A HYBRID APPROACH FOR REVOLUTIONIZING MEDICAL INFORMATION RETRIEVAL USING AI

Hitesh Gehani[1, 2*] and Shubhangi Rathkanthiwar[1]
[1]Ph.D. Research Scholar, Department of Electronics Engineering, Yeshwantrao Chavan College of Engineering,
Nagpur, India
[2]Assistant Professor, School of Computer Science and Engineering, Ramdeobaba University, Nagpur
[1]Professor, Department of Electronics Engineering, Yeshwantrao Chavan College of Engineering, Nagpur, India
hiteshsgehani@gmail.com[1,2*] and svr_1967@yahoo.com[1]

Abstract

Medical question answering systems face significant challenges in providing accurate, timely, and context-aware responses due to the complexity and vast scope of biomedical information. This paper presents a novel hybrid approach that combines BioBERT, a specialized biomedical language model, with the Qdrant vector database to create an efficient and accurate medical question answering system. Our methodology encompasses three key components: (1) comprehensive data preprocessing using KeyBERT for intelligent tagging of the MedQuAD dataset containing 47,457 medical question-answer pairs, (2) finetuning BioBERT with an innovative negative sampling strategy based on semantic tags, and (3) implementation of efficient semantic search using Qdrant vector database for rapid retrieval of relevant responses. The system evaluation demonstrated clear separation between correctly classified positive and negative samples in cosine similarity scores, with misclassifications occurring primarily near the 0.3 threshold. The resulting chatbot system successfully combines domain-specific language understanding with efficient vector-based retrieval to deliver accurate and contextually relevant medical information. This approach addresses the limitations of existing medical question answering systems by providing a scalable, automated solution that maintains high accuracy while reducing dependence on human experts.

Keywords: Biomedical Question Answering, BioBERT, Vector Database, Semantic Search, Natural Language Processing, Healthcare Chatbot.

I. INTRODUCTION

The exponential growth in biomedical literature and the increasing complexity of healthcare information have created significant challenges in accessing and processing medical knowledge effectively With approximately 26.5 million articles in PubMed alone healthcare professionals and researchers face the daunting task of finding precise answers to their queries within this vast repository of information. Traditional methods of information retrieval and manual search processes are becoming increasingly inadequate leading to delayed access to critical medical information and potential impacts on patient care.

The emergence of advanced natural language processing (NLP) technologies, particularly transformer-based models and biomedical language models, presents an opportunity to address these challenges.

While general-purpose question answering systems have shown promise, they often struggle with domain specific terminology and the nuanced nature of biomedical information. Furthermore, existing systems frequently fail to effectively utilize the semantic information within biomedical datasets, resulting in suboptimal performance.

To address these limitations, we propose an integrated approach combining specialized biomedical language models with advanced vector database technologies. Our solution leverages BioBERT, a domain specific language representation model pretrained on extensive biomedical corpora, along with semantic search capabilities for efficient information retrieval. The system aims to provide accurate, context-aware responses by incorporating multiple knowledge sources and implementing sophisticated answer processing techniques.

II. WORK IN THIS AREA

Recent advances in biomedical natural language processing have been largely driven by the development of specialized language models. BioBERT, introduced by Lee et al. [2],[3],[4] has demonstrated significant improvements over general-domain BERT models, achieving state-of- the-art performance in various biomedical text mining tasks. This success has led to numerous adaptations and improvements, including PubMed BERT, which showed superior performance in clinical text classification with an accuracy of 0.9 and an F1-score of 0.8 [5] [6] [7].

