2024 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB) | 979-8-3503-5663-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/CIBCB58642.2024.10702147

The Development of CanPrompt Strategy in Large Language Models for Cancer Care

Noman Ahmad
Faculty of Science
Thompson Rivers University
BC, Canada
nomandogar56@gmail.com

Ehsan Mamatjan
MamatjanLab
Thompson Rivers University
BC, Canada

Tursun Wali
School of Information Technology
Carleton University
ON, Canada
Tursuntuerxunwaili@cunet.ca

Yasin Mamatjan
Faculty of Science
Thompson Rivers University
BC, Canada
ymamatjan@gmail.com

***Abstract*—Background: The recent revolution in Large Language Models (LLMs) is transforming industries, enhancing communication, and reshaping research methodologies. LLMs have found significant applications across various sectors, notably in finance for stock market predictions, and in healthcare, where complex medical data is analyzed for diagnosis at an early stage, improving diagnostic procedures, and personalized treatment planning. In healthcare, where complex medical data is analyzed for diagnosis at an early stage. Despite the immense potential, challenges such as overwhelming Big Data, model hallucinations, and ethical concerns about patient privacy and bias persist. Method: We implemented novel strategies like CanPrompt to mitigate the accuracy and hallucination concerns to ensure responsible deployment. The CanPrompt strategy utilizes prompt engineering combined with few-shot and in-context learning to significantly enhance model accuracy by generating more relevant answers. The models were tested against a specialized dataset from MedQuAD, focusing on cancer, and evaluated using metrics like ROUGE and BERTScore to assess the semantic and syntactic accuracy of generated responses against validated "Gold Answers". Through this approach, the study seeks to outline the potential and limitations of LLMs in improving cancer care. Result: After applying CanPrompt with models Mistral 7x8b, Falcon 40b, and Llama 3-8b, BERTScore results showed Mistral leading with an accuracy around 84%, Falcon slightly lower, and Llama the least, with respective precision scores also reflecting a similar trend. Conclusion: The study demonstrates the promise of LLMs in cancer care through the introduction of CanPrompt.**

***Index Terms*—Large Language Model (LLM), Cancer, prompt, ROGUE score, BERTScore, Mistral 7x8B, Falcon 40b, and Llama 3-8b**

I. Introduction

Artificial Intelligence (AI) in the healthcare sector has seen a boom and is set to change the medical field with increased operational efficiency, early diagnosis, precision in imaging, and the development of drugs. With a set of quadruple aims, AI is applied in the health sector. These aims include reducing cost per capita, improving treatment experience, improving collective health, and improving the work-life of healthcare providers [1]. Currently, Machine Language and Large Language Models (LLMs) have sped up this revolution as generative bots like Generative Pre-trained Transformer (GPT) assess the genome profile, living patterns, and medical history to suggest personalized medical decisions.

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant (RGPIN-2023-05341).

The trust in telehealth increased with the pandemic as humans devised tech-centered ways to treat and diagnose the disease at a distance [2]. However, the integration of AI in the real world, especially in the health sector faces some serious challenges. LLM requires large datasets to perform its functions and produce reliable decisions. When the datasets provided to AI models are of low quality, or lacking diversity and sufficiency, the results and decisions made by AI will eventually be biased, factually incorrect, or dangerous at most [3]. Furthermore, the biggest concern in integrating LLMs into healthcare is that they generate plausible-sounding but false results, commonly known as hallucinations [4]. These issues can be especially dangerous in high-stakes domains such as healthcare [5].

AI uses LLMs to mimic human behavior and make informed decisions using datasets that otherwise require human cognition. In the healthcare sector, deep learning is used for the detection of cancerous lesions that the human eye cannot perceive and hence, the revolution of radiomics. It also includes surgical robots that have been approved for use in the USA since 2000 [6]. AI can be used to study microorganisms and hence the prediction of pandemics [7]. Moreover, cancer research and anticancer drug development have seen a radical transformation with the use of next-generation sequencing (NGS). This allows the use of a tremendous amount of genomics data, which includes the whole-genome, whole-exome, or whole-transcriptome analysis, allowing a peek into cancer biology as well as its treatment [8]. A research study in the UK published its findings related to cancer diagnosis. The findings show a reduction of 5.7% and 9.4% in false positives and false negatives respectively in interpreting mammograms of breast cancer [9].

