Enhancing Automated Medical Question-Answer Systems Using Fine-Tuned Large Language Models

S M Taslim Uddin Raju
Department of Electrical and Computer Engineering
University of Waterloo, Waterloo, ON, Canada
ID: 21093036
smturaju@uwaterloo.ca

Abstract

Automated question and answer (Q&A) systems play a vital role in the medical field by providing accurate information to healthcare professionals and patients. This paper aims to propose an enhanced approach for automated medical Q&A systems using fine-tuned large language models (LLMs). The MedQuAD dataset is utilized to evaluate the proposed method. In the first stage, various natural language processing (NLP) techniques are applied to clean and preprocess the dataset. This study explored both decoder-only models (GPT-2 and Llama2) and encoder-decoder models (Bloom and T5), and fine-tuned them on the MedQuAD dataset. The performance of these models is then compared to determine the most effective LLMs for medical Q&A tasks. The T5 model demonstrates superior performance, achieving a BLEU-4 score of 42.5%, a METEOR score of 36.7%, and a ROUGE-L score of 39.2%, respectively. These results highlight the potential of LLMs to enhance automated medical Q&A systems, providing significant improvements in accuracy and reliability.

Github repository: https://github.com/raju32742/MSCI641_Project-UW-/

Keywords: Automated Medical Q&A, LLMs, NLP Techniques, Medical Diagnostics, Model Fine-Tuning.

1 Introduction

Automated medical question and answering (Q &A) systems have become essential due to the exponential growth of online medical information (Yagnik et al., 2024). It's difficult for both healthcare professionals and patients to stay updated with this vast amount of information. Automated systems can provide summarized, evidence-based knowledge for healthcare professionals, enhancing their decision-making processes. They also offer reliable medical information, allowing them to make informed health decisions for patients.

Recent advancements in Large Language Models (LLMs) have dramatically transformed the potential of these automated systems (Schimanski et al., 2024). LLMs, such as GPT-2 and T5, have grown increasingly sophisticated, learning from vast amounts of text to perform a wide array of natural language processing (NLP) tasks with remarkable accuracy. This development has led to impressive performances in various NLP tasks, such as sentiment analysis, text summarization, and even creative text generation, including poetry and scripts. Their accessibility has increased, allowing professionals from various fields to utilize these models through user-friendly APIs for specialized applications. This versatility makes them ideal tools for tackling the unique challenges of medical Q&A (Hasan et al., 2024).

In the medical field, automated Q&A systems traditionally relied on effective information retrieval techniques that were limited to understanding patient context. With the introduction of LLMs combined with NLP, there's a significant shift towards more personalized responses. NLP helps these models understand and process complex user questions, turning data into valuable understandings. This necessity to tailor responses more closely to individual needs has spurred the integration of LLMs into medical Q&A, pushing the boundaries of what automated systems can achieve in healthcare (Yu et al., 2024). The evolution of LLMs, from decoder-only models (such as GPT-2, Llama) to encoder models like Bloom, shows substantial improvements in handling diverse challenges. Exploring these models further to understand their possibility to enhance generative question-answering in healthcare is imperative (Allaouzi et al., 2019).

The proposed work aims to enhance the performance of automated medical Q&A systems using LLMs. The MedQuAD dataset is used to evaluate the model which includes a wide range of medical questions and answers. Initially, we ap-

ply various NLP techniques to clean and preprocess this dataset, ensuring the data is optimized for model training. Therefore, we experimented with both decoder-only models (GPT-2 and Llama2) and encoder-decoder models (Bloom and T5) and fine-tuned them on the MedQuAD dataset. After fine-tuning, we compare the performance of these models to gain insights into which types of LLMs are most effective for medical Q&A tasks. This comparison is crucial as it helps identify the strengths and weaknesses of each model type, providing clear guidance on the most suitable LLMs for applications in the medical domain. The main contributions are summarised as follows:

- Evaluating general vs. medical-specific LLMs for medical Q&A, using the MedQuAD dataset.
- Fine-tuning domain-specific LLMs enhances performance over general LLMs.
- Comparing decoder-only and encoder-decoder models, providing insights into their suitability for medical applications.

