[[1]](#footnote-0)

Automatic Answering to PICO Clinical Questions using Deep Learning Techniques

Sabah Mohammed, Jinan Fiaidhi and Hashmath Shaik

***Abstract*—Answering clinical questions automatically is becoming a research area that benefit from the new artificial intelligence techniques. Leave this to Jinan and I as we will write it at the end of this research.**

***Keyword*— Clinical Question Answering, PICO questions, Evidence-Based Medicine, Clinical Case Report, Deep Learning, LLM Transformers Summarization, Automatic Q&A.**

# INTRODUCTION

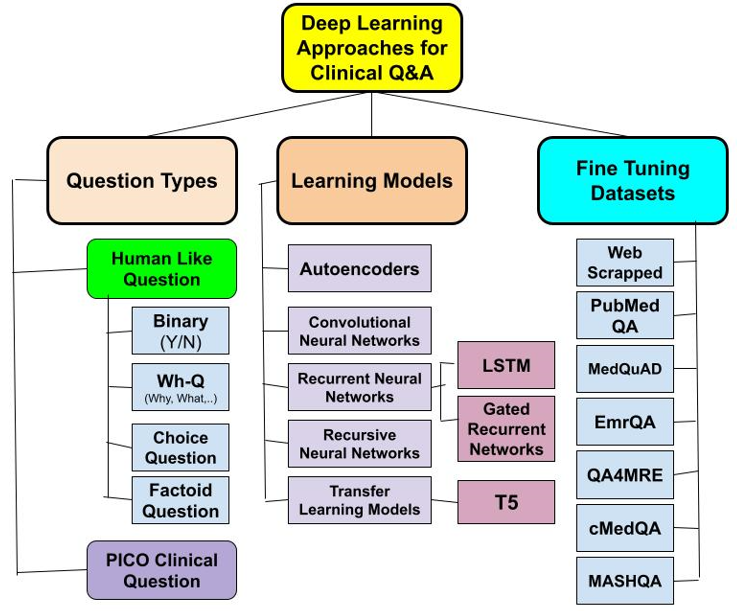
It is common practice in medicine to write a clinical case report when a clinician confronted with an unusual patient presentation. Case reports are the first-line of evidence in the medical literature, and provide medical students and junior doctors with a great opportunity to develop their writing skills [1]. Compiling such clinical case report starts with synthesizing clinical questions around the case that requires answers. Usually clinicians tend to use the PICO format for synthesizing their clinical questions [2] and later to conduct web literature search from medical sound repositories like PubMed or WebMD and go through the medical materials and try summarizing their finding before compiling the final case report [3]. However, this manual process of compiling a clinical case report is time consuming requires specific filtering skills and resources to manage the retrieved information [4]. Skilled physicians may use assistive question answering applications like AskHERMES [5], MiPACQ [6], MEANS [7], MedQA[8] or HONqa [9] to shorten the searching and filtering time, however, these applications hide the details of finding the clinical answers as well as their tested reliability is not acceptable in many cases according to notable scholars [10, 11].

A promising knowledge acquisition solution, however, emerged from the Question Answering (Q&A) research initiative involving deep learning techniques due to the availability of data sources to learn answers to new questions based on training the learning network on previously available domain data of questions and answers [12]. The reported success of Q&A techniques in answering some focused clinical questions based on training information scrapped from the web from sites like WebMD[[2]](#footnote-1), HealthTap[[3]](#footnote-2), eHealthForums[[4]](#footnote-3), patientslikeme[[5]](#footnote-4) and iCliniq[[6]](#footnote-5) encouraged researchers to investigate using this new artificial intelligence Q&A technique for providing more reliable clinical answers [13]. In this article, we are reporting an investigation into using two different deep learning technologies to answer PICO question from sound medical repositories like PubMed. The first investigated technology utilizes Large Language models (LLM) employing transformers like BioBERT and GPT to provide answers to given PICO questions using abstractive summarization and the second technology utilizes deep learning neural technology for Q&A automatic answering that can be trained on relevant Q&A datasets.

