先进的人工智能模型居然不会“ 问诊 ”

作者：writer 来源：科学网

本文原地址：https://www.iikx.com/news/progress/31177.html

先进的人工智能模型居然不会“ 问诊 ”。一项新研究发现，虽然先进的人工智能（AI）模型在专业医学考试中得分很高，但在医生最重要的任务之一——与患者交谈以收集相关医疗信息并提供准确诊断方面，仍然表现不佳。1月2日，相关研究成果发表于《自然-医学》。

图片来源：Just_Super/Getty Images

?

美国哈佛大学的Pranav Rajpurkar说：虽然大型语言模型在多项选择测试中的表现令人印象深刻，但在动态对话中，它们的准确性明显下降，特别是难以进行开放式诊断推理。

当研究人员开发出一种基于模拟医患对话评估临床AI模型推理能力的方法时，这一点变得很明显

。这些患者基于2000个医疗案例，主要来自美国医学委员会的专业考试。

同样来自哈佛大学的Shreya Johri说：模拟患者互动可以评估病史采集技能，这是临床实践的一个关键组成部分。她表示，新的评估基准被称为CRAFT-MD，也反映了现实生活中的情况，即患者可能不知道哪些细节是至关重要的，只有在特定问题提示时才会披露重要信息。

CRAFT-MD基准本身依赖于AI。美国OpenAI公司的GPT-4模型在与正在测试的临床AI的对话中扮演了患者AI的角色。GPT-4还通过将临床AI的诊断与每个病例的正确答案进行比较，帮助对结果进行评分。人类医学专家仔细检查了这些评估。他们还审查了对话，以检查患者AI的准确性，并查看临床AI是否成功收集了相关的医疗信息。

多项实验表明，4种领先的大型语言模型——OpenAI的GPT-3.5和GPT-4模型、美国Meta公司的Llama-2-7b模型和法国Mistral AI公司的Mistral-v2-7b模型，在基于对话的基准测试中的表现比基于书面病例总结进行诊断时差得多。3家公司没有回应置评请求。

例如，当提供结构化的病例摘要并允许从多项选择答案列表中选择诊断时，GPT-4模型的诊断准确性达到了令人印象深刻的82%，而当没有多项选择选项时，其诊断准确率降至49%以下。然而，当它不得不通过模拟的患者对话进行诊断时，准确率降至26%。

在这项研究中，GPT-4模型的表现在测试中是最好的，GPT-3.5模型通常次之，Mistral-v2-7b模型排在第二位或第三位，Llama-2-7b模型通常得分最低。

AI模型在很大程度上也未能收集完整的病史，比如GPT-4模型仅在71%的模拟患者对话中做到了这一点。即使AI模型确实收集了患者的相关病史，它们也并不总是能作出正确的诊断。

美国斯克利普斯研究转化研究所的Eric Topol表示，这种模拟患者对话的方式代表了一种比医学检查更有用的评估AI临床推理能力的方法。

Rajpurkar说，即使一个AI模型最终通过了这一基准，能够根据模拟的患者对话持续作出准确诊断，也并不一定意味着它优于人类医生。他指出，现实世界中的医疗实践比模拟中的更混乱。它涉及管理多名患者、与医疗团队协调、进行身体检查，以及了解当地医疗情况中复杂的社会和系统因素。AI可能是支持临床工作的强大工具，但不一定能取代经验丰富的医生的整体判断。Rajpurkar说。（来源：中国科学报 文乐乐）

相关论文信息：https://doi.org/10.1038/s41591-024-03328-5

作者：Pranav Rajpurkar 来源：《自然—医学》