasoning performance [12]. Yu H et al. constructed an entrepreneurial education question-answering system based on knowledge graph semantic summarization, which improved retrieval and understanding capabilities through text embedding, similarity retrieval and graph convolution, but semantic capture under short fuzzy queries is still insufficient [13]. Xiao J proposed the Edu-VQA (Education Visual Question Answering) framework, which enhances reasoning through multimodal context selection and large language models, but the deep optimization of the fusion of structured knowledge and text is still insufficient [14]. Cao X improved the performance of complex question answering through joint reasoning of language models and knowledge graphs, but there is still a problem of insufficient generalization in the sparse knowledge graph scenario [15].

Although existing research has achieved results in improving the accuracy of question answering and multimodal understanding, it generally faces problems such as inaccurate semantic capture caused by short fuzzy queries from users, failure of multi-hop reasoning caused by knowledge fragmentation, and insufficient fusion of structured knowledge and text, which put forward room for improvement for further research. Currently, DeBERTa and graph neural networks have received widespread attention for their application in knowledge graph question answering and intelligent question answering in the education field. Lu W et al. proposed a natural language question semantic conversion model based on DeBERTa to address the problems of insufficient semantic understanding and inaccurate intent recognition in spatiotemporal queries of natural language. Through DeBERTa encoding and BiGRU (Bidirectional Gated Recurrent Unit) + CRF layers, they realized question intent recognition and query language conversion. On Chinese corpus, the F1 score reached 92.69%, which significantly improved the semantic information extraction accuracy [16]. Li S et al. proposed a latent relation multi-head self-attention joint extraction model based on DeBERTa to improve the accuracy of entity relation extraction in the knowledge graph of the manufacturing field. Through semantic representation, relation extraction and global entity pairing modules, the F1 score exceeded that of existing models [17].

Zhang G et al. proposed relation-aware GAT to address the problem of aimless reasoning in question answering of complex knowledge graphs. Through multi-layer relation-aware attention, they captured different relation path dependencies and improved the Hits@1 by 2.3% and 3.6% respectively on two types of datasets [18]. He P proposed a question-aware GCN (Graph Convolutional) model. The question answering method of the educational knowledge base (GCN) combines the question description, query entity and candidate answer subgraph with Transformer encoding and GCN node update, and outperforms the baseline model on the MOOC question answering dataset [19]. Fang Y et al. proposed the forward reasoning and dual-teacher knowledge distillation method of bilinear GNN (Graph Neural Network) for the false path reasoning problem in educational knowledge graph question answering, and verified the improvement of reasoning ability on the MOOC question answering dataset [20]. In summary, these studies have improved the performance of semantic understanding and graph reasoning through DeBERTa or graph neural networks, but there are still many research gaps in the intelligent question answering system in the field of education, which need further research. Moreover, the text semantic understanding and multi-step reasoning have not been well integrated, the joint optimization is insufficient, and the generalization ability for multi-hop reasoning and sparse long-tail knowledge is limited.

This paper focuses on the construction of an intelligent question answering system for educational scenarios. It designs an end-to-end joint reasoning scheme with DeBERTa-v3-base + GAT as the core, and makes the following contributions: (1) It constructs an end-to-end joint reasoning framework for educational scenarios, which tightly couples the fine-tuned DeBERTa-v3-base text representation with multi-layer multi-head GAT graph reasoning. It constructs a high-quality candidate pipeline for querying the graph through k-hop subgraph extraction, node text/type feature injection and cross reordering, and finally applies it to the system; (2) It designs a multi-task joint training and regularization strategy (including adaptive gradient normalization and adjacency matrix reconstruction constraints) to achieve collaborative optimization between entity recognition, relation classification and triple completion, and enhance the generalization ability for long-tail/sparse knowledge; (3) It realizes a complete verification process from engineer-

ing storage (Neo4j/FAISS) to interpretable evidence chain generation and human teacher review, taking into account interpretability and online retrieval availability, and provides a replicable research and engineering practice path for educational question answering and knowledge navigation scenarios.

II. DEVELOPMENT OF A QUESTION ANSWERING SYSTEM DRIVEN BY EDUCATIONAL KNOWLEDGE GRAPH

In the question-answering system integration phase, this paper constructs an end-to-end workflow, pipelinedly processing user input, knowledge graph retrieval, multi-hop graph reasoning, and answer generation. The specific process includes: first, DeBERTa-v3-base encoding and entity linking are performed on the user query; candidate entities and related subgraphs are retrieved from the knowledge graph; then, information propagation is performed on the subgraphs through multi-layer GAT, fusing entity nodes and relational features, and generating a multi-hop reasoning score for each candidate answer; finally, the optimal answer is selected through a ranking strategy, and interpretability annotation is applied to the answer based on the evidence chain path. A unified interface provides an entire workflow of data interactions between modules so that each stage's output can directly be used by the next stage in real-time to answer questions.

Robustness and efficiency are ensured through asynchronous parallel processing of the query parsing module, graph retrieval module, and inference module. The answer generation module produces the final answer with a corresponding evidence chain by combining the knowledge graph score with the contextually semantically matched output to produce the displayed final answer. Additionally, adaptive strategies are used to determine subjects, question lengths, and levels of colloquialism. By dynamically varying the number of candidate entities, inference layers and attention heads, this system has achieved high precision question answering and multi-hop inference capabilities in complex educational environments. Figure 1 shows the joint reasoning architecture diagram of DeBERTa-v3-base and GAT for educational question answering.

Figure 1 illustrates the joint reasoning architecture for educational question answering based on the DeBERTa-v3-base model and GAT proposed in this paper. The entire process starts with a user query, first passing through the text module to complete word segmentation, encoding, entity recognition, dual-tower vector retrieval, and cross-ranking to ensure accurate semantic capture of short texts and colloquial queries in educational scenarios. The ranked entity candidates are input to the graph module, where a k-hop educational knowledge subgraph highly relevant to the query is extracted through subgraph extraction. In the node feature assembly stage, the text representation and entity type information are combined. Subsequently, a multi-layer, multi-head GAT completes multi-hop information transmission and attention weighting based on relational structure to obtain structured reasoning features. The graph reasoning representation and text semantic representation are deeply integrated in the fusion layer, driving the reader to generate the final answer and evidence chain, and outputting an interpretable reasoning path.

A. EDUCATIONAL ONTOLOGY AND KNOWLEDGE GRAPH CONSTRUCTION

This study defines an entity set E and a relation set R in the educational corpus. Entities include types such as knowledge points (KPs), course chapters, concepts, questions, skills, and resources. Relations include prerequisite relations, coverage relations, examination relations, relevance relations, and resource relations. Data sources include textbook paragraphs, exercise banks, and teacher-student question-and-answer logs. The study uses a fine-tuned DeBERTa-v3-base sequence labeling model for entity extraction [21]. For the input text t_e, the token sequence x is first obtained by tokenizer encoding, and the hidden state sequence is output through DeBERTa, as shown in equation (1).

$$h_t = \mathrm{DeBERTa}(x)_t \tag{1}$$

In equation (1), h_t represents the hidden state sequence. Entity boundary prediction is accomplished using CRF, and the loss function is defined as shown in equation (2) [22]–[24].

$$\begin{aligned} \mathrm{Loss}_{NER} &= -\sum_{i=1}^{B} \log P_\theta \left(y^{(i)} \mid x^{(i)} \right) \\ &= -\sum_{i=1}^{B} \left(\sum_{t=1}^{T_i} C_{y_{t-1}^{(i)}, y_t^{(i)}} + h_t^{(i)} \cdot W_{y_t^{(i)}} \right. \\ &\qquad \left. + \log Z \left(x^{(i)} \right) \right) \end{aligned} \tag{2}$$

In equation (2), C represents the transition matrix; $W_{y_t^{(i)}}$ represents the projection matrix; $Z(x^{(i)})$ represents the CRF normalization factor; B represents the batch size; T_i represents the sequence length, and $y^{(i)}$ represents the label. Relation extraction employs entity pair representation and a multi-classification function. The contextual representation of entity pairs (e_i, e_j) in the text is obtained by DeBERTa sentence vector pooling $v_{ij} = [h_{start}^{(i)}; h_{end}^{(i)}; h_{start}^{(j)}; h_{end}^{(j)}]$, and the relation probabilities are output by Softmax, as shown in equation (3).

$$P(r_k | e_i, e_j) = \frac{\exp(v_{ij}^\top W_k)}{\sum_{k'=1}^{M} \exp(v_{ij}^\top W_{k'})} \tag{3}$$

In equation (3), W_k represents the relation weight vector. The training objective of the relation weight vector is the cross-entropy, as shown in equation (4).

