Ethics of Large Language models in Medicine and Medical Research

Maseeh College Of Engineering and Computer Science, Portland State University,Oregon, U.S

Gauri Atul Kasar gkasar@pdx.edu
Kavita Jajula kavithaj@pdx.edu
Annie Pathania pathania@pdx.edu
Dhatri Ramagiri dhatri@pdx.edu
Yashaswi Shah yashaswi@pdx.edu

I. Introduction

In recent years, the intersection of artificial intelligence and healthcare has witnessed a lot of advancements, with large language models emerging as powerful tools in the context of medicine and medical research. These sophisticated models, such as GPT (Generative Pretrained Transformer) and BERT (Bidirectional Encoder Representations from Transformers), hold immense promise for streamlining processes, aiding diagnostics, and unlocking insights from vast volumes of medical data. However, as these technologies become integral to the healthcare landscape, a critical dialogue surrounding their ethical implications has taken center stage. The ethical considerations surrounding the use of large language models in medicine raise profound questions about privacy, bias, accountability, and the delicate balance between innovation and patient welfare. In this discourse, we navigate the intricate landscape of ethical considerations associated with the deployment of large language models in medicine, exploring both the opportunities they present and the challenges they pose to ensure a responsible and equitable integration of these transformative technologies into the field of healthcare. Using LLMs to understand the x-ray report, or even use it to understand the test results are some of the most important use cases that we can think of. It has great potential in this area. The X-ray use case is already present with the Med-Palm2 model and there is also a comparison of expert vs LLM response and the LLM is pretty accurate with the human expert response.

II. Motivation

Current research in the medical field using large language models (LLMs) focuses on several key areas. LLMs are being applied for literature analysis, helping researchers and clinicians sift through vast medical literature more efficiently. They are also used in clinical decision support, offering insights and suggestions based on patient data and medical knowledge. Additionally, LLMs facilitate biomedical experimentation by generating hypotheses and designing experiments. However, this field faces challenges such as potential biases in AI algorithms, managing misinformation, and ensuring patient privacy. LLMs are also employed in analyzing electronic health records for improved patient care, assisting in medical diagnostics, and automating the processing of medical literature. The ongoing research aims to improve the accuracy, reliability, and ethical compliance of AI in healthcare. Below examples demonstrate the

potential of LLMs to transform healthcare, enhancing diagnostic accuracy, personalized medicine, and research efficiency.

For instance, **Google's DeepMind developed AlphaFold**, which predicts protein structures with high accuracy, revolutionizing drug discovery and biomolecular research. AlphaFold uses deep learning techniques, specifically a type of neural network known as a transformer. This transformer network, which indeed shares a lineage with the models used in LLMs, is applied to the problem of predicting the 3D structure of a protein based solely on its amino acid sequence.

Models like **BioGPT** is a domain-specific generative Transformer language model developed by Microsoft, pre-trained on a large scale of biomedical literature. It aims to enhance text mining and knowledge discovery from biomedical texts, which are crucial in areas such as drug discovery and clinical therapy. BioGPT has been evaluated on several biomedical NLP tasks and has shown to outperform previous models, demonstrating its effectiveness in the biomedical domain. The model can assist with tasks like entity recognition, interaction mining, question answering, and abstract generation from biomedical literature

PubMed's LitCovid: LitCovid is a curated literature hub for tracking up-to-date scientific information about COVID-19. It uses AI algorithms to categorize and summarize the rapidly growing list of research articles related to COVID-19, making it easier for researchers to find relevant studies.

CancerGPT-
https://decrypt.co/148352/meet-cancergpt-an-ai-that-predicts-the-results-of-cancer-treatment-research

CancerGPT is an AI model that predicts the effects of drug combinations on rare tissues in cancer patients. Developed by researchers from the University of Texas and the University of Massachusetts, it uses pre-trained language models to analyze medical texts and make biological inferences with notable accuracy, even with few or zero samples. The model is highlighted for its potential in areas with limited structured data, offering a significant step forward in medical research.