The integration of these models into question answering systems has evolved significantly. Studies have shown that fine-tuning approaches can substantially improve performance on low- resource biomedical NLP applications [8][9][10]. Researches [11][12][13] have achieved remarkable results with an Exact Match score of 83.89 and an F-score of 89.67 using PubMed BERT for biomedical question answering tasks. Recent work with Dense Passage Retrieval frameworks adapted for biomedical domains has demonstrated promising results, achieving an 81% F1 score when evaluated on BioASQ QA datasets using PubMed articles as the knowledge base [1,14]. Question entailment approaches have emerged as an effective strategy for mapping new questions to previously answered similar questions, particularly in medical domains. These approaches have

shown significant improvements over traditional methods [15]. The development of hybrid architectures, which combines Information Retrieval and Seq2Seq models with an attentive Seq2Seq reranking approach, has further advanced the field. These systems have demonstrated superior performance compared to both standalone information retrieval and generation-based models in industrial applications [16].

The foundation of effective biomedical question answering lies in accurate entity recognition and information extraction. Recent work has demonstrated the effectiveness of machine reading comprehension frameworks for biomedical named entity recognition, achieving F1-scores of up to 92.92% on various datasets [1]. The development of tools like BERN has further advanced the field by combining high-performance BioBERT models with probability-based decision rules for overlapping entity identification [17]. Deep Event Mine [18] and related approaches [19] have shown promising results in extracting complex nested events from biomedical texts without relying on external syntactic tools. These advances have been particularly important for processing complex biomedical questions and extracting relevant information from research papers. The evolution of chatbot architectures has seen a shift from simple statistical methods to sophisticated neural network-based approaches [20]. Recent implementations like AliMe Chat

[21] have demonstrated the effectiveness of combining Information Retrieval (IR) and Sequence-to-Sequence (Seq2Seq) models for improved response generation. Studies focusing on healthcare applications [22] have shown that integrating large language models with Multi- BERT configurations can enhance the system's ability to handle complex biomedical queries. This approach has been particularly effective in addressing ambiguity resolution, as demonstrated by the MSQ BioBERT system [23,24,25].

III. PROPOSED METHODOLOGY

The implemented architecture leverages the strengths of BioBERT's domain-specific understanding

and Qdrant's high-performance vector search to create an intelligent and efficient conversational agent for a specialized domain, such as healthcare. As illustrated in Figure 1, the system integrates a user interface built with Stream lit library for input collection and result display, with BioBERT processing and Qdrant search capabilities forming the core components of the architecture.

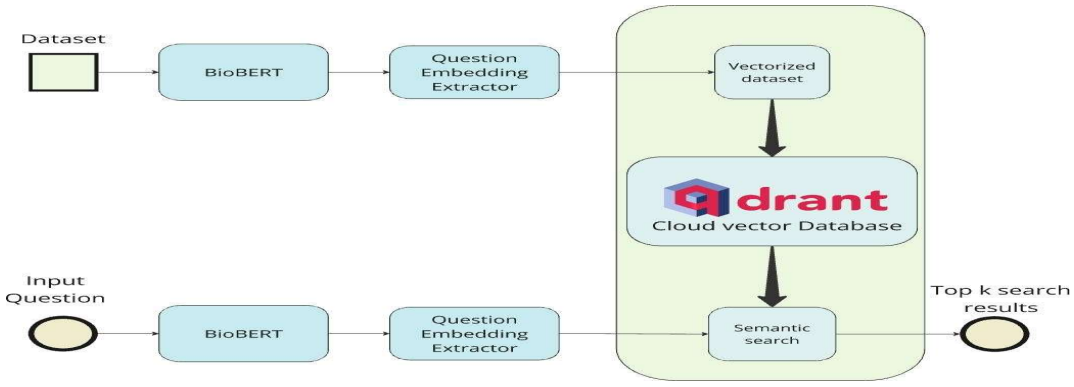


Fig 1. Proposed Methodology

Data availability statement:

To remove any ambiguity, we will revise the Methods and Data Availability sections of the manuscript to explicitly include the following statements:

- All methods were carried out in accordance with relevant guidelines and regulations.
- The study did not involve experiments on human subjects or the collection of primary human data.
- Ethical approval and informed consent were not applicable, as the research used publicly available, non-identifiable datasets.

