Transformers for Pressure Ulcer Management: A Large Language Model Approach in Healthcare

Abstract— **Pressure ulcer management is an area of considerable concern for healthcare professionals, as it is associated with 700,000 patients in the UK every year, costing the National Health Service close to £3.8 million each day. Unfortunately, the training mechanisms presently adopted for healthcare practitioners are rarely accessible or consistent. hence, knowledge gaps lead to poor implementation of treatment strategies. This paper puts forward a proposal about training healthcare professionals using a Large Language Model (LLM)-based question-answering system. We investigate two kinds of transformer-based architectures: BERT, fine-tuned on domain-specific data for extractive QA, and Retrieval-Augmented Generation (RAG), which retrieves and generates contextually relevant responses. The paper discusses dataset preparation, preprocessing, model architectures, hyperparameter tuning techniques, and evaluation strategies using BERTScore and retrieval recall. We compare both models in executing their tasks of accurately informing the medical practitioner in-context. This study illustrates how LLMs can enhance medical education and healthcare decision-making. Our results provide evidence that transformer-based models uniquely facilitate accessibility and real-time training, thereby addressing major obstacles in pressure ulcer management training.**

Keywords— **Pressure Ulcer Management, Large Language Model (LLM), BERT, Retrieval-Augmented Generation (RAG), Medical Education.**

I. Introduction

The most common injuries that occur on the skin and input tissue arise from mechanic stress from prolonged pressure, shear forces, or friction. Pressure ulcers affect hundreds of millions of people across the globe, and even worse, older or very sick patients are more prone to this kind of affliction. In the UK, for example, approximately 700,000 patients are affected yearly by pressure ulcers, with long hospital stays, higher treatment costs, and increased morbidity rates. The National Health Service (NHS) estimates that around £3.8 million is spent daily for pressure ulcer management, indicating the urgent need for better prevention and treatment strategies [1].

The challenge presented, medical knowledge was not garnered consistently among medical practitioners. There were evidence-based guidelines. however, due to different demands at the same time in a hospital environment, one would not have immediate access to this essential information while carrying out clinical decision-making. Old-school teaching in the forms of lectures, textbooks, online courses, etc. simply never addressed-on-any-instance delivering real-time, context-aware learning, leading always to knowledge retention challenges and, eventually, inadequate patient outcomes [2].

It is made more challenging by the extensive and fast-growing quantity of medical literature generated every day, making it difficult for practitioners to remain current with the latest research findings and best practices.

Recent advances in natural language processing (NLP) and machine learning have created the necessaries for large language modeling applications, which have recently given state-of-the-art performance on tasks ranging from automated question-answering (QA), text summarization, and medical knowledge retrieval [3]. Examples of these transformer-based architectures include Bidirectional Encoder Representations from Transformers (BERT) and Retrieval-Augmented Generation (RAG), which have created positivity around directions on medical education and decision support systems [4]. Trained on huge digitized textual corpuses, BERT is intended for extractive QA, where it selects relevant text spans from documents in the response to queries. On the other hand, RAG combines a retrieval-based search mechanism with generative modeling, generating contextually rich and dynamically produced answers integrating external knowledge sources [5].

LLMs in medical training have gained increasing attention. AI-based tutors have been shown to add value with better knowledge retention and diagnostic accuracy than classical training methods [6]. There has been a focus on domain-adapted BERT models for medical QA tasks that have shown improvements in performance over rule-based and statistical methods. Furthermore, retrieval-augmented methods were also proposed to mitigate hallucinations in generative models so that the output responses referring to pressure ulcer management are aligned with factual medical knowledge [1]. These advancements hint that transformer-based models may represent effective, scalable, and precise training options for medical personnel.

In this paper, we propose an LLM-driven QA to train healthcare professionals in the management of pressure ulcers. We discuss two major models: BERT fine-tuned on domain-specific data for extractive QA, and RAG, which retrieves relevant literature and clinical guidelines to generate answers enriched and contextualized with knowledge. Some specific objectives are:

- To evaluate extracting QA BERT model effectiveness for medical education.
- Finding out the merits of RAG for giving dynamic responses enriched with knowledge.
- Comparing the above 2 model-metrics such as BERTScore, BLEU, and retrieval recall.
- To create a working training pipeline for this with optimizations for an RTX 3090 GPU, making it a feasible application in real world.

