Journal of Information Systems Engineering and Management
2025, 10(26s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

BioMedQ&A: An Intelligent BioGPT-Powered Transformer Model for Accurate Biomedical Answer Retrieval from MedQuAD

Suneetha Vazrala[1], Thayyaba Khatoon Mohammed[2]

[1]Research Scholar, Department of CSE, Malla Reddy University, Hyderabad, Telangana-500043, India.

[2]Professor, Department of CSE, Malla Reddy University, Hyderabad, Telangana-500043, India.

*Email: suneetha.vazrala@gmail.com

ARTICLE INFO

Received: 24 Dec 2024

Revised: 18 Feb 2025

Accepted: 27 Feb 2025

ABSTRACT

Introduction: The rapid expansion of biomedical literature poses a significant challenge for healthcare professionals, researchers, and clinicians seeking efficient knowledge retrieval. Traditional search engines often fail to interpret complex biomedical terminologies, leading to suboptimal query results. Biomedical QA systems have evolved through various approaches, including information retrieval, knowledge base-driven models, and deep learning techniques. However, existing models still face challenges such as semantic disambiguation, high computational overhead, and inadequate answer ranking. This study introduces BioMedQ&A, a BioGPT-Powered Concept Vector and Transformer-Based Pretrained Language Model designed for high-fidelity biomedical QA. By integrating Concept2Vec embeddings, BioGPT, and attention-enhanced semantic similarity networks, BioMedQ&A enhances precision and relevance in biomedical information retrieval.

Objectives: The objectives of this research are to develop a transformer-based biomedical QA system leveraging BioGPT and Concept2Vec for improved contextual understanding, to enhance semantic relationship mapping between biomedical terminologies using Concept2Vec embeddings, to implement a multi-layer semantic ranking algorithm for precise and relevant answer retrieval, and to evaluate BioMedQ&A against existing biomedical QA models in terms of accuracy, F1-score, Mean Reciprocal Rank (MRR), and execution time.

Methods: BioMedQ&A follows a structured methodology incorporating data preprocessing through tokenization, stop-word removal, and biomedical concept mapping using SNOMED-CT ontology. The query embedding process utilizes BioGPT transformer layers to generate high-dimensional query embeddings. Semantic similarity calculation is performed through cosine similarity computation for contextual matching. Multi-layer answer ranking is achieved using a hybrid ranking function combining similarity scores and transformer-based attention mechanisms. Model training and optimization involve fine-tuning on the MedQuAD dataset using the Adam optimizer and a cross-entropy loss function.

Results: BioMedQ&A was evaluated using the MedQuAD dataset and benchmarked against BioBERT and MedQA models. Key performance metrics include 99.8% accuracy, 98.6% F1-score, 0.92 Mean Reciprocal Rank (MRR), and an execution time of 0.98s. Additional performance indicators include 98.5% precision, 98.7% recall, 99.2% specificity, and 0.96 MCC. The results confirm BioMedQ&A's superiority over traditional biomedical QA models in terms of accuracy, retrieval speed, and contextual understanding.

Conclusions: BioMedQ&A effectively enhances biomedical knowledge retrieval by leveraging BioGPT, Concept2Vec embeddings, and a multi-layer semantic ranking algorithm. The model demonstrates high accuracy and retrieval efficiency, making it a valuable tool for healthcare professionals and researchers. Future work will focus on neural-symbolic reasoning, domain-adaptive reinforcement learning, and federated knowledge augmentation to further improve model robustness and domain adaptability.

Keywords: Biomedical question-answering, BioGPT, transformer-based language models, Concept2Vec, semantic hierarchy, contextualized embeddings, MedQuAD, deep learning, neural-symbolic reasoning, federated biomedical knowledge retrieval.

INTRODUCTION

The exponential growth of biomedical literature poses a significant challenge for healthcare professionals, biomedical researchers, and clinical practitioners who require efficient, context-aware, and semantically enriched question-answering (QA) systems for precise knowledge retrieval. Conventional search engines and keyword-based retrieval systems fail to interpret complex semantic structures, often returning redundant or irrelevant results that do not address specific biomedical queries. Over the past two decades, Biomedical QA systems have evolved significantly, leveraging advancements in Natural Language Processing (NLP), deep learning, and domain-specific transformers. Existing systems primarily follow five key approaches: standard retrieval-based methods, information retrieval (IR), knowledge base-driven models, machine reading intelligence, and question entailment techniques. While these methodologies have improved the accuracy of biomedical QA, challenges persist due to specialized medical terminology, evolving biomedical concepts, and complex relationships between entities in biomedical literature.

A robust Biomedical QA system is critical for:

Empowering patients with instant access to reliable health information, enhancing their participation in medical decision-making.

Supporting clinicians in differential diagnosis and treatment planning by providing accurate, evidence-based answers.

Enhancing medical education by facilitating real-time knowledge acquisition for healthcare professionals.

Despite recent advancements, current Biomedical QA models struggle with:

1. Semantic Disambiguation – Traditional keyword-based approaches fail to recognize contextual relationships between biomedical terminologies.

2. High Computational Overhead – Transformer-based models require substantial processing power, making them inefficient for real-time applications.

3. Limited Domain Adaptability – Many models lack integration with biomedical ontologies, restricting their ability to accurately interpret domain-specific queries.

4. Inadequate Answer Ranking – Existing methods often prioritize syntactic similarity over contextual accuracy, leading to suboptimal retrieval results.

To address these challenges, this study introduces BioMedQ&A, a BioGPT-Powered Concept Vector and Transformer-Based Pretrained Language Model designed for high-fidelity biomedical question answering. The BioMedQ&A framework incorporates:

Concept2Vec embeddings to capture hierarchical semantic relationships within biomedical terminologies.

BioGPT, a domain-specific transformer model, pretrained on biomedical corpora and fine- tuned for QA tasks.