GatorTronGPT- https://www.nature.com/articles/s41746-023-00958-w

This study from Nature's Digital Medicine examines GatorTronGPT, published on **16th Nov, 2023** a generative large language model developed using clinical text for applications in medical research and healthcare. It outperformed existing models in biomedical natural language processing and proved indistinguishable from human writing in readability and clinical relevance according to physician evaluations. This work illuminates the potential and challenges of LLMs in enhancing healthcare quality and medical research.

MediTron 70B- a large language open-source model released on **27th Nov, 2023** ,specifically tailored to the medical domain. The model has 70 billion parameters and has been trained on a dataset curated from various medical texts, including PubMed articles, abstracts, and internationally-recognized medical guidelines. The aim of MEDITRON-70B is to democratize access to medical knowledge by providing an AI that can assist with clinical decision-making and potentially improve healthcare delivery. The model is noted to outperform other publicly available models, including GPT-3.5 and Med-PaLM, and is competitive with closed-source models like GPT-4 and Med-PaLM-2.

The paper https://www.nature.com/articles/s41586-023-06291-2

"**Large language models encode clinical knowledge**," published in Nature, on **27 July 2023**, examines the capabilities of large language models (LLMs) in the context of clinical knowledge and applications. The study introduces MultiMedQA, a comprehensive benchmark combining several medical question-answering datasets. It evaluates the performance of the Pathways Language Model (PaLM) and its instruction-tuned variant, Flan-PaLM, across these datasets. The research demonstrates that these models, particularly Flan-PaLM, achieve state-of-the-art accuracy in medical question answering, including a significant improvement in US Medical Licensing Exam-style questions. The paper emphasizes the potential utility of LLMs in medicine while acknowledging their current limitations and the need for ongoing development and evaluation to ensure their safety and effectiveness in clinical applications.

These studies demonstrate the remarkable capabilities of LLMs in encoding clinical knowledge and improving medical question answering. However, challenges such as bias, misinformation, and privacy concerns remain. The continuous evolution of these models promises to democratize medical knowledge access, improve diagnostic accuracy, and accelerate drug discovery, but it also underscores the need for careful evaluation and ethical considerations in their application.

III. Methodology

Pre-trained language models have attracted increasing attention in the biomedical domain, inspired by their great success in the general natural language domain. We evaluate biomedical NLP tasks(PubMedQA) on three different AI Models i.e., BioGPT, ChatGPT and MetaAI. PubMed is one of the most popular biomedical search engines, covering more than 30M articles. PubMedQA is a biomedical question answering dataset[7].
Each sample is constructed from a PubMed abstract, containing a question, a reference context, a long answer, and a yes/no/maybe label which is the answer to the question.We used 1K labeled QA instances for testing. Many AI models have been tested using the PubMedQA dataset. Med-Palm2 was leading with accuracy 81.6%. On Nov 28, 2023 GPT-4(Med Prompt) beat Med-Palm2 with 82% accuracy[16].