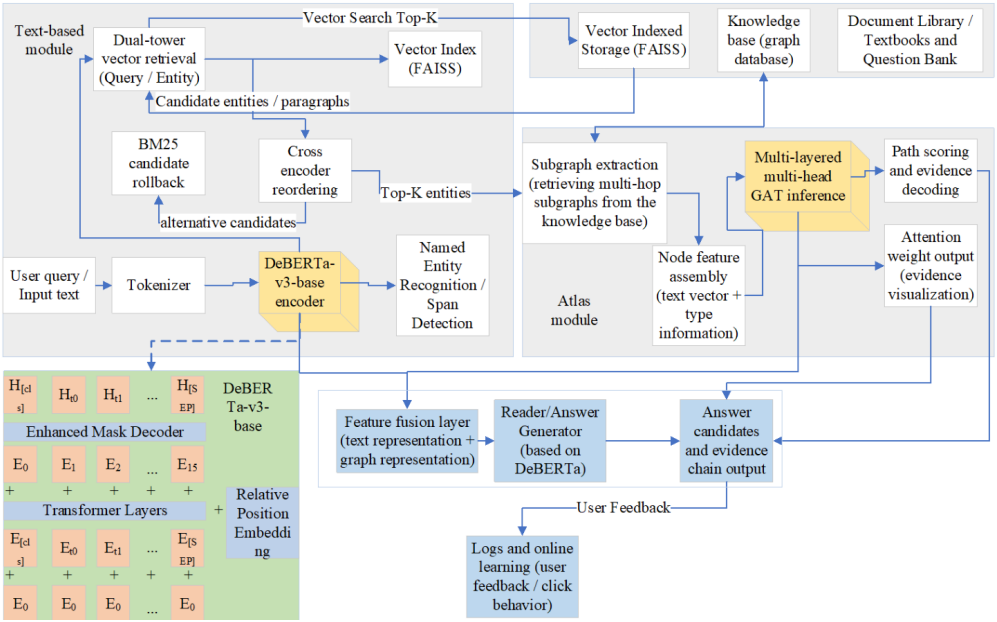

FIGURE 1. DeBERTa-v3-base and GAT's Joint Reasoning Architecture for Educational Question Answering

$$\text{Loss}_{REL} = -\sum_{k=1}^{M}\sum_{(i,j)} y_{ij}^{(k)} \log P(r_k|e_i, e_j) \tag{4}$$

After entity and relation extraction, triples are validated using rule constraints and synonym normalization to generate a preliminary knowledge graph $G = (E, R, TEG)$, where the graph $TEG = (e_i, r_k, e_j)$ is a set of triples. The knowledge graph is stored in the Neo4j graph database, and a vector index $\hat{E}$ is generated and stored in FAISS for subsequent fast candidate retrieval. A graph-based quantitative method is applied to measure the structural properties of the newly constructed educational knowledge graph to help evaluate how well it provides multi-hop reasoning and question-answering coverage. Additionally, this method can assist with understanding how well the different types of nodes are connected and accessible for evidence, and whether the current educational knowledge graph contains adequate structured evidence to support GAT. This method includes counting total nodes of each type, counting the total number of edges broken down by type of relationship, computing the average degree of the KPs, and determining what percentage of the KPs are assigned to questions.

Simultaneously, a 2-hop subgraph example is used to demonstrate the feasibility of transferring evidence from KPs to questions via pre-requisite chains. In Figure 1, the horizontal and vertical axes represent the drawing coordinates of nodes in the schematic layout (for visualization purposes only); the node shapes and colors correspond to node types (dark blue square = Course, blue square = Chapter, orange circle = KnowledgePoint, green rhombus = Question, gray triangle = Resource); the line type and color of the edges represent relationship types (orange solid line = pre_requisite, blue solid line = assessments, gray dashed line = Covered By, gray dotted line = Related To, green dotted line = Resource Of); the thick red line and bold nodes represent the example 2-hop reasoning subgraph (KP1→KP2→KP3→Q3), facilitating intuitive verification of the existence of chain evidence. A schematic diagram of the knowledge graph structure in the education field is shown in Figure 2.

As shown in Figure 2, the total number of nodes is 22 (Course 1, Chapter 3, Knowledge Point 8, Question 6, Resource 4); the total number of edges is 25, of which, by type, are: Covered By = 8, Pre-Requisite = 4, Assess = 6, Related To = 3, and Resource Of = 4. The overall average undirectedness is 2.27, while the knowledge point subset (8 nodes) is touched by edges a total of 32 times, with an average undirectedness of 4.0. 75% of the KPs are directly covered (Assessed) by questions, indicating that most KPs can be directly evaluated using the existing question bank; the remaining 25% (KP4, KP6) are not directly covered by questions, suggesting the existence of long-tail KPs that require expansion of the question mapping. The chain formed by the Pre-Requisite edges (KP1→KP2→KP3→KP4→KP5→KP6) is continuously visible in the graph, enabling access from early KPs to nodes directly supported by questions or resources via 2-hop/3-hop (KP1 can indirectly reach Q3 through KP2 and KP3). This demonstrates the basic feasibility of the graph for multi-hop reasoning. In this example, at least three KPs (KP1, KP2, KP3) form dense sub-chains with an average degree of 5, indicating that nodes acting as "propagation chains" in multi-hop reasoning have higher centrality, making them suitable for GAT to transmit evidence in multi-level aggregation. In summary, the graph currently possesses

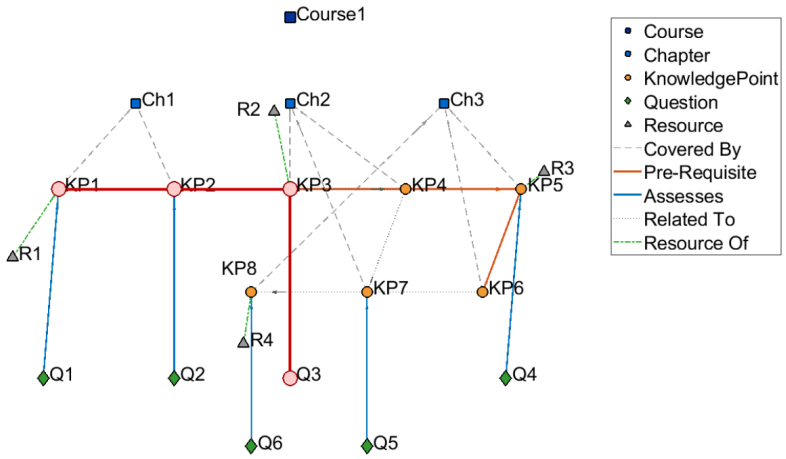

FIGURE 2. Schematic Diagram of Knowledge Graph Structure in the Field of Education

a structural framework for multi-hop reasoning (average KP degree 4.0, 75% of KPs are covered by questions, and there are coherent pre-condition chains), but questions and resources still need to be added for the uncovered KPs (25%) to improve question-answering coverage and long-tail generalization ability.

To perform a quantitative assessment of the near real-time retrieval capability of the end-to-end response time for the core retrieval system and to validate this. The system achieves a balance between computational complexity and retrieval latencies to accomplish this through multiple methods of optimization including: preconstructing a FAISS-based vector index for the entity retrieval; performing asynchronised parallel processing of both the query parsing and graph reasoning modules; and dynamically adjusting the GAT layers and attention heads to retrieve trivial queries. The various optimisation approaches applied throughout the system's integrated pipeline of deep semantic encoding and graph reasoning will allow for millisecond-level responses by the query system, thereby enabling the query system to meet its latency targets of executing near real-time retrieval within the context of real-world educational applications.

B. DEBERTA-V3-BASE TEXT SEMANTIC ENCODING AND ENTITY LINKING

The DeBERTa-v3-base model is a third-generation base version of the DeBERTa series based on the transformer architecture [25]. It achieves the separation modeling of semantic vectors and positional information by applying a decoupled attention mechanism and enhanced relative positional encoding, enabling the model to capture the dependencies between words more accurately [26]. For user query q, it is first encoded as $h_q = \mathrm{DeBERTa}(q)_{[CLS]}$ using DeBERTa-v3-base. For candidate entity vector, entity retrieval uses cosine similarity, as shown in equation (5).

$$\mathrm{sim}(\hat{e}_i|q) = \cos(h_q, \hat{e}_i) = \frac{h_q \cdot \hat{e}_i}{\|h_q\|\|\hat{e}_i\|} \tag{5}$$

In equation (5), $\mathrm{sim}(\hat{e}_i|q)$ represents the cosine similarity. Then, the top-K candidate E_q are selected for cross-ordering. The reordering uses a cross-encoder DeBERTa with input $[q; e_i]$ and concatenated representation z_i, outputting entity link probabilities, as shown in equation (6).

$$P(e_i|q) = \sigma(W_r^\top z_i + b_r) \tag{6}$$

In equation (6), b_r represents the bias; $W_r^\top$ represents the weight matrix; and σ represents the sigmoid function. The reordering training uses binary cross-entropy, as shown in equation (7).

$$\mathrm{Loss}_{EL} = -\sum_{i=1}^{K} [y_i \log P(e_i|q) + (1 - y_i) \log(1 - P(e_i|q))] \tag{7}$$

In equation (7), y_i indicates whether the entity is correct. After the entity links are completed, a subgraph G_q is constructed using candidate entities. The node feature h_{e_i} of the subgraph $h_{e_i} = [e_i; t_i]$ is obtained by concatenating the entity vector e_i with the type embedding t_i.