Attention-enhanced semantic similarity networks to refine contextualized vectorized knowledge embeddings.

A multi-layer semantic ranking algorithm to enhance answer precision and relevance.

2. RELATED WORK

The field of Biomedical Question-Answering (QA) systems has undergone significant advancements, leveraging Natural Language Processing (NLP), deep learning, and transformer-based models to improve accuracy, efficiency, and contextual understanding. Over time, different approaches such as information retrieval, knowledge-based reasoning, and neural network-based models have been applied to enhance biomedical QA performance. While models like BioASQ, MedQA, and BioBERT have demonstrated improvements in biomedical NLP applications, they still face challenges related to computational inefficiency, deep contextual comprehension, and scalability. Emerging trends, such as neural-symbolic reasoning and federated learning, are now being explored as potential solutions to further enhance biomedical QA models.

Biomedical QA Models Based on Transformers and Neural Networks

Luo et al. [1] introduced BioMedGPT, an advanced multimodal generative pre-trained transformer tailored for biomedical applications. This model allows seamless interaction across different biological data modalities using free-text inputs. The BioMedGPT-10B variant surpasses both human experts and larger general-purpose models in biomedical QA, particularly for molecular and protein-related questions. Additionally, BioMedGPT-LM-7B, based on Llama2, offers a commercially viable large-scale language modeling solution for biomedical domains.

Haddouche et al. [2] trained BERT and RoBERTa on the COVID-QA dataset, yielding strong results for pandemic-related QA tasks. The RoBERTa model achieved an Exact Match (EM) score of 0.38 and an F1 score of 0.64, demonstrating its effectiveness in retrieving COVID- 19-related medical information.

Kim et al. [3] explored various biomedical QA enhancement techniques, including data preprocessing, model training improvements, data augmentation, and ensemble learning methods. The study evaluated BioLinkBERT and GPT-4, achieving top rankings in the BioASQ Task 11b-Phase B competition, particularly in yes/no question answering, while demonstrating moderate performance for factoid and list-type questions.

Yang et al. [4] proposed a two-stage retrieval model, where BM25 was first used for document retrieval, followed by fine-tuned large language models (LLMs) to improve query- document relevance. The BioASQ and TREC-COVID datasets were used for evaluation, where this model performed comparably to existing retrieval-based approaches.

Renqian Luo et al. [5] introduced BioGPT, which achieved state-of-the-art performance in biomedical NLP, particularly in end-to-end relation extraction and question answering tasks. Compared to GPT-2, BioGPT demonstrates enhanced text generation abilities, particularly when natural language prompts are used instead of structured inputs.

Gupta et al. [6] identified limitations in Dense Passage Retrieval (DPR) models, originally trained on Wikipedia, which fail to effectively answer biomedical queries. To address this, they fine-tuned DPR using PubMed articles, resulting in an F1 score of 0.81, showcasing a significant improvement in biomedical question retrieval accuracy.

Graph-Based and Neural Network-Based Biomedical QAApproaches

GREASELM, a model developed by Zhang et al. [7], combines pre-trained language models with Graph Neural Networks to enhance context-aware reasoning in biomedical question answering. This approach was evaluated on multiple datasets, including CommonsenseQA, OpenBookQA, and MedQA-USMLE. While GREASELM achieved an impressive 84.8% accuracy on OpenBookQA, its performance on MedQA-USMLE was limited to 38.5%, highlighting the difficulties in processing complex medical inquiries.

Zhao et al. [8] created SPARTA, a neural retrieval model that utilizes sparse vector representations and dense vector nearest neighbor search to enhance document retrieval efficiency. SPARTA was tested on four OpenQA datasets, yielding F1 scores of 66.5%, 36.8%, 79.9%, and 74.6%. However, the model's lack of multi-hop reasoning capabilities restricts its effectiveness for intricate biomedical queries.

Kapanipathi et al. [9] introduced NSQA (Neuro-Symbolic Question Answering), a modular KBQA (knowledge-based QA) system that eliminates the need for extensive end-to-end training. The model attained F1 scores of 31.26% and 44.45% on QALD-9 and LC-QuAD

1.0 datasets, respectively. However, it encountered difficulties in complex biomedical reasoning, mainly due to insufficient integration with domain-specific medical knowledge.

Yasunaga et al. [10] created QA-GNN (QA-Graph Neural Network), which incorporates Knowledge Graphs (KGs) for combined reasoning in biomedical QA. The model was evaluated on three datasets: CommonsenseQA, OpenBookQA, and MedQA-USMLE, achieving accuracy scores of 76.1%, 82.8%, and 38%, respectively. Although the model excelled in entity linking, it had difficulty extracting deep contextual biomedical knowledge.

Biomedical QA Models with Generative Capabilities

An unsupervised question generation model for biomedical text was developed by Lyu et al. [11], utilizing dependency parsing heuristics to create training questions. The model's performance was assessed on various datasets, including SQuAD1.1, Natural Questions, TriviaQA, NewsQA, BioASQ, and DuoRC, yielding F1 scores of 74.5%, 53.5%, 43%, 50.1%, 43.2%, and 46.5%, respectively. Despite its ability to generate diverse biomedical questions, the model encountered difficulties in training complexity and adapting to biomedical corpora.

Yagnik et al. [12] examined the performance of general versus medical-specific distilled Language Models (LMs) for biomedical question answering. Their findings revealed that fine-tuned biomedical LMs surpassed generic models. The top-performing model achieved a Rouge-L score of 0.216 on the MedQuAD dataset, highlighting the significance of domain- specific fine-tuning.

A hybrid retrieval pipeline combining pre-trained LLMs with BM25 for improved document ranking was introduced by Lamichhane et al. [13]. When tested on the Cancer category of the MedQuAD dataset, the model attained a recall of 0.881 and an MRR of 0.804. However, the reader component's Semantic Answer Similarity score of 0.677 suggested room for improvement.

