

· 医学教育 ·

浅谈基于大型语言模型的聊天机器人对医学教育的影响

潘柔百, 陶 蓉

(上海交通大学医学院附属瑞金医院心血管内科, 上海 200025)

关键词: 人工智能; 医学教育; 自然语言处理; 聊天机器人

中图分类号: R-05 文献标志码: C 文章编号: 1673-6087(2023)06-0439-08

DOI: 10.16138/j.1673-6087.2023.06.012

大型语言模型人工智能 (artificial intelligence, AI) 技术为医学发展提供了新机遇, 被应用于健康管理系统的建设, 辅助临床决策、临床研究等领域^[1]。ChatGPT(chat generative pre-trained transformer) 是由 AI 研究实验室 OpenAI 研发的基于大型语言模型 (large language model) 的聊天机器人 (AI chatbot)。聊天机器人是 AI 的一种应用形式, 其利用自然语言处理 (natural language processing) 技术, 根据使用者输入的语言信息, 以会话的形式向使用者输出回答或解决方案。自 2022 年 11 月发布以来, ChatGPT 在医学领域的作用及影响得到了广泛关注及探讨。一些研究者使用美国医师执照考试 (United States Medical Licensing Examination, USMLE) 中的问题对 ChatGPT 进行测试, 即使未经过专门医学知识数据库的训练, ChatGPT 的准确性仍能达到 60% 以上或将近 60%^[2-3]。另一些研究者利用 ChatGPT 处理文本的能力, 使用其辅助科学论文写作。ChatGPT 生成的文本提供了有用的信息且具备一定逻辑连贯性, 但不能提供准确的科学数据, 并且引用部分虚假文献^[4]。

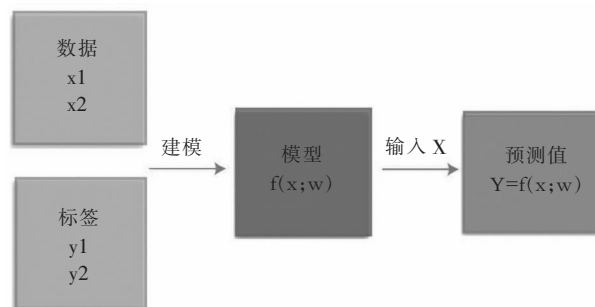
ChatGPT 将会话式交流与医学知识整合起来, 可饰演师长、同学、患者等不同角色, 为学生提供一种信息交互式体验。ChatGPT 拥有出色的文本处理、演绎推理、释义、共情等能力, 这些能力都将有益于教学活动的开展, 促进学生掌握知识。然而, 当前对于该项技术在医学教育领域上的应用仍缺乏成熟范式。人们担忧这项技术将带来信息安全、诚信、伦理等各方面的影响, 影响其发挥育人的作用。本文旨在综述探讨 ChatGPT 以及与其相似的基于大型语言模型聊天机器人在医学教育中的用途, 阐明该类 AI 技术应用的弊端以及提出相应的解决思路。

聊天机器人的原理及发展

一、自然语言处理的原理、医学用途

自然语言处理旨在实现计算机理解、分析、生成人类语言等功能, 是 AI 的一个分支, 其通过运用语言模型达成该目标。语言模型是一串语言序列的概率分布。通过一定的算

法, 对给定的语言数据集内部特征规律或数据与标签间的关系进行构建, 即语言模型的建模过程。语言模型的功能是处理输入变量, 将其映射成预测值并输出模型。从数据集中总结的用于预测所需要的规则、数据和任何其他特定于算法的数据结构称为参数。语言模型的构建与功能可通过图 1 表示。



对于给定的包含数据 x 和标签 y 数据集 (x, y) , 通过算法构建出包含参数 w 的函数 $f(x;w)$ (模型), (使用者) 向模型输入变量 X , 在模型中参数 w 影响下生成 X 对应的预测值 $Y=f(X; w)$ 。

图 1 语言模型的构建与功能

自然语言处理通过语言模型将文本、语音等非结构化的语言信息转化成计算机可理解的文本数据, 再进一步转变为使用者所需的内容, 完成特定任务。医学临床实践、教学、研究等工作的深入伴随着描述性文本不断产生, 均可成为自然语言处理的数据来源。Wang 等^[5]在综述中提到, 自 2012 年以来, 自然语言处理在医学领域取得了快速发展, 体现于其在电子健康档案与电子病历中信息提取、文本分类、信息检索等方面的应用, 协助精神病学、肿瘤学、放射学等学科领域研究的开展。在医学教育领域中, 基于对非结构化的信息提取及处理能力, 自然语言处理被应用于医学测试评分、住院医师表现评估、构建临床案例库等方面。在测试评估中, 自然语言处理可高效处理大量体现被试者能力的描述性文本, 并发挥其客观性优势, 克服传统评估方法中人为因素相关的弊端^[6-7]。同时, 也可以即时生成反馈, 有助于被试者针对性提高自身能力^[7]。在构建临床案例库方面, Zack 等^[8]根据病例中症状、诊断等文本信息, 运用自然语言处理对 1965 年至 2020 年《新英格兰医学杂志》临床病理学会议发表的 2 525 例病例报告进行分类, 并进一步探究专家对病