Through the use of state-of-the-art techniques in NLP, this study aims to close knowledge gaps on pressure ulcer management, presenting a scalable AI learning solution for healthcare practitioners.

The rest of this paper is organized as follows:

Section 2 gives the background and challenges faced in building LLM-based medical QA systems. Section 3 describes the method adopted in developing the system.

Finally, Section 4 concludes this paper and puts forward possible avenues for future work.

II. Background

Pressure ulcer management has always posed a challenge to clinical practice, pushing healthcare professionals to always seek current treatment protocols and preventive measures. Traditional education has always employed static education resources: textbooks, workshops, and online training modules. However, they are typically ineffective in delivering real-time, context-relevant knowledge to healthcare workers and, therefore, delays decision-making and the adopted treatment approaches. The advent of AI-based educational tools and language models could promise a new way of closing such knowledge gaps by enabling on-demand, context facilitated learning experiences that are relevant to specific clinical situations.

A. Current Approche

The existing training programs on pressure ulcer management are majorly lecture-based and delivered by either in-person or online courses. While these approaches provide rigorous structure while disseminating knowledge, they do not possess that interactivity and adaptability leading to considerable loss of practice among healthcare practitioners. Research has demonstrated that AI-enabled solutions, particularly NLP-based systems, are sufficient for enhancing learning experience through personalization, instant query response [7]. However, a significant limitation in AI-based systems has been the failure of contemporary systems to show deep context understanding, which would cause limitation in the accuracy or completeness of responses [8].

Transformer-based models like BERT and Retrieval-Augmented Generation (RAG) are the most preferred in processing so complex medical information and producing outputs that align closely with clinical best practices. And with deep learning-based systems, NLP models will adapt to new medical findings for evidence-based contemporary recommendations for healthcare professionals, unlike traditional rule-based systems [9]. This adaptability is most crucial in management of pressure ulcers, where treatment guidelines change frequently based on new research and clinical trials [10].

B. AI in Medical Education

Artificial intelligence has transformed medical education by enabling interactive and self-directed education programs. These AI systems, including chatbots and virtual assistants, have been shown to provide medical practitioners with real-time guidance that benefits knowledge retention and clinical judgment [4]. NLP models geared toward large-scale medical corpora have demonstrated good performance in question-answering [11], thus assisting healthcare professionals with intelligent tutorial mechanisms.

In fact, BERT-based models have found extensive applications in medical QA systems, thereby achieving state-of-the-art results on several benchmarks, including clinical QA and evidence-based medicine [12]. One key benefit of BERT is its ability to extract information from a very large medical dataset, thus working well for applications requiring evidence-based and precise answers [13]. Extractive models like BERT encounter hurdles when it comes to synthesizing information from multiple sources. thus, retrieval-augmented generative models like RAG may be preferred in such scenarios.

With retrieval-augmented generation models, RAG enhances medical QA systems by joining retrieval-based search with generative text generation so these models can craft answers that are contextually relevant and factually accurate. The application of this method has been reported in retrieval of medical literature and clinical decision support systems where AI is supporting health practitioners in real time [5]. By adding retrieval into an AI system, responses generated remain in link with medical guidelines much more, thus reducing the risk of misinformation, which is one of the major issues in purely generative models [14].

C. Transformer-Based Architectures for Question Answering

The transforming models turn the vocabulary and semantics of NLP: they require a semantic representation and deep contextualization with input including much of the world longer. Traditional machine learning approaches like TF-IDF and LSTMs suffer from a long-distance relationship and ambiguity due to the medical terminologies, limiting their achievements for applications in Clinical domain [15]. Varieties like BioBERT and ClinicalBERT to BERT have preformed in dealing these issues celebrity for improvement performance in health-care related NLP tasks on a very large biomedical corpus [16]

RAGs exemplifies a hybrid that complements a retrieval-based search by generative modeling. that is, it accesses relevant research papers, clinical notes, or treatment guidelines before generating an answer so as to optimize both factuality and context of the response. [17] The retrieval and generation techniques together thus enable RAG to sink the shortcoming of regular QA systems on account of having real-world medical knowledge to ground responses.

One of the largest challenges in making LLM-based QA systems to work in clinical settings is ensuring alignment of the model's outputs, or as they call it predictions, with real-world medical guidelines. Hallucination in generative models has been noted as one of the aspects of risks related to the use of AI systems generating output that looks plausible with respect to the question but is factually incorrect [18]. To solve this problem, RAG has retrieval components that provide external grounding, greatly reducing the odds of generating unreliable answers.