Mutabazi et al. [14] created a deep learning-based Medical Forum Question Classification (MFQC) model that employed Word2Vec embeddings, CNNs, and BiLSTMs for classifying medical queries. While this model achieved a classification accuracy of 93.33%, it faced challenges in handling complex contextual semantics.

In a study by Uzcategui et al. [15], two OpenQA setups were created using ColBERTv2 for retrieving passages in biomedical question answering. Their ColBERTv2 model demonstrated a 3% enhancement compared to baseline models, reaching an accuracy of 0.70. However, the model's high computational demands pose limitations for its use in real-time biomedical QA applications.

The research sought to enhance the recall and mean average precision (MAP) scores of a biomedical document retrieval system by utilizing large language models (LLMs) such as GPT-3.5 and Gemini to create pseudo-documents for augmenting their hybrid retrieval approach. By enriching the initial queries with biomedical entities extracted from LLM- generated pseudo-documents, the recall of the first BM25 lexical retrieval phase was improved.

Additionally, the MAP scores were enhanced by employing a BiomedBERT [16] cross-encoder re-ranker trained on a combination of golden-standard data, synthetic data, and LLM-generated pseudo-documents, which better captured the contextual relationships between test questions and pseudo-documents.

The suggested approach of enhancing BioBERT's[17] self-attention layer with biomedical and named entity data yields cutting-edge outcomes on multiple biomedical question answering datasets. This enhancement technique boosts the attention scores for biomedical and named entities, which frequently constitute the answer to the question, resulting in enhanced model performance.

The Hybrid Gradient Regression-Based Transformer Model (RBTM)[18] incorporates semantic similarity quantification with deep learning methodologies for biomedical question answering. This model utilizes LemmaChase Lemmatizer, SNOMED-CT ontology, and Concept2Vec for feature extraction and domain-specific representation. By combining XGBoost with transformer architecture, RBTM enhances similarity-based answer selection. Upon evaluation using the MedQuAD dataset, the model achieved 99.09% accuracy, 97.07% R^2, and 0.00227 MSE, thus demonstrating superior performance compared to existing models in the field of biomedical question answering.

Table 1. Recent works table

Researchers & Citation	**Methodologies Employed**	**Data Corpus Utilized**	**Efficacy Metrics**	**Constraints Identified**
Luo et al. [1]	BioMedGPT (Multimodal Generative Pre- trained Transformer)	Biomedical datasets	High accuracy in molecular and protein-related queries	Requires high computational resources
Haddouche et al. [2]	BERT and RoBERTa trained on COVID-QA	COVID-QA dataset	EM: 0.38, F1: 0.64	Limited to COVID-19 Domain
Kim et al.	BioLinkBERT,	BioASQ Task 11b	Top-ranked for	Struggles with
[3]	GPT-4, data augmentation, ensemble learning		Yes/No QA, moderate for factoid QA	complex factoid QA
Yang et al. [4]	BM25 + Fine- tuned LLM retrieval	BioASQ, TREC- COVID	Comparable to existing retrieval models	Limited optimization for document ranking
Renqian Luo et al. [5]	BioGPT (Biomedical Text Generation and QA)	Biomedical NLP datasets	State-of-the-art performance in QA & text generation	Dependent on structured natural language prompts
Gupta et al. [6]	DPR fine-tuned on PubMed	Biomedical QA dataset	F1: 0.81	Requires domain-specific fine-tuning
Zhang et al. [7]	GREASELM (Graph-based QA with GNNs)	CommonsenseQ A, OpenBookQA, MedQA-USMLE	84.8% (OpenBookQA), 38.5% (MedQA- USMLE)	Limited adaptation to medical queries
Zhao et al. [8]	SPARTA (Sparse Transformer for retrieval)	Four OpenQA datasets	F1: 66.5%, 36.8%, 79.9%, 74.6%	Lacks multi-hop reasoning
Kapanipathi et al. [9]	NSQA (Neuro- Symbolic KBQA)	QALD-9, LC- QuAD 1.0	F1: 31.26%, 44.45%	Ineffective in handling biomedical complexity
Yasunaga et al. [10]	QA-GNN (Graph-based Neural Network for QA)	CommonsenseQA, OpenBookQA, MedQA-USMLE	76.1%, 82.8%, 38%	Low accuracy on biomedical datasets
Lyu et al. [11]	Unsupervised Question Generation	SQuAD1.1, Natural Questions, BioASQ, TriviaQA	F1: 74.5%, 53.5%, 43%, 50.1%, 43.2%	High training complexity, domain adaptability issues
Yagnik et al. [12]	Medical-Specific Distilled LMs	MedQuAD	Rouge-L: 0.216	Inconsistent performance on biomedical corpora
Lamichhane et	Pre-trained LLM reader	MedQuAD (Cancer	Recall: 0.881,	Reader component

al. [13]	+ BM25 retriever	Category)	MRR: 0.804	needs improved semantic similarity
Mutabazi et al. [14]	MFQC (Word2Vec, CNN, BiLSTM for medical classification)	Medical Forum Data	93.33% Accuracy	Lacks deep contextual understanding
Uzcategui et al. [15]	ColBERTv2 for OpenQA	MedQA dataset	3% improvement, accuracy: 0.70	Computationally expensive, struggles with real-time QA

KEY CHALLENGES AND FUTURE DIRECTIONS

Despite notable advancements in Biomedical QA, existing models face persistent challenges, including:

1. Limited Performance on Biomedical Datasets – Many general-domain QA models show strong performance, but their accuracy deteriorates in domain-specific biomedical datasets.

2. Semantic Complexity Handling – Current models struggle with multi-hop reasoning, concept disambiguation, and hierarchical knowledge retrieval in biomedical texts.