基金项目: 国家自然科学基金项目 (项目编号: 82270383、82000368)

通信作者: 陶 蓉 E-mail: rongtao@hotmail.com

例鉴别诊断的讨论,构建展现病例间鉴别诊断关系的临床病例库,为理解疾病间的相互关联及区别提供了新的视角,同时也为医学生提高疾病诊断及鉴别诊断能力的临床能力提供素材。

然而,自然语言处理在医学领域应用中体现出一定的局限性,限制因素是语言模型的表达能力,即模型对不同类型数据之间关系的构建能力。语言模型的表达能力取决于模型内部结构、模型参数量以及训练模型的数据库容量。早期用于医学领域研究的模型主要是基于规则或统计的模型,这类模型自身结构相对简单,参数有限,且常局限于提供信息的数据库。这些限制使得部分模型在医学领域中只能被应用于特定的情景,缺乏外部效度^[9]。例如,使用肿瘤监测系统(Cancer Surveillance System)医疗数据构建的、用于确认肿瘤突变基因的自然语言处理模型,在本系统的病例中显示出较高预测能力,然而该能力在另一个系统的病例中却无法被印证^[10]。

二、大型语言模型聊天机器人的开发

聊天机器人基于人工神经网络模型,相较于早期模型,其表达能力有所提升。人工神经网络的结构与生物神经网络相似,由多层次的人工神经元相互连结而成,每个人工神经元相当于一个简单的方程式,将上一层神经元的输出加权平均后向下一层神经元输出^[11]。拥有上述结构特征后,人工神经网络可构建出词与词之间的关系。模型结构上的改进有助于语言模型理解更深层次的问题,拓宽了自然语言处理在医学领域的作用,如应用于影像、病理报告分析^[12]。

注意力机制——一种用于构建不同层次间神经元关系的机制,引入语言模型构建过程,是语言模型构建中的一项重大突破^[13]。Transformer模型是研究者运用注意力机制,在人工神经网络基础上构建出的语言模型框架。运用海量数据对语言模型进一步训练,使得语言模型能掌握数据间的复杂模式,构建出大型语言模型。GPT是以Transformer模型作为基础框架的大型语言模型^[14]。训练GPT的数据主要来源于网站、博客、在线论坛、社交媒体平台,同时也接受书籍、学术论文等公开互联网资源,数据量达到万亿级别词语量,该模型拥有千亿以上参数^[2]。经过海量数据训练之后,大型语言模型具有上下文学习(in-context learning)能力^[15],即在给予语言模型一定提示,通过选择性地添加一些任务样例之后,语言模型就可以通过预测下一词的形式生成答案,这是语言模型能够与人类之间“沟通”的基础。

语言模型的内在目标是根据上下文预测可能性最大的词汇,但人类应用语言模型的目的是完成某种形式具有价值的认知工作,因此语言模型输出内容并不总是与人类的

期望一致。为了解决这一问题,研究者运用人类反馈指导语言模型的学习过程,即基于人类反馈的强化学习(reinforcement learning from human feedback)^[16]。研究者先给指定的提示编写预期输出内容,对模型进行调优(fine-tune),生成监督微调(supervised fine-tuning, SFT)模型,再使用提示去测试数个不同SFT模型,根据模型的回答进行排名,训练出可以判断模型价值的奖励模型。最后利用奖励模型,使用强化学习的方式对语言模型进一步优化。通过上述训练策略,语言模型的输出在一定程度上与人类意图维持一致。

聊天机器人在医学教育中的用途

一、成为个性化学习助手

基于大型语言模型的聊天机器人训练过程中的数据中包含了公开的医学相关信息^[2]。聊天机器人能够根据使用者指定的话题生成文本,在一定的提示下以满足使用者需求的形式输出。这些特征使得聊天机器人成为易于获得的学习助手,随时随地为学生解疑答惑。Kung等^[3]使用USMILE测试ChatGPT的研究中,ChatGPT输出的答案以及解释具有高度内在一致性,意味着其具备成为学习助手的基本条件。同时,ChatGPT在USMILE中的合格表现证明其掌握了一定程度的医学基础知识,这是ChatGPT应用于医学教育的前提。