The rest of this paper is summarized as follows: Section II reviews existing literature. Section III outlines the proposed methodology, including preprocessing and LLM models. Section IV details the dataset, evaluation metrics, and results analysis. Section V discusses limitations and future directions. Section VI concludes by summarizing contributions.

2 Literature Reviews

Automated Q&A systems have evolved significantly with the advancement of transformers and large language models (Rasool et al., 2024). Q&A can be categorized into three types based on their input and output dynamics:

- Extractive Q&A: This method uses trained models to retrieve answers directly from a given context, such as text, images, or HTML.
- Open Generative Q&A: Models are provided with a context and tasked with generating an answer in natural language. These models are designed to formulate coherent responses based on the context provided.
- Closed Generative Q&A: This approach does not use any external context. Instead, models generate answers solely from their pre-trained knowledge, independent of any input data.

Yagnik et al. (Yagnik et al., 2024) focused on LLM models, specifically general and medical-specific distilled LLMs, and their application in medical Q&A systems. The study compared the performance of general and medical-specific distilled LLMs for medical Q&A, evaluating the effectiveness of fine-tuning domain-specific LLMs and comparing different families of language models. The findings enhance the application of generative models in medical Q&A, aiding evidence-based healthcare decisions. Zhuang et al. (Zhuang et al., 2024) introduced ToolQA, a dataset for evaluating how well LLMs use external tools in question answering. The dataset, developed through an automated process, includes 13 specialized tools and aims to minimize overlap with pre-training data. Their study identifies the strengths and limitations of current tool-use LLMs, sets new benchmarks, and suggests future improvements.

In another study, Kthe authors (Kriangchaivech and Wangperawong, 2019) developed a machine learning model using transformers to generate questions from Wikipedia passages, moving away from traditional RNN methods. The model was trained on the inverted Stanford Question Answering Dataset (SQuAD), which contains over 100,000 questions. It can produce simple, relevant questions that average eight words in length. Although the high Word Error Rate (WER) suggests that the generated questions differ from those in the original SQuAD, they are mainly grammatically correct and plausible. This indicates a limitation in the similarity of the questions but also confirms the model's effectiveness.

3 Methodology

The proposed architecture of an automated Q&A generation system using fine-tuned LLM models is shown in Figure 1.

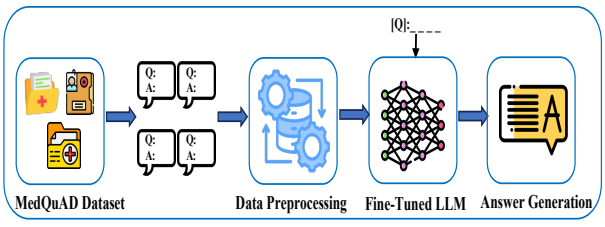


Figure 1: Architecture of automated question-answer systems using fine-tuned LLMs.

3.1 Data Preprocessing

Before we can use our training data with the model, we need to transform it into a format the model can work with and enhance the quality and consistency of inputs.

3.1.1 Natural Language Processing

For this study, both questions and answers were subjected to several **Natural Language Processing** steps:

- **Lowercasing:** All text data, including questions and answers, were converted to lowercase to ensure uniformity and reduce the complexity of case word variations.
- **Lemmatization** The answers were further processed using lemmatization techniques. Lemmatization involves reducing words to their base or dictionary form (lemma), which helps minimize redundancy and improves the model's generalization ability.
- **Additional Preprocessing Techniques:** The NLTK library was used to remove stop words, punctuation, and tokenize the text, refining the content by reducing noise and enhancing model performance.

3.1.2 Concatenate Questions and Answers:

We merged Q&A pairs into a single text sequence to create a generative model. This helped the model understand the connection between questions and answers, enabling it to produce relevant responses.

3.1.3 Truncate Model Input:

Due to varying input and output lengths, we standardized input lengths by truncating them. For the MedQuAD dataset, the limit was 300 tokens for a decoder-only model, while for encoder-decoder models, the question token limit was 64, and the answer token limit was 200. This truncation, primarily targeting the initial portion of answers, optimized content relevance and reduced training time by shortening the input.