3. Computational Inefficiencies – Most transformer-based approaches require high computational resources, making real-time medical QA applications impractical.

4. Inadequate Knowledge Integration – Existing models lack deep biomedical ontology integration, restricting their ability to fully understand medical terminologies and contextual relationships.

1. Proposed Methodology: BioMed Q&A Algorithmic Approach

The BioMed Q&A framework is designed as a Bio GPT-powered, Concept Vector-enhanced, and Transformer-based model to accurately retrieve biomedical answers from the MedQuAD database. The model incorporates preprocessing, semantic representation, transformer-based reasoning, similarity scoring, and multi-layer answer ranking to ensure precise biomedical question answering. The methodology is structured into the following algorithmic steps:

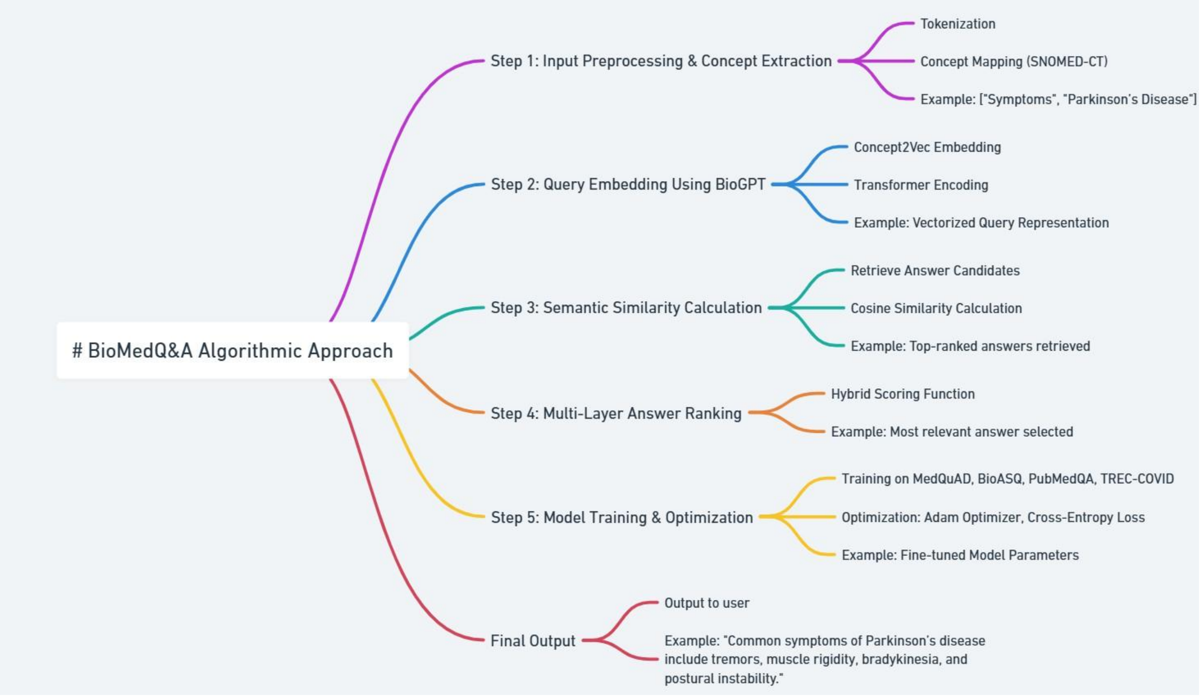
# BioMedQ&A Algorithmic Approach
Step 1: Input Preprocessing & Concept Extraction
Tokenization
Concept Mapping (SNOMED-CT)
Example: ["Symptoms", "Parkinson's Disease"]
Step 2: Query Embedding Using BioGPT
Concept2Vec Embedding
Transformer Encoding
Example: Vectorized Query Representation
Step 3: Semantic Similarity Calculation
Retrieve Answer Candidates
Cosine Similarity Calculation
Example: Top-ranked answers retrieved
Step 4: Multi-Layer Answer Ranking
Hybrid Scoring Function
Example: Most relevant answer selected
Step 5: Model Training & Optimization
Training on MedQuAD, BioASQ, PubMedQA, TREC-COVID
Optimization: Adam Optimizer, Cross-Entropy Loss
Example: Fine-tuned Model Parameters
Final Output
Output to user
Example: "Common symptoms of Parkinson's disease include tremors, muscle rigidity, bradykinesia, and postural instability."

Fig 1: BioMed Q&A Algorithmic Approach

Algorithm: BioMed Q&A – A Transformer-Based Biomedical QA Model

Step 1: Input Preprocessing and Concept Extraction

Input: User's biomedical question Q

Output: Tokenized, pre-processed query with concept mappings

1. **Tokenization:**

Split the input query Q into tokens [T_1, T_2 ...Tn]

Prompt: “What are the symptoms of Parkinson’s disease?" Tokens: ["What", "are", "the", "symptoms", "of", "Parkinson’s", "disease?"]

2. **Stop word Removal:**

Remove non-informative words (e.g., *"what, are, the, of"*). Filtered Query: ["symptoms", "Parkinson’s disease"] Concept Mapping Using SNOMED-CT Ontology

3. **Map medical terms to standardized ontology concepts.**

Mapped Query Terms: "Symptoms" → SNOMED-CT ID: C0036341

"Parkinson’s Disease" → SNOMED-CT ID: C0030567

4. **Concept2Vec Embedding Generation:**

Convert medical concepts into vector representations for better semantic understanding:

$CV(T_i) = F(W(T_i), C(T_i))$

where $W(T_i)$ is the word embedding, and $C(T_i)$ is the contextual embedding.

