Empowering Transformers for Evidence-Based Medicine

Sabah Mohammed, Jinan Fiaidhi and Hashmath Shaik

Abstract— Breaking the barrier for practicing evidence-based medicine rely on effective methods for rapidly identifying relevant evidences from the body of biomedical literature. An important challenge confronted by the medical practitioners is the long time needed to browse, filter, summarize and compile information from different medical resources. Deep learning can help in solving this based on the automatic question answering (Q&A) and transformers. However, Q&A and transformers technologies 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 transformer 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 transformer methods are reaching an acceptable 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. Moreover, we are also reporting experimentations to empower our bootstrapping techniques with patch attentions to the most important keywords in the clinical case and the PICO questions. Our bootstrapped patched with attention is showing relevancy of the evidences collected based on an entropy metrics.

Keyword— Automatic Question Answering, PICO questions, Evidence-Based Medicine, Generative Models, LLM Transformers, Bootstrapping, Attention Patching, Attention Entropy.

I. INTRODUCTION

Clinical practitioners rely on Evidence-Based Medicine (EBM) to provide quality care planning based on the best available evidence from sound medical literature or clinical trials. However, the growing number of medical publications and clinical trials are sharply increasing which makes it extremely difficult to stay updated [1]. The best available practice of collecting clinical evidences is to present clinical questions around the clinical 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 research areas like Question Answering (Q&A) and Generative AI (GenAI) based on transformers which can automatically identify relevant clinical articles or trials based on clinical description and their associated PICO questions [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], HealthTap[3], eHealthForums[4], patientslikeme[5], PubMed[6], Medical Encyclopedia[7] and iCliniq[8] 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

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).

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

(FAQs) portals established by almost every healthcare institution like the CDC.[9] To solve this problem several researchers from the natural language and machine learning 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].
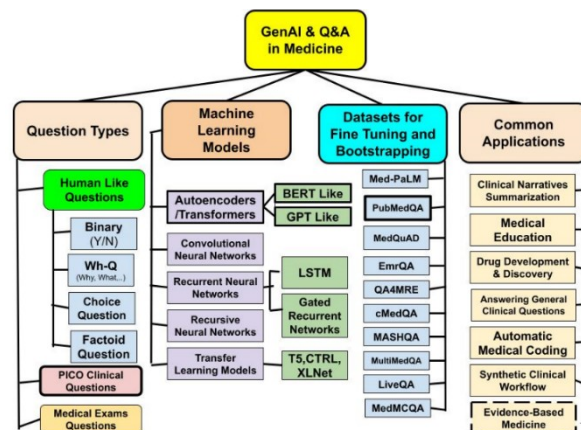

Fig. 1: ML Approaches for Clinical Q&A and GenAI.

In this article we are experimenting towards enhancing the performance of two major LLM models that have been

---

9 https://www.cdc.gov/

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 compared to the BioLinkBERT model. Moreover, our approach introduces another empowerment level by patching attention to the bootstrapping process. It starts with attention visualization followed by patch attention to our bootstrapped models and filtering evidences based on the attention entropy metrics. Figure 2 illustrates our Evidence-Based Transformers and Q&A approach.
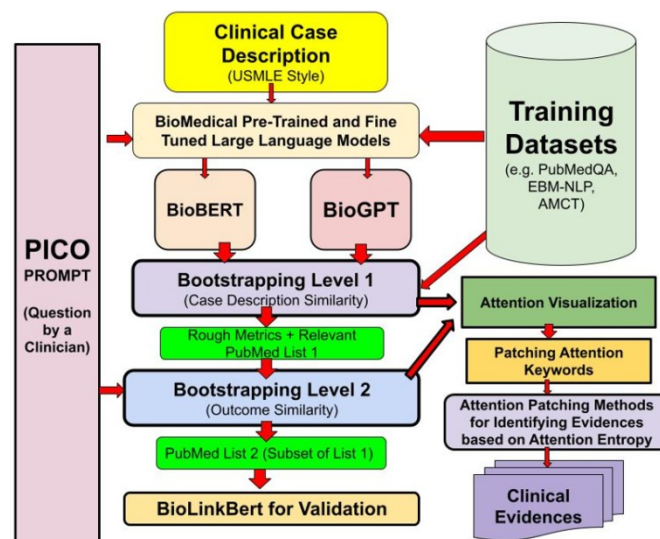

Fig. 2: Evidence-Based Transformers and Q&A Approach.

