谷歌评估AI“问诊”能力

作者：writer 来源：科学网

本文原地址：https://www.iikx.com/news/progress/23681.html

谷歌评估AI“问诊”能力。

谷歌研究和深度思维公司的研究者报道了一个用于评估大型自然语言模型(LLM)能多好地回答医学问题的基准，介绍了一个专精医学领域的大型自然语言模型Med-PaLM。研究表明，在将其临床应用可行之前，还有许多限制要克服。相关研究近日发表于《自然》。

人工智能(AI)模型在医学领域有许多潜力，包括知识检索和支持临床决策。但现有的模型尚不完善，例如可能会编造令人信服的医疗错误信息，或纳入偏见加剧健康不平等。因此需要对其临床知识进行评估。然而这通常依赖有限基准的自动化评估，例如个别医疗测试得分，这可能无法转化为真实世界的可靠性或价值。

为评估LLM编码临床知识的能力，谷歌研究公司的Shekoofeh Azizi和同事探讨了它们回答医学问题的能力。他们提出了一个基准，称为MultiMedQA。它结合了6个涵盖专业医疗、研究和消费者查询的现有问题回答数据集以及HealthSearchQA——这是一个新的数据集，包含3173个在线搜索的医学问题。

研究者随后评估了PaLM(5400亿参数的LLM)及其变体Flan-PaLM。他们发现，在一些数据集中，Flan-PaLM达到了最先进水平的表现。在整合美国医师执照考试类问题的MedQA数据集中，Flan-PaLM超过此前最先进的LLM达17%。不过，虽然FLAN-
PaLM的多选题成绩优良，进一步评估显示，它在回答消费者的医疗问题方面存在差距。

为解决这一问题，研究者使用一种称为设计指令微调的方式进一步调试Flan-PaLM适应医学领域。设计指令微调是让通用LLM适用新的专业领域的一种有效方法。结果产生的模型Med-PaLM在试行评估中表现令人鼓舞。例如，Flan-PaLM被一组医师评分与科学共识一致程度仅61.9%的长回答，Med-PaLM的回答评分为92.6%，相当于医师做出的回答(92.9%)。同样地，Flan-PaLM有29.7%的回答被评为可能导致有害结果，Med-PaLM仅5.8%，相当于医师所作回答(6.5%)。

这一结果表明AI问诊虽然很有前景，但有必要做进一步评估。(来源：中国科学报 冯维维)

相关论文信息：https://doi.org/10.1038/s41586-023-06291-2

作者：Karan Singhal 来源：《自然》