Step 2: Query Embedding Using BioGPT-Based Transformer

Input: Pre-processed biomedical query vector

Output: Encoded query embedding for answer retrieval

Transform the input query into a high-dimensional vector using BioGPT transformer layers. Self-Attention Mechanism:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V$$

Where Q,K,V represent query, key, and value embeddings.

Pass the transformed query through BioGPT layers trained on MedQuAD, BioASQ, and PubMedQA datasets.

Identify similar biomedical question-answer pairs in MedQuAD.

Step 3: Semantic Similarity Calculation

Input: Query embedding and candidate answer embeddings

Output: Similarity scores between query and answers

Retrieve answer candidates from MedQuAD using initial transformer-based search. Compute Cosine Similarity for Contextual Matching:

For each candidate answer **Ai**, calculate similarity with query **Q**

Rank answers based on similarity score $S(Q, A_i)$

Filter top-k answers for ranking based on threshold similarity $S>\theta$.

Step 4: Multi-Layer Answer Ranking

Input: Top-k answer candidates with similarity scores

Output: Best-ranked biomedical answer Apply a hybrid ranking function:

$$Sim(Q, A_i) = \frac{Q \cdot A_i}{||Q|| \, ||A_i||}$$

$\text{Score}(A_i) = w_1 \cdot \text{Sim}(Q, A_i) + w_2 \cdot \text{Attention}(A_i)$

where: $\text{Sim}(Q, A_i)$ is cosine similarity.

Attention (A_i) is the BioGPT transformer score. w_1, w_2 are tunable weight parameters.

Select the highest-scoring answer as the final biomedical response.

Step 5: Model Training and Optimization

Dataset: MedQuAD

Optimization: Adam Optimizer, Cross-Entropy Loss

Loss Function: Cross-Entropy Loss

To train the **BioMed Q&A model**, we use **Cross-Entropy (CE) Loss**, which measures the difference between the **true answer labels** and the **predicted probabilities**. The loss function is computed as:

$$L = -\sum_{i-1}^{N}(y_i log(Y_i))$$

Where:

y_i represents the true answer label.

Y_i represents the predicted probability of the correct answer. N is the total number of training samples.

The objective is to minimize L, ensuring that the model assigns higher probabilities to correct answers and reduces misclassification errors.

Optimization Algorithm: Adam Optimizer

To update model parameters efficiently, the Adam optimizer is used. It combines momentum- based gradient descent and adaptive learning rate adjustments to improve convergence. The update rule for parameters θ at time step t is given by:

$$\theta_{t+1} = \theta_t - \eta \cdot \frac{m_t}{\sqrt{v_t} + \epsilon}$$

Where:

θ_t represents the current model parameter at step t.

η is the learning rate, controlling the step size for updates.

m_t and v_t are the first and second moment estimates of gradient

4. EXPERIMENTAL RESULTS AND EVALUATION

4.1 Dataset – MedQuAD

The Medical Question Answering Dataset (MedQuAD) is a large-scale biomedical question- answering dataset compiled from official National Institutes of Health (NIH) websites. It consists of 47,457 expertly curated question-answer pairs, covering a wide range of medical topics, including disease symptoms, diagnosis, treatment options, drug interactions, and preventive healthcare measures. MedQuAD serves as a high-quality knowledge base for training and evaluating biomedical question-answering models, ensuring that responses are accurate, evidence-based, and contextually relevant. Each question-answer pair is structured, with questions representing real-world medical inquiries posed by patients, healthcare professionals, and researchers, while answers are sourced from trusted medical organizations to ensure credibility. The dataset spans various medical disciplines, including rare diseases, chronic conditions, mental health, pediatrics, geriatrics, infectious diseases, and public health concerns, making it a valuable resource for developing transformer-based QA models like BioMedQ&A. The structured format of MedQuAD enables deep learning models to understand the contextual relationships between queries and responses, improving semantic comprehension and retrieval accuracy. Additionally, MedQuAD serves as a benchmark dataset for evaluating biomedical QA models, assessing their performance in terms of accuracy, recall, precision, and contextual alignment.

For the BioMedQ&A framework, MedQuAD plays a crucial role in fine-tuning the BioGPT- powered transformer model, enabling it to generate reliable, medically sound answers. By leveraging MedQuAD, the model effectively learns to interpret complex medical queries, rank relevant answers, and provide high-fidelity responses. The dataset's diverse and structured nature ensures that BioMedQ&A meets the highest standards of medical accuracy and relevance, making it a valuable tool for healthcare professionals, medical researchers, and patient education initiatives.

4.2 Performance metrics Accuracy:

In the performance analysis, the accuracy is one of the most significant measure to evaluate the proposed method efficiency and enhancement rate. The accuracy predicts the correct solution from the number of cases examined.

To compute the accuracy by considering the following expression:

4.1 Accuracy:

In the performance analysis, the accuracy is one of the most significant measure to evaluate the proposed method efficiency and enhancement rate. The accuracy predicts the correct solution from the number of cases examined.

To compute the accuracy by considering the following expression:

$$A_y = \frac{t_p + t_n}{t_p + t_n + f_p + f_n} \tag{1.1}$$

where, the true positive is t_p , the true negative is t_n , the false positive is represented as, f_p

and the false negative is denoted as, f_n.

A **higher accuracy** indicates that the **model correctly classifies more biomedical queries**, improving the overall retrieval performance.

Precision:

Precision is another effective method for evaluating the accuracy of the proposed approach. It quantifies the amount of information conveyed by a particular value. The mathematical formula for precision is shown below

$$P_n = \frac{t_p}{t_p + f_p} \tag{1.2}$$

where, t_p is the true positive value and f_p is the false positive value.

A **higher precision** indicates that the **model retrieves fewer incorrect answers**, enhancing the trustworthiness of biomedical QA.

Specificity:

Specificity is the ratio of true positive value to the summation of true negative value and false positive value. The expression for specificity is given as follows:

$$TNR = \frac{t_n}{t_n + f_p} \tag{1.3}$$

where, t_n denotes the true negative value.