Thus, BERT and RAG form complementary benefits in medical question answering. Whereas BERT possesses strong capabilities regarding extraction of answers directly from structured medical documents, RAG amplifies fact-based correctness through providing additional context prior to generating responses. In unison, these models create an excellent framework for interactive AI-assisted medical education, allowing practitioners of healthcare to create and access real-time, evidence-based data on pressure ulcer management and other clinical topics.

III. Methodology

the methodology used for this research pertains to the construction, preprocessing, and fine-tuning of a model for pressure ulcer management through question answering. The research will therefore address improving the training of all healthcare practitioners employed at Mersey Care NHS Foundation Trust, focusing on a domain-specific language model.

A. Dataset Constroction and Preprocessing

The dataset that will be used for BERT and RAG will concern only domain-specific resources on pressure ulcer management-particularly medical textbooks, clinical guidelines, and case studies found in Pubmed, Medline, and arXiv. This data will cover all key issues, such as etiology, prevention, and treatment protocols for management of the patient.

The dataset will be pre-processed by tokenizing into smaller units like sentences or paragraphs, which can then easily be fed into each one of these models. Important medical terms such as condition, treatment, body part are identified using Named Entity Recognition (NER), which will also serve both models in identifying relevant data because text normalization will prepare the dataset by purging it of all entries that belong to other than the medical domain and then standardizing all terminology used in the dataset [19].

RAG will also have a retrieval component. thus the dataset will be indexed for this retrieval step. While both models will have a similar preprocessing pipeline, a special focus will be required in structuring the retrieval component of RAG for efficient document retrieval.

B. Model Selection

1) BERT

This project will have the BERT model as its baseline. It is the model which is used for understanding context-based question answering, because it is a bidirectional transformer model. To build BERT in accordance with this task, it will use pre-trained BERT-base and be further trained using the pressure ulcer data before application in real-time scenarios.

Key parameters used in fine-tuning include:

- Learning Rate: A learning rate of 2e-5 will be applied. This is a common learning rate for fine-tuning transformer models.
- Batch Size: A batch size of 16 will also be beneficial in terms of memory use while promoting computation across an RTX 3090 GPU.
- Epochs: The model gets a period of 4 epochs of training to have sufficient time for convergence while preventing overfitting.
- Optimizer: The AdamW optimizer for weight decay will be used, as it has shown the best performance with transformer models [20].

2) RAG

On the contrary, for the RAG model, that is Retrieval-Augmented Generation, the way will be different. One important point of RAG is that it couples retrieval with generation: that is, it allows retrieval from the indexed dataset of relevant documents first in order to create the answer based on the information later. The model will be checked if it can generate an answer from the retrieval of suitable medical content associated with pressure ulcers.

- Hyperparameters: For RAG, the main hyperparameters will include:
- Retriever Learning Rate: The retriever will also be given a learning rate of 2e-5, just like the one used for BERT.
- Generator Learning Rate: The learning rate of the generator will also be at 2e-5 to allow it to be suitably fine-tuned.
- Batch Size: A batch size of 16 will be allocated for training the RAG model, similar to the BERT model [21].

C. Model Evaluation

The evaluation of both models, BERT and RAG, will involve several metrics for determining their effectiveness in providing appropriate answers concerning pressure ulcer management.

- F1-Score: This index computes the precision and recall in the same scores, thereby ensuring that the model answers questions accurately without losing relevant details [22].
- BERTScore: BERTScore was measured by estimating the semantic similarity between the given responses and the expected responses [23].
- BLEU Score: This is what both models will be evaluated on in order to check the fluency of their answers, especially in cases where they will be fed multiple answers or question answering remains a work in progress [24].
- Retrieval Recall (for RAG): This metric will measure the performance of the RAG model with respect to the retrieval of relevant information from the dataset before generating an answer. It will be more concerned on how effective the retrieval mechanism is in providing the needed context for the generator [25].

Additionally, latency tests will be performed on both models in terms of the time taken to produce an answer by both models, so as to be usable in real-time applications, especially in health services where time is of the essence.

D. Justification and Comparison

BERT is chosen for the exceptional understanding of the context and nuances of natural language. BERT is also being applied to many successful NLP tasks, including those requiring the understanding of long-distance dependency context in texts: it may therefore be used for answering medical questions based upon context.

