A Deep Learning Journey in Closed-Domain Medical Question Answering with RNN-Attention and Intelligent Question Expansion

Nayyab Saeed , Anum Yasmin * , Shireen Tahira

Posted Date: 1 December 2025

doi: 10.20944/preprints202512.0093.v1

Keywords: medical question answering (MQA); BioBERT; recurrent neural networks (RNNs); deep learning; question expansion; medical NLP; closed-domain QA systems; UMLS

 doi:10.20944/preprints202512.0093.v1

Article

A Deep Learning Journey in Closed-Domain Medical Question Answering with RNN-Attention and Intelligent Question Expansion

Nayyab Saeed, Anum Yasmin * and Shireen Tahira

Department of Computer Science, International Islamic Universit (IIU), Islamabad, Pakistan

* Correspondence: anum.yasmin@iiu.edu.pk

Abstract

Deep learning-based Medical Question Answering (MQA) systems are transforming access to healthcare information by enabling accurate and timely responses to complex queries. However, existing systems face challenges such as limited large-scale, high-quality medical datasets, inadequate contextual understanding, and difficulties in managing diverse medical terminologies This research proposes a novel closed-domain MQA system that addresses these limitations through innovative methodologies. The system employs BioBERT-based domain-specific embeddings trained on biomedical literature to accurately capture medical terminology, abbreviations, and contextual nuances. To model sequential dependencies in queries, Recurrent Neural Networks (RNNs) are integrated, enabling contextual interpretation across longer text sequences. Additionally, a question expansion mechanism utilizing medical dictionaries and ontologies like UMLS addresses synonymy, ambiguity, and terminological variations, ensuring that diverse medical expressions map to consistent, semantically relevant concepts for precise answer retrieval. Extensive evaluation using metrics such as F1-score, precision, recall, and exact match demonstrates the system's superior performance compared to existing models. The key contributions include improved contextual understanding, better handling of medical terminology, and a scalable framework for future medical NLP applications. This system not only offers a reliable tool for healthcare professionals and patients but also advances the field of intelligent question answering by supporting evidence-based clinical decision-making.

Keywords: medical question answering (MQA); BioBERT; recurrent neural networks (RNNs); deep learning; question expansion; medical NLP; closed-domain QA systems; UMLS

I. Introduction

A. Background

Deep learning-based Question Answering (QA) systems have revolutionized Natural Language Processing (NLP) by offering the ability to comprehend and answer questions with high precision and speed. These systems leverage deep neural networks to interpret the context of questions and retrieve relevant information from structured databases, unstructured texts, images, and knowledge graphs [1]. Over the last decade, substantial advancements in deep learning algorithms and the growing availability of-large-scale datasets have significantly enhanced the reliability and accuracy of QA systems. As a result, QA systems have found widespread applications across sectors such as education, customer service, virtual assistance, and healthcare [2].

Traditional QA systems often struggled with the complexity of natural language, including syntactic ambiguities, diverse terminologies, and intricate sentence structures. However, deep learning-based systems have addressed these challenges using powerful neural architectures capable of understanding context and relationships between words. In particular, closed-domain@QA

 doi:10.20944/preprints202512.0093.v1

systems focus on answering questions within a specific-area of knowledge, offering enhanced precision compared to open-domain systems that utilize general data sources [3]. The accuracy and specificity of closed-domain systems are especially crucial in sensitive domains such as healthcare, where the quality of information can directly impact outcomes [4].

B. Research Motivation and Objectives

Despite significant advancements in medical question answering (MQA) systems, several critical challenges remain unaddressed, limiting their effectiveness and reliability. Existing systems suffer from the scarcity of large-scale, domain-specific datasets, inadequate diversity in medical question types, and the complexity and variability inherent in medical terminologies. Moreover, weak contextual understanding often results in inaccurate or irrelevant answers, which is particularly concerning in sensitive domains such as healthcare where misinformation can have severe consequences. To overcome these limitations, this research aims to develop a robust closed-domain MQA system that leverages comprehensive medical datasets and specialized domain embeddings such as BioBERT to enhance semantic understanding. The proposed approach incorporates Recurrent Neural Networks (RNNs) to better capture sequential-dependencies [in medical queries and employs question expansion techniques using medical dictionaries and thesauri like UMLS and MeSH to handle terminological variations effectively. Central to this effort are several key research questions: How can the deficiency of large-scale medical datasets be addressed to improve system performance? What methodologies best enable the interpretation of diverse and complex medical queries? How can terminological variability be managed to enhance answer accuracy? Can contextual understanding be improved through advanced deep learning architectures? Finally, will the proposed system demonstrate superior accuracy and reliability compared to existing methods? By answering these questions, the study aims to provide a scalable and context-aware MQA framework that supports healthcare professionals and patients in accessing timely and accurate medical information, ultimately helping to enhance the results for patients and medical decision-making.

C. Research Contributions

This study presents a novel closed-domain Medical Question Answering (MQA) system that makes several key contributions to advance the field. First, it integrates a comprehensive and diverse collection of medical datasets, enabling robust training and enhanced generalizability across different medical contexts. Second, the system employs domain-specific@pre-trained embeddings such as@BioBERT, SciBERT, and PubMedBERT, which significantly improve the understanding of specialized biomedical language compared to generic models like BERT and ALBERT. Third, the proposed architecture replaces conventional Convolutional@Neural@Networks (CNNs) with Recurrent#Neural#Networks (RNNs), allowing for better capture of sequential dependencies and contextual relationships inherent in medical queries. Finally, a novel question expansion technique is introduced, leveraging synonym resources from established medical ontologies such as UMLS and MeSH, thereby enhancing*the system's-ability to interpret and respond accurately to diverse medical question formulations. Experimental evaluation demonstrates that the proposed system outperforms existing models in accuracy and contextual relevance, underscoring its potential as a valuable tool for intelligent healthcare applications.

II. Related Work

Research in closed-domain medical question answering (QA) has seen remarkable advancements through the integration of diverse deep learning architectures, domain-specific adaptations, and innovative data handling techniques. Multiple studies have contributed novel

 doi:10.20944/preprints202512.0093.v1

frameworks that improve answer accuracy, relevance, and system robustness in specialized medical contexts.

One influential approach [3] combines Convolutional Neural Networks (CNN) with self-attention mechanisms to enhance understanding of medical queries. The use of a Candidate*Answer Identifier_(CAI) module-and a Question*Expansion (QE) technique significantly improves retrieval accuracy across datasets such as Tesla, California, and COVID-QA. However, challenges regarding small dataset sizes and computational overhead remain, limiting generalizability and real-time deployment.

Large language models (LLMs) tailored for healthcare, such as Med-PaLM 2 [4], show notable benchmark improvements, achieving an 86.5% accuracy on the MedQA dataset. Nevertheless, these models need further exploration concerning clinical deployment safety, ethical considerations, and adaptability to diverse medical conditions.

Similarly, the COBERT system [5] effectively addresses domain-specific information overload during the COVID-19 pandemic by combining DistilBERT and TF-IDF vectorization, but its reliance on COVID-centric data restricts broader applicability. Bias detection and mitigation strategies are essential for improving its reliability.

Addressing language-specific challenges, especially in Arabic [6], ON-LSTM combined with CNN models have enhanced question-answer relevance ranking. However, their adaptability across languages and domains requires further validation. Data augmentation methods like back-translation and word substitution [7] boost performance on limited biomedical datasets but have yet to be tested extensively on complex question types and clinical scenarios.

The integration of structured clinical data with QA frameworks, as seen in emrQA with SQuAD V2.0 [8], provides valuable resources for span-extraction tasks but lacks consideration of dataset biases and nuanced queries. Innovations in retrieval pipelines incorporating temporal and citation factors [9] enhance answer credibility but face scalability challenges.