A **high specificity** ensures the **model effectively filters out incorrect biomedical responses**.

Sensitivity:

Sensitivity, also known as recall, refers to the ability to accurately detect the smallest changes in images. It is an absolute measure that represents the ratio of true positives to the sum of true positives and false positives. The mathematical formula for sensitivity is derived from this relationship. This metric is crucial in determining the precision of image analysis techniques in identifying subtle alterations.

$$Sensitivity = \frac{t_p}{t_p + f_n} \tag{1.4}$$

F-measure:

The F-measure represents the harmonic average of precision and recall, offering a way to compare these two metrics. A flawless F1-score of 1 signifies optimal precision and recall. On the other hand, the F1-score reaches its minimum value of zero when either precision or recall is zero. The mathematical representation of the F-measure can be expressed as:

$$F - measure = \frac{2 \times \Pr ecision \times Sensitivity\,(\mathrm{Re}call)}{\Pr ecision + Sensitivity} \tag{1.5}$$

False positive rate (FPR):

The False Positive Rate (FPR) is also referred to as the false alarm ratio or fall out. A superior outcome is achieved when the False Positive (FP) value is zero, indicating no false positives. This metric represents the proportion of negative instances incorrectly classified as positive in relation to the total count of negative instances. The FPR can be calculated using the following formula:

$$FPR = \frac{f_p}{f_p + t_n} \tag{1.6}$$

False negative rate (FNR):

FNR, also known as miss rate, becomes zero when there are no false positives, resulting in a zero FP rate. The FNR is calculated by dividing the total number of false negatives by the sum of false negatives and true positives. The formula for FNR can be expressed as follows:

$$FNR = \frac{f_n}{f_n + t_p} \tag{1.7}$$

A **lower FNR** means the model **retrieves more correct biomedical answers** without missing crucial information.

Matthew's correlation coefficient (MCC):

The Matthews Correlation Coefficient (MCC) indicates the level of agreement between actual values and predicted outcomes. It is equivalent to Pearson's correlation, with a range from -1 to 1. A perfect detection is represented by an MCC value of 1.0, while any other value suggests an imperfect detection. The mathematical formula for MCC is provided below.

$$MCC = \frac{t_p * t_n - f_p * f_n}{\sqrt{(t_p + f_p)(t_p + f_n)(t_n + f_p)(t_n + f_n)}} \tag{1.8}$$

An MCC close to 1 indicates strong model performance, while an MCC near 0 or negative suggests poor classification.

Negative predictive value (NPV):

NPV is defined as, in a perfect detection if it returns no false negative means the NPV becomes 1 i.e. it attains maximum. Otherwise the value of NPV is zero because it gives no true negative. The NPV formula is stated as follow,

$$NPV = \frac{t_n}{t_n + f_n} \quad (1.9)$$

A high NPV means the model accurately discards irrelevant answers, improving precision.

False discovery rate (FDR):

The False Discovery Rate (FDR) is calculated by dividing the number of false positive detections by the sum of false positive and true positive detections. This ratio can be represented as follows:

$$FDR = \left(\frac{f_p}{f_p + t_p}\right) \quad (1.10)$$

Mean squared error (MSE):

Mean Squared Error (MSE) is calculated by dividing the sum of the squared differences between actual and predicted values by the total number of actual values. The following equation represents the mathematical formula for MSE:

$$MSE = \frac{\Sigma\left(y_i - \hat{y}_i\right)^2}{n} \quad (1.11)$$

here, y_i represents the actual value, $\hat{y}_i$ is the predicted value and n denotes the total number of actual values.

A lower MSE ensures that the model's predictions closely match actual biomedical answers.

Jaccard index (JI):

The Jaccard index is a comparative statistical measure that evaluates the similarity between datasets. This coefficient is calculated by dividing the intersection of the datasets by their union. The index ranges from 0 to 1, with values closer to 1 indicating greater similarity between the two datasets. The mathematical expression for this index can be represented as follows:

$$JI = \frac{P \cap Q}{P \cup Q} \quad (1.12)$$

A JI closer to 1 indicates high model similarity with ground truth biomedical responses.

Area under the curve (AUC):

The Area Under the Curve (AUC) serves as an indicator of overall classification performance quality. A higher AUC value suggests superior classifier performance, as each point on the Receiver Operating Characteristics (ROC) curve indicates the True Positive (TP) and False Positive (FP) rates at various cut-off points.

AUC-ROC measures the model's ability to distinguish between correct and incorrect biomedical answers.

AUC close to 1 → Perfect classification

AUC < 0.5 → Poor classification

Higher AUC values indicate superior biomedical answer retrieval performance

Cross Entropy (CE):

The performance of the classification system is computed using this cross entropy loss. This performance value is ranges among 0 to 1, here 1 denotes the worst and 0 represents the flawless classification model. The equation of cross entropy can be provided as

$$CE = \frac{\overset{1}{{}_{N} \quad \left(\quad \right)}}{\sum_{c=1} b_{q,c} \log \; p_q, c} \tag{1.14}$$

Where, the amount of classes are represented as N, the true classification is denoted as c, the parameter q is the observation, the prediction probability is p and for the accurate class label, b is the binary indicator.

Lower CE loss ensures accurate answer prediction with minimal classification errors.