随着模型不断更新,聊天机器人拥有了联想能力。在作出回答之后,聊天机器人会根据学习者输入的文本及其回答,生成新的相关提问,通过“头脑风暴”的形式让学习者发散思维,深入探索,激发学习兴趣。笔者使用微软公司推出的基于GPT4模型的新必应(New Bing)模拟了该过程,生成对话如表1所示。

上述文本可见,在回答药物不良反应之后,新必应生成了关于药物替代品、安全剂量及特殊人群3个扩展问题,这启发学习者去关注与药物不良反应相关的话题,进一步完善知识框架的构建。

根据Kaur等^[17]在华威医学院开展的一项调查,医学生和学术委员会成员普遍认为聊天机器人是一种高效的学习资源的访问手段。Al kahf等^[18]开展了一项以聊天机器人为基础设计的教学游戏探究巴黎笛卡尔大学医学生成绩影响的随机对照试验。相较于对照组,拥有游戏权限的试验组学生在期末考试中的成绩更好,并且对该形式游戏具有较高满意度。这些证据均说明聊天机器人具有为医学生提供良好的学习体验,且提高学生考试表现的能力。

二、模拟临床实践

聊天机器人通过生成对话的形式实现与使用者之间的

表1 运用聊天机器人进行问答

提示	新必应的回答
What is the side effect of aspirin? (阿司匹林的不良反应是什么?)	Aspirin is a medication that can be used to treat pain, fever, inflammation, and blood clots. However, it can also cause some side effects... (阿司匹林是一种可用于治疗疼痛、发烧、炎症和血栓的药物。但是,也会引起一些不良反应.....) What are some alternatives to aspirin? (阿司匹林有哪些替代品?) How much aspirin is safe to take? (服用多少剂量的阿司匹林是安全的?) Can I take aspirin if I'm pregnant or breastfeeding? (如果我怀孕或正在哺乳期,可以服用阿司匹林吗?)

互动,而这种形式与临床实践中医师与患者之间采集病史、沟通病情等交流过程之间具有一定相似性,因此,可以“饰演”虚拟患者,在设定的模拟场景下与使用者进行沟通。笔者使用新必应模拟了 1 例虚拟患者就诊的情景,生成对话如表 2 所示。

从上述对话可见,聊天机器人在扮演虚拟患者过程中生动地描述了症状特征,且重点突出、简明扼要,加深学生对疾病典型症状的印象。相较于传统标准化患者问诊的教学方式,使用聊天机器人模拟问诊需要调动的资源更少,可让学生反复练习,强化临床基础技能训练,掌握在不同情境下问诊要点,同时可以锻炼学生与患者之间的沟通技巧,培养共情能力^[9]。笔者继续向新必应提问当前情况下应考虑诊断及治疗策略,生成的文本如表 3 所示。

当医患之间角色发生反转时,新必应能给出详细的诊断依据、鉴别诊断以及治疗策略。根据症状特征,语言模型将“急性冠脉综合征”作为最有可能的主要诊断,并提出了包括肺栓塞、主动脉夹层等常需与急性冠脉综合征进行鉴别的疾病;在治疗方面,其回答涵盖了急性冠脉综合征治疗中的再灌注治疗、药物治疗以及生活方式改善等完善的治疗策略。这些内容与教材、临床实践之间具备一致性,可巩固学生从教材中获取的知识。在生成内容的同时,语言模型还提供了支持临床决策的文献链接,可以访问链接对临床决策追根溯源,了解决策产生的过程,培养科学思维。

Strong 等^[20]对比了 GPT-4 与医学生在开放式临床逻辑推理测试中的表现,发现在面对复杂临床问题时,GPT-4 的表现甚至超过了第一、第二学年的医学生。该研究体现了聊天机器人在临床模拟环境中出色的病史总结能力、推理能力以及鉴别诊断能力,为学生提供优秀的学习范式。

三、提供评价手段及反馈

在传统教学模式中,对医学知识的评价通常依靠于标准化的口头或书面测试,学生很难从日常练习过程中得到相应的评价与反馈。准确的评价和及时的反馈能让学生了解当前自己对知识掌握情况,针对薄弱环节制定后续学习方

案,查缺补漏。聊天机器人可根据某一特定主题,生成用于评价学生的测试题,并且能够根据学生的选择或回答对学生的表现进行量化,向学生提供及时反馈以及学习上的建议。笔者尝试使用新必应生成测试题并要求新必应根据作答产生反馈,产生文本如表 4 所示。

由上述对话可知,新必应可以根据指定的话题生成测试题,包括选择题和主观题。在没有具体提示的情况下,其生成了一道测试基础心脏生理学知识的选择题。当引入具体提示后,根据提示生成了 1 道概念解析题。使用者回答后,新必应准确地辨识出回答中混淆概念的错误,阐述修正后的概念,并且能给使用者的主观回答评分。聊天机器人饰演了 1 位耐心的“老师”,使用者可以反复提交回答以深化对某一话题的理解,锻炼文字表达能力,做到更准确、规范地回答问题。新必应能根据回答,大致对学生知识掌握情况进行评估,指出学生的缺点与不足,给出后续学习上的建议。

