Article Review

Large language models encode clinical knowledge et al. (2023) by Karan Singhal author of this paper proposed two LLMs, Pathways Language Model (PaLM) and Flan-PaLM, on a variety of medical question answering tasks. They found that both models were able to achieve state-of-the-art performance on these tasks, but that Flan-PaLM outperformed PaLM by a significant margin.

The article presents a new benchmark called MultiMedQA for evaluating the clinical knowledge of large language models (LLMs). MultiMedQA combines six existing open-sourced medical question answering datasets spanning professional medical exams, research, and consumer queries; such as MedicationQA, PubMedQA, MedQA, MedMcQA, MMLU and a new dataset of medical questions searched online named HealthSearchQA a new free-response dataset of medical questions searched online.

The paper also proposes a human evaluation framework for model answers along multiple axes including factuality, comprehension, reasoning, possible harm, and bias.

The Author evaluated Pathways Language Model (PaLM), a 540-billion parameter LLM, and its instruction-tuned variant, Flan-PaLM, on MultiMedQA. Flan-PaLM achieved state-of-the-art accuracy on every MultiMedQA multiple-choice dataset, including MedQA (US Medical Licensing Exam-style questions). However, human evaluation revealed that Flan-PaLM still has key gaps in its clinical knowledge.

Attempts to assess the model clinical knowledge typically rely on automated evaluations on limited benchmarks. To address these gaps, the authors introduced instruction prompt tuning, a parameter-efficient approach for aligning LLMs to new domains. The resulting model, Med-PaLM, performed encouragingly, but remained inferior to clinicians.

The paper proposed a new approach for evaluating the clinical knowledge of large language models (LLMs). They introduce HealthSearchQA, a dataset of 3,173 commonly searched consumer medical questions. They also pilot a framework for physician and lay user evaluation to assess multiple axes of LLM performance beyond accuracy on multiple-choice datasets.

The authors demonstrated state-of-the-art performance on the MedQA, MedMCQA, PubMedQA and MMLU clinical topics datasets using Flan-PaLM, a 540-billion parameter LLM. They also introduce instruction prompt tuning, a simple, data- and parameter-efficient technique for aligning LLMs to the safety-critical medical domain.

The authors make a significant contribution by introducing a new benchmark for evaluating the clinical knowledge of LLMs. The HealthSearchQA dataset is a valuable resource for researchers in this area. The pilot study of physician and lay user evaluation is also a valuable contribution. This study provides insights into the strengths and weaknesses of LLMs for medical applications. The results demonstrate the potential of LLMs for medical applications. However, they also highlight the need for further research in this area. For example, the authors found that LLMs can sometimes generate harmful or biased responses. This is a serious concern that needs to be addressed before LLMs can be used in clinical settings.

Language is a very important part of medicine. Language enables crucial interactions between clinicians Researchers and patients that can often change lives at the same time human AI collaborations in different kind of forms has been emerging as an important aspect for more useful and reliable AI systems and language has been an important medium for that collaboration however in the medical domain especially in terms of the models that are being used in the real world today a lot of models are kind of narrow single task systems like classifiers or regression models and these models aren’t often always fully utilizing language in the way that we hope for.

At the same time, we have all these recent advances and excitement around foundation models and foundation models especially large language models in particular use language as tool for mediating human AI interaction and that interaction can happen through the outputs of the models in terms of text that they are outputting but it can also happen through feedback that humans can give these models as well. These models can also generalize and be repurposed kind of easily across domains and tasks including the medical domain. This paper showed that this model can encode and retrieve from large scale medical data and they could also manipulate knowledge so not just retrieve knowledge but also use that knowledge to actually form conclusions and do multistep of reasoning, produce explanations and also provide estimation of uncertainty which are all seemingly important in the medical domain and also foundation models in general beyond LLMS can also extend to other modalities beyond text offering for future promise as well as gives us to think about what medical AI can look like in the future.

What’s missing from the literature is the First and foremost thing is safety and reliability these large language models are often influencing producing hallucinations or amplifying biases otherwise misbehaving not reasoning properly which is really important in the medical setting. Second thing there should be benchmarks for evaluating whether these models are containing clinical knowledge in a useful manner. Third is about the safety critical and we cannot rely on automated metrics like accuracies or blue scores on final evaluation. So, we need to think about how to get humans to measure progress and mitigate harms and biases.

