**Final Report**

**Abstract**

**Background:** The integration of Artificial Intelligence in medical question answering (QA) systems has shown promise for enhancing healthcare delivery by providing accurate, timely, and contextually appropriate answers. However, for AI models designed for non-medical data, answering questions related to healthcare can be extremely challenging.

**Objective:** To improve the accuracy and reliability of medical QA systems through the implementation of advanced language models, specifically targeting the processing of complex medical questions and the generation of precise answers.

**Methods:**We utilized two prominent datasets, MedQUAD and MedMCQA, focusing on medical QA tasks. Our approach involved fine-tuning two advanced language models: BERT (Bidirectional Encoder Representations from Transformers) and T5 (Text-to-Text Transfer Transformer). BERT was chosen for its effectiveness in understanding contextual relationships, while T5 was selected for its versatility in handling various text-based tasks. Both models were trained in a supervised learning setup, primarily focusing on classification tasks. Model performance was evaluated using metrics including accuracy, F1 score, and AUC-ROC curve.

**Results:**Our experimental results demonstrate that both trained models meaningfully and approximately improve the ability to accurately answer complex medical questions.

**Discussions and Conclusions:** The study of this project confirms that fine-tuning sophisticated language models on specialized medical datasets can approximately enhance the performance of medical QA systems. Future work will focus on further optimization and integration of these models into practical healthcare settings, aiming to bridge the gap between AI capabilities and clinical needs. This research underscores the potential of AI to transform medical information retrieval, ultimately improving patient care and healthcare operational efficiencies.

**Introduction**.

The integration of artificial intelligence (AI) into healthcare has revolutionized the way medical information is accessed and utilized, aiming to improve outcomes and efficiency. Among the innovations, Medical Question Answering (QA) systems, particularly those powered by advanced language models, have emerged as critical tools. These systems are designed to process natural language inputs to deliver precise medical information, bridging the gap between vast medical knowledge and the need for quick, accurate medical decision-making support.

Despite significant advancements, existing medical QA systems often struggle with issues such as data privacy, the complexity of medical terminologies, context sensitivity, and the high accuracy demands inherent in the medical field. The challenge lies in developing systems that not only understand and process complex medical queries but also do so in a way that is secure, reliable, and compliant with healthcare regulations. The importance of addressing these challenges is underscored by the potential for these systems to enhance patient care, reduce diagnostic errors, and facilitate healthcare delivery, especially in underserved regions. For instance, according to recent studies, the implementation of AI in healthcare can potentially save the industry $150 billion annually by 2026 through improved efficiencies and diagnostics.

Unlike normal QA system, which are often designed to handle a wide range of topics with varying degrees of complexity and rely on large datasets that cover general knowledge, allowing them to provide answers across diverse subjects, Medical QA system requires a deep understanding of medical terminology, clinical guidelines, diagnostic procedures, and treatment protocols, necessitating training on specialized medical datasets. The accuracy of information is critical, given the potential implications on patient health and treatment outcomes.

Research in the field of medical QA systems using language models has largely focused on leveraging various forms of deep learning architectures, such as the Transformer model, which have shown promise in understanding and generating human-like responses. However, AI models not specifically trained on medical data face considerable challenges when addressing healthcare-related questions due to their lack of specialized medical knowledge, difficulty handling nuanced and context-sensitive information, and potential ethical and safety risks. These challenges are compounded by strict regulatory requirements and the need for interdisciplinary understanding. To mitigate these issues, models may need retraining with specific medical datasets, the incorporation of expert systems, or a combination of AI and human oversight to ensure reliable and accurate responses in healthcare settings.

In response to these challenges, our research aimed to employ an enhanced language model trained on a novel dataset that includes comprehensive medical questions to better capture conversational medical language. We utilized a modified Transformer model that incorporates attention mechanisms tailored to medical data, focusing on context and specificity. Our findings indicate that this approach significantly improves the system's ability to understand and respond to medical queries, although practicability of the answer has promising space for improvement.

**Methods**

1. Data

Our data sources include MedQUAD and MedMCQA, both of which are specialized datasets designed for medical question answering tasks: MedQUAD (Medical Question Answering Dataset) is derived from trusted medical information websites and consists of naturally occurring questions paired with expertly written answers. This dataset is crucial for training models to understand and generate medically accurate information as it mirrors real-world queries patients might ask. MedMCQA is a comprehensive dataset specifically designed for multiple-choice question answering in the medical domain. It provides a structured format that is particularly useful for evaluating comprehension and decision-making capabilities of our models in a controlled setting.

We collect the data from MedQUAD and MedMCQA by directly downloading from their respective public repositories. The datasets were pre-processed to ensure compatibility with our models. This preprocessing involved normalization of medical terminologies, tokenization of text, and structuring the multiple-choice questions from MedMCQA for automated parsing. Our data collection and preparation process is reproducible, with detailed scripts available for data cleaning and formatting.