Knowledge graph-based systems such as KG-Rank [10] improve factual grounding and interpretability, yet their adaptability to rapidly evolving medical knowledge and computational demands present obstacles. Hybrid models combining retrieval and LLMs [11] optimize accuracy and efficiency but may restrict generative capabilities needed for complex queries.

Attention-based retrieval with medical entity extraction [12] enables real-time, interactive healthcare QA, yet scalability and handling of novel queries remain concerns. Pediatric disease-focused models leveraging ALBERT, BiLSTM, and knowledge graphs [13] deliver high accuracy, though expansion to broader medical subdomains and real-world applications is needed.

Broader analyses of healthcare QA trends [14] emphasize the shift toward empathetic, personalized conversational agents but lack empirical evaluations. Privacy-preserving QA systems addressing sensitive domains like mental health [15] showcase interpretability with differential privacy, yet the trade-off between privacy and performance requires deeper investigation.

Hierarchical models incorporating mental health signals [16] improve semantic understanding but may be computationally intensive for low-resource environments. Domain-specific pretrained models like BioBERT [17,22] demonstrate strong results on structured datasets, though their applicability to unstructured or multi-step reasoning queries is limited.

Feature-enriched frameworks [18] enhance performance by incorporating syntactic and lexical cues but face scalability and computational cost challenges. Reviews of medical QA systems [19,21] highlight critical gaps such as limited clinical applicability, insufficient user evaluations, and inadequate handling of uncertainties.

Comprehensive overviews of QA datasets [20] aid in understanding system evolution and challenges but focus mainly on English and lack empirical application. There remains a pressing need for multilingual, domain-specific, and clinically validated QA models that integrate evidence-based principles and real-time workflows.

Collectively, these studies underscore significant progress in closed-domain medical QA through innovations in model design, data utilization, and hybrid approaches. However, challenges

 doi:10.20944/preprints202512.0093.v1

such as dataset bias, computational scalability, clinical integration, user trust, and multi-modal support continue to drive future research directions toward more robust, explainable, and practical healthcare QA solutions.

III. Methodology

The methodology aims to build an efficient and adaptable medical Question Answering (QA) system that addresses challenges like complex medical terms, ambiguous queries, and limited annotated data. The system combines a fine-tuned BioBERT model with three key modules: Candidate Answer Identification (CAI), an RNN with self-attention, and Question Expansion (mentioned in Figure 1).

The CAI module uses an enhanced 5W1H framework to identify medical keywords and focus on relevant sentences. The RNN with self-attention captures contextual relationships between questions and answers. The Question Expansion module reformulates unclear queries using a medical dictionary and thesaurus to improve understanding. Each module is individually tested and fine-tuned to ensure precise, reliable, and context-aware answers for the closed-domain medical QA system.

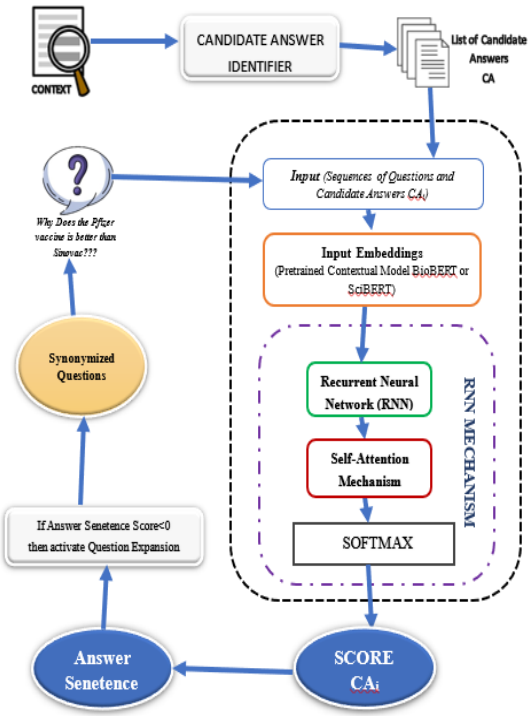

Figure 1. Proposed Solution with RNN-Mechanism and Question Expansion.

A. Dataset Collection

Two datasets are used: **COVID-QA** and **MedQuAD**.

- **COVID-QA** contains question-context-answer triples focused on COVID-19 topics like prevention and treatment. It tests the system's ability to handle specialized medical vocabulary.
- **MedQuAD** includes a wide range of medically curated questions and answers from trusted sources like the U.S. National Library of Medicine, covering general medical topics.

 doi:10.20944/preprints202512.0093.v1

Table 1. Datasets.

Dataset	Domain	Size	Purpose
COVID-QA	COVID-19	2,019 questions	Specialized COVID-19 queries
MedQuAD	General Medicine	16,408 questions	Broad medical topics

Together, these datasets ensure the QA system can handle both specialized and general medical questions effectively.

B. Data Preprocessing and Fine-Tuning

To ensure accurate and context-aware responses in a closed-domain medical QA system, the development process begins with rigorous data preprocessing followed by domain-specific@fine-tuning of the BioBERT model. The overall workflow-is illustrated in-Figure 2.

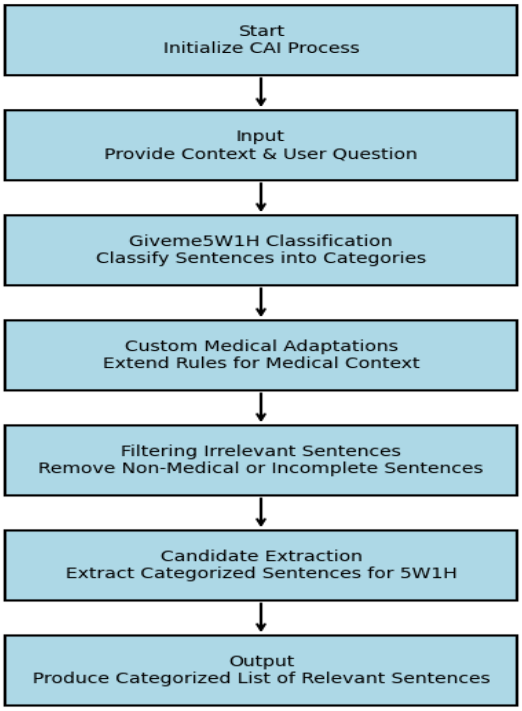

Figure 2. Workflow$of Data Preprocessing and Fine-tuning.

The **preprocessing phase** prepares raw input data by first applying **tokenization** using the BioBERT tokenizer (WordPiece), which effectively breaks down complex medical terms (e.g., *hydroxychloroquine*) into meaningful%subwords. This allows the model to better interpret domain-specific vocabulary. Next, **Part-of-Speech\(POS) tagging** is performed using Stanford CoreNLP to identify the grammatical roles of words and extract critical medical terms like diseases and treatments.

The text is then segmented into individual sentences through **sentence splitting**, which facilitates sentence-level classification in the Candidate Answer Identification (CAI) module. To ensure input consistency, a **text cleaning and normalization** process follows, which includes converting text to lowercase, removing stopwords and special characters, retaining important

 doi:10.20944/preprints202512.0093.v1

numerics (e.g., "10 mg dosage"), and applying **lemmatization** to reduce words to their base forms (e.g., *treatments → treatment*).

In the **fine-tuning=phase**, the pre-trained BioBERT model—originally trained on biomedical corpora such as PubMed and PMC—is fine-tuned on the COVID-QA and

MedQuAD datasets. This step helps the model adapt to domain-specific contexts, enabling it to handle the specialized terminology and complex relationships found in medical queries. By combining systematic preprocessing with targeted fine-tuning, the model becomes highly effective at generating accurate, relevant, and trustworthy answers for real-world medical questions.

C. Candidate Answer Identification (CAI) Module

The Candidate Answer Identification (CAI) module is a core component of the medical Question Answering (QA) system, designed to extract the most relevant sentences from the context that align with the user's query. It uses a customized version of the Giveme5W1H framework, originally built for general question classification. This framework has been significantly tailored for the medical domain by incorporating linguistic patterns and syntactic rules unique to medical texts. The overall workflow-is illustrated in-Figure 3.