III. DEEP LEARNING 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.
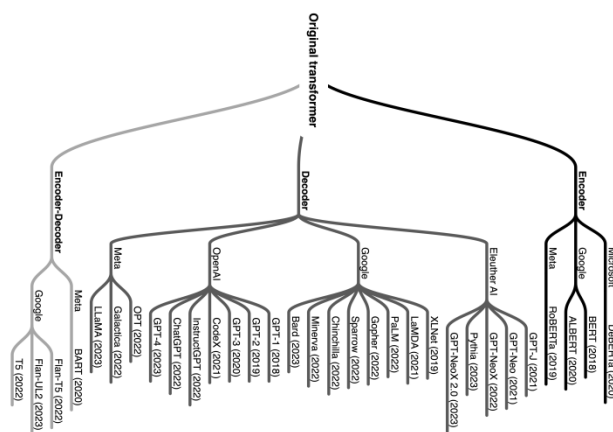
Fig. 3: Variants of the Transformer Model.

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 [28]. Among the notable fine-tuned models of the BERT class is the BioBERT [29] and for the GPT class is the BioGPT [30] both reported to reach the SOTA (State Of The Art) performance in encoding and decoding biomedical data [31]. 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 [32]. For example in a clinical case reported by [33], 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:

Patient = a 69 years man with jaundice

Investigated test = choledochojejunostomy/MRCP

Comparator test result = positive for anti-IgG4 antibody

Outcome= sclerosing cholangitis

However, attempting to answer such a query by using directly a sophisticated GenAI model like Llama 2[10] without bootstrapping will provide only a general answer without providing any reputable evidence on that answer (see figure 4).

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

In order to bootstrap a GenAI model in order to provide evidences from medical literature like PubMed, we are proposing two staged process. 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 bootstrapping to detect similarity to the case description are the ROUGE metrics [36].

Algorithm 1 Bootstrapped Training of BioBERT on PubMed Q&A Dataset - Blind Folded Bootstrapping Technique

```
 1: procedure BOOTSTRAPPING BIOBERT_L1
 2:     Start with a PICO Question
 3:     Acquire Clinical Case Description (USMLE Style)
 4:     Initialize BioBERT with domain-specific weights
 5:     Load PubMed Q&A dataset
 6:     Load BioMedical Pre-Trained and Fine Tuned Large Language Models
 7:     for each epoch in enhanced epochs do
 8:         Tokenize questions and answers using specialized tokenizer
 9:         Forward pass through enhanced BioBERT
10:         if task is multi-choice then
11:             Predict answer using multi-head attention
12:         else
13:             Predict detailed answer using sequence-to-sequence head
14:         end if
15:         Compute advanced loss with true answers
16:         Backpropagate advanced loss
17:         Update BioBERT parameters with adaptive learning rate
18:         Classify sentiment as yes/no/maybe for case description relevancy
19:         if sentiment is "yes" then
20:             Tag article for further processing
21:             Fine-tune BioBERT on the tagged article
22:         else if sentiment is "maybe" then
23:             Tag article for potential relevance
24:             Consider article for later stages of processing
25:         else
26:             Exclude article from the current processing cycle
27:         end if
28:     end for
29:     Evaluate Accuracy and gather Evidences
30:     if performance metric improves then
31:         Store intermediate BioBERT model after bootstrapping
32:     end if
33: end procedure
```

