Generative AI for Evidence-Based Medicine: A PICO GenAI for Synthesizing Clinical Case Reports

Sabah Mohammed
Department of Computer Science
Lakehead University
Thunder Bay, Canada
mohammed@lakeheadu.ca

Jinan Fiaidhi
Department of Computer Science
Lakehead University
Thunder Bay, Canada
jfiaidhi@lakeheadu.ca

ICC 2024 - IEEE International Conference on Communications | 978-1-7281-9054-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICC51166.2024.10622271

Abstract—**Clinical research and practice are generating important new findings at exponential rate which need to be readily available to clinicians. However, clinicians are confronted with serious challenges when they try to seek such information for their evidence-based decision making or to generate new clinical case report. One important challenge is the long time needed to browse, filter, summarize and compile information from different resources. The other important challenge is to identify relevant important evidence-based information resources required to answer clinical questions or support a clinical finding. Artificial intelligence can help in solving both challenges based on the automatic question answering (Q&A) and generative technologies. However, Q&A and generative techniques are not trained to answer clinical queries that can be used for evidence-based practice nor it can respond to structured clinical questioning protocol like PICO (Patient/Problem, Intervention, Comparison and Outcome). This article describes the use of deep learning techniques for Q&A that is based on generative models like BERT and GPT to answer PICO clinical questions that can be used for evidence-based practice extracted from sound medical research resources like PubMed. We are reporting acceptable clinical answers that are supported by findings from PubMed. Our generative methods are reaching state of the art performance based on two staged bootstrapping process involving filtering relevant articles followed by identifying articles that support the requested outcome expressed by the PICO question.**

Keywords—Automatic Question Answering, PICO questions, Evidence-Based Medicine, Clinical Case Report, Generative Models, LLM Transformers, Fine Tuning, Bootstrapping.

I. 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) and Generative AI (GenAI) research initiative involving deep learning techniques due to the availability of data sources to learn answers to new questions or summarize clinical data based on training the learning network on previously available domain data [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 [1], HealthTap [2], eHealthForums [3], patientslikeme [4], PubMed [5], Medical Encyclopedia [6] and iCliniq [7] encouraged researchers to investigate using this new artificial intelligence Q&A technique for providing more evidence-based 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 like BioGPT 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.

II. Deep Learning Technologies for Clinical Q&A and GenAI

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. [8] To solve this problem several researchers from the natural language and machine learning

[1] https://www.webmd.com/
[2] https://www.healthtap.com/
[3] https://www.healthboards.com/
[4] https://www.patientslikeme.com/
[5] https://pubmed.ncbi.nlm.nih.gov/
[6] https://medlineplus.gov/encyclopedia.html
[7] https://www.icliniq.com/
[8] https://www.cdc.gov/

fields developed attempts to provide automated techniques for generate clinical synthetic information [15, 16]. Several notable attempts in this direction brought extended attention to the Q&A and GenAI 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 and bootstrapping started to provide to encouraging results in several common fronts of GenAI and Q&A. Figure 1 illustrates an overview to these attempts. However, the current GenAI and Q&A provide only general help in synthesizing clinical documents like clinical notes summaries, medical education supportive materials, matching patient cases from online resources and answering general clinical questions. However, providing expert-level information that is credible for evidence-based purposes (e.g. providing evidence supporting a clinical case report) is still a challenge [22]. GenAI and Q&A provides general information when they use large language models (LLM) that have the capability to process and understand natural language. These models are trained on massive amounts of text data to learn patterns and entity relationships in the language. Although the LLM models can perform useful language tasks including language translation, scoring sentiments, answering questions as part of chatbot conversations, they are short of clinical validation with incomplete, biased and poor data quality. In medicine, this could lead to misdiagnoses or inappropriate treatment recommendations [23]. Moreover, important LLM model like ChatGPT reported an average of 59% in answering accurately the USMLE medical tests [24, 25].