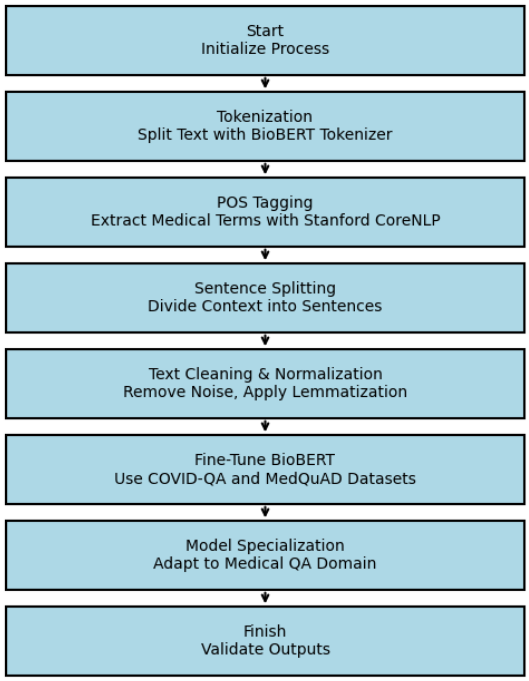

Figure 3. Workflow-of Candidate Answer Identifier Module.

The Giveme5W1H framework classifies sentences into six categories: "What," "Why," "Who," "When," "Where," and "How." These categories help organize and extract structured information from the context. Each sentence is evaluated and classified based on its structure and meaning. For example, sentences containing factual information, such as "The COVID-19 vaccine reduces severe illness," fall under the "What" category. Sentences that explain reasons, like "Vaccination is essential because it reduces the spread of the virus," are grouped under "Why." References to individuals or organizations such as "The WHO recommended vaccination for adults" belong to the "Who" category. Sentences with temporal or locational details, like "The pandemic started in December 2019" or "The virus was first detected in Wuhan, China," fall under "When" and "Where," respectively. The "How" category captures process-based sentences, such as "The vaccine is administered intramuscularly."

To better suit the medical domain, the framework includes custom rules and patterns involving keywords like "treatment for," "caused by," or "symptoms include." These rules ensure that

 doi:10.20944/preprints202512.0093.v1

important medical sentences are identified and prioritized. Additional linguistic patterns, such as "to prevent" or "for reducing," have been added to enhance classification for categories like "What" and "Why."

To maintain high precision, the CAI module filters out irrelevant or incomplete sentences. Sentences that lack medical keywords or do not match any of the 5W1H categories are excluded. Grammatical errors or syntactically incomplete sentences are also removed.

The output is a categorized list of relevant sentences based on the 5W1H labels. For instance, the "What" category might include ["The COVID-19 vaccine reduces severe illness."] and "Why" might include ["Vaccination reduces the spread of the virus."]. This structured approach ensures that only accurate and contextually appropriate information is passed on for the next stage of the QA pipeline.

D. RNN-Attention based Answer Selector

The RNN-Attention based Answer Selector is a crucial module designed to evaluate the-relevance between a user's-question and the list of candidate sentences identified in the previous stage. It processes each question-candidate pair using a combination of BioBERT embeddings, Recurrent Neural Networks (RNNs), and a self-attention mechanism to ensure that the most contextually appropriate answer is selected. The overview of this module is shown in Figure 4.

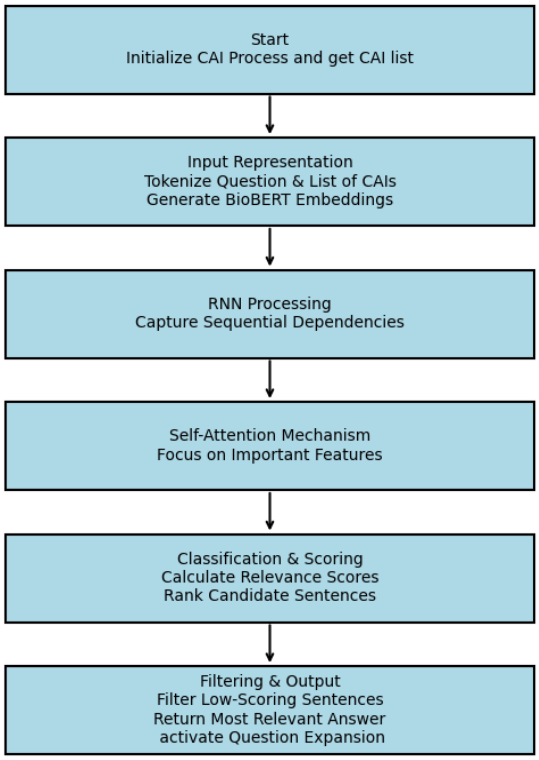

Figure 4. *Workflow of Recurrent Neural Networks (RNNs), and a self-attention layer.*

To begin with, each candidate sentence is paired with the user's question and formatted into a single input sequence. This sequence follows the structure:

$$\mathbf{[CLS]\ q_1, q_2, ..., q_m\ [SEP]\ c_1, c_2, ..., c_n\ [SEP]}$$

where=the tokens $q_1,q_2...,q_m$ represent the question, $c_1,c_2,...,c_n$ represent the candidate sentence, and special tokens [CLS] and [SEP] are used for segment separation and classification. This combined input is then passed through BioBERT, a domain-specific language model trained on biomedical texts.

BioBERT tokenizes the input using WordPiece tokenization, which splits uncommon words into meaningful subunits—for example, "hydroxychloroquine" might be broken into "hydroxy," "chloro," and "quine." After tokenization, BioBERT generates contextual embeddings for each token,

 doi:10.20944/preprints202512.0093.v1

capturing not only the semantic meaning of each word but also its relationship to surrounding words. The embedding corresponding to the [CLS] token plays a vital role, as it summarizes the entire input sequence.

The output embeddings from BioBERT are then fed into a Recurrent Neural Network (RNN) to model sequential dependencies in the text. RNNs process the input one token at a time, maintaining a hidden state that evolves over each time step. The update to the hidden state at time step t_t is calculated using the equation:

$$h_t = f(W_h \cdot x_t + U_h \cdot h_{t-1} + b_h)\%\% \quad (1)$$

Here, x_t is the input at-time t_t, h_{t-1} is the-previous hidden-state, W_h-and-U_h are-weight matrices, b_h is the bias-term, and f is an activation-function-such as ReLU or tanh. This mechanism enables the RNN to encode the sequence meaning by preserving word order and long-range dependencies—essential in medical texts where phrasing matters significantly.

However, RNNs alone may not always identify the most relevant parts of a sequence. To address this, a self-attention layer is applied on top of the RNN output. This layer allows-the-model to assign-varying-degrees of-importance to-different tokens-in the sequence-by computing-attention scores-using query- (Q), key (K), and value (V) matrices. These are-derived as:

$$Q = W_q \cdot H \quad - \quad (2)$$

$$K = W_k \cdot H - \quad (3)$$

$$V = W_v \cdot H - \quad (4)$$

where H is the matrix of RNN outputs and W_q,W_k,W_v are trainable weight parameters. The attention scores are then computed using the dot product of Q and K, scaled by the square root of the key dimension d_k:

$$\text{Attention Scores} = \frac{Q \cdot K^T}{\sqrt{d_K}} \quad (5)$$

These raw scores are normalized using the SoftMax function to generate attention weights:

$$\text{Attention Weights} = softmax\left(\frac{Q \cdot K^T}{\sqrt{d_K}}\right) \quad (6)$$

This step ensures that all weights lie between 0 and 1 and sum up to 1, allowing-the-model-to focus on more-relevant tokens while downplaying less important ones. The final attention output is obtained by multiplying these weights-with the value-matrix:

$$\text{Attention Output} = \text{Attention Weights} \cdot V - \quad (7)$$

This results in a context-aware representation where the most critical information is emphasized.