---

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

```
Algorithm 2 Bootstrapped Training of BioGPT on PubMed Q&A Dataset - Blind Folded Bootstrapping
Technique
 1: procedure BOOTSTRAPPING BIOGPT_L1
 2:    Start with a PICO Question
 3:    Acquire Clinical Case Description (USMLE Style)
 4:    Initialize BioGPT with domain-specific weights
 5:    Load PubMed Q&A dataset
 6:    Load BioMedical Pre-Trained and Fine Tuned Large Language Models
 7:    for each epoch in enhanced epochs do
 8:        Tokenize questions and answers using specialized tokenizer
 9:        Forward pass through enhanced BioGPT
10:        if task is multi-choice then
11:            Predict answer using multi-head attention
12:        else
13:            Predict detailed answer using sequence-to-sequence head
14:        end if
15:        Compute advanced loss with true answers
16:        Backpropagate advanced loss
17:        Update BioGPT parameters with adaptive learning rate
18:        Classify sentiment as yes/no/maybe for case description relevancy
19:        if sentiment is "yes" then
20:            Tag article for further processing
21:            Fine-tune BioGPT on the tagged article
22:        else if sentiment is "maybe" then
23:            Tag article for potential relevance
24:            Consider article for later stages of processing
25:        else
26:            Exclude article from the current processing cycle
27:        end if
28:    end for
29:    Evaluate Accuracy and gather Evidences
30:    if performance metric improves then
31:        Store intermediate BioGPT model after bootstrapping
32:    end if
33: end procedure
```

Table 1 and Table 2 illustrate the use of blindfolded bootstrapping of the two models (BioBERT and BioGPT) using the PubMed Q&A dataset. The accuracy measures of the BioBERT scored 0.732 while for BioGPT scored 0.549.

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

Decision	Precision	Recall	F1-Score	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-Score	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 performance of the blindfolded bootstrapping of 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 correlation with the case description.

In contrast, it appears that the BioBERT model seldom resorts to the 'Maybe' 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 correlation, it isn't always reliable. The 'Maybe' and 'No' labels show subpar performance metrics, indicating that the model may struggle to accurately recognize any correlation when it should be uncertain or negative. The overall accuracy for BioGPT is 0.549, which, although lower than BioBERT, still provides some insight into the model's

capabilities. In summary, both models exhibit unique strengths and weaknesses in their blindfolded bootstrapped performances. BioBERT seems to be more balanced in its predictions, while BioGPT leans heavily towards affirmative answers, even if not always accurate. However, the low correlation performances of both models is expected due to the focus on all the important keywords provided by the case description with no mentioning to possible outcomes like diagnosis.

IV.   ENHANCING THE BOOTSTRAPPING OF Q&A

In this section we are investigating an additional boostrapping to the two fine-tuned LLM models (BioBERT and BioGPT) where the correlation is directed with an attention to the case outcomes (e.g. the diagnosis). For this we are considering 16 cases that have been described with diagnosis from sound clinical cases used in medical training [37]. Algorithms 3 and 4 illustrate this new guided bootstrapping with the clinical outcome provided. The additional bootstrapping considers adding a PICO wrapper that helps to provide the additional information (e.g. outcome of the case) needed for guiding the correlation between the case description and the PubMed searched by the PICO protocol.

```
Algorithm 3 Bootstrapped Training of BioBERT on PubMed Q&A Dataset - Unblinded bootstrapping
Technique
 1: procedure BOOTSTRAPPING BIOBERT_L2
 2:    Initialize enhanced BioBERT from Algorithm 1
 3:    for each article in PubMed Q&A dataset do
 4:        Extract specific sections relevant to case description
 5:        Tokenize sections and obtain embeddings using BioBERT
 6:        if embeddings match case outcome relevancy criteria (e.g., Diagnosis) then
 7:            Tag article with specific case outcome label
 8:            Fine-tune BioBERT on the tagged article
 9:        end if
10:    end for
11:    Evaluate performance on validation set
12:    if accuracy or other metric improves then
13:        Store enhanced BioBERT model with updated weights
14:    end if
15: end procedure
```

```
Algorithm 4 Bootstrapped Training of BioGPT on PubMed Q&A Dataset - Unblinded bootstrapping
Technique
 1: procedure BOOTSTRAPPING BIOGPT_L2
 2:    Initialize enhanced BioGPT from Algorithm 2
 3:    for each article in PubMed Q&A dataset do
 4:        Extract sections relevant to case description
 5:        Tokenize sections and obtain embeddings using BioGPT
 6:        Generate context-aware representation for each section
 7:        if representation matches case outcome relevancy criteria then
 8:            Tag article with specific case outcome label
 9:            Fine-tune BioGPT on the tagged article
10:        end if
11:    end for
12:    Evaluate performance on validation set
13:    if accuracy or other metric improves then
14:        Store enhanced BioGPT model with updated weights
15:    end if
16: end procedure
```