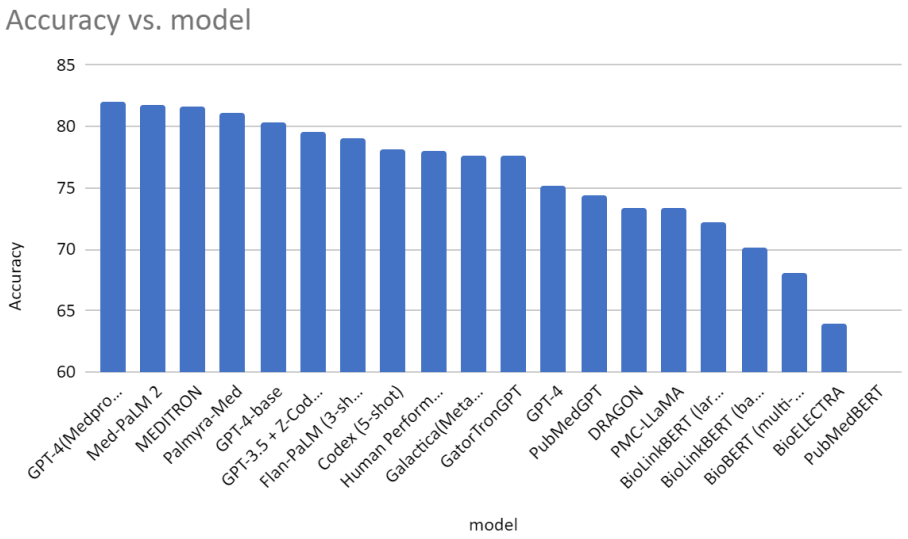

Fig 1: PubMedQA Leaderboard with accuracy

AI models for testing :

We used ChatGPT(GPT 3.5), BioGPT and MetaAI to test the PubMedQA dataset. We asked AI models to reply in yes/no/maybe format for the biomedical questions.

Prompt **:** Answer yes/no/maybe for below biomedical questions

Question : "Is it appropriate to implant kidneys for elderly donors in young recipients?"

Answer: ?

Reply from AI model : Yes

Below are the replies from AI model

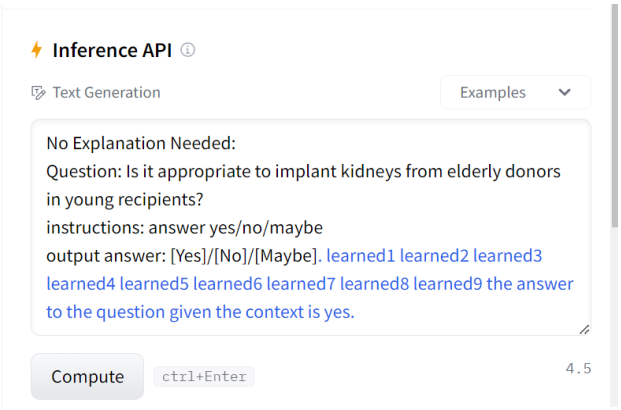

Fig 2: BioGPT prompt and reply for PubMedQA

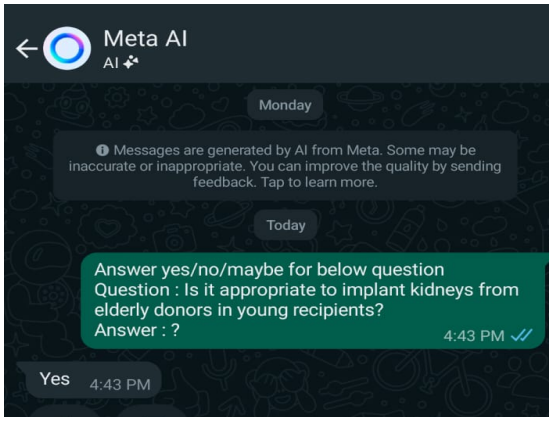

Fig 3: MetaAI prompt and reply for PubMedQA

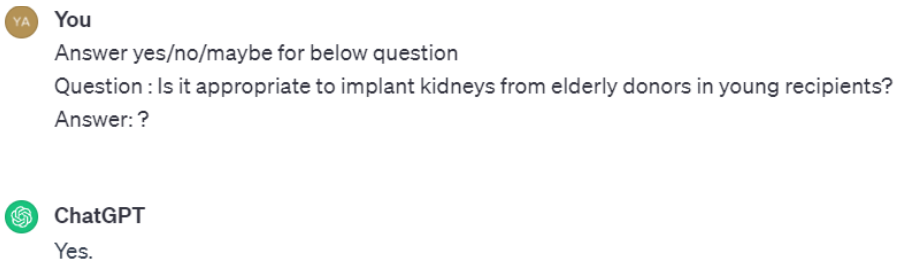

Fig 4: ChatGPT prompt and reply for PubMedQA

IV. Results and Analysis

Dataset for testing Questions : Out of 1k labeled questions from PubMedQA dataset, we used 198 questions splitted equally as yes/no/maybe.
Results : Results Link has all questions and responses of all AI models tested.