C. RELATIONSHIP IDENTIFICATION AND TRIPLE COMPLETION

After entity extraction and linking, this study performs relation recognition on entity pairs (e_i, e_j). First, the context representation of each entity pair in the text is encoded. Given the start and end token indices (s_i, e_i) of entity e_i in the text t_e, it is represented as a pooled vector, as shown in equation (8).

$$v_i = \frac{1}{e_i - s_i + 1} \sum_{t=s_i}^{e_i} h_t \tag{8}$$

In equation (8), v_i represents the vector of entity e_i after pooling. Similarly, the representation v_j of entity e_j is obtained. The entity pair representation is obtained by concatenation and linear mapping to obtain the relation feature vector, as shown in equation (9).

$$v_{ij} = \mathrm{ReLU}(W_{re}[v_i; v_j; v_i \odot v_j] + b_{re}) \tag{9}$$

In equation (9), $\odot$ represents element-level product. The relationship classification probability is output through Softmax, as shown in equation (10).

$$P(r_k|e_i, e_j) = \frac{\exp(u_k^\top v_{ij})}{\sum_{k'=1}^{M} \exp(u_{k'}^\top v_{ij})}, \quad k = 1, \ldots, M \tag{10}$$

In equation (10), M is the total number of relation types and $u_k^\top$ is the relation weight vector. The training loss for relation recognition is cross-entropy, as shown in equation (11).

$$\mathrm{Loss}_{REL} = -\sum_{k=1}^{M} \sum_{(i,j)} y_{ij}^{(k)} \log P(r_k|e_i, e_j) \tag{11}$$

In equation (11), $y_{ij}^{(k)}$ is the label of the entity pair (e_i, e_j) corresponding to the relation r_k. To complete the triplet completion, this study applies graph-end prediction. An adjacency matrix is constructed to represent the current knowledge graph structure, where a relation exists $A_{ij} = 1$, if a relationship exists between known entity pairs (e_i, e_j), and 0 otherwise. Edge representation learning and a bilinear scoring function are used to predict latent relations, as shown in equation (12).

$$\hat{A}_{ij}^{(k)} = \sigma(e_i^\top W_k e_j) \tag{12}$$

In equation (12), W_k is the relation-specific parameter matrix, and the output value $\hat{A}_{ij}^{(k)}$ represents the probability of the existence of the triple. The completion loss adopts the negative sampling binary crossentropy, as shown in equation (13).

$$\mathrm{Loss}_{GRAPH} = -\sum_{(i,j,k)\in TE^+} \log \hat{A}_{ij}^{(k)} - \sum_{(i,j,k)\in TE^-} \log(1-\hat{A}_{ij}^{(k)}) \tag{13}$$

In equation (13), TE^+ is the set of positive triples and TE^- is the set of negative sampled triples.

D. GRAPH FEATURE INJECTION AND GAT INFERENCE

To enhance multi-hop reasoning capabilities, text and type features are injected into nodes in the subgraph. Each node is represented $h_{e_i}^{(0)} = [e_i; t_i; v_i]$. As a type of graph neural network, GAT is used in this paper to update node representations, as shown in equation (14) [27], [28].

$$h_{e_i}^{(l+1)} = \Big\Vert_{d=1}^{H} \sigma\left(\sum_{j\in N(i)} \alpha_{ij}^{(d)} W^{(d)} h_{e_j}^{(l)}\right), \quad l = 0, \ldots, L-1 \tag{14}$$

In equation (14), H is the number of attention heads; $W^{(d)}$ is the linear transformation of the d-th head; and $N(i)$ is the set of neighbors of node i. The attention coefficient is defined as shown in equation (15) [29]–[31].

$$\alpha_{ij}^{(d)} = \frac{\exp(\mathrm{LeakyReLU}(a^{(d)\top}[W^{(d)}h_{e_i}^{(l)} \| W^{(d)}h_{e_j}^{(l)}]))}{\sum_{k\in N(i)} \exp(\mathrm{LeakyReLU}(a^{(d)\top}[W^{(d)}h_{e_i}^{(l)} \| W^{(d)}h_{e_k}^{(l)}]))} \tag{15}$$

In equation (15), $a^{(d)}$ represents the attention vector. Through L-layer iterative updates, the node representation $h_{e_i}^{(L)}$ simultaneously encodes multi-hop graph structural information and textual semantic information. During the inference phase, the updated node representation is used to calculate the entity pair relevance score, as shown in equation (16).

$$\mathrm{sco}(e_i, e_j) = h_{e_i}^{(L)\top} h_{e_j}^{(L)} \tag{16}$$

In equation (16), $\mathrm{sco}(e_i, e_j)$ represents the entity pair relevance score.

E. JOINT TRAINING AND MULTI-TASK LOSS

To simultaneously optimize text semantic understanding and knowledge graph structure reasoning, this paper designs a joint training framework, taking entity recognition, relation recognition, triple completion, and entity linking as multi-task joint training objectives. The overall loss function is defined as shown in equation (17).

$$\begin{aligned}\mathrm{Loss}_{total} = \lambda_1 \mathrm{Loss}_{NER} + \lambda_2 \mathrm{Loss}_{REL} \\ + \lambda_3 \mathrm{Loss}_{GRAPH} + \lambda_4 \mathrm{Loss}_{EL}\end{aligned} \tag{17}$$

In equation (17), Loss_{total} represents the overall loss function. This paper dynamically weights the loss of each task and uses the adaptive gradient normalization (GradNorm) method to adjust λ, so that the gradient contributions of each task are balanced and the joint optimization is not overly dominated by any one task. In order to further enhance the multi-hop reasoning capability, an adjacency matrix reconstruction regularization is added to the graph reasoning part, as shown in equation (18).

$$\mathrm{Loss}_{adj} = \sum_{i,j} \|\hat{A}_{ij} - A_{ij}\|_F^2 \tag{18}$$

In equation (18), $\hat{A}_{ij}$ represents the adjacency prediction updated by GAT, and A_{ij} represents the actual adjacency. The final joint loss is shown in equation (19).

$$\mathrm{Loss}_{joint} = \mathrm{Loss}_{total} + \beta\, \mathrm{Loss}_{adj} \tag{19}$$

In equation (19), β represents the regular weights of the graph structure. During training, the optimizer used was AdamW [32]–[34]. The learning rate was scheduled using cosine annealing [35], [36]. This training strategy ensures stable convergence of the model in a multi-task environment, taking into account the performance of text understanding, entity recognition, relation reasoning, and triple completion.

III. PERFORMANCE VERIFICATION SCHEME FOR INTELLIGENT QUESTION ANSWERING SYSTEM

A. EXPERIMENTAL DATA AND PREPROCESSING

The datasets upon which to conduct these experiments include several unique corpora: First is a collection of 15,000 publicly available high school mathematics and physics textbook paragraphs taken from chapters published by both People's Education Press and Higher Education Press; Second is an exercise-and-answer pair collection consisting of 40,000 publicly available exercise-and-answer pairs taken from the university question database of both Universities' question banks and previous examinations; Third is a set of 12,000 unique question-answer pairs from teachers' and students' question/answer logs that have been posted on MOOC platforms and located on the discussion threads of selected Universities' college-classroom discussion threads; Finally, the expert group assembled reference datasets to create a gold standard through manual annotation of 5,000 examples of entity-relations within the textbooks and Q&A datasets and through manual annotation of 3,000 complex Q&A examples with multi-hop reasoning paths.

The initial knowledge graph constructed based upon the aforementioned corpora contains approximately 22,000 entities and 55,000 triples. The scale uses a small sample of highly annotated graphs previously determined to be experimentally verified, with controllable annotations, for use in a wide range of core knowledge across many disciplines. It focuses specifically on high school core knowledge systems (English, Chinese, math, and physics) by removing marginal, extended, and reproducing content-related information so as to allow for both the logical rationality of knowledge graphs used in our experiment, and their reasoning performance. All data used in the experiment were split up based on "standard experimental" conditions, with 80% of the data included in the training dataset, 10% in the validation dataset, and 10% in the testing dataset. The test dataset included previously unseen entities and relationships to test for long-tail generalization performance.

The preprocessing procedures are made up of several cleaning methods and standardization methods to ensure consistency among datasets for training and evaluation purposes. The first method is removing duplicates/anonimizing the original text; then the data undergoes character normalization and standardization of punctuation. This is followed by sentence segmentation (to obtain stable sentence units), word segmentation (to obtain stable word units), formatting of questions and answers, and splitting multi-turn dialogues into separate question-answer pairs for every turn of the conversation. After that, proofed labelled data is reviewed for consistency in the label; the labelling scheme is block-based, and labelling boundaries are aligned with word segmentation results through character alignment. Entity standardization is achieved by constructing an alias table and a thesaurus. Candidate normalized mappings are generated based on fuzzy string matching and semantic vector retrieval, and manually sampled for verification. A negative sample set is generated for relation and graph training to balance positive and negative distributions, and samples are stratified by subject and entity frequency to ensure comparability of the training, validation, and test sets in terms of subject distribution and entity frequency. All preprocessing results are output as normalized text files, labeled files, and vector index files for direct use by downstream model training and retrieval modules.