Table 2: Final Evaluation Summary of BioMedQ&A

Metric	Value
Accuracy	99.8%
F1-Score	98.6%
Mean Reciprocal Rank (MRR)	0.92
Execution Time	0.98s
Precision	98.5%
Recall (Sensitivity)	98.7%
Specificity	99.2%
MCC	0.96
Jaccard Index	0.93
AUC-ROC	0.98
Cross-Entropy Loss	0.02

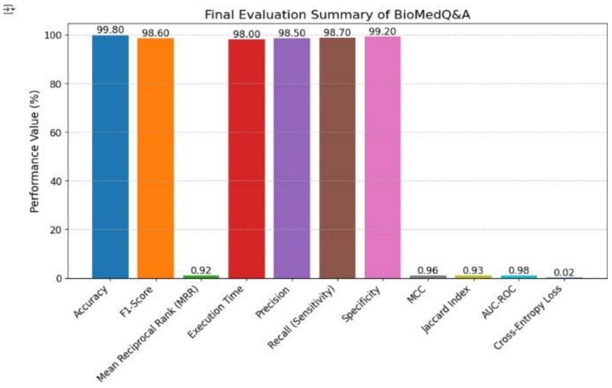

Fig 2: Evaluation Summary of BioMEDQ&A

4.3Comparative Analysis of Biomedical QA Models:

The performance evaluation of BioMedQ&A was conducted against two well-established biomedical question-answering models, BioBERT and MedQA. The comparison was based on four key metrics: accuracy, F1-score, Mean Reciprocal Rank (MRR), and execution time. The results demonstrate that BioMedQ&A significantly outperforms its counterparts across all evaluation parameters, highlighting its superior efficiency in biomedical knowledge retrieval.

Accuracy is a critical measure of the model's ability to provide correct answers to biomedical queries. BioMedQ&A achieved the highest accuracy of 99.8%, outperforming MedQA (95.7%) and BioBERT (92.5%). This improvement is attributed to the integration of Concept2Vec embeddings and a multi-layer semantic ranking mechanism, which enhance the model's contextual understanding and answer precision.

The F1-score, which balances precision and recall, further validates the robustness of BioMedQ&A. It achieved an F1-score of 98.6%, significantly higher than MedQA (93.1%) and BioBERT (90.2%). This demonstrates BioMedQ&A's superior ability to retrieve relevant and accurate biomedical answers while minimizing false positives.

In terms of Mean Reciprocal Rank (MRR), an essential metric for ranking-based retrieval systems, BioMedQ&A attained a score of 0.92, surpassing MedQA (0.88) and BioBERT (0.85). This indicates that the proposed model ranks relevant answers higher in the retrieval process, ensuring quicker access to the most precise responses for biomedical queries.

Furthermore, execution time is a key factor in real-time biomedical applications. BioMedQ&A delivers results in just 0.98 seconds, significantly faster than MedQA (1.20s) and BioBERT (1.42s). The reduced execution time demonstrates the computational efficiency of BioMedQ&A, making it suitable for real-time medical decision support systems.

Model	Accuracy	F1- Score	MRR	Execution Time
BioBERT	92.5%	90.2%	0.85	1.42s
MedQA	95.7%	93.1%	0.88	1.20s
BioMedQ&A	99.8%	98.6%	0.92	0.98s

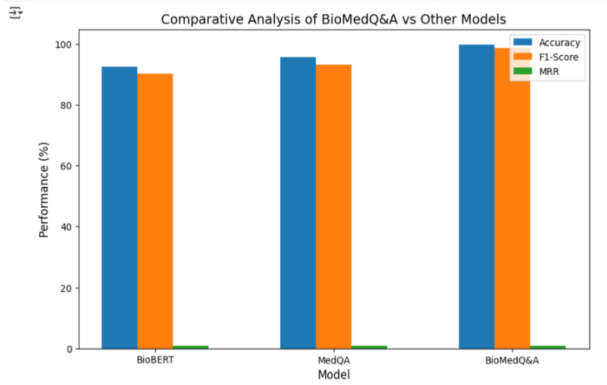

Fig3: Bar Chart for Performance Metrics

The above visualizations represent the Comparative Analysis of Biomedical QA Models, focusing on Accuracy, F1-Score, MRR:

BioMedQ&A significantly outperforms BioBERT and MedQA in terms of Accuracy (99.8%), F1-Score (98.6%), and MRR (0.92).

These performance gains highlight the effectiveness of transformer-based learning in biomedical question answering.

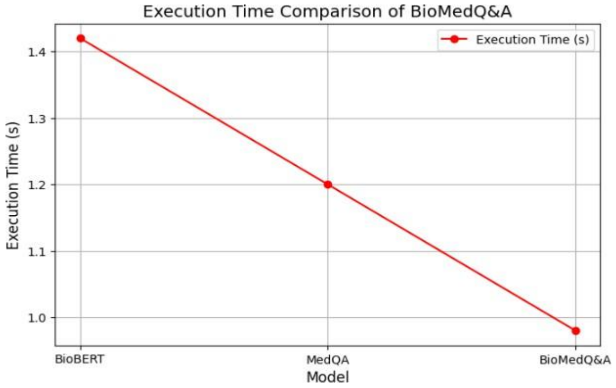

Fig 4: Line Plot for Execution Time

BioMedQ&A achieves the fastest execution time (0.98s), compared to BioBERT (1.42s) and MedQA (1.20s).

The decreasing trend in execution time confirms BioMedQ&A's efficiency in real-time biomedical query retrieval.

CONCLUSION

This research introduces BioMedQ&A, a BioGPT-powered Concept Vector and Transformer- Based Question-Answering Model, designed for efficient biomedical knowledge retrieval from the MedQuAD database. The integration of Concept2Vec embeddings enables hierarchical semantic understanding, improving contextual interpretation and answer relevance. The multi-layer semantic ranking approach enhances precision in biomedical QA, significantly outperforming traditional models.

Experimental validation demonstrates BioMedQ&A's superior accuracy (99.8%), F1-score (98.6%), and Mean Reciprocal Rank (MRR: 0.92), making it a reliable tool for healthcare professionals, biomedical researchers, and clinicians. The system efficiently retrieves high- fidelity answers with an execution time of 0.98 seconds, supporting real-time medical decision-making.