3.1.4 Tokenize and Fine-tune:

After truncating the input, we implemented model-specific tokenization. Following tokenization, the models were fine-tuned with information from our datasets. This fine-tuning aimed to enhance the performance of the generative models on the integrated question-answer sequences, using standardized lengths and tailored tokenization techniques.

3.2 LLM Models

To generate the answer from the given question, the following LLM models have been fine-tuned.

3.2.1 Encoder-Decoder Models Architecture

The encoder-decoder architecture is a popular framework in sequence-to-sequence (Seq2Seq) tasks, where the goal is to transform an input sequence into an output sequence. This architecture is generally utilized in various natural language processing (NLP) tasks, including machine translation, summarization, and question-answering. The encoder processes the input sequence (e.g., a question) and converts it into a context vector or a sequence representing the input information in a latent space. The decoder then takes this context vector and generates the output sequence (e.g., an answer) one token at a time. In this study, T5 (Text-To-Text Transfer Transformer) (Ni et al., 2021) and Bloom (Scaria et al., 2024) LLM are used for generating automated answers to given questions, making them ideal for tasks like question-answering where both input and output are sequences of text. These architectures exploit a transformer-based encoder to understand the input sequence. A transformer-based decoder is then used to generate the output sequence, where attention mechanisms play a crucial role in capturing dependencies within and between the sequences. Figure 2 presents the general architecture of the encoder and decoder models for answer generation.

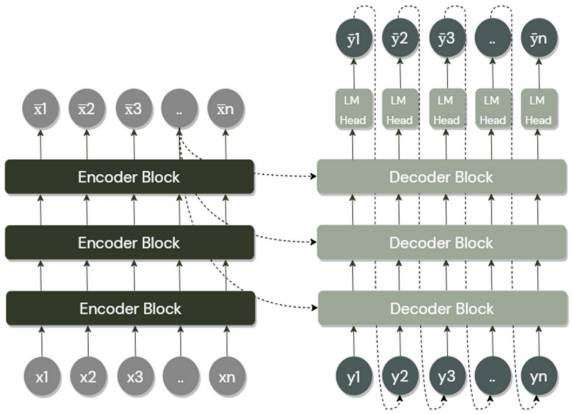


Figure 2: Encoder-d ecoder model architecture for text generation (Touma et al., 2023).

3.2.2 Decoder-Only Models Architecture

The decoder-only architecture is another approach used in generative models, particularly for tasks involving autoregressive text generation. Unlike the encoder-decoder model, the decoder-only model

does not require a separate encoder. Instead, it generates sequences by predicting the next token based on the previous tokens. GPT-2 (Generative Pre-trained Transformer) (Tsai et al., 2021) and Llama2 (Hybl, 2024) are examples of decoder-only architectures. Both GPT-2 and Llama2 models are pre-trained on vast amounts of text data using a causal language modeling objective. During training, the model learns to predict the next word in a sequence, enabling it to generate coherent text. This capability is useful for tasks such as question-answering, where the model generates an answer based on a given question. The main advantage of the decoder-only architecture lies in its simplicity and efficiency, as it eliminates the need for an encoder and allows for faster generation of text sequences. Figure 3 shows the general architecture of only decoder model for answer generation.

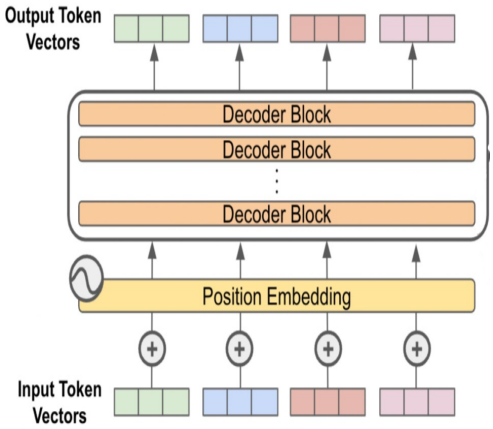


Figure 3: Decoder model architecture for text generation.