GPT-4 has been used for the diagnosis of cancer using different datasets and models. Bhattarai et al. (2023) compared and analyzed the application of GPT-4 with other models including Flan-T5-xl, Flan-T5-xxl, scispaCy, and medspaCy in the identification of phenotypes extracted from Electronic Health Record (EHR) which is not present in structured data [10]. This research focuses on the identification of treatment, stages, and malignancy of cancer, and results show a higher F1 score achieved by GPT-4 in the cancer diagnosis as compared to other models. However, this study employs closed-source

979-8-3503-5663-2/24/$31.00 ©2024 IEEE

LLMs that are costly and may pose privacy risks as the data-collection process cannot be monitored, owing to its closed nature

To address these issues and tackle the challenges found in question-answering systems concerning cancer-related problems, we introduced a technique known as CanPrompt (Fig. 1). This study aims to evaluate different LLM models and build a robust framework for cancer applications. To achieve this goal, we divided the overall objective into the following research questions (RQs):

1) RQ1: Do different techniques show an agreement in cancer-specific questions?
2) RQ2: How much accuracy score can be achieved for predicting/QA with prompt engineering?
3) RQ3: What is the capability/limitation of our LLM model in various cancer questions?

II. Methodology

A. Overview (workflows)

We proposed a new technique known as CanPrompt approach that employs "prompt engineering", leveraging a few illustrative examples (few-shot learning) alongside precise contextuality to clearly outline the method (Fig. 1). In this framework, we used a dataset composed of questions and domain-validated responses (golden answers) that specifically address diverse cancer topics. To assess the effectiveness of our approach, we applied SOTA (State of the Art) open-sourced models to produce answers. These responses were then measured against the golden answers. Our evaluation utilized two key metrics: BERTScore, which evaluates the semantic similarity between the generated text and the golden answer, and the Recall-Oriented Understudy for Gisting Evaluation (ROUGE), which measures the content overlap.

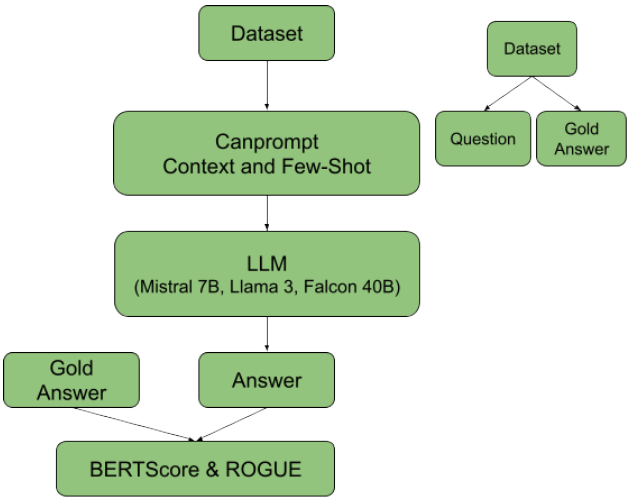

Fig. 1. A robust framework for the CanPrompt prompting method and evaluation of different LLM models for cancer applications.

B. Experimental Setup (data collection)

We used the MedQuad Cancer Q/A Dataset which includes questions varying around diagnoses and treatment. MedQuAD is a large dataset that was developed from 12 NIH websites with approximately 47,457 question-answer pairs, constituting 37 question types including diagnosis, treatment, and side effects [11]. The data contains a set of questions and an annotated answer we call it a gold answer. These gold answers are based on trustworthy medical sources and are expertly crafted to offer precise and sufficient information that directly addresses the questions asked.