B. EXPERIMENTAL DESIGN AND PARAMETER SETTING

To verify the advantages of the DeBERTa-v3-base + GAT model in educational question answering and multi-hop reasoning tasks, the experiment adopted a three-part split (80/10/10) for training/validation/testing, and additionally reserved a subset of unseen entities/unseen relations in the test set for long-tail generalization evaluation. The control groups included: DeBERTa-v3-base + GraphSAGE (Graph Sample and aggregate), DeBERTa-v3-base model, RoBERTa-base, BERT-wwm-ext, BiLSTM-CRF (a traditional baseline for sequence labeling), Word2Vec + LSTM, and BM25 (Best Matching 25). All baselines were trained and evaluated on the same corpus and under the same training/testing split, and the same retrieval candidate pool and evaluation script were used to ensure comparability. The evaluation process includes: candidate retrieval $\rightarrow$ entity linking $\rightarrow$ subgraph extraction $\rightarrow$ inference/generation $\rightarrow$ evaluation metric calculation; for each model, metrics such as question answering EM, question answering F1, precision, recall, path-acc, evidence coverage, and total latency are recorded.

Path-Acc quantifies the level of consistency between the model's reasoning path and the gold standard path created manually. It indicates whether the model uses logically rigorous and educationally compliant paths of knowledge associations to derive an answer. This is important for educational QA. Educational knowledge reasoning is distinctively structured, relying on the hierarchical and associative relationships of knowledge points. Thirdly, unlike simple answer accuracy indicators that focus on just one answer, Path-Acc can also track the model's reasoning logic, which is especially critical in educational scenarios that require an interpretable answer and present an evidence chain to support the answer, and is congruent with the overall system design goal of generating interpretable evidence chains for learning tutoring. Finally, Path-Acc has hierarchical discriminability for difficulty in multi-hop reasoning, with a gradual decline in performance as the number of hops increases. The test set was grouped according to each question type's

TABLE 1. Parameter Settings

Parameter Categories	Value	Illustrate
Text Learning Rate	2.00E-05	AdamW Optimizer
Text Batch	16	Batch Size per Card (with gradient accumulation)
Text Training Elections	3–5	Early Stopping Based on Validation Set
Maximum Sequence Length	512	Text Truncation/Padded Length
Text Dropout	0.1	Transformer Layer Dropout
GAT Layer	2	Number of Graph Network Layers
Number of GAT Heads	8	Number of Multi-Head Attention Heads
GAT Hidden Dimension	256	Total Representation Dimension After Single Head or Concatenation
GAT Learning Rate	1.00E-03	Individual Optimizer or Parameter Set
GAT Dropout	0.1	Layer Dropout
Loss Weights	2.0 / 1.0 / 1.0 / 0.5	Initial Weights, Dynamically Adjustable During Training Using GradNorm
Adjacency Reconstruction Weights	0.1	Adjacency Matrix Reconstruction Regularization Strength
Negative Samples	5	Number of Negative/Positive Samples for Triple Patch Completion
Subgraph Hops	2	Subgraph Expansion Depth
Search Top-K	20	Number of Vector/Text Candidates
Random Seed	42	Reproducible Experiment Results

length, use of colloquialisms, and the general topic of each question in order to measure the model's stability between each of the different distributions of input used for testing. In addition, 300 random samples were assessed by 3 educational professionals on the relevance, adequacy, and accuracy (1–5 point scale) of the model's evidence chain in order to validate multi-hop inference and accountable. The overall latency of the system for both batch and concurrent queries was also analyzed. Each experiment was conducted 30 times in a consistent hardware and software environment, and the average and standard deviation were provided for comparison. Table 1 displays parameters.

IV. MULTIDIMENSIONAL ASSESSMENT OF EDUCATIONAL QUESTION-ANSWERING SYSTEM

A. OVERALL QUESTION-ANSWERING PERFORMANCE ANALYSIS

In the following segment, we will look at the overall performance of varying models within the context of answering questions from knowledge graphs in the field of education based on how well the models measure up on accuracy, completeness, and precision. To evaluate the performance of each of the models, four different metrics (EM, F1, Precision and Recall) have been collected. The overall performance of each of the four metrics for the question-answering models is displayed in Figure 3. The figure depicts the results of 30 repeated experiments on all four metrics across the three models, using a violin plot to represent the probability density and depicting the mean and standard deviation of the results as displayed by the black dots and black lines as indicated on the graphic.

As shown in Figures 3(a) to 3(d), DeBERTa-v3-base+GAT performs best across all four metrics: EM = 0.934 ± 0.018, F1 = 0.949 ± 0.023, Precision = 0.955 ± 0.022, and Recall = 0.940 ± 0.020, demonstrating its superior accuracy and coverage. DeBERTa-v3-base+GraphSAGE and DeBERTa-v3-base models follow closely behind, with EM values of 0.900 and 0.889, and F1 values of 0.926 and 0.909, respectively. RoBERTa-base and BERT-wwm-ext perform relatively well, but slightly worse than the DeBERTa series. The traditional BiLSTM-CRF and Word2Vec+LSTM models exhibit moderate overall performance, with EM values of 0.790 and 0.750, and F1 values of 0.840 and 0.812, respectively. BM25 performed the worst as the baseline model, with an EM score of only 0.604 and an F1 score of only 0.698, demonstrating the significant advantage of deep learning-based Transformer models in question-answering tasks. Differences between metrics also show that Precision is generally slightly higher than Recall, indicating that the model's accuracy in predicting answers is slightly better than its coverage. The performance differences mentioned above are mainly due to model structure and feature representation capabilities. DeBERTa-v3-base+GAT, by combining Transformer encoding with graph structure information from graph neural networks, can better capture relationships between entities and contextual semantics, improving question-answering accuracy and coverage.

B. ASSESSMENT OF MULTI-HOP REASONING ABILITY

To evaluate the multi-hop reasoning ability of different models in knowledge graph question answering in the education field, this section uses two metrics: Path-Acc (path accuracy) and Evidence Coverage (evidence coverage), which measure whether the model correctly identifies the reasoning path and the degree of evidence coverage, respectively. The multi-hop reasoning ability evaluation is shown in Figure 4. Figure 4 shows the performance of each model within a three-hop range: the horizontal axis represents the number of hops, and the vertical axis represents the mean of the two metrics; each scatter point represents a model, the error bars represent the standard deviation, and the horizontal shift is used to distinguish different models with the same number of hops.

As shown in Figure 4(a) Path-Acc, DeBERTa-v3-base+GAT performs best across all hop counts: 0.960 ± 0.008 for 1-hop, 0.945 ± 0.010 for 2-hop, and 0.913 ± 0.013 for 3-hop. In contrast, the traditional statistical method BM25 performs worst, with a 3-hop count of only 0.555 ± 0.022. The DeBERTa-v3-base+GraphSAGE and DeBERTa-v3-base models also outperform other Transformer and sequence models, but are slightly lower than the GAT combination. Figure 4(b) shows that the trend of Evidence Coverage is similar to that of Path-Acc. The coverage rates of DeBERTa-v3-base+GAT in 1-hop, 2-hop, and 3-hop are 0.948 ± 0.009, 0.928 ± 0.011, and 0.895 ± 0.014, respectively,

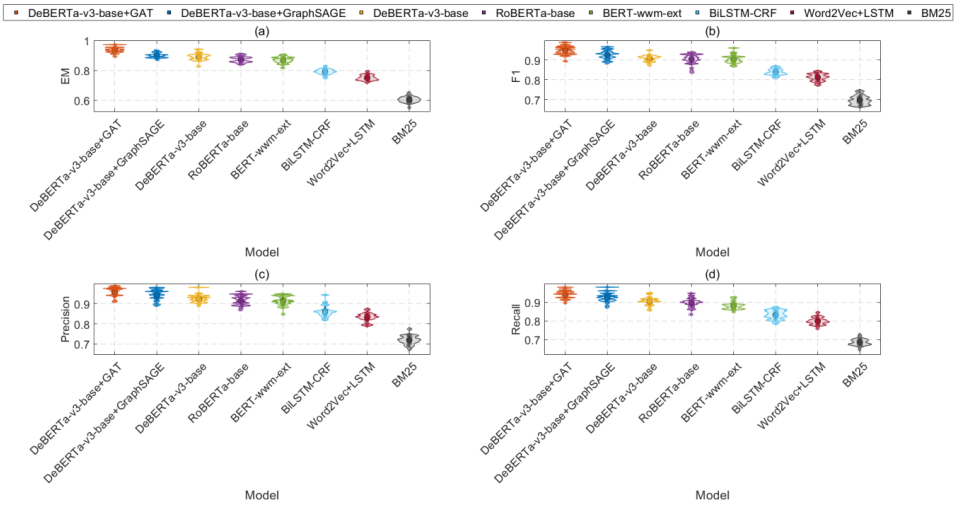

FIGURE 3. Overall Question Answering Performance Analysis

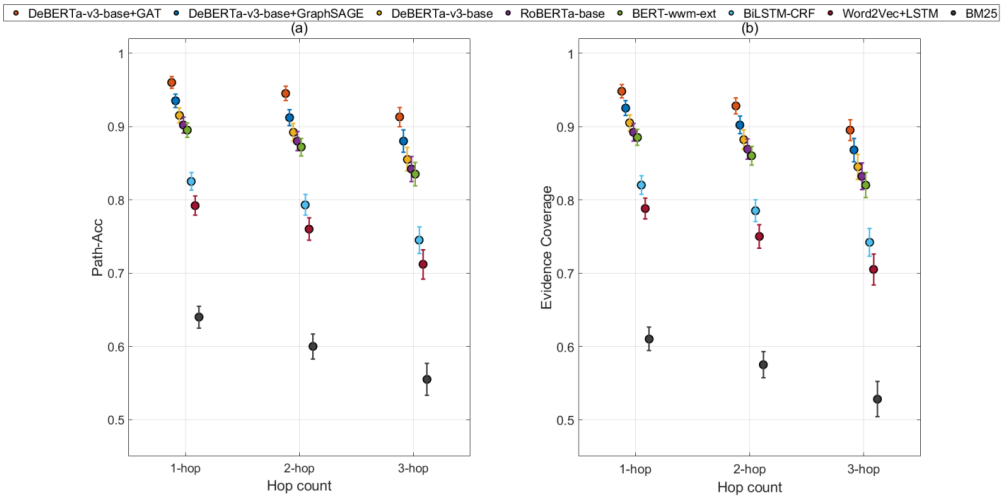

FIGURE 4. Assessment of Multi-hop Reasoning Ability

while BM25 is only 0.528 ± 0.024. This indicates that the combination of deep Transformer and graph neural network can significantly improve evidence coverage in multi-hop inference.