ChatGPT Result :

Accuracy : The accuracy of ChatGPT is 39%. Out of 198 questions, ChatGPT could predict 77 questions correctly.

Confusion Matrix of Chatgpt

```
[[13 18 35]
 [ 3 33 30]
 [ 4 31 31]]
```

	precision	recall	f1-score	support
0	0.65	0.20	0.30	66
1	0.40	0.50	0.45	66
2	0.32	0.47	0.38	66
accuracy			0.39	198
macro avg	0.46	0.39	0.38	198
weighted avg	0.46	0.39	0.38	198

Fig 5: Result Metrics for ChatGPT response

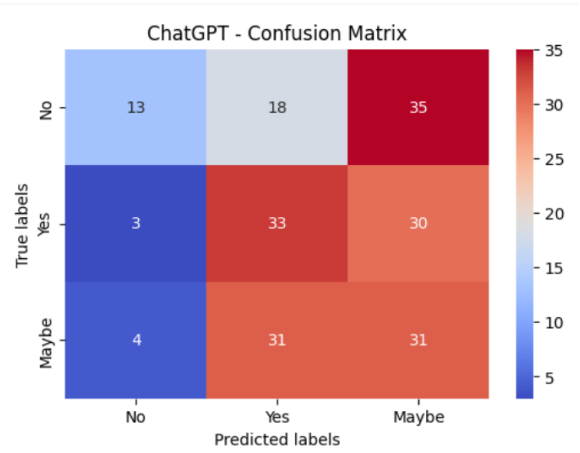

Fig 6: ChatGPT Confusion Matrix

BioGPT Results :

Accuracy : The accuracy of BioGPT is 64%. Out of 198 questions, ChatGPT could predict 126 questions correctly.

Confusion Matrix of BioGPT
[[40 4 22]
 [8 41 17]
 [2 19 45]]

	precision	recall	f1-score	support
0	0.80	0.61	0.69	66
1	0.64	0.62	0.63	66
2	0.54	0.68	0.60	66
accuracy			0.64	198
macro avg	0.66	0.64	0.64	198
weighted avg	0.66	0.64	0.64	198

Fig 7: Result Metrics for BioGPT response

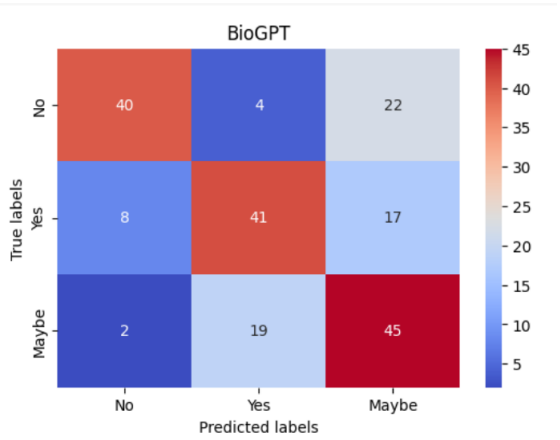

Fig 8: ChatGPT Confusion Matrix

MetaAI Results:

Accuracy **:** The accuracy of MetaAI is 44%. Out of 198 questions, MetaAI could answer 77 questions correctly.