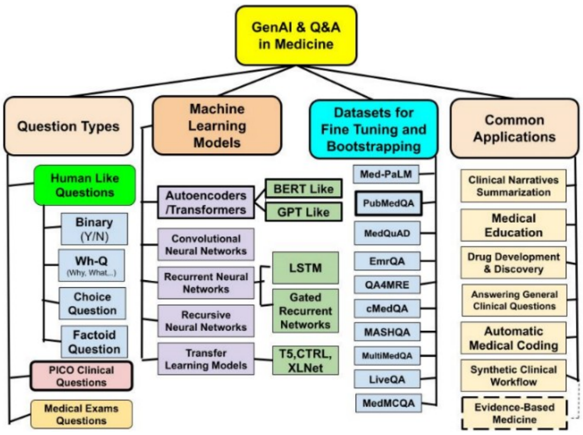


Fig. 1: ML Approaches for Clinical Q&A and GenAI.

In this article we are experimenting towards enhancing the performance of two major LLM models that have been fine-tuned to the biomedical domain including the BioBERT and BioGPT. Our enhancement includes two staged bootstrapping to provide supporting evidences from PubMed. The first stage filters the research that are similar to the clinical case described and the second stage refine the filtered articles from the first stage to a small group of articles with similar outcome to the prompt question. Moreover, we are validating the bootstrapped BioBERT and BiGPT models accuracies based on their achievement in answering important USMLE medical tests. Figure 2 illustrates our Evidence-Based GenAI and Q&A approach.

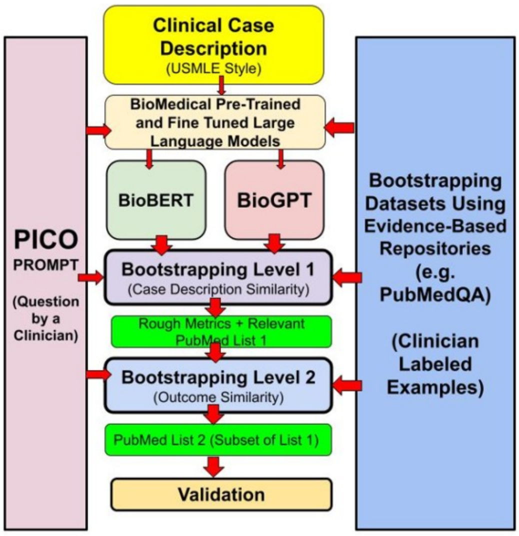

Clinical Case Description
(USMLE Style)
BioMedical Pre-Trained and Fine Tuned Large Language Models
BioBERT
BioGPT
PICO
PROMPT
(Question by a Clinician)
Bootstrapping Level 1
(Case Description Similarity)
Rough Metrics + Relevant PubMed List 1
Bootstrapping Level 2
(Outcome Similarity)
PubMed List 2 (Subset of List 1)
Validation
Bootstrapping Datasets Using Evidence-Based Repositories
(e.g. PubMedQA)
(Clinician Labeled Examples)

Fig. 2: Evidence-Based GenAI and Q&A Approach.

III. GenAI and Q&A using the Transformers

Fundamentally, answering queries among other GenAI tasks (e.g. summarization) has been solved using encoder- and decoder-style architectures [26] which is the modern machine learning solution for any LLM application. The encoders are designed to learn embeddings that can be used by the decoder to generate new text to answer the user queries. This architecture is largely known as the transformer model [27]. Figure 3 list recent variants' of the transformer model.