Overall, as the number of hops increases, the Path-Acc and Evidence Coverage of all models show a decreasing trend, but the decrease of DeBERTa-v3-base+GAT is the smallest, indicating that it still maintains high stability in long-chain inference. These performance differences primarily stem from variations in the models' capabilities in capturing structured knowledge information and modeling relationships. DeBERTa-v3-base+GAT adopts a graph attention mechanism. It effectively models the interactions and relationships between knowledge nodes. This ensures accurate path selection in multi-hop inference. It also guarantees the coverage of key evidence nodes. Although GraphSAGE can also be utilized to obtain information of the graph structure, the aggregation of node information is slightly less effective due to its longer chain propagation, which has resulted in lower Path-Acc and lower Evidence Coverage. The standard Transformer or BiLSTM models do not model their graph structure explicitly and therefore lose many of their intermediate node information as the number of hops increases during the process of multi-hop inferences, causing a significant degradation of overall performance. Additionally, BM25, a text-based matching algorithm, is not intended for logical reasoning across multiple hops, resulting in its worst performance in tasks that require three or more hops to complete.

Based on these results, the Transformer models perform better at providing answers to questions from educational knowledge graphs using multi-hop inference than other methods. Furthermore, these results demonstrate that the length of an inference chain can affect the performance of the models. To visualize the different ways that their model either successfully or unsuccessfully completed the

multistep reasoning task and which key nodes played a role in this success, the analysis created a subgraph of the knowledge graph, along with heat encoding of the nodes. This visualization of successful and unsuccessful paths is provided in figure 5. Figure 5 displays heat values (between 0 and 1) for color and size of each node, where red indicates low confidence and yellow/white indicates high confidence for that model's evidence or level of confidence in that entity node. Gray represents the edges of each node and indicates a knowledge relationship, the type of relationship between nodes is indicated in the middle of each edge. Solid green arrows will indicate typical successful paths, while dotted light red arrows will indicate typical failed paths. The start and finish for the paths will be highlighted with differing shapes to illustrate where important connections exist.

The successful path (Quadratic Eq $\rightarrow$ Discriminant $\rightarrow$ Solution Step 2 $\rightarrow$ Example Problem) as illustrated in Figure 5 has substantially higher heatmap scores than the equivalent nodes within the failed path. For example, Quadratic Eq = 0.98, Discriminant = 0.90. In comparison, the paths associated with Completing Square and Solution Step 1 have much lower heatmap scores (Completing Square = 0.88, Solution Step 1 = 0.76), indicated by dashed lines. This indicates that the model's inference deteriorates or has less confidence through Completing Square and Solution Step 1 associated with Example Problem compared to the successful path. This visualization also demonstrates that the model includes fewer higher-order dependency nodes in stages where knowledge conveys across multiple steps/transformations.

C. ROBUSTNESS UNDER DIFFERENT QUESTION LENGTHS/COLLOQUIALISM LEVELS

In this section, we looked at how robust models are given different complexity of inputs from two angles: the length of the question (Short/Medium/Long) and the amount of colloquialism (Low/Medium/High). In general, the lengths of question indicate to us how adaptable a model will be to semantic span and information density. In contrast, looking at the levels of colloquialism allows us to understand how stable the models are in relation to non-standard expression. The results from the eight models were compared using EM and F1 as performance metrics across various lengths and levels of colloquialism. The results from the robustness comparisons between the different lengths and colloquialism levels are contained in Figure 6. Figures 6(a) and (b) provide detail on the various length group comparisons, whereas Figures 6(c) and (d) provide detail on the various levels of colloquialism comparisons. Each comparison consists of eight model bars, as well as the corresponding standard deviation for each bar, which reflects the model's sensitivity to input type.

As can be observed from Figures 6(a) and 6(b), the EM and F1 scores of all models decrease as the problem length increases from Short to Medium to Long, but the performance gradients differ significantly. DeBERTa-v3-base+GAT consistently maintains the highest performance, achieving EM scores of 0.96 ± 0.018, 0.94 ± 0.020, and 0.90 ± 0.014 respectively, and F1 scores of 0.97 ± 0.017, 0.95 ± 0.013, and 0.92 ± 0.014, demonstrating the smallest performance degradation with increasing problem length. DeBERTa-v3-base+GraphSAGE follows closely (EM $0.94 \pm 0.019 \rightarrow 0.88 \pm 0.019$; F1 $0.95 \pm 0.018 \rightarrow 0.90 \pm 0.014$). Colloquialism has a negative effect on the performance of the models shown in Figures 6(c) and 6(d). Models with low to high levels of colloquialism decrease in performance from low, to medium and then finally high levels of colloquialism. As for the GAT model (Overall score) this decrease from low to high colloquialism is as follows: EM: $0.95 \pm 0.017 \rightarrow 0.90 \pm 0.019$; F1: $0.96 \pm 0.018 \rightarrow 0.91 \pm 0.014$. On the other hand BM25 which is purely statistical has seen the largest drops in performance due to the presence of colloquialisms, specifically high colloquial expressions: EM: $0.60 \pm 0.015 \rightarrow 0.52 \pm 0.013$. As noted earlier, at medium to high levels of colloquialism RoBERTa-base and BERT-wwm-ext had much more consistent performances than models based on word segmentation or vectorized words. Finally, the model structure (GAT) combined with a high quality pre-trained model, performed better than all models tested on the robustness of each of these models using two different tests.

D. PERFORMANCE BY ENTITY TYPE AND DISCIPLINE DISTRIBUTION

This section of research focuses on determining the performance of different entities in a variety of disciplines from an educational perspective. Specifically, this study will look at the prediction ability and recognition ability of a variety of entities from an educational knowledge graph perspective and determine if the performance of these models differ based on the type of knowledge structure present. In this study, the metrics that have been used to compare performance across entities and disciplines are F1 score and Path-Acc. The F1 score reflects the accuracy and recall for the classification of entities, while Path-Acc reflects the correctness of the prediction of path relationships between entities. Figure 7 represents the performance of entities by type across the distribution of disciplines. Figure 7 is presented in a grouped radar chart, and each radar chart represents the performance of a particular discipline. There are five axes on each radar chart, which represent the five different types of entities (Knowledge Point, Course, Chap, Question, Resource). The radial grid scale in the figure is [0.5, 0.7, 0.9], used to help observe the relative magnitudes of each metric.

As can be seen from Figure 7, there are significant differences in the performance of each discipline across different entity types. Figure 7(a) In mathematics, Question has the highest F1 score (0.96) and Path-Acc score (0.93) among all indicators, while Resource has the lowest (F1=0.78, Path-Acc=0.70). Figure 7(b) In physics, Question also has the highest score (F1=0.88, Path-Acc=0.85), while Resource

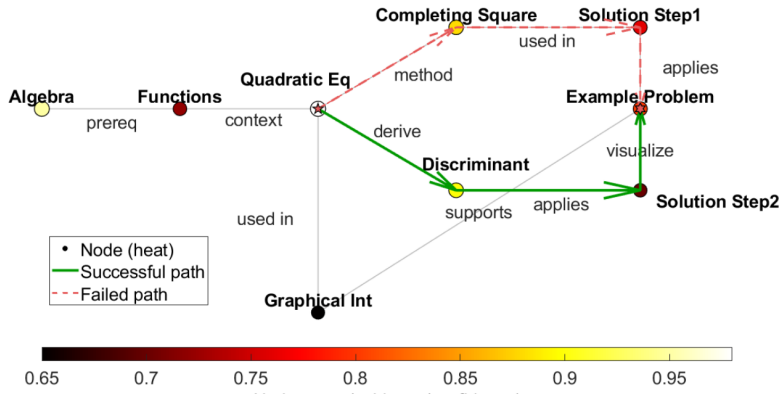

FIGURE 5. Visual Representation of Successful and Failed Paths in a Knowledge Graph Subgraph