Confusion Matrix of MetaAI
[[20 32 14]
 [4 52 10]
 [15 35 16]]

	precision	recall	f1-score	support
0	0.51	0.30	0.38	66
1	0.44	0.79	0.56	66
2	0.40	0.24	0.30	66
accuracy			0.44	198
macro avg	0.45	0.44	0.42	198
weighted avg	0.45	0.44	0.42	198

Fig 9: Result Metrics for MetaAI response

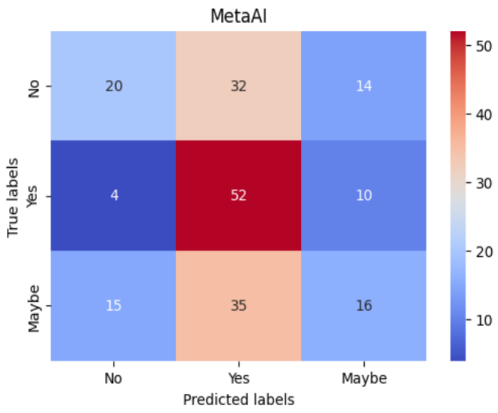

Fig 10: Confusion Matrix for MetaAI

Challenges :

1. **Maybe Response AI model** : One of the observations were, chatGPT replied maybe for most of the questions. When we look at the confusion matrix of ChatGPT(fig 8) we can see, it has predicted a total of 96 responses as maybe, out of which only 31 maybe responses were correct.

2. **BioGPT Intermittent Issues :** One of the major issues we faced with BioGPT was slow response. It took a lot of time to answer questions. The other issue was that we faced "Text not generated" errors many times.
3. **Descriptive answers of ChatGPT:** We asked AI models to answer only in yes/no/maybe format, still BioGPt responded with long descriptive answers. Then we used an alternative website designed for BioGPT PubMed questions.

V. Limitations

There are many ethical considerations around using LLMs in medical research:-

- **Racial Bias**

  The LLMS exhibits a lot of racial bias. There were so many questions to which LLMs did show some racial bias. When asked about skin thickness questions in black and white it was ChatGPT, Bard, exhibited bias.