3.3 Loss Function

The loss for the LLM models is the categorical cross-entropy loss between the generated answer and the actual answer:

$$L_{LLM} = -\frac{1}{T}\sum_{t=1}^{T}\log(p(C_t|C_{<t}, x)) \quad (1)$$

where T is the length of the sequence, $C_{<t}$ is the sequence of tokens before time step t, and $p(C_t|C_{<t}, x)$ is the probability of the correct token C_t given by the LLM models.

4 Results and Discussions

4.1 Dataset

In this work, we evaluate our proposed approach with the MedQuAD dataset (Ben Abacha and Demner-Fushman, 2019). MedQuAD is a comprehensive resource containing 47,457 question and answer (Q&A) pairs sourced from various NIH websites. In this study, we have used 16,410 question and answer pairs. The dataset is well known for its high-quality content, as the authoritative support of the NIH backs the answers. MedQuAD addresses a broad range of medical questions, categorized into 16 different types of diseases such as information, yymptoms, treatment, and inheritance etc. Figure 4 shows the distribution of question types. The answers to the questions in this dataset range from 50 to more than 3,500 words. However, the most critical information typically appears at the beginning of the response, with the remaining content providing additional details. Figure 5 presents the length distribution of answers. We present below one examples of MedQuAD ques-

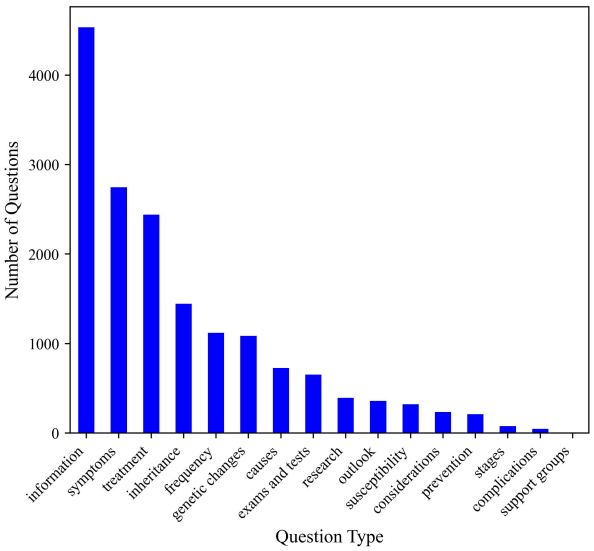


Figure 4: Distribution of question types in MedQuAD.

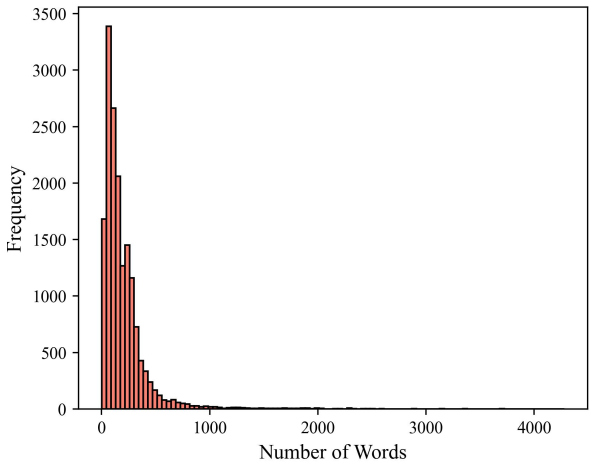


Figure 5: Answer length distribution (words).

tion and answer:

- **qtype:** *susceptibility*
- **Question:** *Who is at risk for Parasites - Cysticercosis?*
- **Answer:** *Cysticercosis is an infection caused by the larvae of the tapeworm, Taenia solium. A person with an adult tapeworm, which lives in the person's gut, sheds eggs in the stool.*

4.2 Evaluation Metrics

We evaluate our proposed method using BLEU, METROR, and ROUGE as evaluation criterias.