DATA COLLECTION AND PREPROCESSING

The dataset employed in this study was obtained from the Med Quad repository and subsequently converted into a CSV file format. The dataset encompasses a diverse array of medical categories, including cancer, senior health, growth hormones and receptors, cardiovascular and respiratory diseases, genetic and rare disorders, disease prevention and control, neurological conditions and stroke, diabetes, and other miscellaneous topics. Notably, MedQuAD comprises 47,457 medical question-answer pairs curated from 12 reputable NIH websites, such as cancer.gov, niddk.nih.gov, GARD, and MedlinePlus Health Topics. The collection covers 37 distinct question types associated with diseases, drugs, medical tests, and other healthcare entities, encompass- ing areas like treatment, diagnosis, and potential side effects.

As demonstrated in Figure 2, the initial preprocessing step involved generating relevant tags for each question-answer pair using the KeyBERT library, a state-of-the-art keyword extraction tool. This tool excels at extracting meaningful tags that encapsulate the key concepts and topics covered within each question-answer pair.

Question	Answer
question_1	answer_1
question_2	answer_2
.....	.....
question_n	answer_n

KeyBERT

Question	Answer	Tags
question_1	answer_1	tags_1
question_2	answer_2	tags_2
.....	.....	.....
question_n	answer_n	tags_n

Subsequently, the extracted tags played a crucial role in creating high-quality negative samples for training the BioBERT-based classification model. It was imperative to ensure that the tags associated with the negative samples did not overlap with the tags of the positive samples. This precautionary measure prevented the model from erroneously treating semantically similar question-answer pairs as

negatives during the training process.

This tagging step significantly enhanced the quality of the dataset by providing additional metadata that could be leveraged throughout the framework. From facilitating the training of the BioBERT model to aiding in the retrieval of relevant prior knowledge, these tags played a pivotal role in optimizing the performance and effectiveness of the entire system.

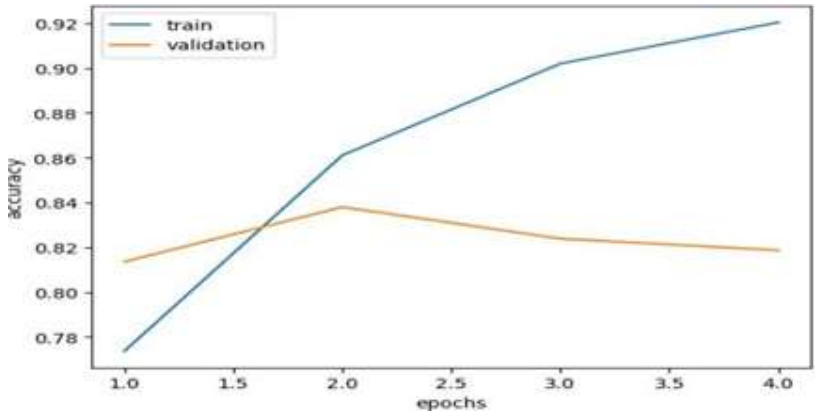
Preprocessing of the question-and-answer data involved decontracting words (e.g., ”won’t” to ”will not”), removing unnecessary symbols, and truncating sequences to a maximum length of 512 tokens. This approach allowed the utilization of the default pretrained weights of BioBERT while ensuring that no critical information was lost. Additionally, an attention mask was passed to the BioBERT model, enabling it to focus on the actual content rather than padding tokens.

BIOBERT FINE-TUNING

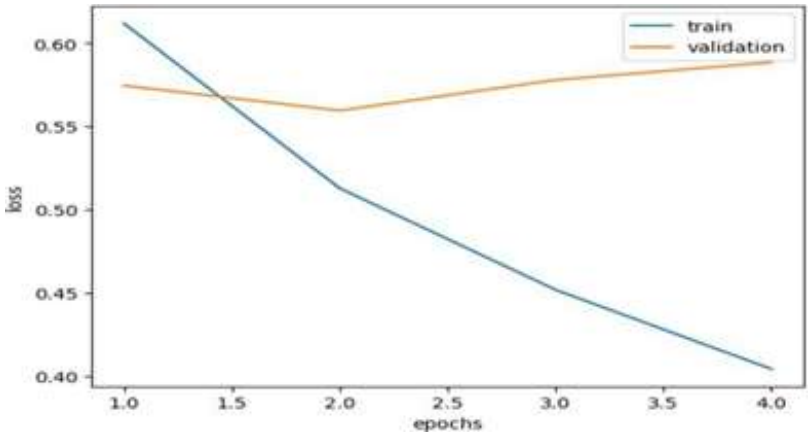
The model architecture consisted of a shared pretrained BioBERT model and two residual blocks with dropout layers for the question-and-answer data, respectively. The inclusion of residual blocks aimed to enable the model to directly pass the BioBERT embeddings when deemed appropriate. The dropout layers were employed to mitigate overfitting during the training process.