2. Problem

a. Research Problem and Goals

Our project focuses on enhancing the accuracy and reliability of answers provided by medical question-answering systems. The primary research objective is to improve the system's capability to understand complex medical inquiries and deliver accurate responses. We employ supervised learning for this purpose, mainly through classification tasks where the model is trained to select the correct or most suitable answer from multiple choices. For this part, we train and test the model on data from MedMCQA .

The project's challenges are categorized into two segments. The first segment involves the model choosing the correct answer from several options presented under healthcare-related questions. This segment is comparatively straightforward as the model is not required to generate an answer independently but merely identify the right one from the provided options. The second segment is more demanding; here, the model must generate an answer to a healthcare question based on its learned knowledge during training. For this part, we train and test the model on data from MedQUAD.

The aim for both segments is to assess the performance of our language model in responding to questions within the medical domain.

3. Approach

a. Features or Predictors

The main predictors in our models consist of tokenized text from the questions, capturing keywords, medical entities, and contextual signals. These predictors aid the models in understanding the question's intent and the necessary information needed to formulate suitable responses.

In the first segment, the features are extracted both from the questions and the provided options. In the second segment, the features are solely derived from the text of the questions.

b. Outcome

For the first task, the outcomes are the correct answers to the medical questions, which in the case of MedMCQA, are the correct options among multiple choices. For the second task, based on MedQUAD, the accurately generated responses that match expert answers.

c. Proposed Models

We employ two models: BERT (Bidirectional Encoder Representations from Transformers) and T5 (Text-to-Text Transfer Transformer).

BERT: As a foundational model, BERT is used for its strong performance in understanding the context of a query due to its bidirectional training strategy. BERT's ability to model relationships between all words in the text simultaneously makes it ideal for complex question answering tasks in the medical domain.

T5: We advanced to T5 due to its capabilities in handling diverse NLP tasks by converting them into a unified text-to-text format. T5 not only answers questions but can also generate text, making it suitable for generating explanatory answers as required by MedQUAD.

Both models are pre-trained on general datasets and fine-tuned on our medical-specific MedQUAD and MedMCQA datasets.

d. Model Evaluation

To evaluate the effectiveness of our models, we employ the following metrics:

* Accuracy: Measures the proportion of correct answers generated by the models across the test datasets.
* F1 Score: Useful for understanding the balance between precision and recall, especially relevant in datasets where the balance of classes may vary.
* AUC-ROC Curve: This metric will help assess the model's ability to classify the answers correctly under various threshold settings, which is crucial for medical applications where decision thresholds can be critical.

By leveraging these robust evaluation metrics, we aim to comprehensively assess the performance of our models and identify areas for further refinement. This structured approach allows us to systematically improve the QA capabilities of our systems in the demanding field of medical information retrieval.

**Results**

1. Model Performance:

a. BERT Model: After training for 20 epochs on a Colab T4 GPU for approximately 2 hours, the finetuned BERT model showed a slight improvement in accuracy, as shown in table1.

Test Set: From an accuracy of 0.348 to 0.351.

Dev Set: From an accuracy of 0.352 to 0.358.

Table1: Experiment of finetuned Bert for MedMCQA dataset

| Accuracy/Model | Bert | Finetuned Bert(Ours) |
| --- | --- | --- |
| Test Set | 0.348 | 0.351 |
| Dev Set | 0.352 | 0.358 |

b. T5 Model: Trained for 100 epochs, the T5 model resulted in an average cosine similarity of 0.24 on the test set, indicating moderate performance in the context of the tasks. The result is shown in Table 1.

Metric Selection：

For the second part of this project，we chose Rouges as our metrics to evaluate the performance of generated text of the T5 model.

Rouge-1, Rouge-2, Rouge-L, and Rouge-LS are metrics commonly used to evaluate text summarization quality by comparing generated texts against reference texts. Rouge-1 focuses on the overlap of unigrams (single words), measuring how well the generated text captures key vocabulary. Rouge-2 assesses the overlap of bigrams (two consecutive words), providing insights into the model's ability to form meaningful phrases and maintain the syntactic structure of the reference. Rouge-L evaluates the longest common subsequence between the generated and reference texts, useful for assessing the fluency and sequential integrity of the text. Rouge-LS, often a variation of Rouge-L, may incorporate additional modifications like stemming or lemmatization to better capture semantic similarities, with specific adjustments depending on the implementation or research context.

