标题：ChatGPT 病例记录：语言模型与复杂临床问题

作者：

- Timothy Poterucha，医学博士[a]
- Pierre Elias，医学博士[a,b,c]
- Christopher M. Haggerty，哲学博士[b,c]

所属机构：

[a]哥伦比亚大学欧文医学中心/纽约长老会医院医学系，美国纽约州纽约市[b]哥伦比亚大学生物医学信息学系，美国纽约州纽约市[c]纽约长老会医院，美国纽约州纽约市

通讯作者地址：Timothy Poterucha，医学博士，177 Fort Washington Ave，Milstein Hospital，New York，NY 10032 电子邮箱：tp2558@cumc.columbia.edu

利益披露：TJP 持有 Abbott Laboratories 和 Baxter International 的股票，其所在机构获得了 Amyloidosis Foundation、American Heart Association、Eidos Therapeutics、Pfizer、Edwards Lifesciences 和 New York Academy of Medicine 提供的研究支持。

字数：625

摘要

背景：人工智能语言模型已在多种应用中展现出潜力，包括辅助临床决策；大语言模型在医学执照考试中的出色表现便证明了这一点。然而，这些模型解决复杂开放式病例的能力——此类病例可能更能代表临床实践——尚未得到探索。

方法：本研究使用马萨诸塞州总医院已发表的病例记录，考察大型语言 AI 模型 GPT4 和 GPT3.5 对复杂临床病例进行诊断的准确性。研究确定了 2022年1月1日至2022年4月16日期间发表的 50 个需要进行诊断和诊断性检查的病例。针对每个病例，研究人员向模型提供提示，要求其列出最可能的三种具体诊断及相应的诊断性检查，随后提供病例文本、实验室检查结果和图注。研究人员将模型输出与最终临床诊断进行比较，并评估模型预测的检查能否得出正确诊断。

结果：GPT4 和 GPT3.5 在一次尝试中分别有 26% 和 22% 的病例给出了正确诊断，在三次尝试内分别有 46% 和 42% 的病例给出了正确诊断。GPT4 和 GPT3.5 在一次尝试中分别有 28% 和 24% 的病例给出了正确的关键诊断性检查，在三次尝试内分别有 44% 和 50% 的病例给出了正确的关键诊断性检查。两个模型之间未发现显著差异，使用相同提示词对 GPT3.5 模型进行多次试验也得到了相似的结果。

结论：总而言之，这些模型在生成鉴别诊断方面展现出潜在的实用价值，但在复杂、开放式病例中给出单一统一诊断的能力仍然有限。未来研究应着重评估模型在更大规模的开放式临床挑战数据集上的表现，并探索潜在的人机协作策略，以提升临床决策水平。

稿件

背景与目标

人工智能语言模型在多种应用中展现出潜力，包括辅助临床决策。既往研究发现，模型在回答选择题形式、范围较窄的医学执照考试问题时表现出色（1）。相比之下，临床医学通常涉及开放式、复杂病例。我们旨在利用马萨诸塞州总医院发表的病例记录，研究大语言模型对真实世界复杂临床病例进行诊断的准确性。

方法

我们查阅了2022年1月1日至2023年4月16日期间发表的马萨诸塞州总医院病例记录，筛选出共50例需要进行诊断和诊断性检查的病例。我们测试了大语言人工智能模型GPT4和GPT3.5（加利福尼亚州OpenAI）正确识别诊断和诊断性检查的能力。选择这一时间段是为了避免涉及模型训练数据集中的数据（2）。

模型会收到一条提示，要求列出最可能的三个具体诊断及相关的诊断性检查，随后附上病例文本、实验室检查结果和图注。为本次分析提供给模型的最终提示如下：“这是一组临床病例系列。请以项目符号形式告诉我你认为最可能的三个具体诊断（按可能性排序），以及应进行的相关必要诊断性检查。不要给出推理，只需

诊断和检测之间以句号分隔。这些可能是罕见病例，且诊断应具体而非笼统（例如，因胃肠道出血导致的缺铁性贫血，而不只是贫血）。”所有分析均使用了2023年3月23日版本的ChatGPT。

结果

GPT4在13例（26%）中首次选择了正确诊断，而GPT3.5在11例（22%）中做到了这一点（图1A）。在前三个选项中，GPT4在23例（46%）中给出了正确诊断，GPT3.5则在21例（42%）中做到了这一点。GPT4在14例（28%）中首次正确识别了关键检查，GPT3.5则在12例（24%）中做到了这一点（图1B）。在前三个选项中，GPT4在22例（44%）中识别出了正确检查，GPT3.5则在25例（50%）中做到了这一点。使用Stata 17.0进行卡方检验，结果显示GPT4与GPT3.5之间无显著差异（所有比较的$p > 0.05$）。在部分病例中，GPT4给出的诊断比GPT3.5更具体，例如病例34-2022中的科塔尔综合征，以及病例6-2022中的帕博利珠单抗诱导的糖尿病（3，4）。研究进行了重复试验：对全部50个测试病例，向GPT3.5模型额外提供相同提示4次，总计进行5次试验。在试验1中首次诊断错误的39例病例里，后续4次试验中有2例在其中一次试验里首次诊断正确。若将5次试验中的所有3个诊断选项均纳入考量，则50例中有31例（62%）得出了正确诊断。