C. CanPrompt

CanPrompt enables the model to contextualize medical information, such as symptoms, treatments, and research findings, within the broader framework of existing knowledge. The aforementioned training strategy is specifically for the task of cancer research and treatment, for few-shot learning, a set of input-output pairs focused on cancer-related data. For instance, pairs could consist of patient genetic profiles along with successful treatment outcomes, or histopathology image inputs correlated with diagnostic annotations. Through this process, the model learns to recognize and predict outcomes based on the patterns inherent in the data, simulating a training session. Once the initial examples are in place, the prompt extends the model to predict learned patterns to the effectiveness of new, experimental cancer treatments based on similar genetic or cellular characteristics found in the training data. This learning paradigm in Transformer models involves dynamic computation of attention weights based on the input sequence. The few-shot examples serve as a form of soft prompt, guiding the model's attention to relevant features and relationships within its vast parameter space. This allows the model to implicitly fine-tune its behavior for the specific task at hand, without altering its underlying weights. This task would make full use of the model's pre-trained knowledge of biological processes, genetic variations, and their impacts on disease progression and treatment response. Furthermore, the model's capacity for in-context learning is leveraged by guiding it to apply these learned insights to analogous scenarios in cancer research, such as predicting patient response to immunotherapy based on their tumor micro-environment or genetic mutations. This method not only teaches the model through specific detailed examples but also utilizes its broader capabilities to address related challenges, ultimately aiming to enhance its performance in the complex field of oncology.

D. SOTA Models: Mistral 7x8B, Falcon 40b, and Llama 3-8b

We used SOTA (State of the Art) open-sourced models such as Mistral 7x8B, Falcon 40b, and Llama 3-8b. We applied model quantization, converting the models to Bfloat16 (16-bit floating point format for machine learning) precision. This not only reduces the computational costs but also makes the models more accessible for broader research and application purposes. Additionally, we used our prompting strategy, "CanPrompt", to test our model on a specialized dataset. This

dataset, consist of 720 questions with gold-standard answers, focuses specifically on cancer research, providing a robust framework for evaluation.

E. Evaluation Techniques

To evaluate the model, we ask a question from the dataset which consists of questions and "gold answers". The gold answers are the ideal answers to the questions asked and the standard against which the model is tested. We are evaluating the model's answer with ROUGE and BERTScore by comparing the model's answer to the gold answer. ROUGE is a set of metrics used to evaluate generated outputs by comparing their similarity to a human's response. A ROUGE-1 score of the number of words in the generated summaries matches those in the reference, assessing word-level accuracy. The ROUGE-2 score examines two-word sequences and syntactic similarity between the generated and reference texts. The ROUGE-L score number of the longest common sub-sequence of words.

BERTScore is another metric for evaluating generated text, however, it does not measure the similarities of text based on n-gram overlaps. BERTScore computes cosine similarities between the contextual embedding of the generated text and the original text's tokens. The BERTScore outputs its result with these values: the f1 score, the recall score, and the precision score. All of these values are between 0 and 1. If the value is closer to 1, then it is more accurate, and if it is closer to 0, then it is less accurate.

III. Results

A. BERTScore

As shown in Fig. 2, Mistral achieved an F1 score of 80%, a recall score of 77%, and a precision score of 84.3%. Falcon scored slightly lower with an F1 score of 79.6%, a recall score of 77.1%, and a precision score of 82.6%. Llama3-8b recorded an F1 score of 79.3%, a recall score of 77.5%, and a precision score of 81.4%.

B. BERTScore Variability: Mistral 8x7B

Fig 3.A, 3.B, and 3.C present histograms of BERTscore metrics—Precision, Recall, and F1-Score—used to evaluate the performance of the Mistral 8x7B model. Precision Scores in Fig. 3.A, which measure the proportion of relevant instances among retrieved ones, range from approximately 0.76 to 0.86, with a common concentration between 0.80 to 0.82, indicating high relevance in the model's retrieved content. Recall Scores depicted in Fig. 3.B span from about 0.675 to 0.85, peaking between 0.75 and 0.775, showing the model's ability to retrieve a significant portion of relevant information, though with more variation than precision. Lastly, the F1-Scores in Fig. 3.C, accounting for both false positives and false negatives and useful in scenarios of class imbalance, vary from roughly 0.74 to 0.88, primary clustering of F1 scores is below the 0.82 mark, suggesting a balanced accuracy between precision and recall for the dataset or evaluations assessed.