4.2.1 BLUE

BLEU (Papineni et al., 2002) is a widely used metric for assessing text quality, primarily in machine translation. It compares captions based on modified n-gram precision to generate BLEU scores. The precision for each $n-$gram size (unigrams, bigrams, trigrams, etc.) $P(n)$ is computed as:

$$P(n) = \frac{Matched(n)}{H(n)} \tag{2}$$

where $Matched(n)$ represents the count of $n-$gram tuples in the hypothesis as well as present in the actual caption, while $H(n)$ denotes the count of $n-$gram tuples in the hypothesis. The brevity penalty ρ is computed as follows:

$$\rho = \exp\left(\min\left(0, 1 - \frac{n}{L_{cap}}\right)\right) \tag{3}$$

where n represents the length of the hypothesis and L_{cap} denotes the length of the actual caption. The BLEU score is then calculated as follows:

$$BLEU = \rho\left(\prod_{i=1}^{N} P(i)\right)^{\frac{1}{N}} \tag{4}$$

where N is the maximum order of $n-$grams.

4.2.2 METEOR

METEOR (Denkowski and Lavie, 2014) relies on unigram alignment between the generated and actual captions. The calculation of unigram precision P_e and recall R_e in METEOR is expressed as follows:

$$P_e = \frac{w_m}{w_t}, R_e = \frac{w_m}{w_r} \tag{5}$$

where w_m is the number of unigrams in the candidate translation as well as present in the actual caption, w_t and w_r are the number of unigrams in the candidate and the actual translation, respectively. The harmonic mean of precision and recall F_μ is computed as:

$$F_\mu = \frac{10P_eR_e}{R_e + 9P_e} \tag{6}$$

The penalty ρ is calculated using:

$$\rho = 0.5\left(\frac{C_k}{u_m}\right)^3 \tag{7}$$

where, C_k represents the number of chunks, and u_m signifies the count of mapped unigrams. Finally, the METEOR score M for a segment is calculated as:

$$M = F_\mu(1 - \rho) \tag{8}$$

4.2.3 ROUGE

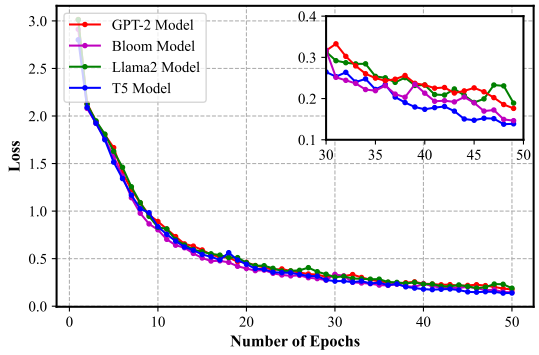
ROUGE (Lin, 2004) measures the frequency of $n-$gram matches in the total of manually annotated descriptions, while other methods account for the frequency in the aggregation of all generated descriptions. ROUGE-N and ROUGE-L are often used in tasks like document summarization and video captioning. The ROUGE-N score can be expressed as follows:

$$ROUGE - N = \frac{\sum_{S \in AS}\sum_{g_n \in S} MatchedCount(g_n)}{\sum_{S \in AS}\sum_{g_n \in S} TotalCount(g_n)} \tag{9}$$

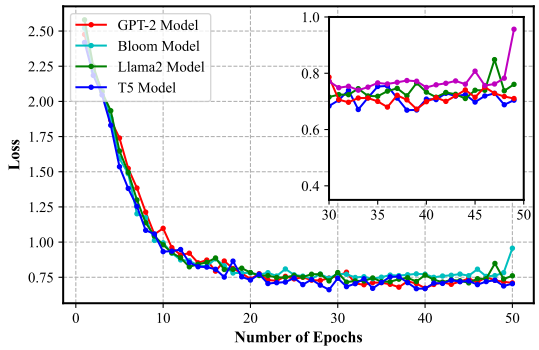
where $MatchedCount(g_n)$ represents the maximum number of $n-$grams that appear both in a candidate summary and a collection of actual summaries.