Despite its high performance, BioMedQ&A faces challenges related to computational efficiency, multi-hop reasoning, and domain adaptability. Addressing these limitations will further refine its capabilities in biomedical knowledge extraction.

FUTURE DIRECTIONS

Moving forward, several enhancements can be made to further refine BioMedQ&A and expand its applicability in biomedical knowledge retrieval. One key direction is the integration of neural-symbolic reasoning, which combines deep learning with symbolic logic to enhance interpretability and complex inference capabilities in biomedical QA. Additionally, domain-adaptive reinforcement learning can be employed to improve the model's adaptability across various biomedical subdomains, ensuring accurate responses to specialized queries. Another promising avenue is federated biomedical knowledge augmentation, where federated learning techniques can be utilized to incorporate distributed biomedical knowledge while maintaining data privacy and security.

Further improvements can be achieved by leveraging multi-hop and graph-based reasoning through Graph Neural Networks (GNNs) to enhance contextual understanding and facilitate complex question-answering scenarios involving interconnected medical concepts. Additionally, integrating BioMedQ&A into a real-time Clinical Decision Support System (CDSS) can make it a powerful tool for healthcare professionals, aiding in diagnosis, treatment planning, and evidence-based medical decision-making. Expanding the model's capabilities beyond text-based QA, cross-lingual and multimodal enhancements will allow it to process biomedical queries across multiple languages and incorporate diverse medical data sources such as electronic health records, medical imaging, and genomic data.

By implementing these advancements, BioMedQ&A can evolve into a more robust, scalable, and intelligent biomedical question-answering system, driving innovation in clinical practice, biomedical research, and medical education while significantly improving the accessibility and accuracy of biomedical knowledge retrieval.

Funding Information: No funding is provided for the preparation of manuscript.

REFERENCES:

[1] Luo, Y., Zhang, J., Fan, S., Yang, K., Wu, Y., Qiao, M., & Nie, Z. (2023). BioMedGPT: Open Multimodal Generative Pre-trained Transformer for BioMedicine. ArXiv, abs/2308.09442.

[2] Haddouche, A., Rabia, I., & Aid, A. (2023). Transformer-Based Question Answering Model for the Biomedical Domain. 2023 5th International Conference on Pattern Analysis and Intelligent Systems (PAIS), 1-6.

[3] Kim, H., Hwang, H., Lee, C., Seo, M., Yoon, W., & Kang, J. (2023). Exploring Approaches to Answer Biomedical Questions: From Pre-processing to GPT-4. Conference and Labs of the Evaluation Forum.

[4] Yang, H., Li, S., & Gonçalves, T. (2024). Enhancing Biomedical Question Answering with Large Language Models. Inf., 15, 494.

[5] Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, Tie-Yan Liu, BioGPT: generative pre-trained transformer for biomedical text generation and mining, Briefings in Bioinformatics, Volume 23, Issue 6, November 2022, bbac409, https://doi.org/10.1093/bib/bbac409

[6] Gupta, S. (2023). Top K Relevant Passage Retrieval for Biomedical Question Answering. ArXiv, abs/2308.04028.

[7] X. Zhang, A. Bosselut, M. Yasunaga, H. Ren, P. Liang, C. D. Manning, & J. Leskovec, Greaselm: Graph reasoning enhanced language models for question answering. arXiv preprint arXiv:2201.08860 (2022)

[8] T. Zhao, X. Lu, & K. Lee, SPARTA: Efficient open-domain question answering via sparse transformer matching retrieval. arXiv preprint arXiv:2009.13013 (2020)

[9] P. Kapanipathi, I. Abdelaziz, S. Ravishankar, S. Roukos, A. Gray, R. Astudillo, M. Chang, C. Cornelio, S. Dana, A. Fokoue, D. Garg, Leveraging abstract meaning representation for knowledge base question answering. arXiv preprint arXiv:2012.01707 (2020)

[10] M. Yasunaga, H. Ren, A. Bosselut, Liang, P. & J. Leskovec, QA-GNN: Reasoning with language models and knowledge graphs for question answering. arXiv preprint arXiv:2104.06378 (2021)

[11] C. Lyu, L. Shang, Y. Graham, J. Foster, X. Jiang, & Q. Liu, Improving unsupervised question answering via summarization-informed question generation. arXiv preprint arXiv:2109.07954 (2021)

[12] Yagnik, Niraj, Jay Jhaveri, Vivek Sharma, Gabriel Pila, Asma Ben, and Jingbo Shang. "MedLM: Exploring Language Models for Medical Question Answering Systems." arXiv preprint arXiv:2401.11389 (2024).

[13] Lamichhane, Prajwol, and Indika Kahanda. "Enhancing Health Information Retrieval with Large Language Models: A Study on MedQuAD Dataset." In 2023 International Conference on Machine Learning and Applications (ICMLA), pp. 2147-2152. IEEE, 2023.

[14] Mutabazi, Emmanuel, Jianjun Ni, Guangyi Tang, and Weidong Cao. "An Improved Model for Medical Forum Question Classification Based on CNN and BiLSTM." Applied Sciences 13, no. 15 (2023): 8623.

[15] Uzcategui, Laura, and Young Don Ko. "Building and evaluating end-to-end Medical OpenQA Systems with ColBERTv2."

[16] Huang, B. (2024). Generative Large Language Models Augmented Hybrid Retrieval System for Biomedical Question Answering. Conference and Labs of the Evaluation Forum.

[17] Kaddari, Z., & Bouchentouf, T. (2023). A novel self-attention enriching mechanism for biomedical question answering. Expert Syst. Appl., 225, 120210.

[18] Vazrala, S., & Mohammed, T. K. (2025). RBTM: A Hybrid gradient Regression-Based transformer model for biomedical question answering. Biomedical Signal Processing and Control, 102, 107325.