Large language models (LLMs) have demonstrated impressive capabilities in natural language understanding and generation, but the quality bar for medical and clinical applications is high. In addition, the authors evaluate Pathways Language Model (PaLM) and its instruction-tuned variant, Flan-PaLM, on MultiMedQA. Using a combination of prompting strategies, Flan-PaLM achieves state-of-the-art accuracy on every MultiMedQA multiple-choice dataset (MedQA, MedMCQA, PubMedQA, and MMLU clinical topics), including 67.6% accuracy on MedQA (US Medical Licensing Examination (USMLE)-style questions), surpassing the prior state of the art by more than 17%.

However, human evaluation reveals key gaps in Flan-PaLM's responses. To resolve this the authors have introduced instruction prompt tuning, a parameter-efficient approach for aligning LLMs to new domains using a few exemplars. The resulting model, Med-PaLM, performs encouragingly, but remains inferior to clinicians.

There are some strengths about the article The authors present a comprehensive evaluation of the clinical knowledge of LLMs. They propose a new benchmark, MultiMedQA, that is more accurate than the existing benchmarks. They have also developed a human evaluation framework that can be used to assess LLM answers along multiple medical axes. Their work provides valuable insights into the strengths and weaknesses of LLMs for clinical applications.

One of the strengths of LLMs is that they can be scaled up to be very large, which can improve their performance on a variety of tasks. The authors of the article found that increasing the size of the LLM from 8B to 540B led to a significant improvement in performance on the MedQA dataset.

The LLMs can be fine-tuned to be specific to a particular task or domain. The authors of the article found that using instruction prompt tuning to fine-tune the LLMs for medical question answering led to further improvements in performance.

However, the authors of the article also pointed out some limitations of LLMs for medical question answering. One limitation is that LLMs can sometimes generate inaccurate or harmful answers. The authors of the article found that this was particularly true for consumer medical question answering datasets, where the questions are often more open-ended and challenging.

Another limitation of LLMs is that they can be biased. The article found that LLMs trained on biomedical corpora can sometimes reflect the biases that are present in those corpora. This can be a problem for medical applications, where it is important to provide accurate and unbiased information.

The article suggests that some of the limitations of LLMs for medical question answering can be addressed by using a combination of techniques, such as scaling, fine-tuning, and bias mitigation. They also suggest that more research is needed to develop LLMs that are specifically designed for medical applications.

There are some existing models that are being developed for LLMs and medical domain.

BioGPT: A large language model trained on a biomedical corpus.

PubMedGPT: A large language model trained on the PubMed medical literature.

Galactica: A large language model trained on a combination of biomedical and general-domain corpora.

These models have shown promising results on a variety of medical question answering tasks. However, there is still more research that needs to be done to develop LLMs that are specifically designed for clinical applications.

The study is limited to a single LLM, which is PaLM. It would be useful to evaluate other existing LLMs on MultiMedQA. The human evaluation was conducted by a small number of experts and clinicians. It would be helpful to conduct a larger-scale human evaluation on the model. The authors did not explore the potential risks of using LLMs in clinical applications, such as the risk of bias or the risk of generating harmful content. The study is limited to two LLMs, so it is not clear how generalizable the results are to other LLMs.

The human evaluation was conducted by a small number of experts, so it would be helpful to conduct a larger-scale evaluation. The authors did not explore the potential of using LLMs for other clinical applications, such as diagnosis and treatment planning.

The authors have shown that comprehension, knowledge recall and reasoning improve with model scale and instruction prompt tuning, suggesting the potential utility of LLMs in medicine. However, the authors also emphasize the importance of both evaluation frameworks and method development in creating safe, helpful LLMs for clinical applications. In summary, the paper presents a new benchmark for evaluating the clinical knowledge of LLMs, and shows that these models can achieve state-of-the-art performance on multiple tasks. However, human evaluation reveals that these models still have some limitations, and the authors call for further research in this area.

The paper concluded that their work demonstrates the potential utility of LLMs in medicine, but that more research is needed to develop safe and helpful LLMs for clinical applications.

Here are some ways to improve the use of LLMs for medical question answering: Scale up the models to be even larger. This will improve their performance on a variety of tasks. Fine-tune the models to be specific to a particular task or domain. This will improve their performance on that task. Use bias mitigation techniques to reduce the biases that are present in the models. Develop models that are specifically designed for clinical applications. This will ensure that the models are able to generate accurate and unbiased information that is relevant to clinical practice.

The limitations of LLMS in medical diagnosing are the MultiMedQA benchmark is not exhaustive, LLMs are not yet at the level of clinician experts on many clinically important sectors. The human evaluation framework is subjective and needs to be improved. The use of LLMs can cause harms that contribute to health disparities and create biases development of procedures for the evaluation of bias and fairness-related harms in LLMs is ongoing. Careful consideration will need to be given to the ethical deployment of this technology.