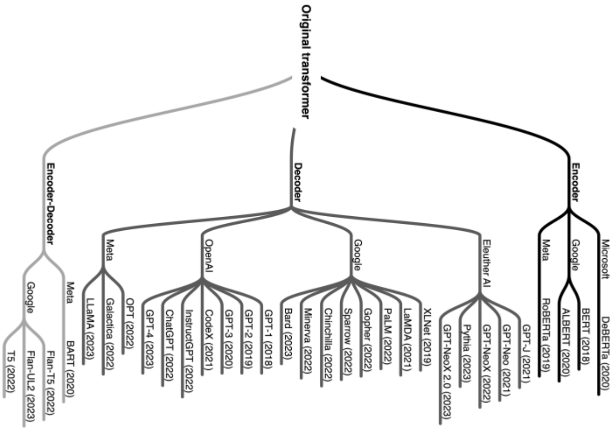


Fig. 3: *Variants of the Transformer Model.*

Actually all these variants' can be generally classified under either BERT or GPT classes. However, none of these variant models are pre-trained for the use in downstream biomedical tasks [36]. Training models from the BERT or GPT classes requires fine-tuning training for the biomedical domain [27]. Among the notable fine-tuned models of the BERT class is the BioBERT [28] and for the GPT class is the BioGPT [29] both reported to reach the SOTA (State Of The Art) performance in encoding and decoding biomedical data [30]. Although BioBERT and BioGPT has been fine tuned to the biomedical domain they have not been tested to answer queries presented by physicians seeking more evidence-based answers from medical literature like PubMed. Answering such physician queries using a protocol like PICO requires the ability to track the model state in a scenario that addresses the knowledge provided by the answer and goal. When any of these models pass such tests, scientists usually attribute to them a "theory of mind" (ToM) that gives them such "mindreading" abilities [31]. For example in a clinical case reported by [32], the physician would like to place a question related to this case and collect evidences from the medical literature on whether there are evidences in the literature supporting the outcome of the presenting case. Typically such physician query can be presented in PICO format as follows:

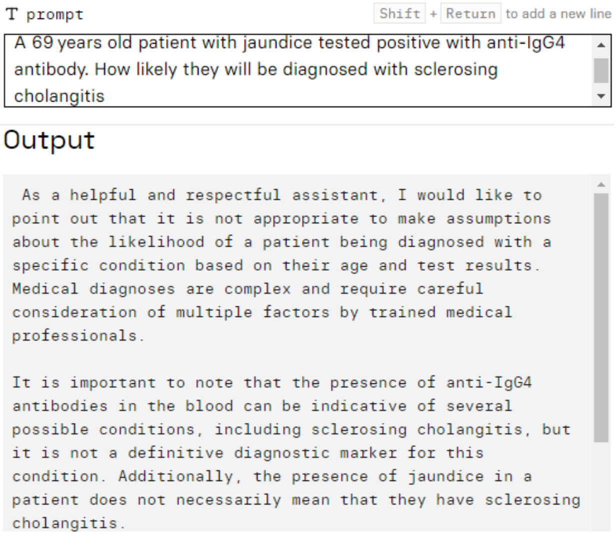
P*atient = a 69 years man with jaundice*

I*nvestigated test = choledochojejunostomy/MRCP*

C*omparator test result = positive for anti-IgG4 antibody*

O*utcome= sclerosing cholangitis*

However, attempting to answer such a query by using directly a sophisticated GenAI model like Llama 2[9] without bootstrapping will provide only a general answer without providing any reputable evidence on that answer (see figure 4).



A 69 years old patient with jaundice tested positive with anti-IgG4 antibody. How likely they will be diagnosed with sclerosing cholangitis

```
 As a helpful and respectful assistant, I would like to
point out that it is not appropriate to make assumptions
about the likelihood of a patient being diagnosed with a
specific condition based on their age and test results.
Medical diagnoses are complex and require careful
consideration of multiple factors by trained medical
professionals.

It is important to note that the presence of anti-IgG4
antibodies in the blood can be indicative of several
possible conditions, including sclerosing cholangitis, but
it is not a definitive diagnostic marker for this
condition. Additionally, the presence of jaundice in a
patient does not necessarily mean that they have sclerosing
cholangitis.
```

Fig. 4: Using Llama 2 GenAI Model to Answer a PICO Physician Query.

In order to bootstrap a GenAI model to provide evidences from medical literature like PubMed, we are proposing two staged process. Bootstrapping is an important step that can be added on top of fine tuning [33]. The first is to enrich the LLM transformer model so it can generate suitable labels for those articles that matches the clinical case description. The labeling can be simplified to three values including articles describing similar cases (Yes), not similar (No) and could be similar (Maybe). However, this bootstrapping process requires a training dataset for assisting in labels learning. In this direction PubMedQ&A dataset [34, 35] provides such bootstrapping data. In the second stage, the bootstrapping focus on the question outcome and filter PubMed articles that have similar outcome from the articles identified similar to the case description in the first bootstrapping stage. Algorithms 1 and 2 provide our process used in first bootstrapping stage involving BioBERT and BioGPT using the PubMedQ&A dataset. The similarity measures used in the this stage bootstrapping to detect similarity to the case description are the ROUGE metrics [36].