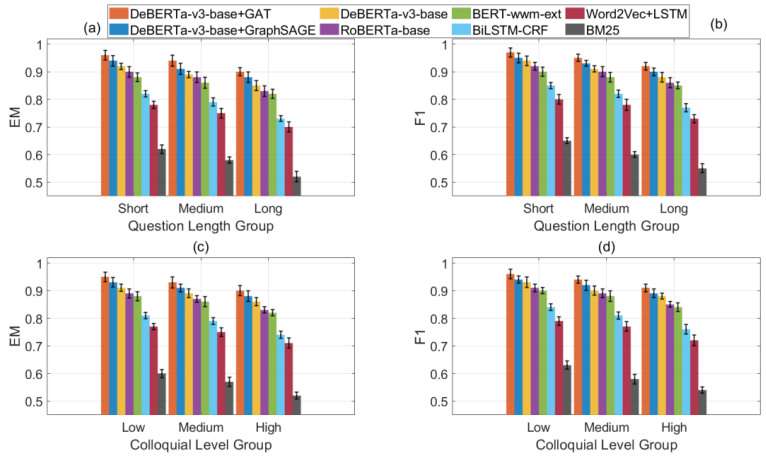

FIGURE 6. Robustness under different question lengths/colloquialism levels

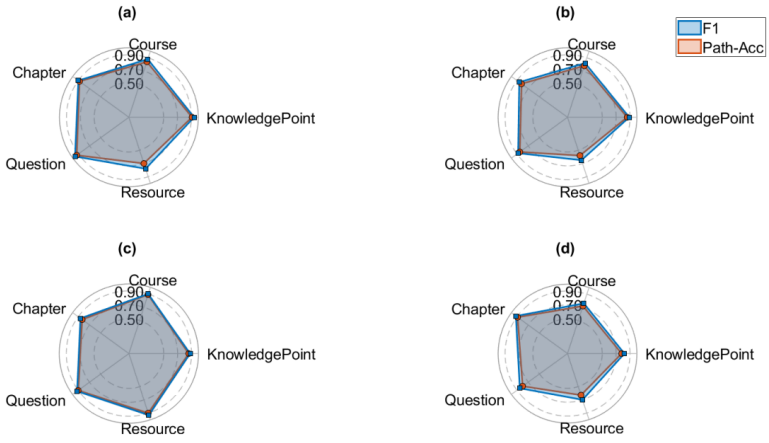

FIGURE 7. Performance by Entity Type and Subject Distribution

Figure 7(a) Performance of Diferent Entty Types under the Discipline of Mathematcs; Figure 7(b) Performance of Diferent Entty Types under the Discipline of Physics; Figure 7(c) Performance of Diferent Entty Types under the English Subject; Figure 7(d) Performance of Diferent Entty Types under the Chinese Subject

has the lowest (F1=0.65, Path-Acc=0.58). Figure 7(c) In English, the performance is relatively balanced, with KnowledgePoint F1=0.88, Path-Acc=0.86, and Resource F1=0.93, Path-Acc=0.90, indicating that the model is relatively stable in English entity recognition and path prediction. Figure 7(d) In language arts, Chapter has the most outstanding F1 score (0.92) and Path-Acc (0.89), while Course and Resource have relatively low performance (Course F1=0.76, Path-Acc=0.72, Resource F1=0.70, Path-Acc=0.63). Overall, Question and Chapter have high predictive performance in most subjects, while Resource and Course have inconsistent predictive performance.

TABLE 2. Results of Operational Efficiency and Delay Analysis

Model	Retrieval Time (ms ± std)	Total Latency (ms ± std)	Throughput (GPUs) (qps)	Throughput (CPUs) (qps)
DeBERTa-v3-base + GAT	50.8 ± 4.0	120.5 ± 7.2	200 ± 12	48 ± 5
DeBERTa-v3-base + GraphSAGE	60.5 ± 5.2	140.3 ± 9.1	170 ± 15	40 ± 6
DeBERTa-v3-base	48.1 ± 3.6	115.2 ± 6.5	210 ± 10	50 ± 4
RoBERTa-base	52.3 ± 4.1	128.7 ± 7.8	190 ± 12	45 ± 5
BERT-wwm-ext	55.7 ± 4.8	132.4 ± 8.0	185 ± 13	43 ± 5
BiLSTM-CRF	35.2 ± 3.1	95.6 ± 5.4	250 ± 15	60 ± 6
Word2Vec + LSTM	22.5 ± 2.4	80.3 ± 4.2	300 ± 20	75 ± 8
BM25	12.1 ± 1.7	45.7 ± 3.0	500 ± 25	120 ± 10

E. OPERATIONAL EFFICIENCY AND DELAY ANALYSIS

Key metrics to evaluate the practicality of knowledge graph question-answering systems for education include the model's efficiency of operation and the speed of response to the end user. In this section, I will compare how well pretrained transformers combined with graph networks perform on queries compared to pure pretrained transformers, deep learning methods, and BM25. I will summarize those results of the efficiency of operation and response times in Table 2. Table 2 gives the efficiency of operation results. The average time and fluctuation for one query is reported for the average Retrieval Time, the total amount of time it takes for the responder to complete a specific request (Total Latency Time), and the throughput, or concurrency of the model for both GPU and CPU processing for a specific number of queries.

As shown in Table 2, BM25 has the shortest retrieval time (12.1 ± 1.7 ms), the lowest total latency (45.7 ± 3.0 ms), and the highest throughput (GPU 500 qps, CPU 120 qps), demonstrating the speed advantage of traditional retrieval methods. Word2Vec + LSTM and BiLSTM-CRF also exhibit low latency and high throughput, at 22.5 ± 2.4 ms / 80.3 ± 4.2 ms and 35.2 ± 3.1 ms / 95.6 ± 5.4 ms, respectively, with throughputs of 300 qps and 250 qps on GPU and respectively. Deep language models combined with graph neural networks, such as DeBERTa-v3-base + GAT, exhibit moderate retrieval time and total latency (50.8 ± 4.0 ms / 120.5 ± 7.2 ms), with throughput of 200 qps on GPU and 48 qps on CPU, slightly lower than the pure DeBERTa-v3-base model (48.1 ± 3.6 ms / 115.2 ± 6.5 ms). GraphSAGE combined with the DeBERTa-v3-base model has the highest total latency (140.3 ± 9.1 ms) and the lowest throughput (170 qps on GPU and 40 qps on CPU), indicating that complex graph network structures increase computational overhead. Other purely pre-trained models (RoBERTa-base, BERT-wwm-ext) perform in the middle, with latency between 128.7 and 132.4 ms and GPU throughput between 185 and 190 qps. Traditional retrieval methods (BM25) require no deep computation and rely primarily on inverted indexes, resulting in extremely low latency and high throughput. Word2Vec + LSTM or BiLSTM-CRF require less computation in sequence modeling and also achieve fast response times. Pre-trained language models combined with graph neural networks require additional computation between text embedding and graph structure propagation, thus increasing retrieval and latency while decreasing throughput. Pure language models, although having a large number of parameters, can be highly parallelized, resulting in slightly lower latency than graph network methods.

TABLE 3. Validation Results of Generalization Ability for Long-tail/Sparse Knowledge

Model	EM (mean ± std)	F1 (mean ± std)	Path-Acc (mean ± std)
DeBERTa-v3-base + GAT	0.72 ± 0.03	0.78 ± 0.02	0.81 ± 0.03
DeBERTa-v3-base + GraphSAGE	0.68 ± 0.04	0.74 ± 0.03	0.77 ± 0.04
DeBERTa-v3-base	0.63 ± 0.05	0.70 ± 0.04	0.72 ± 0.05
RoBERTa-base	0.60 ± 0.06	0.67 ± 0.05	0.70 ± 0.05
BERT-wwm-ext	0.58 ± 0.06	0.65 ± 0.04	0.68 ± 0.06
BiLSTM-CRF	0.50 ± 0.05	0.57 ± 0.04	0.60 ± 0.05
Word2Vec + LSTM	0.42 ± 0.07	0.50 ± 0.06	0.55 ± 0.06
BM25	0.30 ± 0.05	0.35 ± 0.04	0.40 ± 0.05

F. VERIFICATION OF GENERALIZATION ABILITY FOR LONG-TAIL/SPARSE KNOWLEDGE

This experiment tested how well the model generalized to long-tailed or sparse knowledge by reserving unseen entities and relations in the test set, and comparing different models on three evaluation metrics: EM, F1 score, and Path Accuracy. The EM metric looks at how accurately a model can match the correct answer, while the F1 score combines the precision and recall of the model. However, Path Accuracy measures how accurately the model predicts the relationships between entities based on the relationship path connecting them to the first entity being tested. Table 3 provides evidence of the models' overall performance and consistency in making predictions regarding entities and relationships based on their testing of long-tailed and sparse knowledge in terms of the evaluation metrics used.

As illustrated in Table 3, the most effective results were achieved by the combination of the DeBERTa-v3-base and GAT models across all metrics (EM = 0.72 ± 0.03; F1 = 0.78 ± 0.02; Path-Acc = 0.81 ± 0.03). The model's improved performance indicates that the addition of graph-based information greatly enhanced its ability to generalize to unseen entities and relationships. Following closely behind, the DeBERTa-v3-base + GraphSAGE model also performed very well overall, achieving slightly lower results than GAT (EM = 0.68 ± 0.04; F1 = 0.74 ± 0.03; Path-Acc = 0.77 ± 0.04). Text-only models achieved moderate levels of performance–for instance, the DeBERTa-v3-base model achieved an EM of 0.63 ± 0.05; the EM scores for the RoBERTa-base and BERT-wwm-ext models were both < 0.70, indicating that solely relying on text information has a limitation in predicting sparse structured knowledge. In contrast, traditional systems and retrieval techniques produced the poorest performance–the EM scores for both the BiLSTM-CRF and Word2Vec+LSTM were only 0.50 and 0.42 respectively, while BM25 produced an even lower EM score of 0.30.

