# Article information:

大型语言模型编码临床知识 - PubMed  
<https://pubmed.ncbi.nlm.nih.gov/37438534/>

# Article summary:

1. 本文介绍了一个名为MultiMedQA的基准测试，用于评估大型语言模型（LLMs）在临床应用中的知识。该基准测试结合了六个现有的医学问答数据集，并提出了一个人工评估框架，包括事实性、理解力、推理能力、潜在危害和偏见等多个方面。

2. 文中评估了一种名为PaLM的5400亿参数LLM及其经过指令调整的变体Flan-PaLM在MultiMedQA上的表现。通过采用多种提示策略，Flan-PaLM在每个多项选择数据集上都取得了最先进的准确率，包括在MedQA上达到了67.6%的准确率，超过了之前最先进技术水平17%以上。

3. 然而，人工评估揭示出模型存在关键缺陷。为解决这个问题，作者引入了指令提示调整方法，一种高效对齐LLMs到新领域的方法。结果显示，经过指令调整后的模型Med-PaLM表现良好，但仍然不及临床医生。研究还发现，随着模型规模和指令提示调整的改进，理解力、知识回忆和推理能力得到了提高，表明LLMs在医学领域具有潜在的实用性。人工评估揭示了当前模型的局限性，强调了评估框架和方法开发在创建安全、有益的临床应用LLMs中的重要性。

# Article rating:

Appears moderately imbalanced: The article provides some useful information, but is missing several important points or pieces of evidence that would be required to present the discussed topics in a balanced and reliable way. You are encouraged to seek a more balanced perspective on the presented issues by exploring the provided research topics and looking at different information sources.

# Article analysis:

这篇文章主要介绍了大型语言模型在临床知识编码方面的应用。文章提到了一个新的基准测试集MultiMedQA，该测试集结合了六个现有的医学问题回答数据集，并介绍了一种人工评估框架来评估模型答案的准确性、理解力、推理能力、潜在危害和偏见等多个方面。此外，文章还评估了一种名为PaLM的大型语言模型及其经过调整的变体Flan-PaLM在MultiMedQA上的表现。通过使用不同的提示策略，Flan-PaLM在每个多项选择数据集上都取得了最先进的准确率，包括在MedQA（美国医疗执照考试样式问题）上达到了67.6%的准确率，超过先前最先进技术17%以上。然而，人工评估揭示出关键差距。为了解决这个问题，作者引入了指令提示调整方法，这是一种使用少量示例将LLMs与新领域对齐的参数高效方法。结果显示，经过调整后的模型Med-PaLM表现令人鼓舞，但仍然不如临床医生。作者还展示了随着模型规模和指令提示调整的改进，理解力、知识回忆和推理能力也得到了提高，这表明LLMs在医学领域具有潜在的实用性。人工评估揭示了当前模型的局限性，强调了评估框架和方法开发在创建安全、有益的LLMs用于临床应用中的重要性。

从文章内容来看，它提供了一些关于大型语言模型在临床知识编码方面的有趣发现，并介绍了一种新的基准测试集和人工评估框架。然而，在进行批判性分析时，我们可以注意到以下几点：

1. 潜在偏见及其来源：文章没有明确提及作者或研究团队可能存在的潜在偏见或利益冲突。这可能导致读者对研究结果产生怀疑，并质疑其客观性和可靠性。

2. 片面报道：文章主要关注大型语言模型在临床知识编码方面的优势和潜力，但未充分探讨其局限性和风险。这种片面报道可能会给读者带来误导，并忽略了其他可能存在的问题。

3. 无根据的主张：文章声称通过调整模型的指令提示可以提高其在临床知识编码方面的表现，但没有提供足够的证据来支持这一主张。缺乏实验证据可能使读者对该方法的有效性产生怀疑。

4. 缺失的考虑点：文章没有充分讨论大型语言模型在临床应用中可能面临的伦理、隐私和安全等问题。这些是使用这些模型时需要认真考虑的重要因素，但文章未对其进行深入探讨。

5. 所提出主张的缺失证据：尽管文章声称大型语言模型在临床知识编码方面具有潜在实用性，但并未提供足够的证据来支持这一主张。缺乏实际应用案例或定量数据可能使读者对该主张产生怀疑。

6. 未探索的反驳：文章没有探讨大型语言模型在临床知识编码方面可能存在的挑战或批评观点。这种未探索反驳可能导致读者对研究结果形成不完整或片面的理解。

7. 宣传内容和偏袒：文章过于强调大型语言模型在临床知识编码方面的优势，可能存在宣传内容和偏袒的倾向。这可能会导致读者对研究结果的客观性产生怀疑，并质疑其真实性。

8. 是否注意到可能的风险：文章没有充分讨论大型语言模型在临床应用中可能带来的潜在风险。这种缺乏对潜在风险的关注可能使读者对该技术的可行性和安全性产生担忧。

9. 没有平等地呈现双方：文章主要关注大型语言模型的优势和潜力，但未平等地呈现其他观点或方法。这种不平等可能导致读者对整个领域的理解产生偏差。

综上所述，尽管这篇文章提供了一些关于大型语言模型在临床知识编码方面的有趣发现，但在进行批判性分析时需要注意其中存在的潜在偏见、片面报道、无根据的主张、缺失的考虑点、所提出主张的缺失证据、未探索的反驳、宣传内容和偏袒等问题。为了更全面客观地评估大型语言模型在临床应用中的潜力和限制，需要进一步研究和讨论。

# Topics for further research:

* Potential biases and conflicts of interest
* One-sided reporting
* Lack of evidence for the claim about adjusting model prompts
* Failure to discuss ethical
* privacy
* and security considerations
* Lack of evidence for the potential practicality of large language models in clinical knowledge encoding
* Failure to explore challenges or criticisms of large language models in clinical knowledge encoding
* Promotion and bias in the article
* Failure to address potential risks
* Lack of equal presentation of opposing viewpoints
  By using these key phrases in a Google search
* users can find additional information and perspectives on the topics not covered in the article
* starting from point

# Report location:

<https://www.fullpicture.app/item/3df862240bcadf0696efe05afded4169>