After obtaining the embeddings for both the question and answer by passing them through their respective residual blocks, the cosine similarity between the two embeddings was calculated. The model was then trained using mean squared error loss, computed between the expected cosine similarities (-1.0 and 1.0) and the cosine similarities predicted by the model. The best-performing model was selected based on its performance on both the training and validation datasets, after experimenting with various thresholds to classify positive and negative samples.

The model’s training progress is visualized in Figure 3, where (a) shows the improvement in accuracy over epochs, while (b) displays the corresponding loss reduction. These graph demonstrate the model’s successful convergence during the fine-tuning process, with both metrics showing stable improvement over time.



(a)

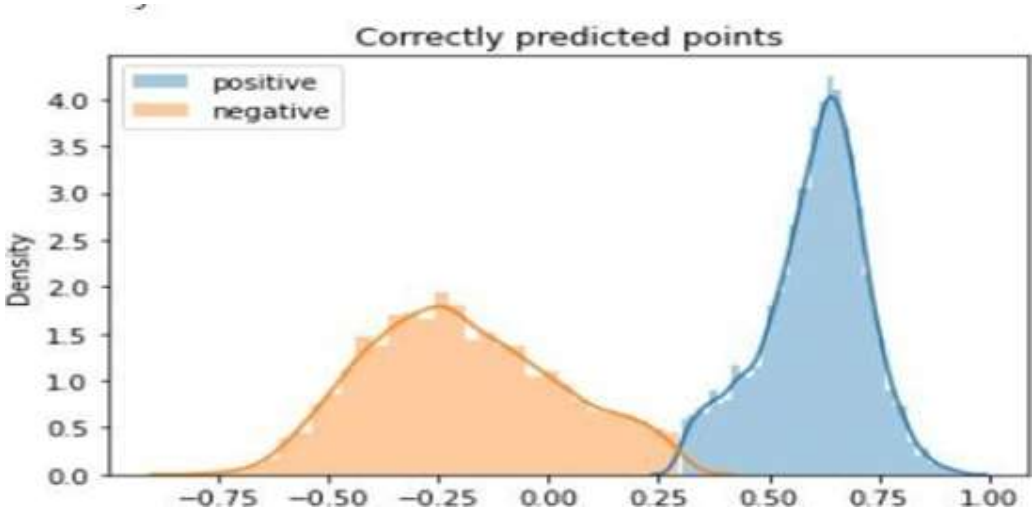


(b)