To understand this intuitively, consider the example where the question is *"What does the vaccine prevent?"* and the candidate sentence is *"The vaccine prevents severe illness."* The attention mechanism will assign higher weights to words like "prevents" and "severe illness" as they are closely related to the question. Less relevant tokens such as "the" may receive lower weights. This ensures that the model focuses on meaningful information for answer selection.

In summary, this module combines the strengths of BioBERT's contextual embeddings, RNN's ability to handle sequence information, and self-attention's capacity to highlight relevant content. This layered approach enables the model to align the user's question with the most appropriate candidate sentence, enhancing the system's-ability to deliver precise and contextually accurate answers in the medical domain.

 doi:10.20944/preprints202512.0093.v1

E. Question Expansion (QE) Module

Some user questions may be unclear or lack sufficient domain-specific details, making it difficult for the system to retrieve accurate answers. To solve this, the Question Expansion (QE) module reformulates the original question by identifying key keywords and replacing them with more relevant synonyms from the medical domain. Using resources like WordNet and the Unified-Medical Language-System- (UMLS), the module ensures reformulations are contextually accurate for medical queries. The overview of this module is shown in Figure 5.

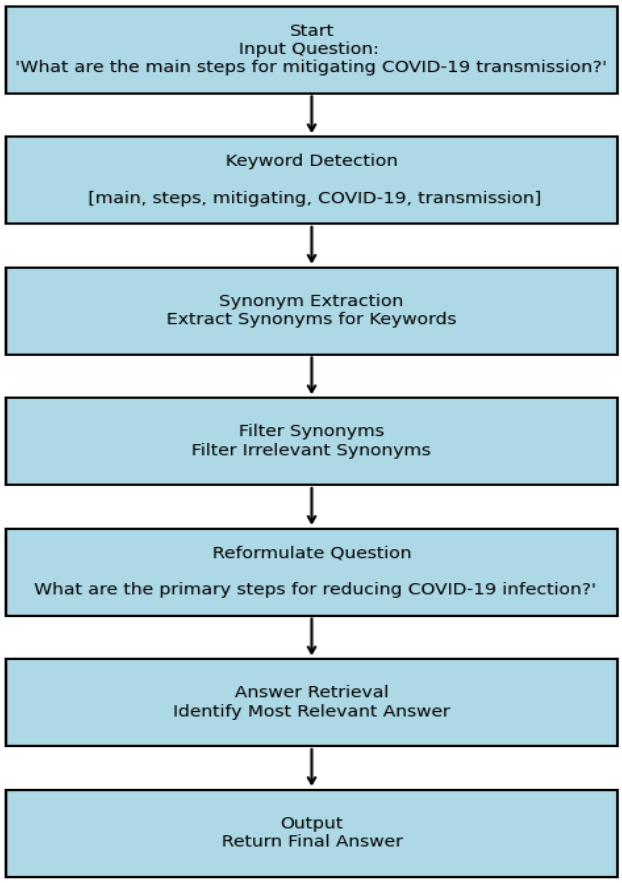

Figure 5. Workflow of Question Expansion module.

The QE module activates only when needed — if the confidence score of the candidate answer falls below a threshold θ\theta, indicating ambiguity or low certainty. For example, if the question "What are the steps to reduce COVID-19 spread?" yields a low-confidence answer, QE triggers to refine it.

First, the module detects keywords using Part-of-Speech (POS) tagging and filters out stopwords, keeping important terms like nouns, verbs, and adjectives. For example, in the question "What are the main steps for mitigating COVID-19 transmission?" the keywords are ["main," "steps," "mitigating," "COVID-19," "transmission"].

Next, synonyms for these keywords are retrieved from WordNet and UMLS, ensuring replacements match the original grammatical category. Examples include "main" → ["primary," "major"], "mitigating" → ["reducing," "alleviating"], and "transmission" → ["infection," "contagion"].

To retain only relevant synonyms, the module applies semantic similarity filtering using BioBERT embeddings. The similarity between the original keyword QQ and a synonym SS is calculated as the cosine similarity of their embeddings:

 doi:10.20944/preprints202512.0093.v1

$$Similarity = \frac{Emb(Q) \cdot Emb(S)}{||Emb(Q)|| \cdot ||Emb(S)||} \quad (9)$$

Synonyms with similarity scores above a threshold α\alpha are kept. For example, "infection" with 0.87 similarity is retained, while "dispatch" with 0.45 is discarded.

The question is then reformulated by substituting keywords with selected synonyms while preserving grammar. For instance, "What-are-the main- steps for- mitigating COVID-19 transmission?" becomes "What are the primary steps for reducing COVID-19 infection?"

Each reformulated question is re-evaluated by the QA system. If its confidence score exceeds θ\theta, the process stops and the answer is selected; otherwise, the original best answer is returned.

This QE process improves handling of vague or ambiguous medical queries, leveraging linguistic and domain-specific resources to deliver precise and relevant answers.

In summary, the proposed system integrates candidate answer identification using BioBERT embeddings, RNNs, and self-attention to effectively align questions with relevant answers. The Question Expansion module further enhances accuracy by reformulating unclear queries with domain-specific synonyms. This layered approach ensures precise understanding and retrieval of medical answers, making the system robust and reliable for closed-domain medical question answering.

IV. Experiment Setup

The experimental setup was carefully designed to address the specific challenges of medical question answering by leveraging advanced natural language processing techniques tailored to the biomedical domain. The data preparation phase involved comprehensive preprocessing, including tokenization and cleaning of both actual question-answer pairs and candidate answer datasets to ensure consistency and compatibility. BioBERT embeddings were utilized to generate rich contextual representations that capture the nuanced semantics of medical terminology, significantly improving the system's ability to distinguish subtle differences in meaning within medical queries.

The core model architecture comprised a bidirectional*Long*Short-Term*(LSTM) network integrated with a self-attention*mechanism. This configuration enabled the model to effectively capture sequential dependencies and focus on the most relevant elements within both the questions and candidate answers. By highlighting key terms critical for accurate interpretation, the self-attention layer enhanced the model's performance, particularly in complex, multi-sentence scenarios common in medical queries.

To further improve the system's handling of ambiguous or incomplete questions, a Question Expansion module was incorporated. This module reformulates queries by replacing keywords with domain-specific synonyms sourced from WordNet and the Unified Medical Language System (UMLS). Synonyms were filtered through semantic similarity scoring using BioBERT embeddings to ensure relevance, thereby producing reformulated questions that better align with medical context and terminology.

Training and evaluation were conducted in a GPU-enabled environment with techniques such as dropout regularization to mitigate overfitting and adaptive learning rate scheduling to optimize convergence. The model was validated on unseen datasets and subjected to cross-validation to ensure robustness and generalizability. Results indicated that BioBERT embeddings were responsible for approximately 53% of the system's accurate predictions. The self-attention mechanism significantly improved answer selection by focusing on critical input features, while the Question Expansion module enhanced performance on vague or ambiguous queries.

Overall, this comprehensive experimental framework demonstrated strong accuracy and reliability in retrieving relevant answers within a closed-domain medical QA system. The integration of domain-specific embeddings, attention mechanisms, and query reformulation establishes a solid foundation for future work, including expanding datasets, refining query expansion, and incorporating multilingual capabilities to extend system applicability.

 doi:10.20944/preprints202512.0093.v1

V. Results & Discussion

This study evaluates the performance of two models, CA-AcdQA and CA-ExpQA, using the COVID-QA and MedQuAD datasets to address the challenges of biomedical Question Answering (QA). The CA-AcdQA model was tested with multiple variants based*on*pre-trained*language models, including*BERT, ALBERT, SciBERT, and BioBERT, while the CA-ExpQA model utilized BioBERT along with specialized enhancements such as the Candidate Answer Identification (CAI) module and the Question Expansion (QE) module.