In contrast, RAG is selected because of its ability to fetch external knowledge through the retrieval mechanism, making it very useful when large domain-specific datasets are involved and composing documents, as in medical literature. By retrieving related documents or passages and generating answers on such basis, RAG aims to produce well-informed and accurate responses, especially in a highly specialized domain such as management of pressure ulcers.

While the main difference between the models is their working functionality, with BERT producing the answers entirely on training data and a query input while RAG enhances the model further by going for relevant knowledge retrieval from an external corpus prior to producing an answer. This retrieval-based scheme benefits these models in answering domain-specific queries where knowledge has to be exhaustive and current at the same time.

Both of these models are suitable for the task in question. however, performance comparison between the models will certainly shed light on the best effective one for answering questions regarding pressure ulcer management. This implies that by evaluating contextually relevant accurate-answer-producing capability, the best model may be chosen according to application needs.

IV. Conclustion and Future Work

This research investigates the development of a domain-specific question-answering system regarding pressure ulcers. For the development of a broad-specific dataset, textbooks on medical subjects, clinical guidelines, and peer-reviewed research papers/records were included. Data preprocessing consisted of tokenization, named entity recognition (NER), and text normalization (for RAG-a supplementary task- indexation for efficient retrieval). This planting ground for the models was necessary for them to be able to process input and generate context-relevant output, i.e., contextually relevant answers.

Fine-tuning was done on the entire list of datasets with the BERT model used as a sturdy baseline for good contextual understanding against the RAG model that added retrieval to counter low response accuracy with external knowledge sources. The models were evaluated against several performance metrics which include F1-score, BERTScore, BLEU score, and retrieval recall for precision answer generation and fluency during the evaluation process. Time-latency tests were also performed to check for the feasibility of real-time applications in healthcare. Nevertheless, BERT presented a strong foundation in completing linguistic tasks. however, RAG seemed promising in dynamically retrieving domain-specific information which would massively improve response accuracy.

The next phase will focus on implementing the functional prototype of the question-answering system, integrating the most effective model based on performance results. The chosen model will therefore be made available either through a web-based or API solution to enhance accessibility to healthcare professionals. Model quantization and batch processing are optimization approaches envisaged to further enhance inference speed and computational efficiency. An intuitive interface will be developed for seamless interaction, likely in the form of a chatbot or structured query system geared toward use in medical applications.

Further prospective improvements will include the linking of real-time medical databases to update the system on the latest research and clinical guidelines. Usability, reliability, and practical effectiveness will be tested and validated with the help of healthcare practitioners. Future improvements may also involve an increase in the dataset volume, adopting multi-turn conversation abilities, and transformer-based model domain-specific fine-tuning which will allow the precision of the system to be improved for use in clinical decision-making.

This will contribute to the evolution of the AI-based medical question-answering systems, hence proving the capability of NLP technologies in fostering knowledge reach and supporting clinical decision making in pressure ulcer management.

References

[1] Y. Shi, S. Xu, Z. Liu, T. Liu, X. Li, and N. Liu, “MKRAG: Medical Knowledge Retrieval Augmented Generation for Medical Question Answering,” 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:263136303

[2] A. K. Lahiri and Q. Hu, “AlzheimerRAG: Multimodal Retrieval Augmented Generation for PubMed articles,” ArXiv, vol. abs/2412.16701, 2024, [Online]. Available: https://api.semanticscholar.org/CorpusID:274982505

[3] Z. Zhan, J. Wang, S. Zhou, J. Deng, and R. Zhang, “MMRAG: Multi-Mode Retrieval-Augmented Generation with Large Language Models for Biomedical In-Context Learning,” ArXiv, vol. abs/2502.15954, 2025, [Online]. Available: https://api.semanticscholar.org/CorpusID:276576064

[4] J. Lee et al., “BioBERT: a pre-trained biomedical language representation model for biomedical text mining,” CoRR, vol. abs/1901.08746, 2019, [Online]. Available: http://arxiv.org/abs/1901.08746

[5] P. S. H. Lewis et al., “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” CoRR, vol. abs/2005.11401, 2020, [Online]. Available: https://arxiv.org/abs/2005.11401

[6] Y. Zhang et al., “DialoGPT: Large-Scale Generative Pre-training for Conversational Response Generation,” CoRR, vol. abs/1911.00536, 2019, [Online]. Available: http://arxiv.org/abs/1911.00536