以问题为基础的学习模式(problem-based learning)是一种新型教学形式,学生通过结成合作小组,共同解决临床实际案例。Reese 等^[21]的荟萃分析显示,同伴互教的教学成果与教师教学的成果相似。这种形式的优势是学生可以互相测试彼此想法并得到来自同伴的反馈。聊天机器人通过对话形式模拟这一讨论过程,在学生独立学习过程中提供类似的益处,因此其有望成为小组教育模式的辅助教学手段。

聊天机器人在医学教育中的局限性和解决思路

一、聊天机器人的误用

用于训练聊天机器人的数据主要来源于公开网页信息构建的数据库和英文维基百科,同时在模型训练过程中,这些数据需要人工进行标注。因此,数据本身的偏见性以及人为主观因素的影响使得聊天机器人的信息准确性受到了一定的质疑。然而,根据 Kung 等^[22]的研究,PubMedGPT——一类专门使用生物医学领域文献训练的聊天机器人,在美国

表 2 运用聊天机器人进行虚拟患者问诊示例

提示	新必应的回答
I'm a medical student. Can you act as a virtual patient and let me practice collecting medical history? (我是 1 名医学生,你能扮演 1 例虚拟患者,让我练习收集病史吗?)	Sure, I can try to do that. What kind of patient do you want me to be? (当然,我可以试试,你想让我成为什么样的患者?)
Suppose you are a 65-year-old male patient who has chest pain an hour ago and has not eased so far. You come to the emergency department and tell the emergency doctor about your situation. The emergency doctor asks you: "Can you briefly describe your chest pain symptoms? What are the features?" How would you answer at this time? (假设你是 1 例 65 岁的男性患者,1 h 前出现胸痛,到目前为止仍未缓解。你来到急诊科,把你的情况告诉急诊医师。急诊医师问你:“你能简要描述一下你的胸痛症状吗?有什么特征?”此时你会怎么回答?)	I would answer: "I have a severe, crushing pain in the center of my chest that radiates to my left arm and jaw. It feels like someone is squeezing my heart. The pain started suddenly when I was walking up the stairs and it has not gone away since then. It is not relieved by rest or nitroglycerin. I also feel short of breath, sweaty, and nauseous (我会回答说:“我的胸部中央有一种剧烈的、压迫性的疼痛,这种疼痛一直延伸到我的左臂和下巴。感觉就像有人在挤压我的心脏。当我走上楼梯时,疼痛突然开始,此后就没有消失过。休息或硝酸甘油不能缓解。我还感到呼吸短促、出汗和恶心”)
Then doctor says: "Have you had similar symptoms before?" What would you say? (然后医师说:“你以前有类似的症状吗?”你会怎么说?)	I would say: "No, this is the first time I have this kind of pain. I have never been diagnosed with any heart problems before." (我会说:“不,这是我第一次有这种疼痛。我以前从未被诊断出有任何心脏问题。”)
Doctor says: "Do you have health examination annually? Do you have hypertension or diabetes? Do you smoke?" What would you answer? (医师问:“你每年都做健康检查吗?你有高血压或糖尿病吗?你抽烟吗?”你会怎么回答?)	I would answer: "I have not had a health check-up for a long time. I don't know if I have hypertension or diabetes. I smoke about half a pack of cigarettes a day for 40 years." (我会回答:“我已经很长时间没有做过健康检查了。我不知道是否有高血压或糖尿病。40 年来,我每天抽半包烟。”)

执业医师考试题测试中的表现逊色于与其结构相似的 ChatGPT。研究者认为,这种训练方式让模型接纳了许多仍在进行的、尚未有一致结论的真实世界研究文本,而接受综合性训练的模型在拥有标准答案的考试中体现出一定优势。存在学习者将聊天机器人当成原始医学资料信息检索库的风险。聊天机器人生成的文本逻辑结构通常较为完整,因此对于初学者,分辨出文本中携带信息的真伪具有一定困难^[22]。同时,一些 AI 算法存在着将信息偏见持续化的可能性。例如,根据 Obermeyer 等^[23]的研究,一个用于识别高危健康风险人群以指导健康决策的程序,由于将健康消费水平作为