Both models were*evaluated using*Exact Match (EM) *and*F1 Score, which*measure the precision and recall of retrieving accurate and comprehensive answers. The comparison offers valuable insights into how well these models handle domain-specific challenges such as clinical terminology, ambiguity in queries, and reasoning in biomedical contexts.

The results, presented in Table 2, highlight the importance of domain-specific pre-trained models in achieving high performance in biomedical QA tasks. Among the CA-AcdQA variants, CA-AcdQA (BioBERT) demonstrated the best performance on the*COVID-QA dataset, achieving an EM*score of 73.2 and*an F1*score of 83.8. This performance advantage is attributed to*BioBERT's pre-training*on*large-scale*biomedical*corpora, enabling it to effectively capture the complexities of medical terminology and contextual relationships.

Table 2. Performance Comparison of CA-ExpQA (BioBERT) Against CA-AcdQA Variants.

MODEL	DATASET	EVALUATION METRICES	
		EM	F1-SCORE
CA-AcdQA* (BERT)	COVID-QA	55.6	76.4
CA-AcdQA* (ALBERT)	COVID-QA	57.5	78.8
CA-AcdQA* (SciBERT)	COVID-QA	68.8	80.6
CA-AcdQA (BioBERT)	COVID-QA	73.2	83.8
CA-ExpQA (BioBERT)	COVID-QA	87.0	89.0
CA-ExpQA (BioBERT)	MEDQUAD	85.3	89.0

In contrast, the CA-ExpQA (BioBERT) model, which integrates the QE and CAI modules, significantly outperformed the CA-AcdQA variants. On*the*COVID-QA*dataset, it achieved an EM of 87.0 and*an*F1 score of 89.0. This improvement is largely due to the QE module's ability to resolve ambiguities in user questions by reformulating unclear or incomplete queries with domain-relevant synonyms. Additionally, the CAI module improved the system's ability to select*the*most*relevant sentences from the context, enhancing both precision and recall.

On the MedQuAD dataset, which features a broader range of medical queries, the CA-ExpQA (BioBERT) model also excelled, achieving an EM of 85.3 and F1*score of 89.0. These*results demonstrate the generalizability of the CA-ExpQA model across diverse biomedical datasets and its effectiveness in handling complex medical questions. By comparison, the CA-AcdQA (BioBERT) variant showed lower performance on this dataset, further highlighting the value of the additional enhancements in the CA-ExpQA model.

 doi:10.20944/preprints202512.0093.v1

The first bar chart (Figure 6) visualizes the Exact Match (EM) scores for different variants of the CA-AcdQA model and the enhanced CA-ExpQA model on the COVID-QA dataset. Among the CA-AcdQA variants, BioBERT achieves the highest EM score of 73.2, outperforming BERT, ALBERT, and SciBERT, due to its biomedical-specific pre-training. The CA-ExpQA (BioBERT) model further improves the EM score to 87.0, demonstrating the strong impact of the QE and CAI modules in improving context understanding and relevance.

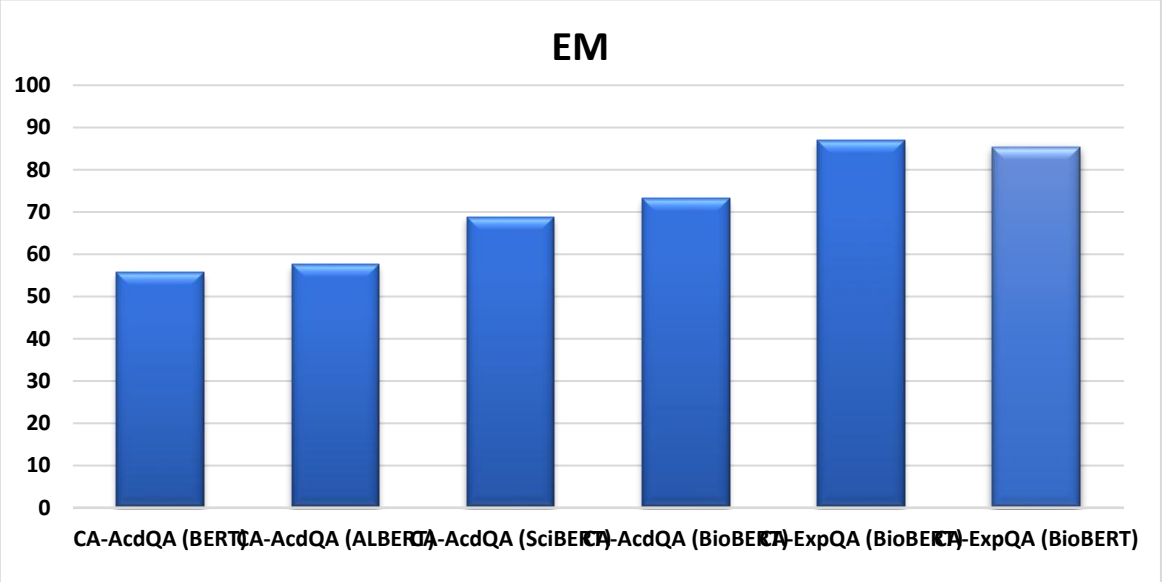

Figure 6. Performance Comparison of CA-ExpQA(BioBERT) model against CA-AcdQA by Exact Match.

The second bar chart (Figure 7) presents the F1 scores of the same models on the COVID-QA dataset. Once again, CA-AcdQA (BioBERT) leads among its variants with an F1 score of 83.8, but the CA-ExpQA (BioBERT) model outperforms it with a score of 89.0. This performance gain reflects the effectiveness of combining domain-specific language modelling with QA-specific enhancements.

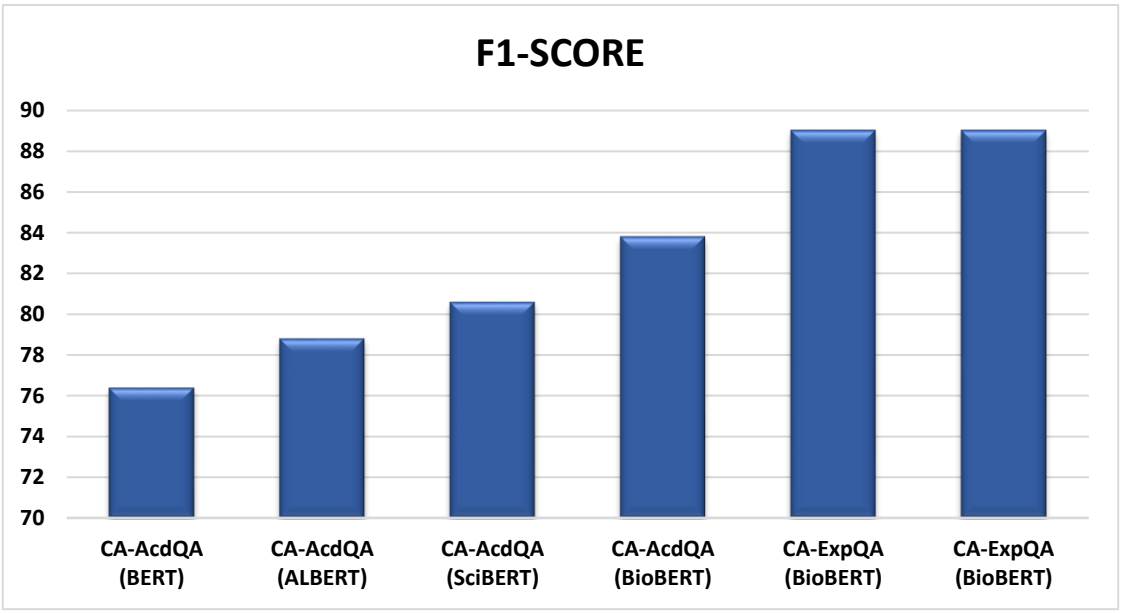

Figure 7. Performance Comparison of CA-ExpQA(BioBERT) model against CA-AcdQA by F1 Score.

 doi:10.20944/preprints202512.0093.v1

A. Results of Candidate Answer Identifier Module:

The results highlight a comprehensive comparison of the Candidate Answer Identifier (CAI) component's performance across different models and datasets. The evaluation focuses on F1 scores for six question categories: What, Where, When, Why, Who, and How. Two primary models are analyzed: CA-AcdQA, which employs a general-purpose CAI framework in English, and CA-ExpQA, which incorporates a customized medical-specific CAI module. This comparison offers insights into the value of domain-specific adaptations in biomedical question-answering tasks.