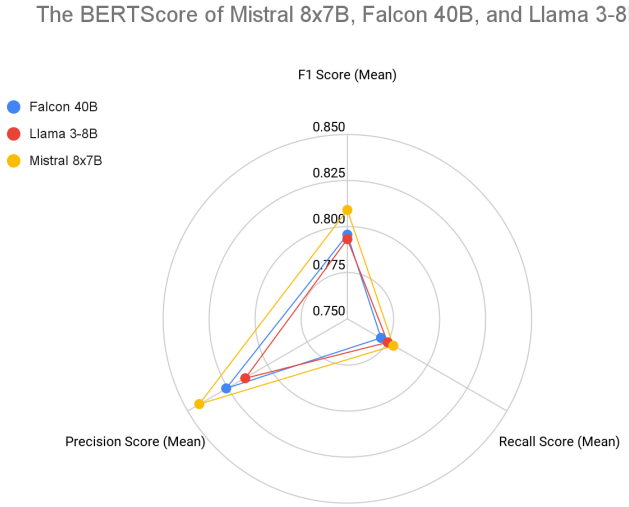

Fig. 2. The performance scores of Mistral 8x7B, Falcon 40B, and Llama 3-8B on the BERTScore metrics and their comparisons.

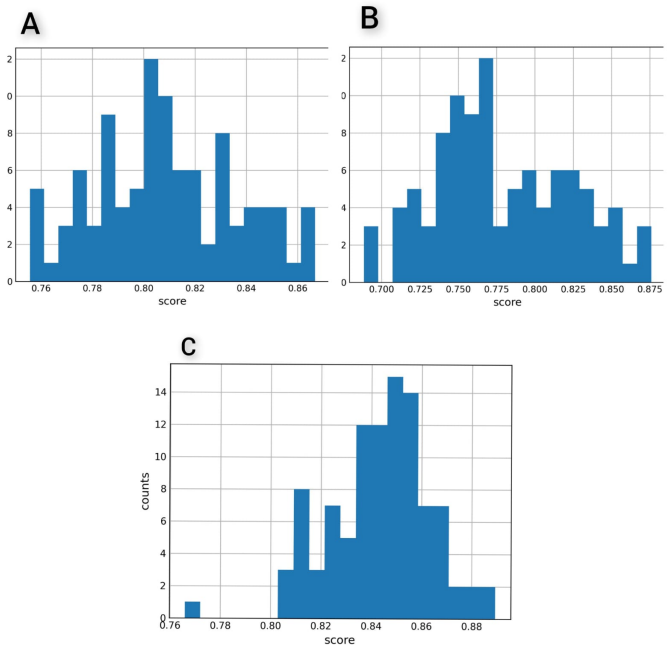

Fig. 3. The histogram illustrates the BERTScore performance of Mistral 8x7B on a cancer dataset, showcasing its superior performance. Mistral 8x7B achieved the highest average Precision scores of 0.843 (A), Recall scores of 0.779 (B), and F1 scores of 0.809 (C). The x-axes indicate the score values, while the y-axes represent the count of occurrences, providing insight into the distribution and central tendency of these evaluation metrics on the dataset.

C. BERTScore Variability: Falcon 40B

As shown in Fig. 4.A, 4.B, and 4.C, the Precision Score in Fig. 4.A shows similar trends in scores concentrated around 0.8 to 0.85, indicating that some of the accuracies are very high for "Gold Answer" but slightly lower than Mistral. The recall scores in Fig. 4B have a wider distribution, ranging from about 0.65 to 0.85. There are multiple peaks, indicating several clusters where recall scores tend to congregate. Lastly, the F1 score histogram (Fig. 4.C) shows a clear peak around 0.8, which indicates that the F1 scores are mostly high and there is less variation in F1 scores than in recall scores. The shape is somewhat asymmetrical, leaning to the right, suggesting that while many of the F1 scores are clustered around a central value, there is a tail towards the higher scores.