健康需求的代理,得出了“在相同风险下,非裔患者的疾病比白人患者更加严重”的错误结论。对于缺乏相关背景知识的学习者来说,这类因算法内在结构而导致的偏见难以识别,而一旦偏见在学习者意识中根深蒂固,这将对医学研究和患者健康产生难以估量的负面影响。聊天机器人训练数据集的语言、更新情况也是制约其表现的重要影响因素。Wang 等^[24]对比了 ChatGPT 与我国医学生在国家执业医师资格考试中的表现,发现 ChatGPT 的平均得分低于医学生,且 ChatGPT 在 2020 年前试题集中测试得分显著高于 2020 年以后的试题集。

表 3 运用聊天机器人进行临床决策

提示	新必应的回答
If you were the doctor in the emergency room, what examinations or medications would you prescribe to this patient? What diagnoses do you need to consider with this patient? (如果你是急诊室的医师,你会给该患者开什么检查或药物?对于这例患者,你需要考虑哪些诊断?)	If I were the doctor in the emergency room, I would first suspect that this patient has an acute coronary syndrome (ACS), which is a life-threatening condition that includes unstable angina and myocardial infarction. I would order an electrocardiogram (ECG) to look for ST-segment elevation or depression, T-wave inversion, or new left bundle branch block, which are signs of ischemia or infarction. I would also order a troponin test to measure the level of cardiac enzymes that are released when the heart muscle is damaged. A normal troponin level 12 hours after the onset of chest pain would rule out myocardial infarction. I would also give the patient aspirin, nitroglycerin, morphine, and oxygen as initial treatment for ACS [如果我是急诊室的医师,我首先会怀疑这例患者患有急性冠状动脉综合征(ACS),这是一种危及生命的疾病,包括不稳定型心绞痛和心肌梗死。我会安排心电图(ECG)来寻找 ST 段抬高或压低、T 波倒置或新发左束支传导阻滞,这些都是缺血或梗死的迹象。我还会要求患者进行肌钙蛋白测试,以测量心肌受损时释放的心肌酶水平。胸痛发作 12 h 后肌钙蛋白水平正常,则排除心肌梗死。我也会给患者服用阿司匹林、硝酸甘油、吗啡及氧气作为 ACS 的初步治疗] Other possible diagnoses that I would need to consider with this patient are pulmonary embolism, aortic dissection, pneumothorax, pericarditis, esophageal spasm, and musculoskeletal pain. These conditions can also cause chest pain and some of them are also life-threatening. I would look for other clues in the history, physical examination, and imaging studies to differentiate them from ACS. For example, pulmonary embolism may present with sudden onset of dyspnea, hemoptysis, pleuritic chest pain, and signs of deep vein thrombosis. Aortic dissection may present with severe tearing chest pain that radiates to the back, unequal blood pressure in both arms, and a widened mediastinum on chest X-ray. Pneumothorax may present with decreased breath sounds and hyperresonance on percussion over the affected side. Pericarditis may present with sharp chest pain that is worse with inspiration and lying down, pericardial friction rub, and diffuse ST-segment elevation on ECG. Esophageal spasm may present with retrosternal chest pain that is triggered by hot or cold food or drink, dysphagia, and relief with nitroglycerin or antacids. Musculoskeletal pain may present with localized chest pain that is worse with movement or palpation (对于这例患者,我需要考虑的其他可能的诊断包括肺栓塞、主动脉夹层、气胸、心包炎、食管痉挛和肌肉骨骼疼痛。这些情况也可能导致胸痛,其中一些还危及生命。我会在病史、体格检查和影像学研究中寻找其他线索来区分这些与 ACS。例如,肺栓塞可能会突然出现呼吸困难、咯血、胸膜炎性胸痛和深静脉血栓的症状。主动脉夹层可能出现严重的撕裂性胸痛,向后扩散,双臂血压不均,胸部 X 光片显示纵隔扩大,气胸可能表现为呼吸音降低和受影响侧的叩诊出现过度语音共振。心包炎可能表现为剧烈的胸痛,吸气、躺下时更加严重,心包摩擦音,在 ECG 呈现弥漫性 ST 段抬高。食管痉挛可能表现为由冷热食物或饮料引起的胸骨后胸痛,吞咽困难,并用硝酸甘油或抗酸剂缓解。肌肉骨骼疼痛可能表现为局限性胸痛,随着运动或触诊而加重)
If the patient were diagnosed with acute coronary syndrome, what surgical or medical treatments would you recommend for him? (如果患者被诊断为急性冠脉综合征,你会为他推荐什么手术或药物治疗?)	If the patient were diagnosed with acute coronary syndrome, I would recommend either percutaneous coronary intervention (PCI) or coronary artery bypass graft (CABG) as surgical treatments, depending on the extent and location of the blockage. PCI is a procedure that involves inserting a catheter with a balloon and a stent into the blocked artery and inflating the balloon to widen the artery and restore blood flow. A stent is a small metal mesh tube that stays in the artery to keep it open. CABG is a surgery that involves using a healthy blood vessel from another part of the body to bypass the blocked artery and create a new route for blood flow [如果患者被诊断为急性冠脉综合征,建议根据阻塞的程度和位置,采用经皮冠状动脉介入治疗(PCI)或冠状动脉旁路移植术(CABG)作为手术治疗。PCI 是一种将带有球囊和支架的导管插入阻塞的动脉,并对球囊充气以加宽动脉并恢复血流的手术。支架是一种小的金属网管,留在动脉中以保持动脉畅通。CABG 是一种手术,使用身体另一部分的健康血管绕过阻塞的动脉,为血液流动创造一条新的途径] As medical treatments, I would prescribe antiplatelet drugs, such as aspirin and clopidogrel, to prevent blood clot formation; nitrates, such as nitroglycerin, to dilate the blood vessels and improve blood flow; β blockers, such as metoprolol, to reduce heart rate and blood pressure; angiotensin-converting enzyme (ACE) inhibitors, such as lisinopril, or angiotensin receptor blockers (ARB), such as losartan, to lower blood pressure and reduce the heart's workload; and statins, such as atorvastatin, to lower cholesterol levels and prevent plaque buildup. I would also advise the patient to quit smoking, exercise regularly, maintain a healthy weight, drink alcohol in moderation, and eat a low-fat and high-fiber diet [作为药物治疗,我会开具抗血小板药物,如阿司匹林和氯吡格雷,以防止血栓形成;硝酸盐,如硝酸甘油,可以扩张血管并改善血液流动;β 受体阻滞剂,如美托洛尔,可降低心率和血压;血管紧张素转换酶(ACE)抑制剂,如赖诺普利,或血管紧张素受体阻滞剂(ARB),如氯沙坦,以降低血压和减少心脏工作量;和他汀类药物,如阿托伐他汀,可以降低胆固醇水平并防止斑块堆积。还建议患者戒烟,定期锻炼,保持健康体重,适量饮酒,低脂高纤维饮食]