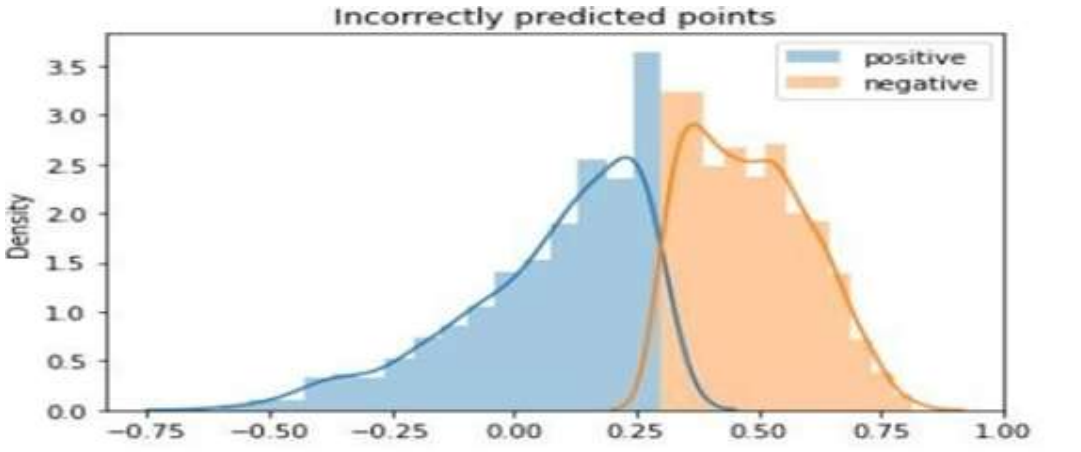
Figure 3. Fine-tuning results (a) epoch vs accuracy and (b) epoch vs loss

MODEL EVALUATION ON VALIDATION DATA

The cosine similarities returned by the model for the validation dataset were compared for correctly classified negative and positive points, as well as for misclassified points that were originally negative or positive. The evaluation results, presented in Figure 4, provide crucial insights into the model's performance. Figure 4(a) shows the distribution of correctly classified cosine similarities, while Figure 4(b) illustrates the cases where misclassification occurred. The analysis revealed that correctly classified negative and positive points maintained clear separation, while misclassified points clustered near the 0.3 threshold.



(a)



(b)

Figure 4. (a) Correctly classified cosine similarities (b) Incorrectly classified cosine similarities

These observations indicated that when the model differentiated correctly, it ensured a clear cosine separation between positive and negative points. However, when the model failed to classify correctly, the misclassification occurred with only a small margin from the threshold.

SEMANTIC SEARCH WITH QDRANT

To enable efficient semantic search, the question-answer pairs in the dataset were transformed into high-dimensional vector representations. The project utilized a specialized question extractor model, created using the fine-tuned BioBERT model, which was trained to encode questions into meaningful vector embed- dings. This question extractor model was applied to the entire dataset, generating a vectorized representation for each question, with the corresponding answers associated with these question vectors.

The Qdrant vector database was employed to store and index the question vectors generated in the previous step. Qdrant is designed to efficiently manage high-dimensional vector data and enable fast similarity based retrieval. The project integrated the Qdrant database as a core component of the chatbot system, allowing for seamless storage and querying of the vectorized dataset. The dataset was ingested into the Qdrant database, with the question vectors serving as the primary index for efficient retrieval.

When a user query was received, it was first encoded into a vector representation using the same question extractor model employed for the dataset. The chatbot system then performed a semantic search within the Qdrant database, leveraging the vector similarity between the user's query and the stored question vectors. Qdrant's powerful search capabilities, including support for fuzzy and multilingual search, were utilized to retrieve the most relevant question-answer pairs from the indexed dataset. The top-k most similar results were then returned by the Qdrant search, and the corresponding answers were presented to the user as the chatbot's response. The semantic search process was designed to be efficient and scalable, allowing the chatbot to provide timely and accurate responses, even as the underlying dataset grew in size and complexity.

By integrating the Qdrant vector database into the chatbot system, the project leveraged the power of semantic search to enable intelligent and efficient retrieval of relevant answers, tailored to the specific needs and context of the healthcare or life sciences domain.

Figure 5. presents the high-level architecture of the Qdrant-based semantic search implementation. This diagram illustrates how the system processes incoming queries, transforms them into vector representations, and efficiently retrieves relevant responses from the indexed dataset.

