Whitney Kenner U0777962
CS6017 7/21/23

I read the paper *Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine* by Stefan Harrer. This paper does an excellent job of briefly breaking down how large language models (LLMs) are built, discussing the pertinent risks and dangers present in the current models, and then introduces solutions as well as risk mitigation strategies for current and future development. This article has a focus on LLMs as tools in medicine and healthcare but still covers the broad range of uses of LLMs. 3 major limitations in the content generated by LLMs is layed out. The first is that the model has been trained indiscriminately on content from the internet which includes facts and truths as well as blatant lies and misinformation. This leads to the second limitation which is that the model has no way to assess the truthfulness or accuracy of the content created or relay this to the user. The third is a reliability and reproducibility problem in that it can generate a wildly varied response to the same prompt. The article then covers a number of ethical, technical and cultural approaches to the (re-)design of generative AI models that would increase the safety and efficacy of their use in a clinical setting.

The first two described approaches around accountability and fairness of the companies producing these models. Particularly when it comes to accountability, I see the benefit in companies having a responsibility to their users to make it clear the limitations and use cases of their product. Without an understanding of the underlying functionality of generative AI (which identifies patterns and structures and generates responses based on statistical likelihood of placement, NOT based on any measurement of truthfulness), I believe many people have a fundamental misunderstanding of what it is and how reliable the generated content is. At best, this can result in misinformation being spread and at worst could risk peoples lives in cases of inaccurate information being used as medical advice. As part of the fairness approach suggested in the article, some sort of oversight to reduce model biases is needed (such as an ethics oversight committee). I think this approach would be useful and necessary during the development of more specific and limited models, such as a model suggested in the article for helping physicians and healthcare professionals with documentation. This would require a massive time investment and expertise to plan what information and sources should be included in the training data as well as reviewing the content after the model has been trained. This would become increasingly challenging as the scope of the model is scaled; the number and variety of experts would become too vast to effectively fund or manage. In my opinion, creating smaller, well defined and audited, robustly managed generative AIs with a limited scope of “expertise” would make the management of fairness and accountability easier to establish and lessons learned from these well developed niche models can be used for larger scale models in the future.

Data privacy and selection is discussed as one of the concerns in ethical generative AI models for use in the healthcare sector. Under HIPAA, an individual's medical history is protected, which creates a need to build an AI model that ethically and legally trains on information not protected under HIPAA *and* does not store individual patient information in any meaningful way that risks being compromised. In addition to this complexity, this is such a new field that the legal landscape is not established for how to safely and ethically incorporate AI into protected fields such as healthcare.

Overall, I agreed with the message of this article. I think the growth of LLMs is an exciting new field, but I also think because it is so new there is a lack of oversight, planning, and mitigation processes for the spread of misinformation. I think the healthcare sector could greatly benefit from having a highly trained generative AI for documentation tasks that could enable healthcare providers to have more patient time and less paperwork time. And then in the future, these smaller scale models and the lessons learned in developing them, could be used in broader, more general AI applications.