4.3 Training Phase and Hyperparameters

The MedQuAD dataset is partitioned in a ratio of approximately 6:2:2, with 9,843 entries for training, 3,282 for validation, and 3,282 for testing. To address the task of medical question-answering, we utilized a combination of four large language models (LLMs): GPT-2 and Llama2, which are primarily decoder models, and Bloom and T5, which function as both encoders and decoders. These models were fine-tuned on the MedQuAD dataset to better align with the specific requirements of medical information processing. The fine-tuning process involved adjusting the models using the following hyperparameters: a learning rate of 0.001, a weight decay of 0.001, and a batch size of 4 for both the training and evaluation phases. The training was set to run for 50 epochs to ensure adequate learning with less overfitting. **Logging**



(a)



(b)

Figure 6: Loss curve representation on the MedQuAD dataset using proposed fine-tuned LLM model. (a) Training loss (b) Validation loss.

was configured to capture updates every ten steps into a designated log directory within the output directory to monitor the training progress effectively. The Trainer interface from the Hugging Face Transformers library was employed to manage the training and evaluation process. This setup utilized a **DataCollatorForLanguageModeling** for batching, which does not mask tokens, fitting our need for direct sequence-to-sequence modeling. The tokenizer functions of our chosen LLMs facilitated the conversion of text into a format suitable for model processing, while their sequence-to-sequence generation capabilities were essential for producing coherent and contextually appropriate answers. Figure 6 shows the loss curve of fine-tuned LLM models on the training and validation set of MedQuAD dataset.

4.4 Fine-tuned LLM Models Performance

The QA generation performance was evaluated using two categories of models: only decoder models and combined encoder-decoder models, specifically tuned on the MedQuAD dataset. Initially, the decoder-only models, GPT-2 and Llama2, were assessed. These models are designed to generate text based on the input they receive without a distinct separation of encoding and decoding processes. In our experiments, GPT-2 achieved a BLEU-1 score of 40.2% and a BLEU-4 score of 36.4%, while Llama2 showed slightly lower performance with BLEU-1 at 38.0% and BLEU-4 at 35.4%. The ROUGE scores followed a similar pattern, indicating robust capabilities in generating medically relevant answers. On the other hand, the encoder-decoder models, particularly T5, demonstrated superior performance, reflecting the effectiveness of their architectural design. T5, which operates by encoding the input text into a meaningful representation and then decoding it to generate the output, achieved the highest scores across all metrics: a BLEU-1 score of 44.2%, a BLEU-4 of 42.5%, and a ROUGE-L of 39.2%. Bloom, another encoder-decoder model, also performed well, though it was slightly behind T5. The superior performance of these models can be attributed to their ability to better understand the context and semantics of the questions, thereby generating more accurate and contextually relevant answers.

Figure 7 presents a violin plot analysis of the BLEU@4 scores for the models GPT-2, Bloom, Llama2, and T5, based on evaluations conducted on a test set comprising 3,282 cases. T5 shows the highest scores, suggesting precise alignment with reference answers, while Bloom and GPT-2 display broader score distributions. In Figure 8, a 3D bar chart is used to display the ROUGE-1, ROUGE-2, and ROUGE-L scores for each model. This visual representation helps in comparing the performance of each model across different metrics, highlighting how each model handles overlap and sequence order in their generated text. The box plot in Figure 9 compares the METEOR scores of GPT-2, Bloom, Llama2, and T5, providing insights into the statistical distribution of scores, including medians, quartiles, and outliers. T5 shows the highest median score and a compact interquartile range, indicating strong and consistent performance.

5 Limitations and Future Works

In this study, we faced several resource and computational limitations that restricted the scope of experiments on the fine-tuned models. Specifically, constraints on computational resources inhibited our ability to conduct extended training with increased epochs and larger batch sizes. Such enhancements could improve model performance by

Table 1: Performance metric of our proposed method for answer generation on test set of MedQuAD dataset.

Model	BLUE-1(%)	BLUE-4(%)	ROUGE-1(%)	ROUGE-2(%)	ROUGE-L(%)	METEOR(%)
GPT-2	40.2	36.4	39.5	30.3	36.6	35.5
Bloom	39.3	35.5	41.4	27.9	33.0	32.0
Llama2	38.0	35.4	39.0	28.9	37.2	35.1
T5	**44.2**	**42.5**	**43.1**	**33.2**	**39.2**	**36.7**

* Note: Bold color denotes the best performing model.