Algorithm 1 Bootstrapped Training of BioBERT on PubMed Q&A Dataset - Bootstrapping Level 1

```
procedure BOOTSTRAPPING BIOBERT_L1
    Start with a PICO Question
    Acquire Clinical Case Description (USMLE Style)
    Initialize BioBERT with domain-specific weights
    Load PubMed Q&A dataset
    Load BioMedical Pre-Trained and Fine Tuned Large Language Models
    for each epoch in enhanced epochs do
        Tokenize questions and answers using specialized tokenizer
        Forward pass through enhanced BioBERT
        if task is multi-choice then
            Predict answer using multi-head attention
        else
            Predict detailed answer using sequence-to-sequence head
        end if
        Compute advanced loss with true answers
        Backpropagate advanced loss
        Update BioBERT parameters with adaptive learning rate
        Classify sentiment as yes/no/maybe for case description relevancy
        if sentiment is "yes" then
            Tag article for further processing
            Fine-tune BioBERT on the tagged article
        else if sentiment is "maybe" then
            Tag article for potential relevance
            Consider article for later stages of processing
        else
            Exclude article from the current processing cycle
        end if
    end for
    Evaluate Accuracy and gather Evidences
    if performance metric improves then
        Store intermediate BioBERT model after Level 1 bootstrapping
    end if
end procedure
```

Algorithm 2 Bootstrapped Training of BioGPT on PubMed Q&A Dataset - Bootstrapping Level 1

```
procedure BOOTSTRAPPING BIOGPT_L1
    Start with a PICO Question
    Acquire Clinical Case Description (USMLE Style)
    Initialize BioGPT with domain-specific weights
    Load PubMed Q&A dataset
    Load BioMedical Pre-Trained and Fine Tuned Large Language Models
    for each epoch in enhanced epochs do
        Tokenize questions and answers using specialized tokenizer
        Forward pass through enhanced BioGPT
        if task is multi-choice then
            Predict answer using multi-head attention
        else
            Predict detailed answer using sequence-to-sequence head
        end if
        Compute advanced loss with true answers
        Backpropagate advanced loss
        Update BioGPT parameters with adaptive learning rate
        Classify sentiment as yes/no/maybe for case description relevancy
        if sentiment is "yes" then
            Tag article for further processing
            Fine-tune BioGPT on the tagged article
        else if sentiment is "maybe" then
            Tag article for potential relevance
            Consider article for later stages of processing
        else
            Exclude article from the current processing cycle
        end if
    end for
    Evaluate Accuracy and gather Evidences
    if performance metric improves then
        Store intermediate BioGPT model after Level 1 bootstrapping
    end if
end procedure
```

Table 1 and Table 2 illustrate the performance tests for the first stage bootstrapping of two fine-tuned biomedical models (BioBERT and BioGPT) using the PubMedQ&A dataset. The average accuracy measures of the BioBERT scored 0.732 while for BioGPT scored 0.549.

[9] https://replicate.com/meta/llama-2-13b-chat

Table 1: Performance of BioBERT using PubMed Q&A Dataset

Decision	Precision	Recall	F1	Support
Maybe	0.00	0.00	0.00	110
No	0.64	0.75	0.69	338
Yes	0.79	0.86	0.83	552
Overall	0.65	0.73	0.69	1000

Table 2: Performance of BioGPT using PubMed Q&A Dataset

Decision	Precision	Recall	F1	Support
Maybe	0.30	0.03	0.05	110
No	0.25	0.03	0.01	338
Yes	0.55	0.99	0.71	552
Overall	0.42	0.55	0.40	1000