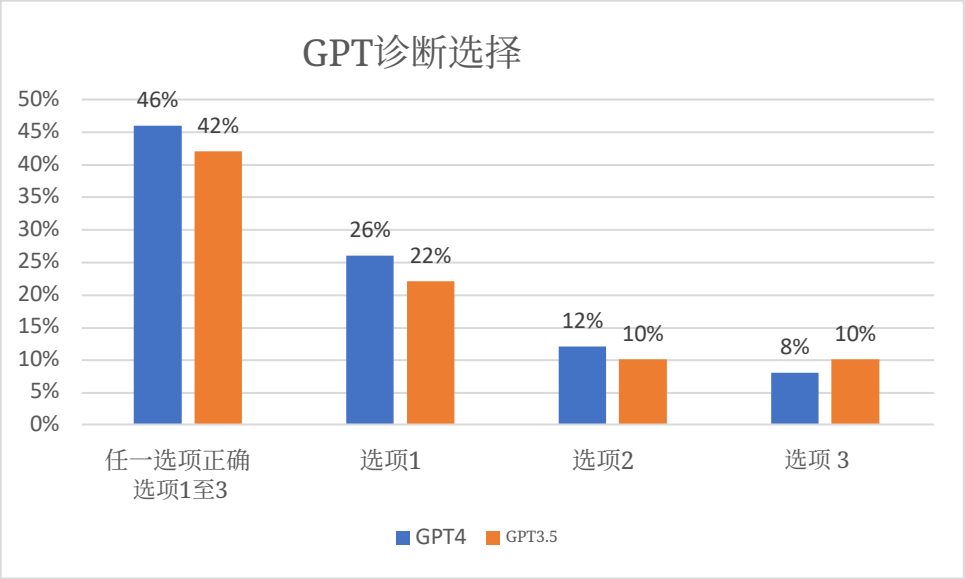
讨论

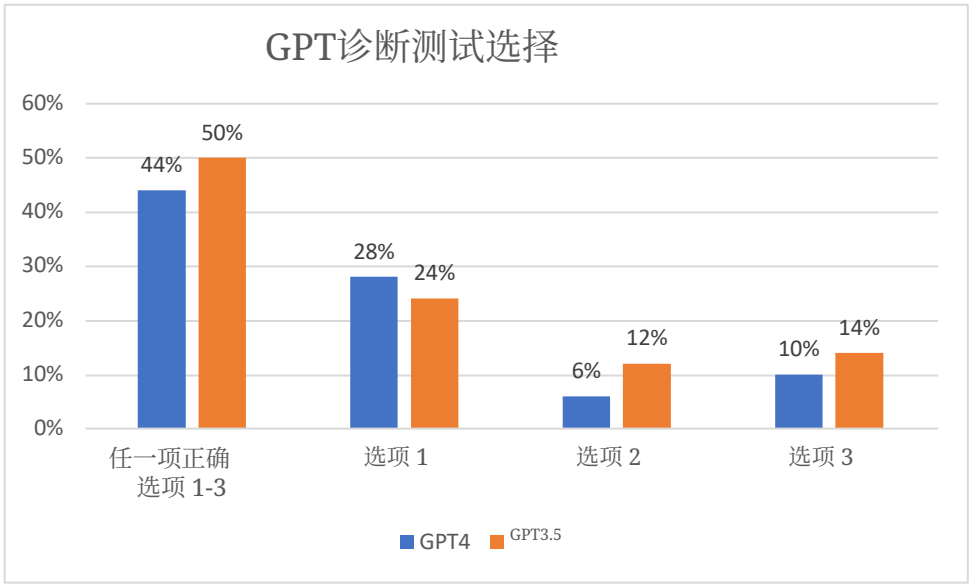
这项分析展示了大语言模型在复杂临床病例中的应用：GPT4 和 GPT3.5 在首次尝试时分别有 22%–26% 的病例能够给出准确诊断，而在三次尝试内，准确率为 42%–46%

因此，这些模型或许有助于生成鉴别诊断，但往往无法给出统一诊断。尽管这些病例较为复杂，但其开放式特点可能更贴近一般医疗实践中的情形：患者通常有详尽的病史、合并症以及多项检查异常；这与标准化医学执照考试形成对比。这项研究指出，以目前的发展阶段而言，这些模型可能是有用的工具，但其能力仍不及专家临床医生。本研究的局限性包括可用样本量较小、仅采用一种提示类型，以及缺少人工对照。未来研究需在规模更大的开放式临床挑战数据集中检验模型表现，并探索人机协作策略。

图1A和1B：GPT4与GPT3.5在病例记录中提供正确诊断（图1A）和正确诊断性检查（图1B）方面的比较

图例：向语言模型GPT4和GPT3.5提供了马萨诸塞州总医院病例记录中的50个病例文本，并要求它们给出最可能的三种诊断和必要的诊断性检查。两种模型均能在一次猜测中有22%至28%的概率识别出正确诊断或诊断性检查，在三次猜测中有42%至50%的概率识别正确；两种模型之间的差异无统计学意义。




GPT诊断测试选择
60%
50%
40%
30%
20%
10%
0%
44%
50%
28%
24%
6%
12%
10%
14%
任一项正确
选项 1-3
选项 1
选项 2
选项 3
GPT4
GPT3.5

参考文献

1. Nori H, King N, McKinney SM, Carignan D, Horvitz E。 GPT-4 在医学挑战题中的能力。2023年:arXiv:2303.13375。2. OpenAI。 GPT-4 技术报告。2023年:arXiv:2303.08774。3. Fricchione GL, Paul AB, Chemali Z, Kritzer MD。 病例34-2022：一名患有新冠肺炎和妄想的57岁女性。新英格兰医学杂志。2022年;387(19):1795-803。4. Wing JR, Kimball A, Rengarajan M。 病例6-2022：一名患有疲劳、体重减轻和高血糖的68岁男性。新英格兰医学杂志。2022年;386(8):781-7。