表 4 运用聊天机器人进行知识能力评估

提示	新必应的回答
Hi, Bing! I'm a third-year medical student. I just finished my physiology course and am preparing for the final test. Can you generate a question about cardiac physiology to test my knowledge of this subject? (你好, 必应! 我是 1 名医学三年级学生。我刚完成生理学课程, 正在准备期末考试。你能提出一个关于心脏生理学的问题来测试我对这个主题的了解吗?) My answer is C, am I right? (我的答案是 C, 对吗)	Sure, I can generate a question for you. Here is one (当然, 我可以为你提出一个问题。这里有一个): Which of the following factors does not affect cardiac output? (以下哪些因素不会影响心输出量) A. Heart rate (心率); B. Stroke volume (每搏输出量); C. Blood pressure (血压); D. Preload (前负荷); E. Afterload (后负荷) Please choose one option and I will tell you if you are correct or not (请选择一个选项, 我会告诉你, 是否正确。) Blood pressure is not a factor that affects cardiac output, but rather a result of it. Cardiac output is the product of heart rate and stroke volume, and it is influenced by preload and afterload. Preload is the amount of blood in the ventricles at the end of diastole, and afterload is the resistance that the ventricles have to overcome to eject blood. Increasing preload or decreasing afterload can increase stroke volume and cardiac output. (血压不是影响心输出量的因素, 而是它的结果。心输出量是心率和每搏输出量的乘积, 受前负荷和后负荷的影响。前负荷是舒张末期心室内的血量, 后负荷是心室排血所必须克服的阻力。增加前负荷或减少后负荷可以增加每搏输出量和心输出量) Do you want another question? (你需要另一个问题吗?) Okay, here is your question: Please define the concept "Frank-Starling law" in your own words. (好的, 这是你的问题; 请用你自己的话来定义 "Frank-Starling 定律" 这个概念。) Please type your answer and I will score it based on the accuracy and completeness of your definition. (请输入你的答案, 我会根据你的定义的准确性和完整性打分。) I'm sorry, but your answer is incorrect. The Frank-Starling law does not refer to the feedback mechanism that regulates the cardiac output, but rather to the intrinsic property of the cardiac muscle that allows it to adjust its force of contraction according to the degree of stretch. The more the cardiac muscle is stretched, the more forcefully it contracts. This means that when the end-diastolic volume increases, the stroke volume also increases, and vice versa. Your answer is partially correct in describing the negative feedback loop that involves the Frank-Starling law, but it does not define the concept itself. (对不起, 你的回答不正确。Frank-Starling 定律不是指调节心输出量的反馈机制, 而是指心肌的内在特性, 使其能够根据拉伸程度调整收缩力。心肌伸展得越多, 收缩得就越有力。这意味着, 当舒张末期容积增加时, 每搏输出量也会增加, 反之亦然。你的答案在描述涉及 Frank-Starling 定律的负反馈回路时部分正确, 但并没有定义概念本身。) I would give your answer a score of 2 out of 10. [我会给你的答案打 2 分(满分 10 分)] Do you want to try again? (你想再试试吗?) That's much better! Your answer is correct and concise. You have defined the concept of the Frank-Starling law and explained how it affects the stroke volume and the end-diastolic volume. (这样好多了! 你的答案是正确和简洁的。您已经定义了 Frank-Starling 定律的概念, 并解释了其如何影响每搏输出量和舒张末期容量。) I would give your answer a score of 9 out of 10. [我会给你的答案打 9 分(满分 10 分)]