The performance outcomes are summarized in Table 3. The CA-AcdQA (BERT) and CA-AcdQA (ALBERT) models utilize generalized linguistic rules from the Giveme5W1H framework, originally intended for standard English queries. These models show consistent but relatively modest performance, especially on the COVID-QA dataset. They perform slightly better in categories like Where and When, with F1 scores ranging from 76.7 to 79.7. However, their effectiveness is limited when handling medical-specific terminology and reasoning due to the lack of tailored mechanisms for biomedical language.

Table 3. Performance*Comparison* (F1 score) of CA-ExpQA model for each*question*category with*proposed Candidate Answer*Identifier* (Customized Giveme5W1H with medical format) with CA-AcdQA .

MODEL	DATASET	F1 score for each question category					
		What	Where	When	Why	Who	How
CA-AcdQA (BERT) Cust. Giveme 5W1H CAI(English)	COVID-QA	75.3	76.7	76.9	74.9	76.6	77.7
CA-AcdQA (ALBERT) Cust. Giveme 5W1H CAI(English)	COVID-QA	78.9	78.6	79.7	77.2	78.5	79.9
CA-ExpQA (BioBERT) Cust. Giveme 5W1H CAI(Medical)	COVID-QA	87.5	96.0	96.0	94.0	98.8	80.5
CA-ExpQA (BioBERT) Cust. Giveme 5W1H CAI(Medical)	MEDQUAD	96.6	99.7	99.8	99.5	96.4	97.0

In contrast, the CA-ExpQA (BioBERT) model, integrated with a medical-specific CAI module, demonstrates a substantial performance improvement across all categories. On the*COVID-QA dataset, this model achieves an F1 score of 98.8 for Who questions and 96.0 for both Where and When questions. These improvements result from the module's ability to align closely with biomedical

 doi:10.20944/preprints202512.0093.v1

contexts and effectively manage domain-specific vocabulary and complex reasoning. While the F1 score for How questions (80.5) is slightly lower, this is expected given the multi-step reasoning and procedural depth typically involved in such questions. The performance on the MedQuAD dataset is even more impressive.

The CA-ExpQA model achieves F1 scores exceeding 96.0 in all categories, with near-perfect scores of 99.7 for Where and 99.8 for When questions. This demonstrates the model's scalability and robustness in handling a wide variety of biomedical questions. Its ability to maintain consistent performance across different datasets reflects the strength of domain-specific linguistic rules and the use of BioBERT embeddings.

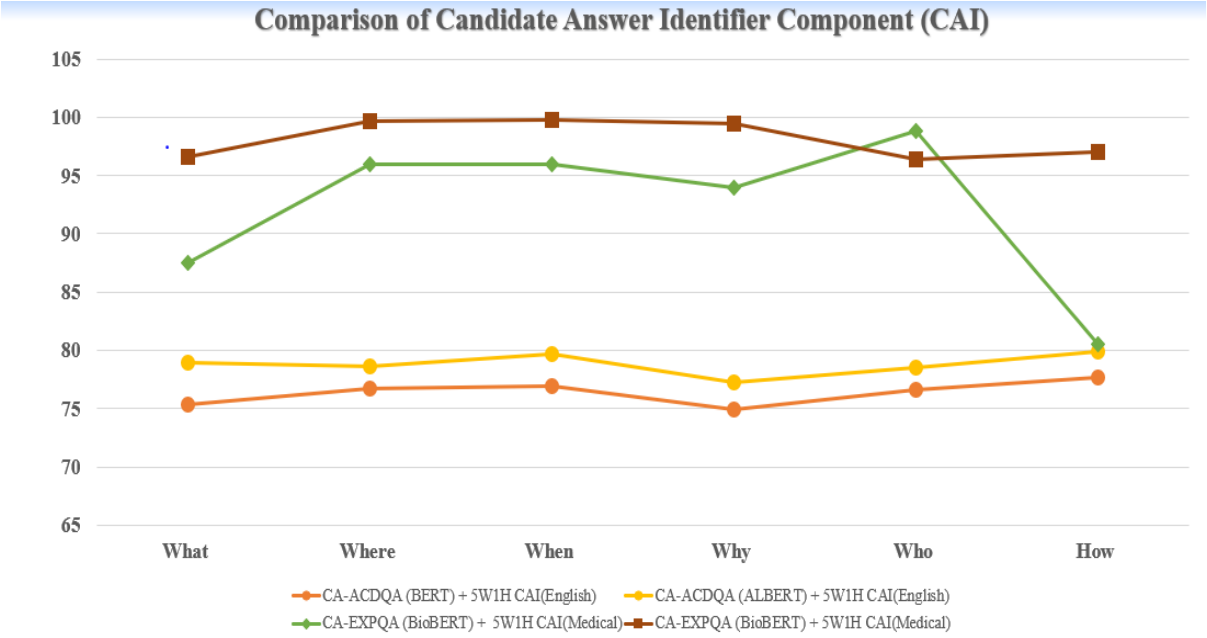

Figure 8. Graphical analysis of Performance Comparison (F1 score) of CA-ExpQA model for each*question*category with*proposed Candidate*Answer*Identifier (Customized*Giveme5W1H with medical format) with CA-AcdQA.

The graphical analysis in Figure 11 further emphasizes the advantage of using the medical-specific CAI module in CA-ExpQA. Its performance consistently exceeds that of CA-AcdQA across all categories, with especially large margins—between 15 to 20 F1 points—in *Who*, *Where*, and *When* questions. On the MedQuAD dataset, the CA-ExpQA model's curve shows minimal variation, reaffirming its adaptability to a wide range of medical topics.

In conclusion, the results validate the effectiveness of integrating a customized CAI module tailored to biomedical data. The CA-ExpQA model, supported by medical-specific rules and BioBERT embeddings, offers a significant improvement in accuracy and reliability over general-purpose models. This domain-aware approach addresses the unique challenges of biomedical question answering, making CA-ExpQA a robust solution for practical applications in the medical domain.

B. Results of RNN-Attention Module:

The results in Table 4 compare the*performance of*RNN-based and CNN-based*mechanisms for candidate answer identification in biomedical question answering (QA). Two datasets, COVID-QA and MedQuAD, were used for evaluation, with performance measured using Exact*Match (EM) and F1-score*metrics. In this comparison, the CNN-based mechanisms are implemented in the CA-AcdQA models using BERT and ALBERT, while the CA-ExpQA model incorporates an RNN-based mechanism, offering insights into the relative strengths and limitations of each architectural choice.

 doi:10.20944/preprints202512.0093.v1

The CA-AcdQA models demonstrate moderate performance on the COVID-QA dataset, with EM scores ranging from 55.6 to 57.5 and F1 scores between 76.4 and 78.8. These results reflect the limited capacity of CNN mechanisms to capture sequential dependencies in text. While CNNs are effective in identifying localized patterns, they struggle to represent long-range contextual relationships, which are essential in understanding complex biomedical queries. This architectural limitation impacts the overall ability of CNN-based models to reason effectively across multiple tokens in medical texts.