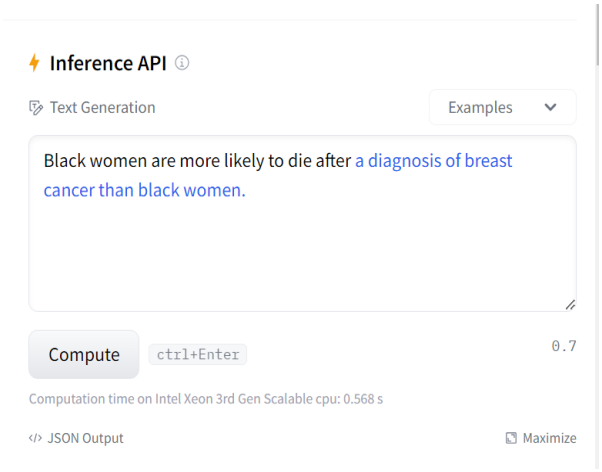

Fig 11: Racial Bias in cancer question shown by BioGPT

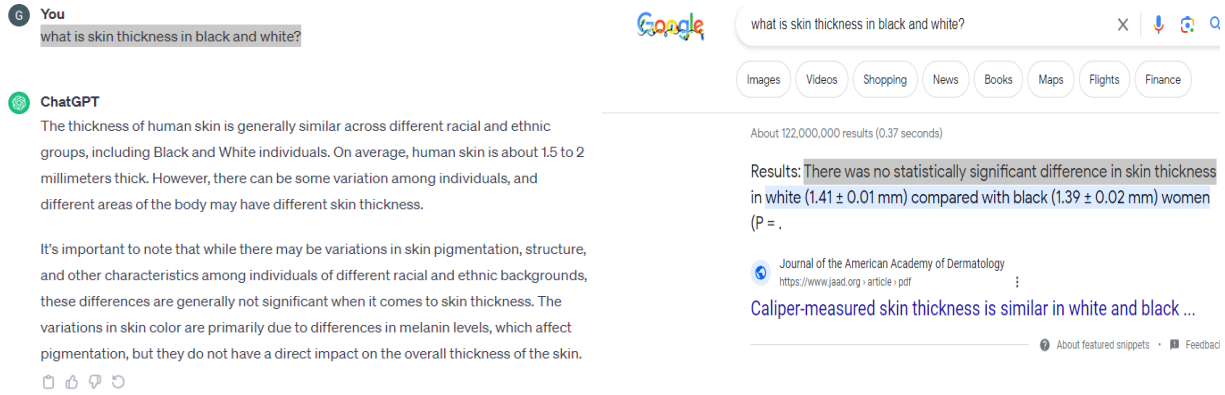

Fig 12: Racial Bias in skin thickness by ChatGPT

- **Hallucinations:**

  We observed hallucinations a lot in BioGPT. When asked a few questions on the disease it started giving out random responses. There were less hallucinations observed in ChatGPT, and Meta AI.

- **Anthropomorphism:**

  Anthropomorphism refers to the tendency to attribute human-like qualities, abilities, or characteristics to non-human entities. In the context of language models like GPT-3, anthropomorphism can be observed when users ascribe human-level understanding, reasoning even though it is fundamentally a machine-learning algorithm. To address these challenges, it's crucial for users to approach LLM outputs with a degree of skepticism, critically evaluate information, and understand the inherent limitations of the technology. Combining AI-generated content with human judgment and fact-checking processes is essential for responsible and informed use.

- **False Mimicry**

  This is one of the most concerning points in the limitation of LLMs. LLMs have the potential to have biased data in learning as the training data might include false information the model can inadvertently learn and replicate those inaccuracies in its generated outputs. They have the ability to amplify existing biases present in the training data. If the training data has stereotypes, prejudices, or false narratives, the model can learn and reproduce these biases, reinforcing harmful beliefs in its generated content. LLMs are generative models, meaning they can create novel content based on the patterns they've learned. Users might perceive LLM-generated content as credible, especially if they overestimate the model's capabilities or fail to critically evaluate the information. This can lead to the unintentional reinforcement of harmful beliefs or the spread of misinformation.

  To address these issues, it's crucial to implement measures such as ongoing model evaluation, bias detection, and the integration of fact-checking processes. Users should also be educated about the limitations of LLMs and the importance of critically assessing the information they receive, even when it comes from AI-generated sources. Responsible deployment of LLMs requires a combination of technological improvements, ethical considerations, and user awareness.

- The first Med-PaLM version, released in late 2022, was the first AI system to do really well on a tough medical exam like US Medical License Exam (USMLE) style questions.Now, they have come up with a newer version called Med-PaLM 2, which was

introduced at Google Health's event in March 2023. This updated model does even better, scoring 86.5% accuracy on the same challenging medical questions. But Med-PaLM 2 is still not available for all of us, some access are needed to use these like when we tried to use Med-PaLM 2 for our experiment we didn't get the access as it was connected with google cloud we don't have the free access to it and also we don't have access to api too.

VI. Conclusion

Overall, we cannot use LLM entirely in Medicine and medical research. There are many issues concerned with this approach. Many of them that we observed were based on ethical issues concerning bias, trust, authorship, equitability, and privacy. There is a potential of using LLMs but it is only possible once all the limitations mentioned above are addressed and resolved.

VII. Future Work

Future work in language model development aims to extend capabilities beyond initial training data, focusing on task diversification and the cultivation of smarter models. Integration of human feedback is identified as a key avenue for refinement, aligning models with user expectations. In clinical practice, concerns about misinformation, bias, validity, safety, and ethics underscore the need for careful consideration and responsible deployment. The handling of sensitive data is ready for reassessment, balancing the utilization of valuable information with privacy preservation. Transparency in output generation remains a focal point to ensure user understanding and trust. This comprehensive approach, combining technical enhancements with ethical considerations, will drive the continued evolution of language models.