[7] E. Alsentzer et al., “Publicly Available Clinical BERT Embeddings,” CoRR, vol. abs/1904.03323, 2019, [Online]. Available: http://arxiv.org/abs/1904.03323

[8] A. Vaswani et al., “Attention Is All You Need,” CoRR, vol. abs/1706.03762, 2017, [Online]. Available: http://arxiv.org/abs/1706.03762

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” CoRR, vol. abs/1810.04805, 2018, [Online]. Available: http://arxiv.org/abs/1810.04805

[10] S. Nerella et al., “Transformers and large language models in healthcare: A review,” Artif Intell Med, vol. 154, pp. 102900–102900, 2023, [Online]. Available: https://api.semanticscholar.org/CorpusID:259316437

[11] Y. Liu et al., “RoBERTa: A Robustly Optimized BERT Pretraining Approach,” CoRR, vol. abs/1907.11692, 2019, [Online]. Available: http://arxiv.org/abs/1907.11692

[12] K. Clark, M.-T. Luong, Q. V Le, and C. D. Manning, “ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators,” CoRR, vol. abs/2003.10555, 2020, [Online]. Available: https://arxiv.org/abs/2003.10555

[13] S. Wiegreffe, E. Choi, S. Yan, J. Sun, and J. Eisenstein, “Clinical Concept Extraction for Document-Level Coding,” CoRR, vol. abs/1906.03380, 2019, [Online]. Available: http://arxiv.org/abs/1906.03380

[14] T. Wolf et al., “HuggingFace’s Transformers: State-of-the-art Natural Language Processing,” CoRR, vol. abs/1910.03771, 2019, [Online]. Available: http://arxiv.org/abs/1910.03771

[15] D. Bahdanau, K. Cho, and Y. Bengio, “Neural Machine Translation by Jointly Learning to Align and Translate,” CoRR, vol.

abs/1409.0473, 2014, [Online]. Available: https://api.semanticscholar.org/CorpusID:11212020

[16] K. Huang, J. Altosaar, and R. Ranganath, “ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission,” CoRR, vol. abs/1904.05342, 2019, [Online]. Available: http://arxiv.org/abs/1904.05342

[17] C. Raffel et al., “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer,” CoRR, vol. abs/1910.10683, 2019, [Online]. Available: http://arxiv.org/abs/1910.10683

[18] S. M. T. I. Tonmoy et al., “A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models,” ArXiv, vol. abs/2401.01313, 2024, [Online]. Available: https://api.semanticscholar.org/CorpusID:266725532

[19] G. Xiong, Q. Jin, X. Wang, M. Zhang, Z. Lu, and A. Zhang, “Improving Retrieval-Augmented Generation in Medicine with Iterative Follow-up Questions,” Pac Symp Biocomput, vol. 30, pp. 199–214, 2024, [Online]. Available: https://api.semanticscholar.org/CorpusID:271600473

[20] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “Bleu: a Method for Automatic Evaluation of Machine Translation,” in Annual Meeting of the Association for Computational Linguistics, 2002. [Online]. Available: https://api.semanticscholar.org/CorpusID:11080756

[21] T. B. Brown et al., “Language Models are Few-Shot Learners,” CoRR, vol. abs/2005.14165, 2020, [Online]. Available: https://arxiv.org/abs/2005.14165

[22] B. Saha, U. Saha, and M. Z. Malik, “QuIM-RAG: Advancing Retrieval-Augmented Generation With Inverted Question Matching for Enhanced QA Performance,” IEEE Access, vol. 12, pp. 185401–185410, 2025, [Online]. Available: https://api.semanticscholar.org/CorpusID:274630309

[23] H. Yu, A. Gan, K. Zhang, S. Tong, Q. Liu, and Z. Liu, “Evaluation of Retrieval-Augmented Generation: A Survey,” ArXiv, vol. abs/2405.07437, 2024, [Online]. Available: https://api.semanticscholar.org/CorpusID:269758033

[24] A. Radford et al., “Learning Transferable Visual Models From Natural Language Supervision,” CoRR, vol. abs/2103.00020, 2021, [Online]. Available: https://arxiv.org/abs/2103.00020

[25] D. Ru et al., “RAGChecker: A Fine-grained Framework for Diagnosing Retrieval-Augmented Generation,” ArXiv, vol. abs/2408.08067, 2024, [Online]. Available: https://api.semanticscholar.org/CorpusID:271874517