PsychBench：评估大型语言模型心理健康模拟的流行病学准确性

来源: arXiv
日期: 2026-04-19
DOI: -
链接: https://arxiv.org/abs/2604.17359v1

【中文标题】
PsychBench：评估大型语言模型心理健康模拟的流行病学准确性

【研究背景】
随着大型语言模型（LLM）在临床培训、研究和心理健康工具中的应用日益增多，但其人群水平的有效性尚未得到充分验证。

【研究方法】
本研究引入了PsychBench，这是首个对LLM患者模拟进行流行病学审计的工具。研究人员评估了来自四个前沿模型（GPT-4o-mi
Flash、GLM-4.7）的28,800个患者档案，这些档案与NHANES和NESARC-III基准在120个交叉群体中进行比较。

【主要发现】
研究发现，模型在产生临床合理的个体方面表现出色，但在代表其来源人群方面存在偏差。模型的方差压缩从14%（GLM-4.7）
尾部。尽管重测相关系数高于r = 0.90，但36.66%的案例在运行之间跨越了诊断阈值。症状相关矩阵在人口统计学群体中存在差异
偏差具有系统性和不对称性。对于大多数群体，模型高估了抑郁严重程度3.6到6.1点（Cohen d = 1.13到1.91），这与在具有高基础
性，方向相反：模型仅捕捉到8%到46%的记录少数群体压力增加，产生了-5.42的残留值（d = -1.55）。模型还将易怒归咎于黑人
种族化和性别化的假设。这些模式在美国和中国架构中重复出现，表明失败与当前的训练范式有关，而不仅仅是孤立的实施。

【临床意义】
对于大多数用户，LLM心理健康工具存在将普通压力病理化的风险；对于跨性别用户，算法抹去了真正的需求。患者看起来是对