We decided to test our outcome guided algorithms using a clinical case from [37: Cardiothoracic Case No. 5] for a patient that we will name John Doe:

A 65-year-old woman arrives to the ED complaining of chest pain. Her past medical history includes hypertension, atherosclerosis, and coronary artery disease. She underwent a coronary artery bypass graft (CABG) 3 weeks ago for three-vessel disease. She reports that her chest pain worsens with inspiration and lessens when leaning forward. A friction rub is heard on auscultation. ECG shows global ST elevation.

The corresponding PICO query for the above patient:

- o **P (Patient/Problem):** A 65-year-old woman with a history of hypertension, atherosclerosis, coronary artery disease, and recent coronary artery bypass graft (CABG) for three-vessel disease, presenting to the ED with chest pain that worsens with inspiration and alleviates when leaning forward, accompanied by a friction rub on auscultation and global ST elevation on ECG.
- o **I (Intervention):** Evaluation and management of suspected post-cardiac surgery pericarditis.
- o **C (Comparison):** Usual care or other differential diagnoses management like acute coronary syndrome management.
- o **O (Outcome):** Relief of chest pain, resolution of ECG changes, prevention of complications like constrictive pericarditis or cardiac tamponade, and improvement in overall patient's clinical status.

The corresponding PubMed query generated by our guided bootstrapping:

> *("postoperative pericarditis" OR "pericarditis post CABG") AND ("management" OR "treatment") AND ("ST elevation" OR "chest pain")*

Since this case describes a chest pain, clerks may use differential diagnosis to identify the case according to the following option list [42]:

Nonischemic cardiovascular
        Aortic dissection
        Myocarditis
        Pericarditis
        Hypertrophic cardiomyopathy
        Stress cardiomyopathy
Chest wall/musculoskeletal
        Cervical disk disease
        Costochondritis
        Herpes zoster
        Neuropathic pain
        Rib fracture
Pulmonary
        Pneumonia
        Pulmonary embolus
        Tension pneumothorax
        Pleurisy
Gastrointestinal
        Cholecystitis
        Peptic ulcer disease
            Nonperforating
            Perforating
        Gastroesophageal reflux disease
        Esophageal spasm
        Bocrhaavc syndrome (esophageal rupture with mediastinitis)
        Pancreatitis
Psychiatric
        Depression
        Anxiety disorder/panic attack
        Somatization and psychogenic pain disorder

This approach can be used in a teaching a learning clinical setting but in practice seeking evidence to prove an option

requires further tests and investigations. What is more practical in clinical setting is to use the physician intuition to narrow the options into more likely relevant to the case described. In asking a physician from our local TBRHSC[11] we may end with a shorter list that may point to different options for the diagnosis of this case:

1. **Postoperative infection:** Given the patient's recent surgery, there is a risk of developing an infection, which could present with chest pain that worsens with inspiration and improves with leaning forward. The presence of a friction rub on auscultation suggests inflammation or fluid in the chest cavity.

2. **Sternal wire infection:** As a complication of CABG surgery, the sternal wire used to close the sternum can become infected, leading to chest pain, swelling, and redness at the incision site. The patient's symptoms and signs are consistent with this possibility.

3. **Pneumonia:** The patient's history of hypertension and atherosclerosis increases her risk for developing pneumonia, especially if she has been immobile or oxygen-deprived post-surgery. The chest pain that worsens with inspiration and the presence of a friction rub suggest pneumonia as a possible diagnosis.

4. **Pulmonary embolism:** Although the patient has a history of coronary artery disease, the sudden onset of chest pain and shortness of breath raises the suspicion of pulmonary embolism. The global ST elevation on ECG supports this possibility.