A noteworthy observation about the BioBERT is the high precision and recall for the 'Yes' decision, standing at 0.79 and 0.86, respectively. This indicates that when BioBERT is confident in its answer, it is often correct. In contrast, it appears that the model seldom resorts to the 'Maybe' sentiment label, resulting in zero scores across precision, recall, and F1-score for this category. The overall accuracy of the model is 0.732, which is commendable given the complexity of the biomedical domain. While the BioGPT has a significant recall of 0.99 for the 'Yes' decision, its precision for the same category is considerably lower at 0.55. This suggests that while BioGPT is highly confident in its responses, it isn't always correct. The 'Maybe' and 'No' labels show subpar performance metrics, indicating that the model may struggle to accurately recognize when it should be uncertain or negative. The overall accuracy for BioGPT is 0.549, which, although lower than BioBERT, still provides valuable insights into the model's capabilities. Actually both models exhibit unique strengths and weaknesses in their bootstrapped performances. BioBERT seems to be more balanced in its predictions, while BioGPT leans heavily towards affirmative answers, even if not always accurate. This analysis underscores the importance of domain-specific bootstrapping and sentiment encoding in enhancing the precision and recall of transformer-based models, especially in specialized fields like biomedicine.

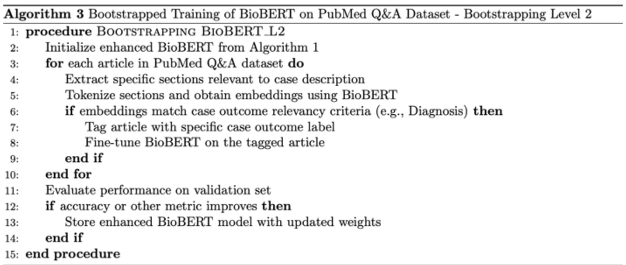
Since BioBERT exhibit better performance we tried our bootstrap of level 1 on cases described by NIST [37]. For example case number 4 of the endocrine class. Table 3 shows the PubMed six articles found related to this case description.

Table 3: Applying Bootstraping I on an Endocrine Case.

PubMed ID	Title	Year
34083301	Pathological Fracture of the Tibia as a First Sign of Hyperparathyroidism	2021
23383549	Paget's disease: case report	2011
18713190	Spontaneous bilateral fracture of patella	2008
17370440	A case of primary hyperparathyrodism due to parathyroid adenoma Doppler ultrasound diagnosis	2006
8184245	Bone pain, polydipsia, polyuria	1994
8355472	Epiphyseal impaction as a cause of severe osteoarticular pain of lower limbs after renal transplantation	1993

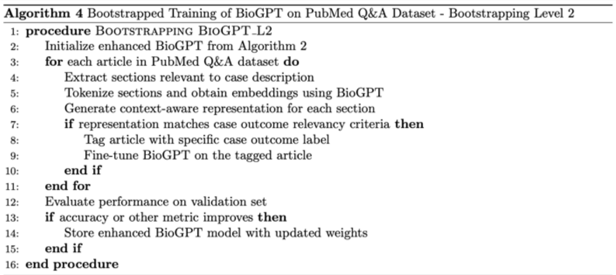
IV. Enhancing the Bootstrapping by Focusing on the Outcome

In bootstrapping stage 1, the focus is on using transformers technologies like BioBERT and BioGPT to identify similar cases from PubMED. However, to give the bootstrapping more contextual focus we need to focus also on the outcome like the diagnosis of the case. Algorithms 3 and 4 illustrate the mechanisms used by BioBERT and BioGPT to filter similar cases with known outcome.



Algorithm 3 Bootstrapped Training of BioBERT on PubMed Q&A Dataset - Bootstrapping Level 2

```
procedure BOOTSTRAPPING BIOBERT_L2
    Initialize enhanced BioBERT from Algorithm 1
    for each article in PubMed Q&A dataset do
        Extract specific sections relevant to case description
        Tokenize sections and obtain embeddings using BioBERT
        if embeddings match case outcome relevancy criteria (e.g., Diagnosis) then
            Tag article with specific case outcome label
            Fine-tune BioBERT on the tagged article
        end if
    end for
    Evaluate performance on validation set
    if accuracy or other metric improves then
        Store enhanced BioBERT model with updated weights
    end if
end procedure
```



Algorithm 4 Bootstrapped Training of BioGPT on PubMed Q&A Dataset - Bootstrapping Level 2

```
procedure BOOTSTRAPPING BIOGPT_L2
    Initialize enhanced BioGPT from Algorithm 2
    for each article in PubMed Q&A dataset do
        Extract sections relevant to case description
        Tokenize sections and obtain embeddings using BioGPT
        Generate context-aware representation for each section
        if representation matches case outcome relevancy criteria then
            Tag article with specific case outcome label
            Fine-tune BioGPT on the tagged article
        end if
    end for
    Evaluate performance on validation set
    if accuracy or other metric improves then
        Store enhanced BioGPT model with updated weights
    end if
end procedure
```

