EHR Question and Answering for Surgical Notes: A Language Model Study

David M. Lee
University of Mississippi Medical Center
dmlee@umc.edu

Ahmad P. Tafti, PhD
University of Pittsburgh
tafti.ahmad@pitt.edu

Hamidreza Moradi, PhD
North Carolina A&T State University
hmoradi@ncat.edu

Motivation

EHR often contain a vast amount of information, making it challenging for medical professionals to **quickly extract essential patient data**. The introduction of Language Models (LM) and Large Language Models (LLM) have enabled the development of sophisticated models capable of understanding and extracting relevant information from written language. By leveraging these advanced language models, we aim to streamline the process of **information retrieval from EHRs**, facilitating **well-informed clinical decisions based on patient backgrounds**.

Objective

- Assessing the performance of diverse **pre-trained** language models.
- Investigating the encountered during the **fine-tuning** process, focusing on domain adaptation and unstructured data complexities.
- Comparing the **accuracy and efficiency** of models concerning the volume of training data required for optimal performance.

Dataset

The dataset used in this study comprises of **1,200 patient** records extracted from the EHR of the University of Mississippi Medical Center (**UMMC**) from 2016 to 2021. The study has received approval from the Institutional Review Board (IRB). The dataset consists of **surgical notes** for Total Knee Arthroplasty (TKA) procedures, providing detailed information about each patient's surgeries (CPT codes: 27446, 27447, 27486, 27487, and 27488).

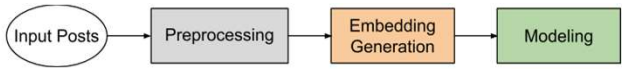
Considered Questions

- Laterality of the surgery (Right, Left, Bilateral)
- Constraint type (Posterior-Stabilized and Cruciate-Retaining)
- Presence of Patellar Resurfacing
- Implant model (DePuy ATTUNE/SIGMA, Smith and Nephew LEGION, Conformis iTotal, and DJO EMPOWR)

Pre-Processing

Prior to embedding, the data is cleaned and preprocessed.

- Entire data is converted to lowercase where required by uncased embeddings utilized.
- Extraneous spaces are removed. Depending on the embeddings' requirements, non-alphanumeric characters are removed.



Models

- BERT: With masked language model approach utilizing only the encoder part of the Transformer architecture
- DistilBERT: A compact version of BERT, emphasizing model compression and speed without sacrificing performance using distillation technique.
- BioBERT: Fine-tuned on biomedical literature, capturing domain-specific semantics and terminologies.
- BioClinicalBERT: Extends the domain specificity of BioBERT to clinical narratives and medical records. It aids in information extraction and knowledge discovery within clinical contexts.
- PubMedBERT: Tailored for biomedical literature, specifically from PubMed. It is optimized to domain-specific relationships found in research papers.
- RoBERTa: Retrained BERT with additional data, larger batch sizes, and longer training times. It omits the next sentence prediction task and introduces dynamic masking.
- GPT2: Employs the decoder part of the Transformer architecture. The decoder processes the input sequence autoregressively, generating one token at a time. primarily designed for generative tasks.
- LLaMA LoRA: A family of LLMs developed by Meta AI, ranging from 7B to 65B parameters. LoRA stands for Low-Rank Adaptation. It is a technique for fine-tuning LLMs with limited computational resources.

We utilized tokenizers specific by each model as detailed in their documentations.

Results

Model	Max. Overall Accuracy	Min. Training Data Req.
DistilBERT	97.48	0.7
BERT	**98.05**	0.4
BioBERT	95.76	0.2
BioClinicalBERT	95.88	0.3
PubMedBERT	97.48	**0.2**
RoBERTa	96.91	0.6
GPT2	97.71	0.3
LLaMA LoRA (LLM)	90.72	0.3

Conclusion

- BERT demonstrating the highest overall exact match accuracy of 98.05%, indicating its superior performance in accurately answering questions with no domain specific pre-training.
- Similar accuracy can be achieved by utilizing only 20% of training data using PubMedBERT model.
- GPT based model provides a good balance of required training data and accuracy.

Future Works

- Domain-specific Fine-tuning: Developing tailored fine-tuning strategies addressing surgical lexicon challenges for improved model adaptability.
- Prompt Engineering: Utilizing prompt engineering techniques like Zero-shot and few-shot to elicit desired outputs from large language models (LLMs).

References

[1] J. Milstein, A. Holmgren, P. Kralovec, et al. Electronic health record adoption in us hospitals: the emergence of a digital "advanced use" divide. Journal of the American Medical Informatics Association, 24(6):1142–1148, 2017.