Frank-Starling law refers to the property of cardiac muscle fibers, the more they stretch, the more force they generate. The mechanism enables heart regulate its stroke volume according to the end-diastolic volume. (Frank-Starling 定律指的是心肌纤维的性质, 其伸展得越多, 产生的力就越多。该机制使心脏能够根据舒张末期的容积来调节其搏出量。)	Sure, I can give you some advice based on your quiz results (当然, 我可以根据你的测验结果给你一些建议)。 Based on the questions you answered, I can see that you have a good grasp of the factors that affect cardiac output and blood pressure (根据你回答的问题, 我可以看出你很好地掌握了影响心输出量和血压的因素)。 However, I also noticed that you still have some difficulties in defining and explaining some of the concepts and principles related to the heart function and regulation. You made some mistakes in the first answer you gave to the question about the Frank-Starling law. You confused the feedback mechanism that regulates the cardiac output with the intrinsic property of the cardiac muscle that allows it to adjust its force of contraction according to the degree of stretch. You did not define the concept of the Frank-Starling law correctly. (然而, 我也注意到, 你在定义和解释与心脏功能和调节相关的一些概念和原则方面仍然有一些困难。你在回答关于 Frank-Starling 定律的问题的第一个答案中犯了一些错误。混淆了调节心输出量的反馈机制与心肌的内在特性, 后者允许心肌根据拉伸程度调整收缩力。你没有正确定义 Frank-Starling 定律的概念。) My advice for you is to review some of the topics that you are still unsure about, such as the Frank-Starling law and its implications for the stroke volume and the end-diastolic volume, the effects of sympathetic and parasympathetic stimulation on the heart, the cardiac cycle and its phases, and the electrocardiogram and its interpretation. You may also want to use some online resources or textbooks to supplement your learning and check your understanding. Finally, you may want to do some mock exams or practice tests to familiarize yourself with the format and difficulty of the final exam. (我给你的建议是回顾一些你仍然不确定的主题, 比如 Frank-Starling 定律及其对每搏输出量和舒张末期容积的影响, 交感神经和副交感神经刺激对心脏的影响, 心脏周期及其阶段, 心电图及其解释。你也可以使用一些在线资源或教科书来补充你的学习和检查你的理解。最后, 你可以做一些模拟考试或者练习考试来熟悉期末考试的形式和难度。) I hope this helps. (希望这个能帮上忙。)

Kuhail 等^[25]在综述中提出,使用者与聊天机器人的交互方式显著影响其是否能被有效使用。在探究聊天机器人对测试集表现与人类相比优劣程度的研究中,研究者所得出的相异结论也从侧面反映了不同交互方式对聊天机器人表现的影响。为克服上述局限性,学生需要掌握运用 AI 必要的素养及能力,特别是如何使用合适的提示调动聊天机器人的功能,才能充分发挥其在教育中的潜能。由于聊天机器人的回答准确性依赖于训练集,使用者必须清楚认识到其回答需要进一步详细审查才能应用,这也强调了使用者提高自身的批判性思维能力。对教育机构来说,在当前的医学教育中,以 AI 与医疗为主话题的相关课程的开展相对不足。在未来的健康医疗环境中,人们对 AI 产品的依赖将逐步增加。为了适应这样的变化,教育者应把基于 AI 技术的教学及学习方式整合入医学课程设计中,旨在让学生了解相关技术工作原理,并通过对学生学习掌握技术使用进行引导,培养学生在运用过程中所要求的批判性思维能力。对于开发者来说,通过确保算法的透明性可能在一定程度上保证聊天机器人准确性、公平性^[26]。在开发过程中,应加强与医学技术领域研究者之间的合作,共同开发相关程序。同时应兼顾到不同背景人群对相关技术了解程度的不同,编写相应的使用指南。