5. **Myocardial infarction (MI):** The patient's history of CAD and recent CABG surgery increase the likelihood of MI, particularly given the chest pain that worsens with inspiration and the ST elevation on ECG. However, the presence of a friction rub and the patient's recent surgical procedure may point more towards postoperative complications rather than MI.

6. **Pericarditis:** The patient's chest pain that worsens with inspiration and lessens when leaning forward, along with the presence of a friction rub, are consistent with pericarditis, an inflammatory condition affecting the pericardium surrounding the heart. If this pain is persistent then it can be called *acute pericarditis*.

Indeed, the three surgeons (Areg Grigorian, Paul N. Frank and Christian de Virgilio) from the Department of Surgery, Harbor-UCLA Medical Center, Torrance, CA USA determined the diagnosis as *acute pericarditis* [37]. The reasoning behind their diagnosis is that this inflammation occurs in the pericardial sac accompanied by pericardial effusion following post-MI (termed Dressler's syndrome), chest radiation, or recent heart surgery. Patients present with pleuritic chest pain that lessens when leaning forward, friction rub heard on auscultation, global ST elevation, and PR depression. In order to test the accuracy of our models (BioBERT and BioGPT) we decided to extend these models by adding a softmax function to determine the relevancy of the searched PubMed articles to the case description. The softmax function report kind of sentiment the searched

---

[11] Thunder Bay Regional Health Science Center (TBRHSC)

PubMed article compared to the given case description [43]. Figure 5 illustrate our Softmax Sentiment Model. The details of testing our Blind-Folded BioBERT and BioGPT in providing relevant PubMed articles that can be used for evidence based medicine is provided in Tables 3 and 4. We are only showing the sentiment for the first 5 PubMed articles.
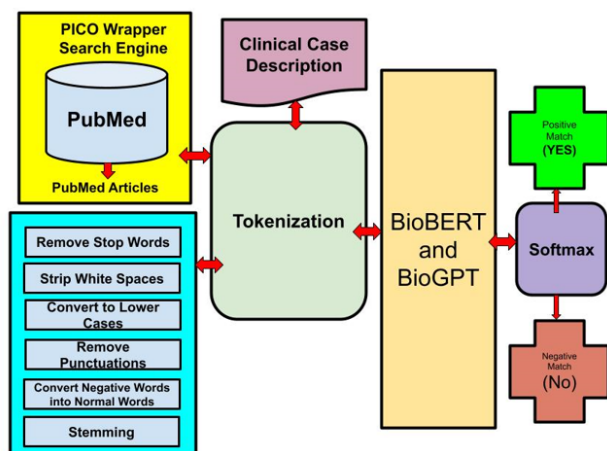

Fig. 5: Sentiment Model between Fetched PubMed Articles and the Case Description.

Table 3: BioBERT Relevancy to the Cardiothoracic Case No. 5

IDX	PubMed ID	Softmax	Evidence Title
1	23388234	no	Acute pericarditis after percutaneous coronary intervention mimicking inferolateral ST-elevation myocardial infarction
2	17921916	yes	Diagnostic value of biohumoral markers of necrosis and inflammation in patients with right ventricular myocardial infarction
3	34067941	no	Acute Pericarditis after Percutaneous Coronary Intervention: A Case Report
4	35018605	yes	CVIT expert consensus document on primary percutaneous coronary intervention (PCI) for acute myocardial infarction (AMI) update 2022
5	32856192	no	Pericarditis and Post-cardiac Injury Syndrome as a Sequelae of Acute Myocardial Infarction

Table 4: BioGPT Relevancy to the Cardiothoracic Case No. 5

IDX	PubMed ID	Softmax	Evidence Title
1	23388234	yes	Acute pericarditis after percutaneous coronary intervention mimicking inferolateral ST-elevation myocardial infarction
2	17921916	no	Diagnostic value of biohumoral markers of necrosis and inflammation in patients with right ventricular myocardial infarction
3	34067941	yes	Acute Pericarditis after Percutaneous Coronary Intervention: A Case Report
4	35018605	yes	CVIT expert consensus document on primary percutaneous coronary intervention (PCI) for acute myocardial infarction (AMI) update 2022
5	32856192	no	Pericarditis and Post-cardiac Injury Syndrome as a Sequelae of Acute Myocardial Infarction

The BioGPT model shows better relevancy compared to the BioBERT using our softmax function. However, we decided to extend our testing by comparing the results the state of art LLM model that has been widely used to target relevant PubMed articles given case description like the BioLinkBERT [39]. Table 5 illustrates the use of BioLinkBert in the Cardiothoracic Case No. 5 mentioned in [37]. Table 5 proves that sound models like BioLinkBert provide irrelevant PubMed articles based on given case description.