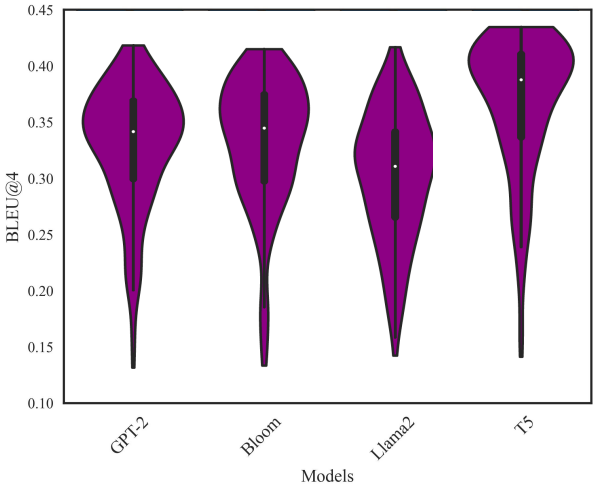


Figure 7: Evaluation of BLEU@4 score on the test set consisting of 3,282 cases.

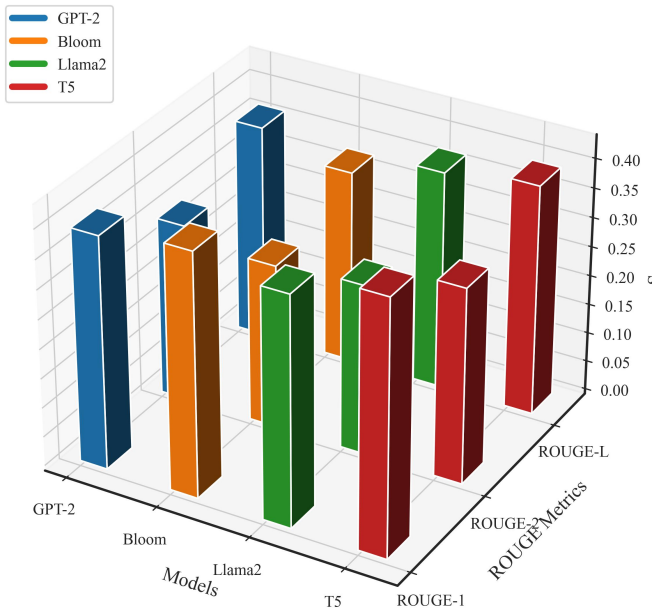


Figure 8: ROUGE-score of all models on test dataset.

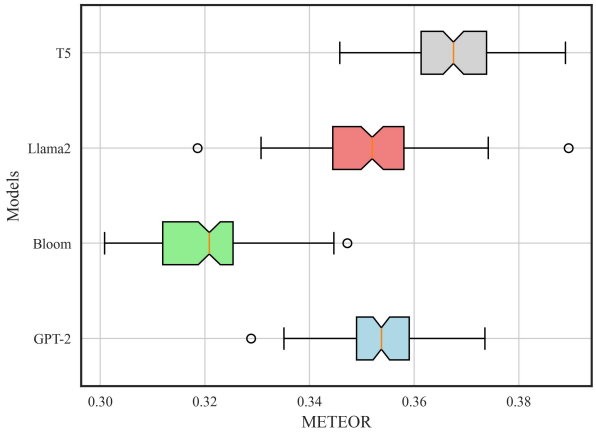


Figure 9: METEOR-score of all models on test dataset.

allowing more comprehensive learning phases. The study also highlights several promising directions for future research to expand upon the existing findings:

- **Advanced Models Fine-Tuning**: Our project currently utilizes distilled versions of models like GPT-2, BLOOM, and T5 due to the unavailability of more refined, open-source versions of GPT-3 and GPT-4. Future studies could benefit from testing and fine-tuning these advanced models. The use of GPT-3 and GPT-4 could enhance the models' understanding of context and reduce the generation of incorrect information, known as hallucination. Access to these more powerful models might lead to more accurate and reliable answers in medical question-answering systems.
- **Dataset Enhancement:** Data augmentation has proven to be an effective technique for improving model performance. Future work could focus on further processing, augmenting, and summarizing the dataset to create a more uniform and cohesive collection of data, which would likely yield more consistent results.