# Deep Learning Technologies for Clinical Q&A

Automatic Question Answering (Q&A) approaches represent systems for retrieving correct and relevant answers to the questions asked by human in natural language [14]. In healthcare it comes as an attempt to overcome the shortcoming in providing the required informational need through the legacy clinical Frequently Asked Questions (FAQs) portals established by almost every healthcare institution like the CDC.[[7]](#footnote-6) To solve this problem several researchers from the natural language and machine learning fields developed attempts to provide automated techniques for general question similarity [15, 16]. Several notable attempts in this direction brought extended attention to the Q&A field such as the development IBM Watson DeepQA [17], the availability several Q&A open benchmarks and datasets [18] (e.g. SQuAD, TriviaQA, BoolQ, PICO, WikiQA, HotportQA, NaturalQuestions, QuAC, CoQA, ELI5, Sharc, MS MAARCO, TWEETQA and NEWSQA) and the growing field of chatbots [19]. However, do not generalize well to the medical domain [20] and do not consider the standard framework for asking clinical questions like the PICO protocol [21].

Interestingly several recent deep learning models with fine tuning to medical domain started to provide encouraging results where clinical expert can use it to automatically answer clinical queries. Figure 1 illustrates tan overview to these attempts. It is also important that several deep learning approaches are also used in several processes leading to the question answering including extracting the important question components, encoding it, indexing it, retrieving answer matches, ranking/re-ranking the answer matches and calibrating the results as well as generating more meaningful result answer [22].



**Fig. 1:** Deep Learning Approaches for Clinical Q&A.

Moreover, there are two major regimes noted at the literature for solving the automatic Q&A, the first employs summarization techniques [23] and the second learns from benchmarks of previous questions and answers to generate answers to newly asked questions [24]. In this article we are experimenting with both techniques to generate answers to PICO clinical questions. The conclusion section will compare those two approaches as well as to discuss whether a hybrid of these two approaches may be more advantages compared to each regime separately conducted [25].

# Q&A using the Transformers Summarization

3.1 Dataset Background

Nature of the PubMed Q&A Dataset:

The PubMed Q&A dataset, serving as the bedrock of our research, is a comprehensive collection of questions and answers harvested from the vast domain of medical literature. Unlike more generic datasets, it encapsulates the depth, diversity, and specificity of medical queries, which range from basic diagnostics to intricate molecular biology questions.

Selection Rationale:

The decision to utilize the PubMed Q&A dataset was driven by its inherent challenge. Medical literature, characterized by its complexity and technical vernacular, demands a higher cognitive understanding from NLP models, pushing them beyond the boundaries set by more generic texts.

3.2 Transformer Architectures: An Overview

BERT - Bi-directional Encoders for Rich Understanding:

BERT, or Bi-directional Encoder Representations from Transformers, is renowned for its capability to analyze text bidirectionally (considering both left and right context in all layers). This holistic understanding of context, combined with its deep layers of transformers, makes it adept at a variety of NLP tasks.

GPT-2 - A Generative Leap:

On the other hand, GPT-2, or Generative Pre-trained Transformer 2, shines in generating coherent, diverse, and contextually pertinent text sequences. Its ability to predict subsequent words in a sequence has rendered it highly effective in text generation tasks, making it an interesting contender in our research.

3.3 Embarking on the Summarization Task

Objective and Need for Summarization:

With the volume of medical literature expanding daily, the need for concise summaries has become paramount. Through summarization, we aimed to distill the essence of extensive medical cases, ensuring that the resulting shorter texts retained critical information while eliminating redundancy.

Model Training Strategy:

Both BERT and GPT-2 were subjected to the medical cases, with the training regimen focusing on generating summaries. Continuous feedback loops, along with gradient descent algorithms, steered the models to enhance their summarization capabilities progressively.

3.4 Transition to Q&A: Fine-tuning on Summaries

Training Approach:

Having trained the models for summarization, they were then acquainted with the Q&A portion of the dataset. The training data was structured such that each question was paired with the summary (generated in the previous step) as context, and the model's task was to produce or identify the correct answer.