Table 5: BioLinkBERT Relevancy to the Cardiothoracic Case No. 5

| IDX	PubMed ID	Softmax	Evidence Title
1	23388234	yes	Acute pericarditis after percutaneous coronary intervention mimicking inferolateral ST-elevation myocardial infarction
2	17921916	no	Diagnostic value of biohumoral markers of necrosis and inflammation in patients with right ventricular myocardial infarction
3	34067941	no	Acute Pericarditis after Percutaneous Coronary Intervention: A Case Report
4	35018605	maybe	CVIT expert consensus document on primary percutaneous coronary intervention (PCI) for acute myocardial infarction (AMI) update 2022
5	32856192	no	Pericarditis and Post-cardiac Injury Syndrome as a Sequelae of Acute Myocardial Infarction

V.   BOOTSTRAPPING WITH ATTENTION PATCHING

In order to enhance the blindfolded bootstrapping we decided to by add an additional attention layer. Our

attention patching approach uses two stages. The first stage attempts to visualize the attention heat map given the case description. For this purpose we are using the attention visualize library described by [40]. Once the heat map is set to identify balanced attention with the parameter Predict = 50% then we can see the focus of the model like the BioBERT on the most important attention keywords like (hypertension, bypass, artery, friction, heard). Figure 6 (a, b and c) illustrates the use of the attention visualization using the Case No. 5 of Cardiothoracic Case in [37].
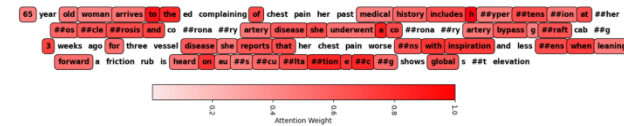
(a) Attention Visualization of BioBERT model
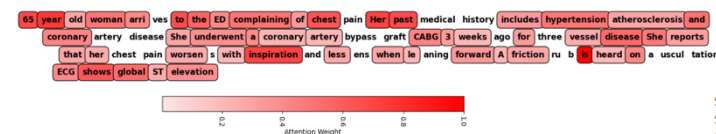
(b) Attention Visualization of BioGPT model

(c) Attention Visualization of BioLinkBERT model
Fig. 6: Attention Visualization and Attention List Extraction

Algorithm 5 illustrate the attention patching that we used for the three LLM models (BioBERT, BioGPT and BioLinkBERT).

```
Algorithm 5 Detailed Algorithm for Implementing and Analyzing Patched Attention in BioBERT, BioGPT,
and BioLinkBERT
 1: Import necessary libraries for NLP, data processing, and visualization
 2: procedure LOADMODEL(modelName)
 3:     Load and return the tokenizer and pre-trained model for 'modelName'
 4:     modelName can be 'BioBERT', 'BioGPT', or 'BioLinkBERT'
 5: end procedure
 6: procedure CUSTOMCOLORMAP(startColor, endColor)
 7:     Generate a linear segmented colormap transitioning from startColor to endColor
 8:     This colormap is used for attention heatmaps and other visualizations
 9: end procedure
10: procedure PATCHATTENTION(model, patchProcedure)
11:     Modify the attention layers in 'model' using 'patchProcedure'
12:     'patchProcedure' alters attention weights based on specific criteria
13:     Ensure that the patched model retains its original architecture
14: end procedure
15: procedure PREPROCESSTEXT(text)
16:     Clean and preprocess 'text' for NLP tasks (e.g., tokenization, normalization)
17:     Return the preprocessed text ready for model input
18: end procedure
19: procedure ANALYZEATTENTION(modelOutputs, tokenizer)
20:     Extract and analyze attention weights from 'modelOutputs'
21:     Use 'tokenizer' to map weights back to textual tokens
22:     Return a structured representation of attention analysis
23: end procedure
24: procedure VISUALIZEATTENTION(analysisData, colormap)
25:     Create visualizations (e.g., heatmaps) of attention analysis using 'analysisData'
26:     Apply 'colormap' for visual consistency and interpretability
27:     Save or display the visualizations
28: end procedure
29: procedure IDENTIFYRELEVANTPUBMED(texts, tokenizer, model, patchProcedure)
30:     for each text in texts do
31:         text ← PREPROCESSTEXT(text)
32:         Apply PATCHATTENTION(model, patchProcedure) to the model
33:         modelOutputs ← Model predictions and attention weights for 'text'
34:         analysisData ← ANALYZEATTENTION(modelOutputs, tokenizer)
35:         VISUALIZEATTENTION(analysisData, CustomColormap)
36:         Use the Siamese Neural Network to Measure Entropy between the PubMed article and the Case
        Description
37:     end for
38: end procedure
39: procedure BOOTSTRAPSAMPLE(dataset, numIterations, patchProcedure)
40:     for i from 1 to numIterations do
41:         sample ← Randomly sample with replacement from 'dataset'
42:         PROCESSTEXTS(sample, tokenizer, model, patchProcedure)
43:     end for
44: end procedure
```