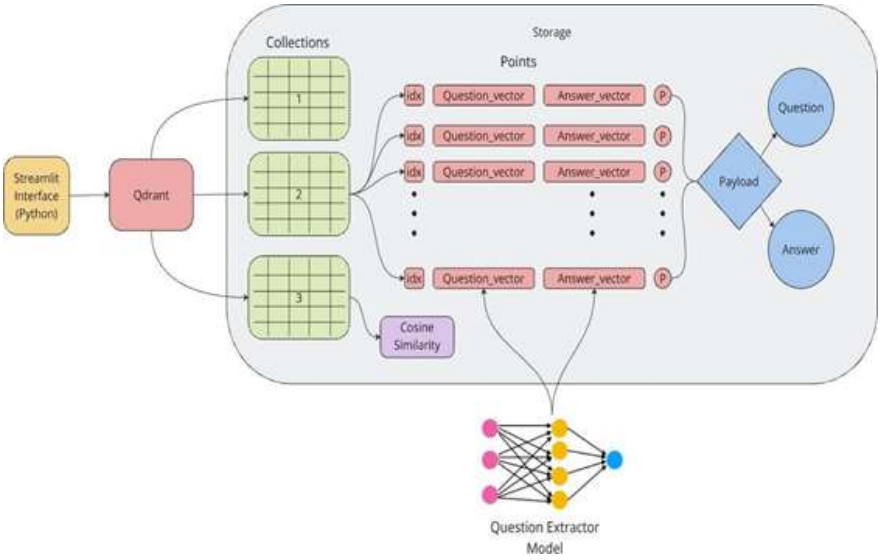


Figure 5. High level view of Qdrant architecture

IV. RESULT AND DISCUSSION

The developed chatbot system has demonstrated impressive capabilities in providing accurate and efficient responses within the healthcare or life sciences domain. One of the key results is the successful fine-tuning of the BioBERT model, leveraging its pretrained knowledge on biomedical literature, to create a powerful question-answering model tailored to the target domain.The utilization of KeyBERT to generate informative tags for the dataset enabled more effective negative sampling during the fine-tuning process, resulting in a more robust and accurate model.The integration of Qdrant, a high-performance vector database, facilitated efficient storage and retrieval of relevant answers based on semantic similarity, ensuring real-time responses to user queries.

The model evaluation on the validation data provided insights into the model's performance. The cosine similarities returned by the model for correctly classified negative and positive points were well-separated, indicating the model's ability to distinguish between positive and negative samples effectively. However, for misclassified points, the cosine similarities were in close proximity to the threshold of 0.3, which yielded the best training and validation accuracy.Overall, the chatbot system demonstrated the ability to provide accurate and timely responses to domain-specific questions, enhancing the overall user experience and reducing the burden on human experts.

Figure 6. UI Screenshot

The practical implementation of this system is demonstrated through the user interface shown in Figure 6, which provides an intuitive platform for users to interact with the chatbot and receive medical information in real-time. The interface demonstrates the seamless integration of all system components, from query input to answer retrieval and display.

V. FUTURE SCOPE

The future development of biomedical question answering systems encompasses several promising re- search directions. A key priority is expanding multi-language support [6], which would democratize access to biomedical information across diverse populations and geographical regions. Alongside this, enhanced context processing capabilities are needed to better handle multi-round interactions in dialog systems [21], particularly when dealing with complex medical queries. The integration of advanced vector database architectures presents opportunities for improved storage and retrieval of high-dimensional biomedical data [8], potentially revolutionizing response generation speed and accuracy. Dataset development remains crucial, with a particular focus on addressing low-resource areas within biomedical NLP [12], building upon the success of initiatives like PubMedQA[24]. In terms of model architecture, research is trending toward hybrid approaches that combine multiple pretrained models [22], leveraging advances in both general and domain-specific language models. Sophisticated question expansion methods utilizing multiple knowledge bases [25] are being explored to enhance the system's comprehension and processing of complex medical queries. Dataset-aware multi-task learning approaches are advancing the field of biomedical named entity recognition and normalization [26], while new techniques for handling ambiguous medical queries [23] are particularly relevant for web-based medical information retrieval. Document retrieval systems specifically designed for biomedical question answering are being optimized [3], incorporating advanced ranking and filtering mechanisms. Additionally, the development of more sophisticated semantic search capabilities for large clinical ontologies [10] is improving systems' ability to handle varying vocabularies and terminology. These advancements collectively contribute to the broader goal of creating more accurate, efficient, and user-friendly systems for biomedical information access and processing, ultimately enhancing healthcare delivery and research capabilities.