[2] E. Sagheb, T. Ramazanian, A. Tafti, et al. Use of natural language processing algorithms to identify common data elements in operative notes for knee arthroplasty. The Journal of arthroplasty, 36(3):922–926, 2021.

EHR Question and Answering for Surgical Notes: A Language Model Study

David M. Lee
University of Mississippi Medical Center
dmlee@umc.edu

Ahmad P. Tafti, PhD
University of Pittsburgh
tafti.ahmad@pitt.edu

Hamidreza Moradi, PhD*
North Carolina A&T State University
hmoradi@ncat.edu

***Abstract*—This study assesses the precision of employing a pre-trained language model for question and answering tasks in Electronic Health Record (EHR) data. EHR often contain a vast amount of information, making it challenging for medical professionals to quickly access essential patient data. The introduction of Large Language Models (LLM) has enabled the development of sophisticated models capable of understanding and extracting relevant information from written language. By leveraging these advanced language models, we aim to streamline the process of information retrieval from EHRs, facilitating well-informed clinical decisions based on patient backgrounds.**

***Index Terms*—Language Model; Deep Learning; NLP; LLM;**

I. Introduction

The widespread adoption of EHR in the United States has eased access to patient charts and historical visits [1]. However, lengthy and detailed clinical notes in these records present challenges for clinicians seeking specific procedure-related details. Natural Language Processing (NLP) offers a solution to analyze EHR data efficiently. The unstructured nature of these documents, along with complex technical language, abbreviations, and misspellings, poses obstacles for AI model development. Recent advances in language models show promising results in various NLP tasks due to extensive training on vast amount of data. In this study, we evaluate the accuracy of pre-trained language models for question and answering tasks in EHRs. Specifically, we aim to answer questions related to knee arthroplasty operative notes, focusing on data element extraction. The exploration of different language models enhances question and answering capabilities, aiding medical professionals in retrieving crucial knee arthroplasty information.

II. Matterials and Methods

The dataset used in this study comprises of 1,200 patient records extracted from the EHR of the University of Mississippi Medical Center (UMMC) from 2016 to 2021. The study has received approval from the Institutional Review Board (IRB). The dataset consists of surgical notes for Total Knee Arthroplasty (TKA) procedures, providing detailed information about each patient's surgeries, including the laterality of the surgery (right, left, bilateral), constraint type (posterior-stabilized and cruciate-retaining), presence of patellar resurfacing, and implant model (DePuy ATTUNE/SIGMA, Smith and Nephew LEGION, Conformis iTotal, and DJO EMPOWR). The accuracy of trained language models in answering the questions related to the aforementioned surgical details is compared against the gold standard registry data [2].

We evaluate the performance of various language models in question and answering tasks. The evaluated models are listed in Table 1 and includes examples of pre-trained language models and a LLM, trained on extensive text data. The dataset was split into 80% for training and 20% for testing. Subsequently, the models were trained on a range of training dataset sizes, starting from 10% up to 100% by 10% increments. The overall exact match accuracy of each model was then calculated on the test set to assess their effectiveness in extracting specific data elements from knee arthroplasty operative notes.

III. Evaluations

The results varied across the models, with BERT demonstrating the highest overall exact match accuracy of 98.05%, indicating its superior performance in accurately answering questions. However, very similar accuracy can be achieved by utilizing only 20% of training data using PubMedBERT model. Moreover, we can observe that LLMs with high capabilities did not show very great results, possibly due to the difficulty of fine-tuning these models for medical data, as they were not originally trained on such specific domain information.

TABLE I: Language Models, Highest Achievable Accuracy, and Minimum Training Data Required to Attain.

Model	Max. Overall Accuracy	Min. Training Data Req.
DistilBERT	97.48	0.7
BERT	98.05	0.4
BioBERT	95.76	0.2
BioClinicalBERT	95.88	0.3
PubMedBERT	97.48	0.2
Roberta	96.91	0.6
GPT2	97.71	0.3
LLaMA LoRA (LLM)	90.72	0.3

*Corresponding author.

References

[1] J. Milstein, A. Holmgren, P. Kralovec, et al. Electronic health record adoption in us hospitals: the emergence of a digital "advanced use" divide. *Journal of the American Medical Informatics Association*, 24(6):1142–1148, 2017.

[2] E. Sagheb, T. Ramazanian, A. Tafti, et al. Use of natural language processing algorithms to identify common data elements in operative notes for knee arthroplasty. *The Journal of arthroplasty*, 36(3):922–926, 2021.