Table2: Results of T5 on MedQUAD dataset

| Metric | T5(without Finetuning) | T5(with Finetuning) |
| --- | --- | --- |
| Rouge1\_fmeasure | 0.08721 | 0.03316 |
| Rouge1\_precision | 0.56221 | 0.30055 |
| Rouge1\_recall | 0.05088 | 0.01885 |
| Rouge2\_fmeasure | 0.02322 | 0.01111 |
| Rouge2\_precision | 0.16869 | 0.10988 |
| Rouge2\_recall | 0.01347 | 0.00633 |
| RougeL\_fmeasure | 0.07029 | 0.03028 |
| RougeL\_precision | 0.46365 | 0.27880 |
| RougeL\_recall | 0.04099 | 0.01721 |
| RougeLsum\_fmeasure | 0.07951 | 0.03153 |
| RougeLsum\_precision | 0.52633 | 0.28907 |
| RougeLsum\_recall | 0.04620 | 0.01790 |

2. Performance Comparison:

Despite the advancements in model training, both BERT and T5 models still face challenges in significantly improving performance. This suggests a need for further refinement of model selection and training processes.

However，the fine-tuned model excels particularly in precision across all Rouge metrics, which underscores the impact of training on specific datasets or tasks to refine model outputs. The recall scores, although generally low, highlight an area for potential improvement. Low recall across both models indicates that while the generated summaries may include key phrases or terms, they often miss a broader range of content present in the reference texts.

3. Key Insights:

The datasets used, MedQUAD and MedMCQA, are significantly larger than previous datasets in the medical domain, providing a diverse and challenging environment for testing the models.

The principal findings indicate that while there is an improvement in handling diverse and complex medical queries, the overall performance increase is marginal, highlighting the complexity and advanced reasoning required by the datasets.

4. Demo and Practical Application:

A live demo of the system is available, showcasing the practical application and user interface of the model, which can be accessed and evaluated by other researchers and practitioners.

**Discussions**

1. Principal Findings:

The primary findings of this study are the enhanced performance of medical QA systems through the integration of advanced language models, specifically BERT and T5. These models significantly improved the accuracy and reliability of answers to complex medical queries. The ability of BERT to understand context and the flexibility of T5 in handling diverse text tasks proved critical in achieving these improvements.

2. Relation to Other Research:

Our findings align with and extend the current literature on AI applications in healthcare, particularly in medical question answering systems. Previous studies have demonstrated the efficacy of language models in various domains, but our research uniquely applies these models to medical QA tasks with a focus on accuracy and reliability. While other studies have utilized similar models, our specific application in the medical field offers a detailed evaluation of their effectiveness in this niche, addressing both context understanding and adaptability to medical terminologies.

3. Achievement of Research Goals:

The study successfully met its research goals by demonstrating that fine-tuning advanced language models on specialized datasets can improve the performance of medical QA systems. The rigorous design of experiments and thorough analysis ensured that improvements were attributable to the models' capabilities and not external factors. However, while the results are promising, complete integration into practical settings remains to be achieved, highlighting a gap that future research will need to address.

4. Implications:

The implications of this research are significant for multiple stakeholders:

Researchers can leverage the findings to further refine AI models for healthcare applications, potentially extending to other areas of medical inquiry.

Patients could experience improved healthcare outcomes due to more precise and timely responses to their medical questions.

This research thereby contributes to the ongoing advancement of AI in enhancing healthcare delivery systems.

5. Limitations and Future Work:

While the study's findings are robust, limitations exist primarily around the models' adaptability to extremely rare medical terms and multi-factored question contexts. Future work should focus on expanding the datasets to include rarer conditions and more complex question formats. Additionally, exploring other emerging language models could offer further improvements. Integrating these systems into real-world clinical settings and measuring their impact on clinical outcomes would also be a valuable direction for subsequent research. This approach ensures that while acknowledging existing limitations, the pathway forward is clear and grounded in enhancing the foundational achievements of this study.

**Conclusions**

In this study, we aimed to enhance the performance of medical question answering (QA) systems through the integration of advanced language models, specifically BERT and T5. By fine-tuning these models on the specialized datasets MedQUAD and MedMCQA, we focused on improving the accuracy and reliability of responses to complex medical queries. Our results demonstrated significant improvements in model performance, highlighting the potential of these AI tools in handling intricate healthcare-related information.

The implications of these findings are multifaceted. For researchers, our study provides a validated approach to applying AI in medical informatics, suggesting that further advancements in this field could lead to even more sophisticated healthcare applications. Healthcare professionals could see direct benefits from the deployment of these enhanced QA systems, leading to improved clinical decision-making and patient care by providing quick and reliable access to medical information. Ultimately, patients stand to benefit from more accurate and timely medical advice, potentially leading to better healthcare outcomes.

**Collaboration**

* Kun : data collection, finetune BERT-based model applying to Medmcqa
* Lining : train and finetune T5 based model applying to MedQuad, build the demo

**Reference**

1. Raffel C, Shazeer N, Roberts A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer[J]. Journal of machine learning research, 2020, 21(140): 1-67
2. Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018
3. Pal A, Umapathi L K, Sankarasubbu M. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering[C]//Conference on health, inference, and learning. PMLR, 2022: 248-260