VI. CONCLUSION

This research successfully demonstrates the effectiveness of combining specialized language models

with vector database technology in creating a robust medical question answering system. The integration of BioBERT's domain-specific knowledge with Qdrant's efficient vector search capabilities has resulted in a system that can effectively process and respond to complex medical queries while maintaining high accuracy and performance. Through the implementation of an intelligent tagging system using KeyBERT, the project achieved enhanced quality of training data, which significantly improved the overall model performance. The development of an effective fine-tuning approach for BioBERT demonstrated particular success, achieving clear separation between positive and negative samples in cosine similarity scores.

The efficient integration of the Qdrant vector database proved crucial for enabling real-time semantic search and response retrieval, creating a scalable system capable of handling complex medical queries while reducing the burden on human experts. The system's architecture demonstrates that combining specialized language models with vector databases can effectively address the fundamental challenges of medical question answering, including the need for domain-specific understanding and rapid information retrieval. The clear separation in cosine similarities for correctly classified samples validates the robustness of our approach, while the clustering of misclassifications near the 0.3 threshold provides valuable insights for future improvements.

This work makes a significant contribution to the broader field of biomedical natural language processing by demonstrating a practical approach to building domain-specific question answering systems. The results suggest promising directions for future research, including opportunities to expand language support, enhance context processing capabilities, and incorporate more sophisticated semantic search techniques. These advancements would further improve the system's effectiveness in healthcare applications, ultimately working toward the goal of more accessible and accurate medical information retrieval systems.

Funding Declaration:

This research did not receive any specific grant or funding from public, commercial, or not-for-profit agencies.

REFERENCE

[1] Riccardo Bellazzi, Large Language Models and the return of knowledge engineering, International Journal of Medical Informatics,2026,106461, ISSN 1386-5056, ttps://doi.org/10.1016/j.ijmedinf.2026.106461.

[2] C. Sun, Z. Yang, L. Wang, Y. Zhang, H. Lin, and J. Wang, "Biomedical named entity recognition using bert in the machine reading comprehension framework," *Journal of Biomedical Informatics*, vol. 118, p. 103799, 05 2021.

[3] N. Kanodia, K. Ahmed, and Y. Miao, "Question answering model based conversational chatbot using bert model and google dialog flow," in *2021 31st International Telecommunication Networks and Applications Conference (ITNAC)*, 2021, pp. 19–22.

[4] H. Bolat and B. S¸ en, "Document retrieval system for biomedical question answering," *Applied Sciences*, vol. 14, no. 6, 2024. [Online]. Available: https://www.mdpi.com/2076-3417/14/6/2613

[5] S. Jeong, C. G. Kim, and T. K. Whangbo, "Question answering system for healthcare information based on bert and gpt," 03 2023, pp. 348–352.

[6] E. Xygi, A. Andriopoulos, and D. Kouts Mitropoulos, "Question answering chatbots for biomedical re- search using transformers," 02 2023, pp. 025–029.

[7] S. Cheon and I. Ahn, "Fine-tuning bert for question and answering using pubmed abstract dataset," in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2022, pp. 681–684.

[8] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: a pre- trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, p. 1234–1240,Sep. 2019. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/btz682Y.

[9] Han, C. Liu, and P. Wang, "A comprehensive survey on vector database: Storage and retrieval technique, challenge," 2023. [Online]. Available: https://arxiv.org/abs/2310.11703

[10] N.-Y. Tung, H.-W. Hu, T.-W. Chang, Y.-M. Hu, and H.-M. Lin, "Multi-model comparison for classifica- tion of medical records using the biobert models," in *2022 IEEE 4th Eurasia Conference on Biomedical Engineering, Healthcare and Sustainability (ECBIOS)*, 2022, pp. 221–224.