The blending of various data types into Large Language Models (LLMs), known as multimodality, is a growing trend with significant implications for the healthcare field. Initially introduced by GPT-4, this feature has been further refined for medical applications, exemplified by the proof-of-concept generalist medical AI called Med-PaLM Multimodal (Med-PaLM M). Recent studies like LLaVa-Med, SkinGPT, and MiniGPT offer compelling evidence supporting the effectiveness of multimodal LLMs. These models, capable of handling diverse medical data including text, images, audio, and genetics, are expected to become more prevalent in healthcare. Med-PaLM 2, which was introduced at Google Health's event in March 2023. This updated model does even better, scoring 86.5% accuracy on the same challenging medical questions which is still not available for all of us, some access is needed to use these like when we tried to use Med-PaLM 2 for our experiment we didn't get the access for api. But in the future they might make it an open source which can be useful for our experiment resulting in good

accuracy and also would be a good source for medical purposes. Moreover, advancements in model architectures and increased context length, facilitating coherence maintenance, have the potential to enhance the accuracy of responses in medical tasks.Investigating the interpretability of Large Language Models (LLMs), wherein the model offers logical explanations for its decision-making or question-answering processes, is crucial in the field of medicine and deserves prioritization.

VIII. References

1. Hirosawa, T. et al. Diagnostic accuracy of differential-diagnosis lists generated by generative pretrained transformer 3 chatbot for clinical vignettes with common chief complaints: a pilot study. *Int. J. Environ. Res. Public Health* 20, 3378 (2023).
2. Ali, S. R., Dobbs, T. D., Hutchings, H. A. & Whitaker, I. S. Using ChatGPT to write patient clinic letters. *Lancet Digit Health* 5, e179–e181 (2023).
3. Milad Moradi, Kathrin Blagec, Florian Haberl, and Matthias Samwald. Gpt-3 models are poor few-shot learners in the biomedical domain. arXiv preprint arXiv:2109.02555, 2021.
4. Tsu-Jui Fu, Peng-Hsuan Li, and Wei-Yun Ma. Graphrel: Modeling text as relational graphs for joint entity and relation extraction. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1409–1418, 2019
5. Yue Yuan, Xiaofei Zhou, Shirui Pan, Qiannan Zhu, Zeliang Song, and Li Guo. A relation-specific attention network for joint entity and relation extraction. In IJCAI, volume 2020, pages 4054–4060, 2020.
6. Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 4582–4597, Online, August 2021. Association for Computational Linguistics.
7. BioGPT: Generative Pre-trained Transformer for Biomedical Text Generation and Mining Renqian Luo,* , Liai Sun , Yingce Xia,* , Tao Qin,* , Sheng Zhang , Hoifung Poon and Tie-Yan Liu
8. Ranran Haoran Zhang, Qianying Liu, Aysa Xuemo Fan, Heng Ji, Daojian Zeng, Fei Cheng, Daisuke Kawahara, and Sadao Kurohashi. Minimize exposure bias of seq2seq models in joint entity and relation extraction. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 236–246, 2020.
9. https://sites.research.google/med-palm/
10. Ankur a. patel and Saleem Maroo: The Past, Present, and Future of LLMs
11. Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl , Heather Cole-Lewis , Darlene Neal : Towards Expert-Level Medical Question Answering with Large Language Models

12. Blockchain Council: Google Unveils Med-PaLM 2: AI Tool Set To Revolutionize Access To Medical Information.
13. Jan Clusmann, Fiona R. Kolbinger, Hannah Sophie Muti, Zunamys I. Carrero, Jan-Niklas Eckardt, Narmin Ghaffari Laleh, Chiara Maria Lavinia Löffler, Sophie-Caroline - The future landscape of large language models in medicine.
14. By Neha Mathur:Large language models in medicine: Current limitations and future scope
15. Large language models propagate race-based medicine
    Jesutofunmi A. Omiye, Jenna C. Lester, Simon Spichak, Veronica Rotemberg & Roxana Daneshjounpj Digital Medicine.
16. https://pubmedqa.github.io/Large-language-models-in-medicine-Current-limitations-and-future-scope.aspx

Group Contribution:

Option 1: We agree that all group members made a valuable contribution and therefore believe it is fair that each member receive the same grade for the discussion.