By contrast, the CA-ExpQA model, which utilizes an RNN-based mechanism along with BioBERT embeddings and self-attention, significantly outperforms the CNN-based counterparts. On*the*COVID-QA dataset, it achieves an*EM*score*of 93.1 and*an*F1*score of 95.0. This dramatic improvement is primarily due to the RNN's ability to model the sequential nature of language, allowing it to better capture context and relationships between terms. The performance indicates that the model is more effective in processing complex medical queries that involve reasoning across multiple steps.

On the MedQuAD dataset, which includes a wider range of medical topics and query types, the CA-ExpQA model continues to perform strongly, achieving an EM score of 87.1 and an F1 score of 92.3. This consistent performance across datasets reflects the adaptability and robustness of the RNN-based mechanism in handling diverse biomedical content. In contrast, the CNN-based models show lower performance, further emphasizing the architectural limitations of CNNs in generalizing to different types of medical data.

Table 4. The*effect*of*RNN*Mechanism*on*the*proposed model on*all*medical*datasets.

*MODEL	DATASET	EVALUATION METRICES	
		*EM	F1-SCORE
CA-ACDQA (BERT) With CNN-Mechanism	COVID-QA	55.6	76.4
CA-ACDQA (ALBERT) With CNN-Mechanism	COVID-QA	57.5	78.8
CA-EXPQA (BioBERT) With RNN-Mechanism	COVID-QA	93.1	95.0
CA-EXPQA (BioBERT) With RNN-Mechanism	MEDQUAD	87.1	92.3

The graphical representation in Figure 9 visually underscores the performance gap between CNN-based and RNN-based mechanisms. The CNN-based models show minimal variance between configurations such as BERT and ALBERT. In contrast, the RNN-based CA-ExpQA model exhibits a significant leap in both EM and F1 scores, particularly on the COVID-QA dataset. This clearly illustrates the importance of capturing long-term dependencies and preserving contextual relevance in biomedical QA, which are effectively addressed by the RNN architecture.

 doi:10.20944/preprints202512.0093.v1

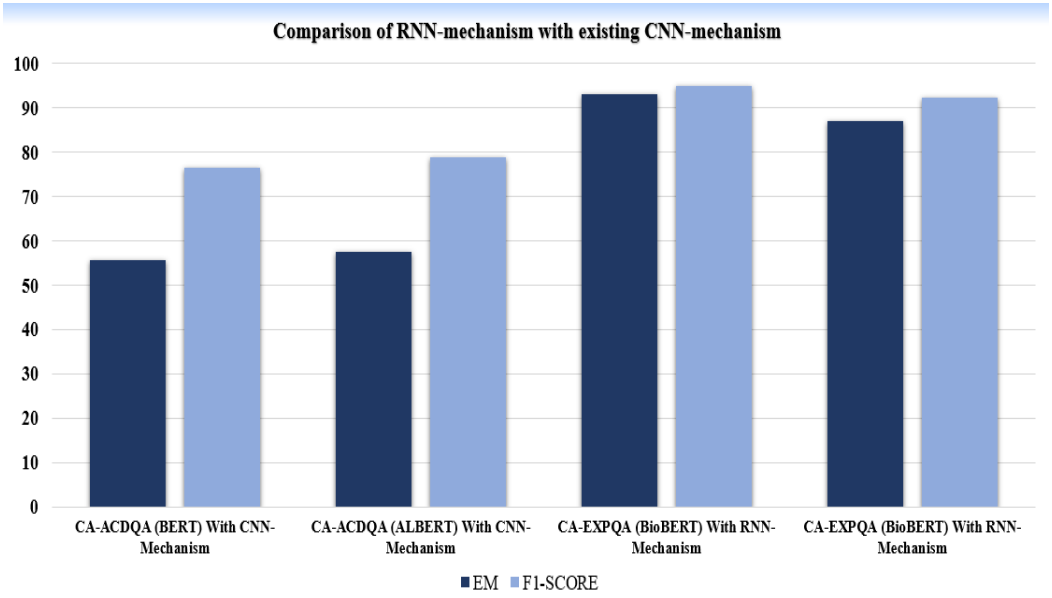

Figure 9. Comparison of RNN-mechanism with existing CNN-mechanism.

In conclusion, the comparative results validate the effectiveness of using RNN-based mechanisms over CNN-based approaches for candidate answer identification in biomedical QA tasks. The ability of RNNs to model sequential dependencies, combined with domain-specific embeddings and self-attention mechanisms, enables the CA-ExpQA model to achieve high accuracy and reliability. These findings highlight the critical role of advanced, context-aware architectures in specialized domains such as biomedical question answering, where precision and depth of understanding are essential.

C. Results of Question Expansion Module:

The comparative analysis in *Table 5 examines the performance of the*Question*Expansion*(QE) module across different models and datasets, specifically assessing the impact of using a general-purpose English thesaurus versus a medical thesaurus. The evaluation relies on two standard metrics, Exact*Match* (EM) *and*F1-score, to quantify how effectively QE module enhances answer retrieval by resolving ambiguities in user queries.

Table 5. The effect of Question Expansion (EQ) tailored with medical thesaurus on all medical datasets.

MODEL	DATASET	EVALUATION METRICES	
		EM	F1-SCORE
CA-ACDQA (BERT) QE with English thesaurus	COVID-QA	55.6	76.4
CA-ACDQA (ALBERT) QE with English thesaurus	COVID-QA	57.5	78.8
CA-EXPQA (BioBERT) QE with medical thesaurus	COVID-QA	77.6	83.0
CA-EXPQA (BioBERT) QE with medical thesaurus	MEDQUAD	75.7	81.2

The CA-AcdQA models, which incorporate QE with an English thesaurus, exhibit moderate performance on the COVID-QA dataset. These models achieve EM scores between 55.6 and 57.5, and

 doi:10.20944/preprints202512.0093.v1

F1 scores ranging from 76.4 to 78.8. While these results are consistent, they also highlight a key limitation: the use of a general-purpose thesaurus is insufficient for capturing the nuance and specificity of medical terminology. As a result, the models often fail to reformulate queries in ways that align effectively with domain-specific content.

In contrast, the CA-ExpQA model, which integrates a QE module tailored with a medical thesaurus, demonstrates notable performance improvements. On the COVID-QA dataset, the CA-ExpQA model*achieves*an*EM*score*of*77.6 and an*F1-score*of*83.0. This substantial gain reflects the ability of domain-specific expansions to resolve semantic ambiguities and enhance the match between reformulated queries and the relevant content within the dataset. The model's use of BioBERT embeddings further supports this alignment by providing rich contextual understanding suited for biomedical text.

A*similar*pattern*is*observed*on*the*MedQuADdataset, *where the CA-ExpQA*model*achieves*an EM*of 75.7*and*an*F1-score*of*81.2. Given that MedQuAD includes a broader variety of medical topics and question types,*these*results*underscore the*robustness*and adaptability*of the medical thesaurus*across*diverse scenarios. In comparison, the CA-AcdQA models, constrained by their reliance on general-purpose language tools, lack the specialization necessary to generalize effectively to complex medical contexts.

The graphical representation in Figure 10 further illustrates the performance disparity between models using an English thesaurus and those employing a medical thesaurus. The CA-ExpQA model consistently achieves the highest EM and F1 scores, with a 4–6% improvement in F1 performance over the CA-AcdQA models. This performance margin reinforces the importance of utilizing domain-specific resources when reformulating questions in specialized fields like biomedicine.