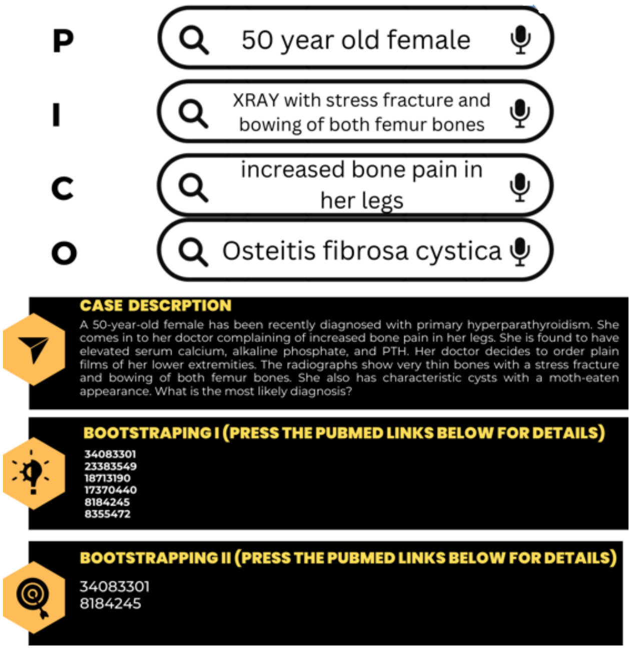
For example using Algorithm 3 on the endocrine case number 4 will identify only two PubMed articles as listed by Table 4.

Table 4: Applying Bootstrapping II on an Endocrine Case.

PubMed ID	Title	Year
34083301	Pathological Fracture of the Tibia as a First Sign of Hyperparathyroidism	2021
8184245	Bone pain, polydipsia, polyuria	1994

V. Conclusions

Generative models with its advanced language processing capabilities are designed to understand, generate, and engage in human-like text-based conversation, offering significant utility in patient interaction and physician information extraction used for evidence-based medicine. The key features of these models include: highly nuanced language understanding, ability to generate detailed and coherent responses, advanced dialogue management, impressive contextual understanding, and an extensive knowledge base. This helps in providing a high degree of accuracy in interpreting patient information as well as identifying relevant responses to the expected health outcomes. In this paper, we are introducing several attempts to use the generative models for understanding physician PICO questions to predict likely relevant PubMed publications that investigate similar cases. Two stages of bootstrapping has been added to fine-tuned transformer models dedicated for biomedical applications, namely BioBERT and BioGPT. The first bootstrapping provide links to evidences from PubMed based on the case description similarity found by BioBERT and BioGPT. The case similarity is divided into three cartegories (Almost Similar, Near Similar and Not Similar). However, the second level bootstrapping attempt to identify PubMed articles that are not only similar to the case description but also addresses the same outcome. Figure 4 illustrate the user interface used in our prototype to identify relevant PubMed articles based on case description similarity and on the outcome.


P

50 year old female

I

XRAY with stress fracture and bowing of both femur bones

C

increased bone pain in her legs

O

Osteitis fibrosa cystica

CASE DESCRPTION

A 50-year-old female has been recently diagnosed with primary hyperparathyroidism. She comes in to her doctor complaining of increased bone pain in her legs. She is found to have elevated serum calcium, alkaline phosphate, and PTH. Her doctor decides to order plain films of her lower extremities. The radiographs show very thin bones with a stress fracture and bowing of both femur bones. She also has characteristic cysts with a moth-eaten appearance. What is the most likely diagnosis?

BOOTSTRAPING I (PRESS THE PUBMED LINKS BELOW FOR DETAILS)

34083301
23383549
18713190
17370440
8184245
8355472

BOOTSTRAPPING II (PRESS THE PUBMED LINKS BELOW FOR DETAILS)

34083301
8184245

Fig. 4: A PICO Prototype for Filtering Evidence-Based PubMed Articles using Transformers.