Challenges Addressed:

Given the inherent ambiguities and multifaceted nature of medical literature, the models were often presented with questions that could have multiple plausible answers. The fine-tuning was thus meticulously designed to ensure that models discerned nuances and were adept at handling such challenges.

3.5 Evaluation: Metrics and Outcomes

Metrics Selection Rationale:

Evaluation metrics are crucial in gauging model performance objectively. F1 scores offer a balanced measure, weighing both precision (correctness of information provided) and recall (completeness of information provided). Meanwhile, the Exact Match (EM) score quantifies the instances where the model's answer perfectly aligns with the ground truth.

Performance Overview:

BERT, with its robust contextual understanding, achieved an F1 score of 0.0500, but couldn't register a positive EM score, standing at 0. GPT-2, while demonstrating its prowess in text generation, yielded an F1 score of 0.03 and, similarly, an EM score of 0. These metrics, while modest, shed light on the intricate challenges posed by medical literature and offer avenues for further research and improvement.

# Q&A using Automatic Q&A

In this section you will need to use the techniques described by the ToDo of your new MITACS to arrive at answers to the same clinical case used in the previous section and then generalize this approach for answering more clinical questions with other cases. You will need additional measures to accuracy besides F1 like EM (Exact Match).

# Conclusion

In the conclusion you will need to compare the two approaches as well as considering if hybrid of both will be a better option.

Acknowledgment

The first and second authors acknowledge the financial support to this research project from MTACS Accelerates Grant (IT22305-2020) and the first author NSERC DDG Grant (DDG-2021-00014).