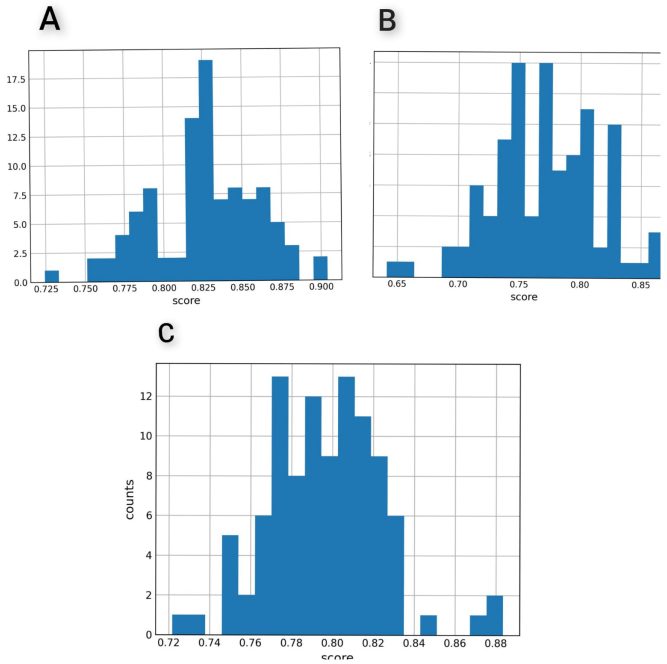

Fig. 4. The histogram illustrates the BERTScore performance of Falcon 40B on a cancer dataset, showcasing slightly lower performance than Mistral 8x7B's. Falcon 40B achieved the highest average Precision scores of 0.826 (A), Recall scores of 0.771 (B), and F1 scores of 0.796 (C). The x-axes indicate the score values, while the y-axes represent the count of occurrences, providing insight into the distribution and central tendency of these evaluation metrics on the dataset.

D. BERTScore Variability: Llama3-8b

BERTScore in Llama3 seems to be lower in comparison to both models evaluated, the distributed variation of Precision Score (Fig. 5.A) ranges from 0.775 to 0.8, suggesting that the majority of the evaluated text has a precision score in this range. Fig. 5.B (recall) variation ranges from 0.75 to 0.775. It shows a wider spread, implying that the model may not consistently retrieve or generate all the relevant information across the board. Lastly, the F1 score (Fig. 5C), being a balance of the two, indicates that while the model has a reasonable trade-off between precision and recall shows a peak around 0.78.

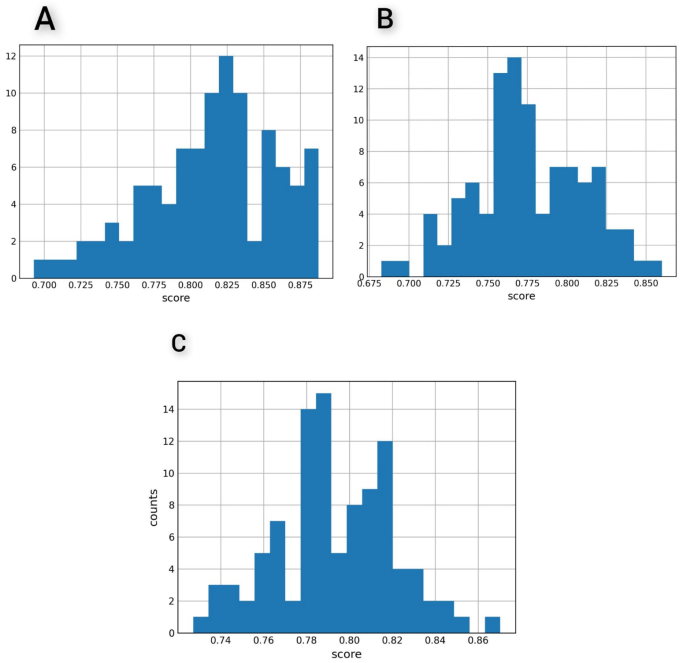

Fig. 5. The histogram displays the BERTScore performance of Llama3-8b on a cancer dataset, highlighting its results. Llama3-8b achieved Precision scores of 0.814 (A), Recall scores of 0.775 (B), and F1 scores of 0.793 (C) among the models evaluated. The x-axis represents the score values, while the y-axis indicates the count of occurrences, offering insight into the distribution and central tendency of these evaluation metrics on the dataset.