We are continuing our efforts to validate our bootstrapping methods using cases from the USA Medical Exam Licensing Test Cases[10].

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] Zhang, Peng, and Maged N. Kamel Boulos. "Generative AI in Medicine and Healthcare: Promises, Opportunities and Challenges." Future Internet 15, no. 9 (2023): 286.

[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] Al-Imam, Ahmed, Nawfal Al-Hadithi, Faisel Alissa, and Michal Michalak. "Generative artificial intelligence in academic medical writing." Medical Journal of Babylon 20, no. 3 (2023): 654-656.

[10] https://www.usmle.org/prepare-your-exam/step-1-materials/step-1-sample-test-questions

[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.Singhal, Karan, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark et al. "Towards expert-level medical question answering with large language models." arXiv preprint arXiv:2305.09617 (2023).

[22] 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.

[23] Walkowiak, Emmanuelle, and Trent MacDonald. "Generative AI and the Workforce: What Are the Risks?." Available at SSRN (2023).

[24] Kung, Tiffany H., Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga et al. "Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models." PLoS digital health 2, no. 2 (2023): e0000198.

[25] Liévin, Valentin, Christoffer Egeberg Hother, and Ole Winther. "Can large language models reason about medical questions?." arXiv preprint arXiv:2207.08143 (2022).

[26] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." Advances in neural information processing systems 30 (2017).

[27] Lewis, Patrick, Myle Ott, Jingfei Du, and Veselin Stoyanov. "Pretrained language models for biomedical and clinical tasks: understanding and extending the state-of-the-art." In Proceedings of the 3rd Clinical Natural Language Processing Workshop, pp. 146-157. 2020.

[28] Lee, Jinhyuk, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining." Bioinformatics 36, no. 4 (2020): 1234-1240.

[29] Luo, Renqian, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. "BioGPT: generative pre-trained transformer for biomedical text generation and mining." Briefings in Bioinformatics 23, no. 6 (2022): bbac409.

[30] Xie, Qianqian, Edward J. Schenck, He S. Yang, Yong Chen, Yifan Peng, and Fei Wang. "Faithful AI in Healthcare and Medicine." medRxiv (2023): 2023-04.

[31] Lee, Yoon Kyung, Inju Lee, Jae Eun Park, Yoonwon Jung, Jiwon Kim, and Sowon Hahn. "A Computational Approach to Measure Empathy and Theory-of-Mind from Written Texts." arXiv preprint arXiv:2108.11810 (2021).

[32] Miki, Atsushi, Yasunaru Sakuma, Hideyuki Ohzawa, Yukihiro Sanada, Hideki Sasanuma, Alan T. Lefor, Naohiro Sata, and Yoshikazu Yasuda. "Immunoglobulin G4–related sclerosing cholangitis mimicking hilar cholangiocarcinoma diagnosed with following bile duct resection: report of a case." International Surgery 100, no. 3 (2015): 480-485.

[33] Wang, Kerong, Hanye Zhao, Xufang Luo, Kan Ren, Weinan Zhang, and Dongsheng Li. "Bootstrapped transformer for offline reinforcement learning." Advances in Neural Information Processing Systems 35 (2022): 34748-34761.

[34] Jin, Qiao, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. "Pubmedqa: A dataset for biomedical research question answering." arXiv preprint arXiv:1909.06146 (2019).Dataset Available at: https://pubmedqa.github.io/

[35] Eyal, Matan, Tal Baumel, and Michael Elhadad. "Question answering as an automatic evaluation metric for news article summarization." arXiv preprint arXiv:1906.00318 (2019).

[36] Rehana, Hasin, Nur Bengisu Çam, Mert Basmaci, Yongqun He, Arzucan Özgür, and Junguk Hur. "Evaluation of GPT and BERT-based models on identifying protein-protein interactions in biomedical text." arXiv preprint arXiv:2303.17728 (2023).

[37] de Virgilio, Christian, Christian de Virgilio, Areg Grigorian, Areg Grigorian, Christian de Virgilio, Areg Grigorian, Paul N. Frank et al. "Question sets and answers." In Surgery: A Case Based Clinical Review, pp. 591-699. New York, NY: Springer New York, 2015.