6 Conclusion

This study demonstrates the effectiveness of fine-tuned LLMs in enhancing automated medical Q&A systems. By utilizing the MedQuAD dataset and applying various NLP techniques for data preprocessing, the study evaluates both decoder-only models (GPT-2 and Llama2) and encoder-decoder models (Bloom and T5). Therefore, fine-tuned the LLM

modwls, we were able to significantly enhance the accuracy and relevance of generated answers. The encoder-decoder model such as T5, emerged as the most effective, outperforming other models in generating precise and informative responses. This underscores the importance of choosing the right model architecture and fine-tuning techniques for specific applications in the medical domain.

Acknowledgements

We express our deep appreciation to Instructor Prof. Olga Vechtomova for her invaluable guidance and support during this project. Her expert advice and insightful feedback were pivotal in directing our research and contributing to its success. We also thank Teaching Assistant Gaurav Sahu for his technical assistance and valuable advice, which played a significant role in the development of our project.

References

Imane Allaouzi, Mohamed Ben Ahmed, and Badr Benamrou. 2019. An encoder-decoder model for visual question answering in the medical domain. In *CLEF (working notes)*.

Asma Ben Abacha and Dina Demner-Fushman. 2019. A question-entailment approach to question answering. *BMC bioinformatics*, 20:1–23.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380.

ASM Hasan, Md Alvee Ehsan, Kefaya Benta Shahnoor, and Syeda Sumaiya Tasneem. 2024. *Automatic question & answer generation using generative Large Language Model (LLM)*. Ph.D. thesis, Brac University.

Matous Hybl. 2024. Comprehensive question and answer generation with llama 2.

Kettip Kriangchaivech and Artit Wangperawong. 2019. Question generation by transformers. *arXiv preprint arXiv:1909.05017*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith B Hall, Daniel Cer, and Yinfei Yang. 2021. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. *arXiv preprint arXiv:2108.08877*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Zafaryab Rasool, Stefanus Kurniawan, Sherwin Balugo, Scott Barnett, Rajesh Vasa, Courtney Chesser, Benjamin M Hampstead, Sylvie Belleville, Kon Mouzakis, and Alex Bahar-Fuchs. 2024. Evaluating llms on document-based qa: Exact answer selection and numerical extraction using cogtale dataset. *Natural Language Processing Journal*, page 100083.

Nicy Scaria, Suma Dharani Chenna, and Deepak Subramani. 2024. Automated educational question generation at different bloom's skill levels using large language models: Strategies and evaluation. In *International Conference on Artificial Intelligence in Education*, pages 165–179. Springer.

Tobias Schimanski, Jingwei Ni, Mathias Kraus, Elliott Ash, and Markus Leippold. 2024. Towards faithful and robust llm specialists for evidence-based question-answering. *arXiv preprint arXiv:2402.08277*.

Roudy Touma, Hazem Hajj, Wassim El-Hajj, and Khaled Shaban. 2023. Automated generation of human-readable natural arabic text from rdf data. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(4):1–13.

Danny CL Tsai, W Chang, and S Yang. 2021. Short answer questions generation by fine-tuning bert and gpt-2. In *Proceedings of the 29th International Conference on Computers in Education Conference, ICCE*, volume 64.

Niraj Yagnik, Jay Jhaveri, Vivek Sharma, Gabriel Pila, Asma Ben, and Jingbo Shang. 2024. Medlm: Exploring language models for medical question answering systems. *arXiv preprint arXiv:2401.11389*.

Haoran Yu, Chang Yu, Zihan Wang, Dongxian Zou, and Hao Qin. 2024. Enhancing healthcare through large language models: A study on medical question answering. *arXiv preprint arXiv:2408.04138*.

Yuchen Zhuang, Yue Yu, Kuan Wang, Haotian Sun, and Chao Zhang. 2024. Toolqa: A dataset for llm question answering with external tools. *Advances in Neural Information Processing Systems*, 36.