The heart of the attention patching is a function that augment attention to the LLM model and find similarity using an entropy function designed as Siamese neural network. Algorithm 6 illustrate our entropy function used to identify relevant PubMed articles after parching the extra attention keywords.

```
Algorithm 6 Siamese Network for Comparing Patient History and PubMed Articles
 1: Preparation:
 2: Load Bootstrapped Model.
 3: Initialize tokenizer and SiameseNetwork.
 4: procedure CLEANTEXT(text)
 5:     Remove punctuation, lowercase, and other cleaning steps.
 6:     return cleaned text.
 7: end procedure
 8: procedure TOKENIZE(sentences)
 9:     tokens ← Tokenize and clean input sentences using CLEANTEXT.
10:     return tokens
11: end procedure
12: procedure COMPUTESIMILARITY(history, articles)
13:     Initialize an empty list all_scores.
14:     for each sentence h in history do
15:         Initialize an empty list scores.
16:         for each sentence a in articles do
 :             score ← Model inference with h and a.
 :             Append score to scores.
 :         end for
 :         Append max(scores) to all_scores.
 :     end for
22:     Calculate overall_entropy as mean of all_scores.
23:     return overall_entropy
24: end procedure
25: procedure MAIN
26:     Load dataset from specified path.
27:     Preprocess the dataset by cleaning and tokenizing.
28:     for each article in dataset do
29:         Prepare patient history and article text.
30:         Tokenize texts using TOKENIZE.
31:         entropy ← COMPUTESIMILARITY(tokenized_history, tokenized_articles)
32:         Print entropy.
33:     end for
34: end procedure
35: Execute MAIN
```
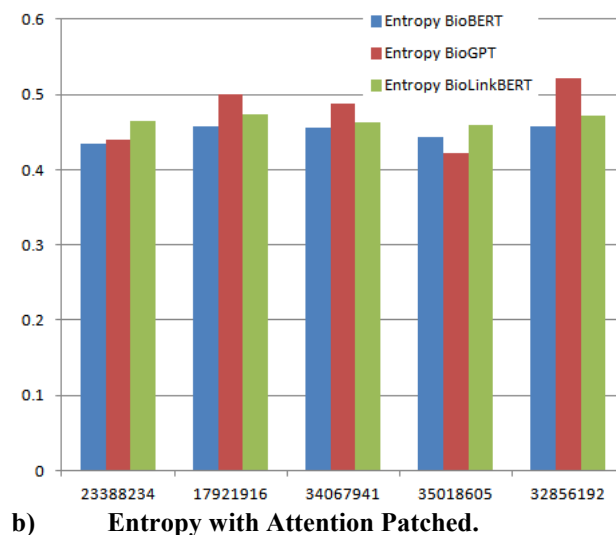
Based on the entropy function, the importance of attention patching become clear. Figure 7 a and b illustrates using the entropy function to identify relevant pubMed articles without and with attention patching.
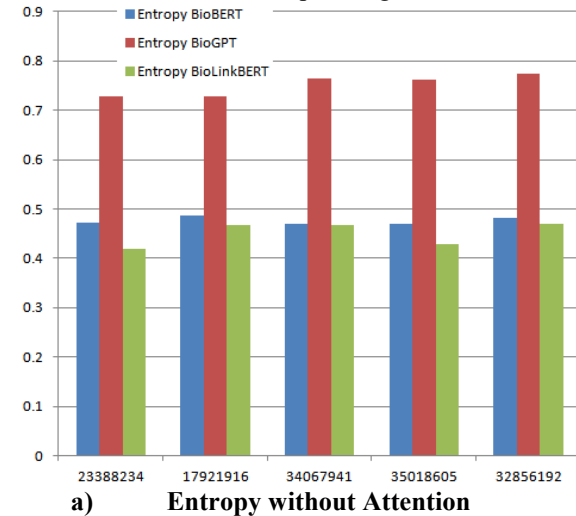
a)   **Entropy without Attention**