References

1. Ganesan, Prasanth. "How to write case reports and case series." International Journal of Advanced Medical and Health Research 9, no. 1 (2022): 55-58.
2. Leonardo, R. "PICO: model for clinical questions." Evid Based Med Pract 3, no. 115 (2018): 2.
3. Lacasse, Miriam, Valérie Lafortune, Lynsey Bartlett, and Jessica Guimond. "Answering clinical questions: What is the best way to search the Web?." Canadian Family Physician 53, no. 9 (2007): 1535-1536.
4. EbEll, Mark H. "How to find answers to clinical questions." American family physician 79, no. 4 (2009): 293-296.
5. Cao Y, Liu F, Simpson P, et al. . AskHERMES: An online question answering system for complex clinical questions. J Biomed Inform 2011; 44 (2): 277–88.
6. Cairns BL, Nielsen RD, Masanz JJ, et al. . The MiPACQ clinical question answering system. AMIA Annu Symp Proc 2011; 2011: 171–80.
7. Abacha AB, Zweigenbaum P. MEANS: A medical question-answering system combining NLP techniques and semantic web technologies. Inform. Process. Manag. 2015; 51 (5): 570–94.
8. Zhang, Xiao, Ji Wu, Zhiyang He, Xien Liu, and Ying Su. "Medical exam question answering with large-scale reading comprehension." In Proceedings of the AAAI conference on artificial intelligence, vol. 32, no. 1. 2018.
9. Wong, Wilson, John Thangarajah, and Lin Padgham. "Health conversational system based on contextual matching of community-driven question-answer pairs." In Proceedings of the 20th ACM international conference on Information and knowledge management, pp. 2577-2580. 2011.
10. Schwartz, Diane G., June Abbas, Richard Krause, Ronald Moscati, and Shravanti Halpern. "Are internet searches a reliable source of information for answering residents' clinical questions in the emergency room." In Proceedings of the 1st ACM International Health Informatics Symposium, pp. 391-394. 2010.
11. Ni, Yuan, Huijia Zhu, Peng Cai, Lei Zhang, Zhaoming Qui, and Feng Cao. "CliniQA: highly reliable clinical question answering system." In Quality of Life through Quality of Information, pp. 215-219. IOS Press, 2012.
12. Abdallah, Abdelrahman, Mahmoud Kasem, Mohamed A. Hamada, and Shaymaa Sdeek. "Automated question-answer medical model based on deep learning technology." In Proceedings of the 6th International Conference on Engineering & MIS 2020, pp. 1-8. 2020.
13. Faris, Hossam, Maria Habib, Mohammad Faris, Alaa Alomari, Pedro A. Castillo, and Manal Alomari. "Classification of Arabic healthcare questions based on word embeddings learned from massive consultations: a deep learning approach." Journal of Ambient Intelligence and Humanized Computing (2022): 1-17.
14. Dwivedi, Sanjay K., and Vaishali Singh. "Research and reviews in question answering system." Procedia Technology 10 (2013): 417-424.
15. Braun, Daniel, Adrian Hernandez Mendez, Florian Matthes, and Manfred Langen. "Evaluating natural language understanding services for conversational question answering systems." In Proceedings of the 18th annual SIGdial meeting on discourse and dialogue, pp. 174-185. 2017.
16. Sarrouti, Mourad, and Said Ouatik El Alaoui. "A machine learning-based method for question type classification in biomedical question answering." Methods of Information in Medicine 56, no. 03 (2017): 209-216.
17. D. Ferrucci, E. Brown, J. Chu-Carroll, J. Fan, D. Gondek, A. A. Kalyanpur, et al., "Building Watson: An overview of the DeepQA project", AI Mag., vol. 31, pp. 59, 2010.
18. Cambazoglu, B. Barla, Mark Sanderson, Falk Scholer, and Bruce Croft. "A review of public datasets in question answering research." In ACM SIGIR Forum, vol. 54, no. 2, pp. 1-23. New York, NY, USA: ACM, 2021.
19. Quarteroni, Silvia, and Suresh Manandhar. "A chatbot-based interactive question answering system." Decalog 2007 83 (2007).
20. McCreery, Clara H., Namit Katariya, Anitha Kannan, Manish Chablani, and Xavier Amatriain. "Effective transfer learning for identifying similar questions: matching user questions to COVID-19 FAQs." In Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining, pp. 3458-3465. 2020.
21. Athenikos, Sofia J., and Hyoil Han. "Biomedical question answering: A survey." Computer methods and programs in biomedicine 99, no. 1 (2010): 1-24.
22. Wang, Shuohang, Mo Yu, Jing Jiang, Wei Zhang, Xiaoxiao Guo, Shiyu Chang, Zhiguo Wang, Tim Klinger, Gerald Tesauro, and Murray Campbell. "Evidence aggregation for answer re-ranking in open-domain question answering." arXiv preprint arXiv:1711.05116 (2017).
23. Suleiman, Dima, and Arafat Awajan. "Deep learning based abstractive text summarization: approaches, datasets, evaluation measures, and challenges." Mathematical problems in engineering 2020 (2020): 1-29.
24. Osman, Alexander FI, and Nabil M. Maalej. "Applications of machine and deep learning to patient‐specific IMRT/VMAT quality assurance." Journal of Applied Clinical Medical Physics 22, no. 9 (2021): 20-36.
25. Savery, Max, Asma Ben Abacha, Soumya Gayen, and Dina Demner-Fushman. "Question-driven summarization of answers to consumer health questions." Scientific Data 7, no. 1 (2020): 322.

1. \*Research supported by NSERC and MITACS.

   **Sabah Mohammed** is a full Professor of Computer Science with the Lakehead University, Thunder Bay Ontario P7B 5E1, Canada (phone 807-343-8777; fax: 807-7667243; e-mail: mohammed@lakeheadu.ca).

   **Jinan Fiaidhi** is a full Professor of Computer Science with the Lakehead University, Thunder Bay Ontario P7B 5E1, Canada (e-mail: jfiaidhi@lakeheadu.ca).

   **Hashmath Shaik** is a Graduate Student with Department of Computer Science, Lakehead University, Thunder Bay, Ontario P7B 5E1, Canada (e-mail: hshaik@lakeheadu.ca). [↑](#footnote-ref-0)
2. https://www.webmd.com/ [↑](#footnote-ref-1)
3. https://www.healthtap.com/ [↑](#footnote-ref-2)
4. https://www.healthboards.com/ [↑](#footnote-ref-3)
5. <https://www.patientslikeme.com/> [↑](#footnote-ref-4)
6. https://www.icliniq.com/ [↑](#footnote-ref-5)
7. https://www.cdc.gov/ [↑](#footnote-ref-6)