TABLE 4. Ablation Test

Model	EM	F1	1-hop Path-Acc	2-hop Path-Acc	3-hop Path-Acc
Full Model	0.934 ± 0.018	0.949 ± 0.023	0.960 ± 0.008	0.945 ± 0.010	0.913 ± 0.013
W/O Cross-ranking	0.867 ± 0.021	0.882 ± 0.024	0.915 ± 0.011	0.872 ± 0.015	0.826 ± 0.018
DeBERTa-v3-base → BERT-wwm-ext	0.859 ± 0.023	0.871 ± 0.025	0.908 ± 0.012	0.865 ± 0.016	0.818 ± 0.019
W/O Entity Type Embedding	0.892 ± 0.020	0.905 ± 0.022	0.936 ± 0.009	0.903 ± 0.012	0.869 ± 0.015
GAT→GCN	0.885 ± 0.021	0.898 ± 0.023	0.929 ± 0.010	0.895 ± 0.013	0.852 ± 0.017
W/O GradNorm+Regular	0.876 ± 0.022	0.891 ± 0.024	0.922 ± 0.011	0.883 ± 0.014	0.838 ± 0.018

G. TEACHER SATISFACTION AND EVIDENCE QUALITY ASSESSMENT

To fully assess the quality of various types of knowledge graph question answering models in educational contexts and their influence on teacher's overall level of satisfaction, this section utilizes "teacher's satisfaction rating" as its primary means of evaluating these models. Three separate dimensions are used for quantitative analysis: relevance, sufficiency, and accuracy. Comparison will illustrate significant variations between model scores as well as the degree of variance in the accuracy and completeness of knowledge graph evidence chain creation. Figure 8 contains a visual representation of the evidence chain and the teacher's satisfaction rating by each of the four models where the horizontal axis indicates the score range (1–5 points), while the vertical axis displays the four knowledge graph question-answering models.

The highest score across the three dimensions was obtained by DeBERTa-v3-base+GAT which scored an average of 4.590 (Relevance), 4.558 (Sufficiency), and 4.559 (Correctness) across all three dimensions. These averages indicate that the generated evidence chain provided by this system is the most complete, most accurate and most relevant evidence chain to the question being asked. DeBERTa-v3-base+GraphSAGE scored an average between 4.304 and 4.323, slightly behind that of the approach described in this paper. The DeBERTa-v3-base model, which is a pure Transformer model, performed moderately well, obtaining average scores of 4.002 (Relevance), 4.037 (Sufficiency), and 3.981 (Correctness). The overall performance of RoBERTa and BERT-wwm-ext is significantly lower than what has been observed for the DeBERTa models, as indicated by the widely varying mean scores (3.465–3.917). All three traditional methods were substantially below 3.1, with BM25 in particular obtaining only a mean score of roughly 2.222 to 2.237. These results demonstrate that traditional approaches for generating evidence chains are significantly less effective than their deep learning counterparts, based on the level of evidence chain generation and teacher satisfaction. Overall, Figure 8 clearly presents the ranking of model performance. The ridge width indicates the concentration of the score distribution; high-scoring models are more concentrated, while low-scoring models are more dispersed, reflecting differences in model stability. To verify the impact of each module on the overall model, an ablation experiment was designed, and the results are shown in Table 4.

Table 4 presents the ablation experiment results of the core modules of the DeBERTa-v3-base + GAT model, with the full model as the benchmark and five ablation variants constructed by single-variable removal or replacement, and the results quantified by EM, F1 and 1/2/3-hop Path-Acc. The full model achieves the optimal performance across all indicators, with EM at 0.934 ± 0.018, F1 at 0.949 ± 0.023, and 1/2/3-hop Path-Acc reaching 0.960 ± 0.008, 0.945 ± 0.010 and 0.913 ± 0.013 respectively. All ablation variants show varying degrees of performance decline compared with the full model. The decline trend of each variant fully reflects the irreplaceable contribution of each core module to the model's performance, and the performance drop is more obvious in multi-hop reasoning indicators, especially 3-hop Path-Acc, indicating that each module plays a more critical role in complex long-chain reasoning tasks.

V. CONCLUSION

This paper constructs an end-to-end joint reasoning framework based on DeBERTa-v3-base and multi-layer multi-head GAT. Candidate entities are obtained through entity recognition, vector retrieval, and reordering. Relationship-aware multi-hop propagation is performed in the k-hop subgraph, and graph reasoning and text semantics are combined in the fusion layer to achieve interpretable answer and evidence chain generation. The system achieves good question-answering accuracy and multi-hop reasoning performance on textbook, question bank, and teacher-student question-and-answer data, verifying the effectiveness of deep coupling between semantic representation and graph reasoning. Limitations include the influence of annotation quality on graph construction, insufficient stability of sparse domain reasoning, and efficiency issues caused by multi-hop graph expansion. Future research could explore incremental graph construction and more efficient structured reasoning mechanisms.

AUTHOR CONTRIBUTIONS

Zhuxue Bai designed, collected and analyzed the data, and drafted the manuscript. Qi Mu conducted the study, critically revised the manuscript for important intellectual content.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

FUNDING

1. This work was supported by the Jilin Provincial Higher Education Teaching Reform Research Project (Ideological and Political Education Special Project), entitled Research and Practice on Network Education Models in Universities of Jilin Province from the Perspective of Integrated Media (No.20250PSPBCU009J).

2. This work was supported by the Industry-University Cooperation Collaborative Education Project of the Ministry of Education, entitled Research on the Innovation of Teacher Training System for Practical Teaching of Ideological and Political Courses in Colleges and Universities (Certificate No.2512113542), undertaken by Changchun Humanities and Sciences College.

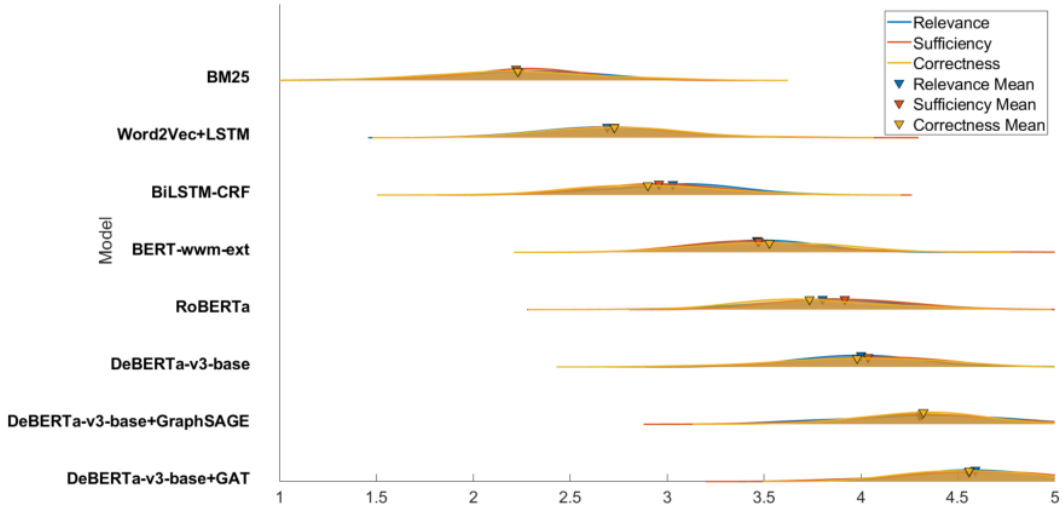

FIGURE 8. Teacher Satisfaction and Evidence Chain Score

3. This work was supported by the Industry-University Cooperation Collaborative Education Project of the Ministry of Education, entitled Research on Teacher Competence Construction for Ideological Risk Identification and Values Guidance Empowered by Artificial Intelligence (Certificate No.251214104), undertaken by Changchun Humanities and Sciences College.

4. This work was supported by the Planning Project of China Association for Non-Government Education (No. CANFZG24084), entitled Research on the Construction of Ideological and Political Course Teachers in Private Colleges and Universities – A Case Study of Jilin Province.

REFERENCES

[1] C. Yi, R. Zhu, and Q. Wang, "Exploring the interplay between question-answering systems and communication with instructors in facilitating learning," Internet Research, vol. 32, no. 7, pp. 32-55, 2022, doi: 10.1108/INTR-08-2020-0459.

[2] G. Lee and X. Zhai, "Realizing visual question answering for education: GPT-4V as a multimodal AI," TechTrends, vol. 69, no. 2, pp. 271-287, 2025, doi: 10.1007/s11528-024-01035-z.

[3] E. Bernasconi, D. Redavid, and S. Ferilli, "Enhancing Personalized Learning with a Context-Aware Intelligent Question-Answering System and Automated Frequently Asked Question Generation," Electronics, vol. 14, no. 7, pp. 1481-1506, 2025, doi: 10.3390/electronics14071481.

[4] Y. Cheng, "Application of a neural network-based visual question answering system in preschool language education," IEIE Transactions on Smart Processing & Computing, vol. 12, no. 5, pp. 419-427, 2023, doi: 10.5573/IEIESPC.2023.12.5.419.