[11] D.-H. Ngo, M. Kemp, D. Truran, B. Koopman, and A. Metke-Jimenez, "Semantic search for large scale clinical ontologies," Jan. 2022.

[12] D. Dimitriadis and G. Tsoumakas, "Word embeddings and external resources for answer processing in biomedical factoid question answering," *Journal of Biomedical Informatics*, vol. 92, p. 103118, 02 2019. A. Ben Abacha and D. Demner-Fushman, "A question-entailment approach to question answering," BMC Bioinformatics, vol. 20, no. 1, Oct. 2019. [Online].Available: http://dx.doi.org/10.1186/ s12859-019-3119-4

[13] W. Yoon, J. Lee, D. Kim, M. Jeong, and J. Kang, Pre-trained Language Model for Biomedical Question Answering, 03 2020, pp. 727–740.

[14] D. Kim, J. Lee, C. So, H. Jeon, M. Jeong, Y. Choi, W. Yoon, M. Sung, and J. Kang, "A neural named entity recognition and multi-type normalization tool for biomedical text mining," IEEE Access, vol. PP,

[15] pp. 1–1, 06 2019.

[16] H.-L. Trieu, T. T. Tran, K. N. A. Duong, A. Nguyen, M. Miwa, and S. Anania Dou, "Deep Event Mine: end-to-end neural nested event extraction from biomedical texts," Bioinformatics, vol. 36, no. 19, pp. 4910–4917, 06 2020. [Online]. Available: https://doi.org/10.1093/bioinformatics/btaa540

[17] J. Bjo¨rne and T. Salakoski, "Biomedical event extraction using convolutional neural networks and dependency parsing," in Proceedings of the BioNLP 2018 workshop, D. Demner-Fushman, K. B. Cohen,

[18] S. Ananiadou, and J. Tsujii, Eds. Melbourne, Australia: Association for Computational Linguistics,

[19] Jul. 2018, pp. 98–108. [Online]. Available: https://aclanthology.org/W18-2311

[20] S. Ayanouz, B. Anouar Abdelhakim, and M. Benhmed, "A smart chatbot architecture based nlp and machine learning for health care assistance," 04 2020.

[21] M. Qiu, F.-L. Li, W. Siyu, X. Gao, Y. Chen, W. Zhao, H. Chen, J. Huang, and W. Chu, "Alime chat: A sequence to sequence and rerank based chatbot engine," 01 2017, pp. 498–503.

[22] C. Qian, X. Shi, S. Yao, Y. Liu, F. Zhou, Z. Zhang, J. Akram, A. Braytee, and A. Anaissi, "Optimized biomedical question-answering services with llm and multi-bert integration," 2024. [Online]. Available: https://arxiv.org/abs/2410.12856

[23] M. Guo, M. Guo, E. T. Dougherty, and F. Jin, "Msq-biobert: Ambiguity resolution to enhance biobert medical question-answering," in Proceedings of the ACM Web Conference 2023, ser. WWW '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 4020–4028. [Online]. Availablhttps://doi.org/10.1145/3543507.3583878

[24] Q. Jin, B. Dhingra, Z. Liu, W. Cohen, and X. Lu, "PubMedQA: A dataset for biomedical research question answering," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP- IJCNLP)*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 2567–2577. [Online]. Available: https://aclanthology.org/D19-1259

[25] I. Gabsi, H. Kammoun, A. Wederni, and I. Amous, "Biobert for multiple knowledge-based question expansion and biomedical extractive question answering," in *Computational Collective Intelligence: 16th International Conference, ICCCI 2024, Leipzig, Germany, September 9–11, 2024, Proceedings, Part I*. Berlin, Heidelberg: Springer-Verlag, 2024, p. 199–210. [Online]. Available: https://doi.org/10.1007/978- 3-031-70816-9 16

[26] M. Zuo and Y. Zhang, "Dataset-aware multi-task learning approaches for biomedical named entity recognition," *Bioinformatics*, vol. 36, no. 15, pp. 4331–4338, 05 2020. [Online]. Available: https://doi.org/10.1093/bioinformatics/btaa515