b)   Entropy with Attention Patched.

Fig 7: Testing BioBERT, BioGPTand BioLinkBERTwithout and with Attention Patching for Case No. 5 of Cardiothoracic Case in [37].

VI.   CONCLUSION

Generative models that are based on transformers are designed to understand, generate, and engage in human-like text-based conversation as well as to answer queries by identifying relevant documentations. In evidence based practice this capability is offering significant utility in answering physician queries used for evidence-based medicine. The key features of the transformer models include: highly nuanced language understanding, ability to generate detailed and coherent responses, advanced dialogue management, impressive contextual understanding, identifying relevant materials, and summarization based on extensive trained 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. We have introduced two transformer based bootstrapping techniques to identify PubMed relevant articles based on clinical case description (Level 1) and clinical case description with attention patching (Level 2). The transformer models used in the two bootstrapping were the BioBERT and the BioGPT. We verified their performance with BioLinkBERT and find their performance is comparable and better in some cases. Moreover we empowered the attention parching with better similarity function than the Softmax used earlier in Level 1 to using an Entropy function designed as Siamese Neural Network.

ACKNOWLEDGMENT

REFERENCES

[1]   Bastian, Hilda, Paul Glasziou, and Iain Chalmers. "Seventy-five trials and eleven systematic reviews a day: how will we ever keep up?." PLoS medicine 7, no. 9 (2010): e1000326.

[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] Stylianou, Nikolaos, and Ioannis Vlahavas. "Transformed: End-to-End transformers for evidence-based medicine and argument mining in medical literature." Journal of Biomedical Informatics 117 (2021): 103767.

[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.

[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] 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.

[38] Hashmath Shaik. "QL4POMR" GitHub repository ,2023. https://github.com/lukecage0/QL4POMR/blob/main/Sept-Dec/Data/extracted_abstracts/patient-2-articles.xlsx

[39] Hall, Karl, Chrisina Jayne, and Victor Chang. "A Transformer-Based Framework for Biomedical Information Retrieval Systems." In International Conference on Artificial Neural Networks, pp. 317-331. Cham: Springer Nature Switzerland, 2023.

[40] Falaki, Ala Alam, and Robin Gras. "Attention Visualizer Package: Revealing Word Importance for Deeper Insight into Encoder-Only Transformer Models." arXiv preprint arXiv:2308.14850 (2023).

[41] Github 2023, https://github.com/lukecage0/QL4POMR/tree/main/Sept-Dec/Data/Tables

[42] Kumar, Amit, and Christopher P. Cannon. "Acute coronary syndromes: diagnosis and management, part I." In Mayo Clinic Proceedings, vol. 84, no. 10, pp. 917-938. Elsevier, 2009.

[43] Durairaj, Ashok Kumar, and Anandan Chinnalagu. "Transformer based Contextual Model for Sentiment Analysis of Customer Reviews: A Fine-tuned BERT." International Journal of Advanced Computer Science and Applications 12, no. 11 (2021).