[5] S. Ehsani and J. Liu, "Elevating Textual Question Answering with On-Demand Visual Augmentation," ACM Transactions on Multimedia Computing, Communications and Applications, vol. 21, no. 10, pp. 1-25, 2025, doi: 10.1145/3729231.

[6] H. A. Alawwad, A. Alhothali, U. Naseem, A. Alkhathlan, and A. Jamal, "Enhancing textual textbook question answering with large language models and retrieval augmented generation," Pattern Recognition, vol. 162, no. 1, pp. 1-11, 2025, doi: 10.1016/j.patcog.2024.111332.

[7] Q. Zhong, Y. Jiang, and H. Du, "Intelligent Q&A system for distance education in universities based on Chinese word segmentation technology," International Journal of Continuing Engineering Education and Life Long Learning, vol. 34, no. 6, pp. 562-576, 2024, doi: 10.1504/IJCEELL.2024.143312.

[8] T. T. Aurpa, R. K. Rifat, M. S. Ahmed, M. M. Anwar, and A. B. M. S. Ali, "Reading comprehension based question answering system in Bangla language with transformer-based learning," Heliyon, vol. 8, no. 10, pp. 1-11, 2022, doi: 10.1016/j.heliyon.2022.e11052.

[9] K. Nassiri and M. Akhloufi, "Transformer models used for text-based question answering systems," Applied Intelligence, vol. 53, no. 9, pp. 10602-10635, 2023, doi: 10.1007/s10489-022-04052-8.

[10] F. Dartiko, M. Yusa, A. Erlansari, and S. A. Basha, "Comparative analysis of transformer-based method in a question answering system for campus orientation guides," INTENSIF: Jurnal Ilmiah Penelitian dan Penerapan Teknologi Sistem Informasi, vol. 8, no. 1, pp. 126-143, 2024, doi: 10.29407/intensif.v8i1.21971.

[11] Y. Ba, "Advancing Educational Management with the ATT-MR-WL Intelligent Question-Answering Model," Journal of Web Engineering, vol. 23, no. 7, pp. 973-1002, 2024, doi: 10.13052/jwe1540-9589.2373.

[12] Y. Zhu, "A knowledge graph and BiLSTM-CRF-enabled intelligent adaptive learning model and its potential application," Alexandria Engineering Journal, vol. 91, no. 1, pp. 305-320, 2024, doi: 10.1016/j.aej.2024.02.011.

[13] H. Yu, E. Wang, Q. Lang, and J. Wang, "Intelligent retrieval and comprehension of entrepreneurship education resources based on semantic summarization of knowledge graphs," IEEE Transactions on Learning Technologies, vol. 17, no. 1, pp. 1198-1209, 2024, doi: 10.1109/TLT.2024.3364155.

[14] J. Xiao and Z. Zhang, "EduVQA: A multimodal Visual Question Answering framework for smart education," Alexandria Engineering Journal, vol. 122, no. 1, pp. 615-624, 2025, doi: 10.1016/j.aej.2025.03.005.

[15] X. Cao and Y. Liu, "Relmkg: reasoning with pre-trained language models and knowledge graphs for complex question answering," Applied Intelligence, vol. 53, no. 10, pp. 12032-12046, 2023, doi: 10.1007/s10489-022-04123-w.

[16] W. Lu, D. Ming, X. Mao, J. Wang, Z. Zhao, and Y. Cheng, "A DeBERTa-Based Semantic Conversion Model for Spatiotemporal Questions in Natural Language," Applied Sciences, vol. 15, no. 3, pp. 1073-1093, 2025, doi: 10.3390/app15031073.

[17] S. Li, J. Cao, J. Yang, Y. He, and P. Wang, "DPRM: DeBERTa-based potential relationship multi-headed self-attention joint extraction model," PLoS One, vol. 20, no. 8, pp. 1-26, 2025, doi: 10.1371/journal.pone.0329120.

[18] G. Zhang, J. Liu, G. Zhou, K. Zhao, Z. Xie, and B. Huang, "Question-directed reasoning with relation-aware graph attention network for complex question answering over knowledge graph," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 32, no. 1, pp. 1915-1927, 2024, doi: 10.1109/TASLP.2024.3375631.

[19] P. He and J. Chen, "Enhancing question answering in educational knowledge bases using question-aware graph convolutional network," Journal of Intelligent & Fuzzy Systems, vol. 45, no. 6, pp. 12037-12048, 2023, doi: 10.3233/JIFS-233915.

[20] Y. Fang, J. Deng, F. Zhang, and H. Wang, "An intelligent question-answering model over educational knowledge graph for sustainable urban living," Sustainability, vol. 15, no. 2, pp. 1139-1157, 2023, doi: 10.3390/su15021139.

[21] Y. Jeong and E. Kim, "Scideberta: Learning deberta for science technology documents and fine-tuning information extraction tasks," IEEE Access, vol. 10, no. 1, pp. 60805-60813, 2022, doi: 10.1109/ACCESS.2022.3180830.

[22] D. Han, Z. Wang, Y. Li, X. Ma, and J. Zhang, "Segmentation-aware relational graph convolutional network with multi-layer CRF for nested named entity recognition," Complex & Intelligent Systems, vol. 10, no. 6, pp. 7893-7905, 2024, doi: 10.1007/s40747-024-01551-8.

[23] Y. Qiu, L. Dong, W. Zhang, H. Xing, and J. Huang, "A diffusion enhanced CRF and BiLSTM framework for accurate entity recognition," Scientific Reports, vol. 15, no. 1, pp. 1-25, 2025, doi: 10.1038/s41598-025-04036-x.

[24] W. Khan, A. Daud, K. Shahzad, T. Amjad, A. Banjar, and H. Fasihuddin, "Named entity recognition using conditional random fields," Applied Sciences, vol. 12, no. 13, pp. 6391-6408, 2022, doi: 10.1016/j.procs.2020.03.431.

[25] Q. Z. Lim, C. P. Lee, K. M. Lim, J. X. Ng, E. K. H. Ooi, and N. K. N. Loh, "Multilingual question answering for Malaysia history with transformer-based language model," Emerging Science Journal, vol. 8, no. 2, pp. 675-686, 2024, doi: 10.28991/ESJ-2024-08-02-019.

[26] J. Zhu, "A multi task learning framework using DeBERTa and BWO optimization for enhancing long term english vocabulary memory," Discover Computing, vol. 28, no. 1, pp. 1-32, 2025, doi: 10.1007/s10791-025-09690-3.

[27] Y. Miao, W. Cheng, S. He, and H. Jiang, "Research on visual question answering based on GAT relational reasoning," Neural Processing Letters, vol. 54, no. 2, pp. 1435-1448, 2022, doi: 10.1007/s11063-021-10689-2.

[28] Z. Li, Z. Cao, P. Li, Y. Zhong, and S. Li, "Multi-hop question generation with knowledge graph-enhanced language model," Applied Sciences, vol. 13, no. 9, pp. 5765-5779, 2023, doi: 10.3390/app13095765.

[29] Y. Li, W. Li, and L. Nie, "Dynamic graph reasoning for conversational open-domain question answering," ACM Transactions on Information Systems (TOIS), vol. 40, no. 4, pp. 1-24, 2022, doi: 10.1145/3498557.

[30] P. Zhu, Y. Yuan, and L. Chen, "Electra-based graph network model for multi-hop question answering," Journal of Intelligent Information Systems, vol. 61, no. 3, pp. 819-834, 2023, doi: 10.1007/s10844-023-00800-5.

[31] A. A. Yusuf, C. Feng, X. Mao, R. Ally Duma, M. S. Abood, and A. H. A. Chukkol, "Graph neural networks for visual question answering: a systematic review," Multimedia Tools and Applications, vol. 83, no. 18, pp. 55471-55508, 2024, doi: 10.1007/s11042-023-17594-x.

[32] M. Reyad, A. M. Sarhan, and M. Arafa, "A modified Adam algorithm for deep neural network optimization," Neural Computing and Applications, vol. 35, no. 23, pp. 17095-17112, 2023, doi: 10.1007/s00521-023-08568-z.

[33] P. Zhou, X. Xie, Z. Lin, and S. Yan, "Towards understanding convergence and generalization of AdamW," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 46, no. 9, pp. 6486-6493, 2024, doi: 10.1109/TPAMI.2024.3382294.

[34] I. V. Modoranu, M. Safaryan, G. Malinovsky, E. Kurtic, T. Robert, P. Richtarik, et al., "Microadam: Accurate adaptive optimization with low space overhead and provable convergence," Advances in Neural Information Processing Systems, vol. 37, no. 1, pp. 1-43, 2024, doi: 10.52202/079017-0001.

[35] C. Zhang, Y. Shao, H. Sun, L. Xing, Q. Zhao, and L. Zhang, "The WuC-Adam algorithm based on joint improvement of Warmup and cosine annealing algorithms," Mathematical Biosciences and Engineering, vol. 21, no. 1, pp. 1270-1285, 2024, doi: 10.3934/mbe.2024054.

[36] O. V. Johnson, C. Xinying, K. W. Khaw, and M. H. Lee, "ps-CALR: periodic-shift cosine annealing learning rate for deep neural networks," IEEE Access, vol. 11, no. 1, pp. 139171-139186, 2023, doi: 10.1109/ACCESS.2023.3340719.