E. ROGUE Score

As shown in Fig. 6.A, Mistral demonstrated superior performance with ROUGE-1 scores of approximately 26%, ROUGE-2 scores of 8.2%, and ROUGE-L scores of 15%. In Fig. 6.B, Falcon followed with ROUGE-1 scores of 21.7%, ROUGE-2 scores of 7.3%, and ROUGE-L scores of 12.29%. Fig. 6.C Llama 3B had the lowest scores, with ROUGE-1 at around 19.5%, ROUGE-2 at 5.9%, and ROUGE-L at approximately 12.5%.

IV. Discussion and Conclusion

LLMs have the potential to dynamically change the medical landscape by speeding up the tests, reducing costs, and ensuring patient safety. However, LLMs are not yet thoroughly investigated in the cancer field. In this study, we introduced CanPrompt, a prompting strategy to improve the SOTA model's accuracy with cancer-related questions.

We evaluated the performance of these LLMs using precision, recall, and F1 score metrics. The findings indicated a high level of concurrence among open-source LLMs, achieving an accuracy rate of up to 84%. This demonstrates the significant potential of LLMs for healthcare applications, suggesting they could be pivotal in enhancing diagnostic processes, personalizing patient care, and improving the efficiency of medical research. In fields like cancer medicine, where errors may have fatal consequences, hallucinations can be dangerous

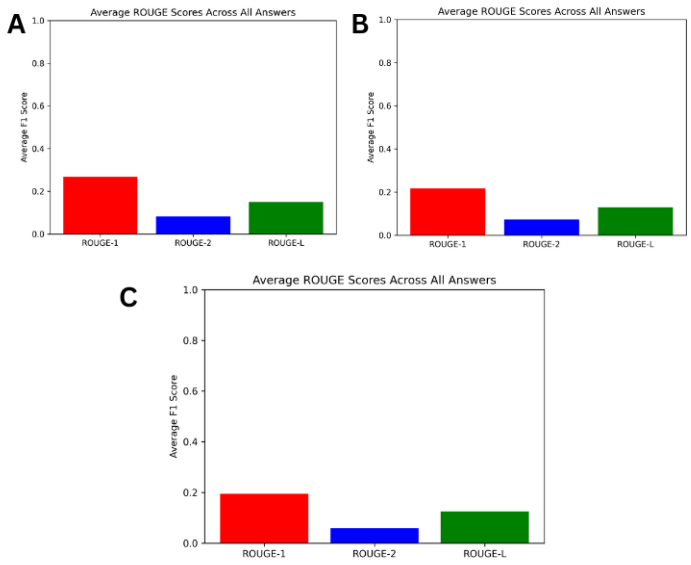

Fig. 6. The scores of Mistral 8x7B (A), Falcon 40B (B), and Llama 3-8B (C) on the ROGUE metrics.

and weaken the trust of healthcare professionals and cancer patients. Fine-tuning the model with reliable data, specific to the topic of the discussion, is proven to help developers reduce hallucinations [12]. However, this requires a significant amount of time and resources, making it difficult for smaller organizations to implement.

However, previous studies [10] used closed-sourced LLM which shows higher precision involves higher costs, particularly regarding usage fees, which are charged per token processed. Furthermore, using closed-source models for sensitive applications, like processing patient data, raises significant privacy concerns. Since the model and its operations are not open for review, it's harder to verify how the data is handled, and stored, or if it is inadvertently retained. In closed-source environments, fine-tuning is challenging flexibility or capability to integrate and manipulate this data due to restricted access to the model's internal workings or training procedures.

CanPrompt shows promising results by leveraging prompt engineering techniques which significantly improves the model understanding of specific questioning tasks. It also reduces the issue of hallucination by providing a context of the specific problem to it. By utilizing prompt engineering techniques tailored for oncology-related applications, the model can be fine-tuned to effectively comprehend and process intricate medical data, terminology, and patient details. For example, in a healthcare setting, CanPrompt could enhance the AI's capability to accurately interpret and respond to detailed questions about various cancer types, available treatments, expected outcomes, and recent research. By equipping the AI with carefully designed prompts that include specific contextual details about a patient's medical history or a particular cancer type, the system can deliver more accurate and relevant responses. This enhancement is particularly valuable for oncologists in identifying early-stage cancers or in devising customized treatment strategies that reflect the unique genetic profile of a patient's cancer.