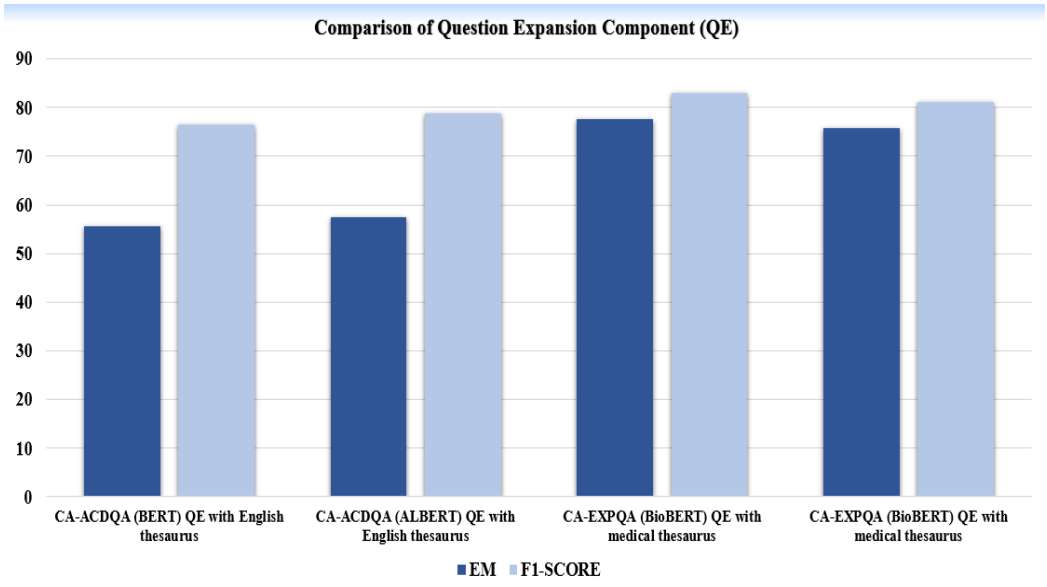

Figure 10. The visual effect of Question Expansion (EQ) tailored with medical thesaurus on all medical datasets by finding EM and F1 Score.

In conclusion, the integration of a question expansion module grounded in a medical thesaurus significantly enhances the performance of biomedical QA models. By addressing the inherent ambiguity in user queries and ensuring more precise contextual alignment, the CA-ExpQA model demonstrates superior accuracy and relevance. These findings highlight the necessity of domain-aware QE mechanisms for achieving robust and reliable performance in biomedical question answering systems.

 doi:10.20944/preprints202512.0093.v1

VI. Conclusion & Future Work

This research presents a significant advancement in Medical Question Answering (MQA) systems by overcoming existing limitations and achieving superior performance over current models. The system is trained on diverse and comprehensive medical datasets,*enabling*it*to deliver*accurate*and*contextually*relevant*answers across a wide range of medical queries, from general to highly specialized.

A key contribution is the use of BioBERT embeddings, which provide deep domain-specific understanding of medical language. Replacing CNNs with RNNs allows the model to better capture sequential dependencies and contextual nuances inherent in medical questions. Additionally, the incorporation of a question expansion module, leveraging medical thesauri like UMLS, enhances the system's ability to interpret complex terminology and ambiguous queries.

Performance evaluation through F1 scores and exact match metrics demonstrates the system's improved accuracy and reliability. These results highlight its practical value for healthcare professionals and patients, offering precise and meaningful medical information.

Future work includes expanding domain coverage to areas such as oncology and neurology, integrating larger and more varied datasets—including clinical records—to enhance generalizability, and adding multilingual support to reach broader populations. Further enhancements may involve personalized responses tailored to user history, voice-based interaction, and explainable AI features to increase transparency and user trust.

Together, these improvements will advance the system into a versatile, intelligent medical QA platform that supports clinical decision-making and empowers users with accessible, trustworthy health information worldwide. This research thus lays a robust foundation for ongoing innovation in medical question answering technology.

References

1. Saeed, N., Ashraf, H., & Jhanjhi, N. Z. (2023). Deep Learning-Based Question Answering System (Survey). Preprints, 2023121739.
2. Mutabazi, E., Ni, J., Tang, G., & Cao, W. (2021). A Review on Medical Textual Question Answering Systems Based on Deep Learning Approaches. Applied Sciences, 11(12), 25456.
3. Kia, M. A., Garifullina, A., Kern, M., Chamberlain, J., & Jameel, S. (2022). Adaptable Closed-Domain Question Answering Using Contextualized CNN-Attention Models and Question Expansion. IEEE Access.
4. Jili Qian et al., "A Liver Cancer Question-Answering System Based on Next-Generation Intelligence and the Large Model Med-PaLM 2," International Journal of Computer Science and Information Technology, 2024
5. Jafar A. Alzubi et al., "COBERT: COVID-19 Question Answering System Using BERT," Arabian Journal for Science and Engineering, 2023
6. Husamelddin A.M.N Balla et al., "Arabic Medical Community Question Answering Using ON-LSTM and CNN," ICMLC, 2022.
7. Dimitris Pappas et al., "Data Augmentation for Biomedical Factoid Question Answering," 2022.
8. Jimenez Eladio and Hao Wu, "emrQA-msquad: A Medical Dataset Structured with the SQuAD V2.0 Framework," Beijing Institute of Technology, 2024.
9. Juraj Vladika and Florian Matthes, "Improving Health Question Answering with Reliable and Time-Aware Evidence Retrieval," Technical University of Munich, 2024.
10. Rui Yang et al., "KG-Rank: Enhancing Large Language Models for Medical QA with Knowledge Graphs and Ranking Techniques," 2024.
11. Chengyu Huang and Feng Yao, "Research on Question-Answering System of Children's Disease Based on ALBERT," ISAIMS, 2023.
12. Yu-Hsuan Chang et al., "Interactive Healthcare Robot Using Attention-Based Question-Answer Retrieval and Medical Entity Extraction Models," IEEE Journal of Biomedical and Health Informatics, 2023.

 doi:10.20944/preprints202512.0093.v1

13. Shrutikirti Singh and Seba Susan, "Healthcare Question-Answering System: Trends and Perspectives," 2023.
14. Prateek Chhikara et al., "Privacy-Aware Question-Answering System for Online Mental Health Risk Assessment," University of Southern California, 2023.
15. Yun Zhao et al., "JMS-QA: A Joint Hierarchical Architecture for Mental Health Question Answering," IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2024.
16. Baivab Das and S. Jaya Nirmala, "Improving Healthcare Question Answering System by Identifying Suitable Answers," 2022 IEEE MysuruCon.
17. Gezheng Xu et al., "External Features-Enriched Model for Biomedical Question Answering," BMC Bioinformatics, 2021.
18. Gregory Kell et al., "Question Answering Systems for Health Professionals at the Point of Care—A Systematic Review," Journal of the American Medical Informatics Association, 2024.
19. Prateek Chhikara et al., "Privacy Aware Question-Answering System for Online Mental Health Risk Assessment," 2023.
20. Singh, S., & Susan, S. (2023). Healthcare Question–Answering System: Trends and Perspectives. Springer.
21. Xu, G., Rong, W., Wang, Y., Ouyang, Y., & Xiong, Z. (2021). External Features Enriched Model for Biomedical Question Answering. BMC Bioinformatics.
22. Yagnik, N., Jhaveri, J., Sharma, V., & Pila, G. (2024). MedLM: Exploring Language Models for Medical Question Answering Systems. arXiv.
23. Das, B., & Nirmala, S. J. (2022). Improving Healthcare Question Answering System by Identifying Suitable Answers. IEEE MysuruCon.
24. Pergola, G., Kochkina, E., Gui, L., Liakata, M., & He, Y. (2021). Boosting Low-Resource Biomedical QA via Entity-Aware Masking Strategies. arXiv.
25. Peng, K., Yin, C., Rong, W., Lin, C., Zhou, D., & Xiong, Z. (2022). Named Entity Aware Transfer Learning for Biomedical Factoid Question Answering. IEEE/ACM Transactions on Computational Biology and Bioinformatics.
26. Pappas, D., Malakasiotis, P., & Androutsopoulos, I. (2022). Data Augmentation for Biomedical Factoid Question Answering