In our evaluation, we used ROUGE1 to evaluate the accuracy of the generated text and compared it with a golden answer. ROUGE evaluates the text by comparing words in the reference text with generated text but it does not account for synonyms or paraphrased content unless the same words are used, which is crucial for evaluating the generating text. Furthermore, it primarily measures the overlap of n-grams (words and sequences of words) between the generated text and a set of reference text. This method can fail to capture the quality of a summary that uses different vocabulary to express the same content, leading to potentially misleading scores.

A. *Conclusion and future work:*

Building upon the promising results of our CanPrompt strategy, particularly with the Mistral 7x8B model, we propose to develop a specialized, fine-tuned LLM specifically tailored for oncological applications. This aims to leverage the robust performance demonstrated by Mistral 7x8B in our initial studies and further enhance its capabilities in the domain of cancer care. Our research evaluated multiple open-source models, including Mistral 7x8B, Falcon 40b, and Llama 3-8b, providing a comprehensive comparison of their performance in cancer-related tasks.

The proposed fine-tuning process will involve curating a comprehensive, cancer-specific dataset, encompassing a wide range of oncological information, including but not limited to diagnostic criteria, treatment protocols, genetic markers, and emerging research findings. This dataset will be validated by oncology experts to ensure its accuracy and relevance to current clinical practices. By focusing on the Mistral 7x8B model, which showed superior performance in our initial evaluations, we aim to create a highly specialized tool for cancer care applications.

While our study provides valuable insights into the performance of several prominent open-source LLMs (Mistral 7x8B, Falcon 40b, and Llama 3-8b), it is important to acknowledge that the landscape of language models is rapidly evolving. Our research, though comprehensive within its scope, does not encompass all available LLMs, particularly the most recent releases or highly specialized models. The performance of these unexplored models in oncological applications remains unknown and could potentially offer different or improved results. This limitation underscores the need for ongoing comparative studies as new models emerge, to ensure that the most effective tools are being leveraged for cancer care applications.

REFERENCES

[1] C.J. Kelly, et al. "Key challenges for delivering clinical impact with artificial intelligence." *BMC medicine* 17 (2019): 1-9.

[2] S. Patil, H. Shankar, "Transforming healthcare: harnessing the power of AI in the modern era." *International Journal of Multidisciplinary Sciences and Arts,* 2.1 (2023): 60-70.

[3] A. Aldoseri, K.N. Al-Khalifa, and A. Hamouda. "Re-thinking data strategy and integration for artificial intelligence: concepts, opportunities, and challenges." *Applied Sciences* 13.12 (2023): 7082.
[4] Z. Ji, et al. 2023. "Survey of hallucination in natural language generation". *ACM Computing Surveys*, 55(12):1–38.
[5] S. Tian, et al. 2024. Opportunities and challenges for chatgpt and large language models in biomedicine and health. *Briefings in Bioinformatics*, 25(1):bbad493.
[6] T. Davenport and R. Kalakota. "The potential for artificial intelligence in healthcare." *Future healthcare journal* 6.2 (2019): 94.
[7] S. Alowais, et al. "Revolutionizing healthcare: the role of artificial intelligence in clinical practice." *BMC Medical Education*, 23.1 (2023): 689.
[8] R. Perez-Lopez, et al. "A framework for artificial intelligence in cancer research and precision oncology." *NPJ Precision Oncology* 7.1 (2023): 43.
[9] S.M. McKinney, et al. "International evaluation of an AI system for breast cancer screening." *Nature* 577.7788 (2020): 89-94.
[10] K. Bhattarai, et al. "Leveraging GPT-4 for Identifying Cancer Phenotypes in Electronic Health Records: A Performance Comparison between GPT-4, GPT-3.5-turbo, Flan-T5 and spaCy's Rule-based Machine Learning-based methods." *bioRxiv*, (2023): 2023-09.
[11] A.B. Abacha, et al. "Overview of the MEDIQA 2019 shared task on textual inference, question entailment and question answering." *Proceedings of the 18th BioNLP Workshop and Shared Task*. 2019.
[12] L. Ouyang, et al. "Training language models to follow instructions with human feedback." *arXiv, Cornell University*, 4 March 2022.