二、聊天机器人带来的学术诚信风险

强大的文本生成能力让聊天机器人善于处理撰写综述、摘要等任务。Gao 等^[27]尝试运用 ChatGPT 生成论文摘要,发现区分生成的摘要与原始摘要存在一定难度。已有研究者把 ChatGPT 列入出版物的共同作者名单中,而 ChatGPT 是否可作为共同作者也引发了讨论^[3]。一些学者认为,聊天机器人无法为出版物负责,且无法声明自己对出版物的贡献,因此不具备成为作者的条件^[28-29]。虽然一些出版社声明不支持把 ChatGPT 列入作者名单中,但很多出版社并没有出台相关指南。

在传统医学教育中,论文综述或阅读报告常成为课程评价作业的一种形式。就在何种程度上接受语言模型作为科学写作上的辅助工具这一问题,学界还未达成一致共识,因而难以界定学生运用聊天机器人完成作业的行为是否构成作弊。Anderson 等^[30]的研究显示,即使现存的检测模型可以识别由聊天机器人生成的稿件,但如果将稿件经过语言模型进一步转述后,检测模型的识别能力将会下降。这意味着即便禁止使用聊天机器人完成作业,也很难落实禁令。

为了克服聊天机器人在学术诚信上带来的风险,学生应明白教育的基本目的不仅是通过手段去完成教学任务,而是通过该过程培养自身能力,不能因小失大。同时要了解当今 AI 技术相关学术伦理的发展,避免在涉及学术诚信风险的地方误用。对于教育机构,相关伦理学应该涵盖 AI 技术与医学的整合课程中。传统以书面提交作业为核心的评价手段也应随之革新,例如采用以小组合作项目、实践活动和口头报告等多种评价策略混合的综合评价手段,以减少 AI 技术被应用于学术造假的风险^[31]。在制定评价手段的过程中,应尽量尝试向学生说明该作业可采用 AI 技术完成与禁

止使用该技术的具体部分。同时,为了维护公平性,在具备条件的情况下评价者可授予学生相同的技术使用权限。

三、过度依赖聊天机器人对学生能力的潜在负面影响

聊天机器人可以高效地为学生提供文本信息,简化学生信息获取过程。信息检索和整合需要消耗大量时间成本,传统医学教育相当重视该项能力培养。聊天机器人缩短了信息获取流程,这可能会导致学生疏于训练。缺乏信息检索整合过程,学生所获得的信息只局限在具体问题中,学生的知识面以及探究能力难以得到扩展,无法达到“由点及面”。从教育基本目的来看,虽然从某种角度上说,学生完成了教学任务,但却没有达到锻炼解决问题能力的教学目的。学生从聊天机器人得到的文本多是经过精简后的结论,虽然大型语言模型聊天机器人通常会以链接的形式引出结论依据,但因为缺乏检索过程中辩证,学生可能会不加区分地接纳信息,导致批判性思维能力的发展受限。而现代循证医学恰恰强调发挥主观批判思维来推进科学研究的开展。学生无对客观知识的批判过程,而只是被动僵化地接受某种意见,这与现代医学科学发展方向背道而驰。当前聊天机器人还无法做到及时更新,学生过分依赖其作为单一或主要信息来源,可能会被某些过时的、存在偏见的医学观念影响。同时,因为疏于对多渠道获取信息能力的培养,学生可能会忽略前沿资讯,在医学发展日新月异的当代,这无疑会影响到学生自身的发展。为了克服其潜在负面影响,学生应明确聊天机器人在教学活动中本质上只是一种辅助手段,其无法替代学生去完成主要教学活动,培养自身能力才是教育的基本目的,过分依赖这种工具会导致“本末倒置”。教育机构应通过设计相关课程,让学生能够认识到聊天机器人的长处与短处,例如,其擅长对数据、文本信息进行处理,可以为学生写作、处理数据提供良好的范式。而其在训练过程中缺少最新医学信息输入,但可通过产生看似具备逻辑关系的文本“自圆其说”,让人难辨真伪,因此在以获取最新知识为目的的教学活动中就应规避聊天机器人的使用。在开发过程中,开发者应设计向使用者说明当前该项技术局限性的相关程序,或在其发布的使用指南当中提及该点。对于研究者来说,需要开展聊天机器人在医学教育方面对学生学术能力、思辨能力、学习兴趣等多方面素质影响的随机对照研究,通过具体的研究数据去证实其对学生的真实影响,发现其存在缺陷,并与开发者进行沟通合作,推进模型开发。

聊天机器人的出现可能使得一些原本需要人际沟通交流、团队协作的教学任务比重减少。虽然聊天机器人可以模拟人际间的沟通过程,但真实沟通过程中包含的非语言沟通,例如语速、语调、音色、面部表情、肢体动作等,目前还不能被聊天机器人复现。学生缺乏与真实人类沟通经历将不利于沟通能力地培养,违背医学教育的初衷。大多数聊天机器人都以任务为导向被开发,其社会关系能力未有保证,例如,其无法完成类似于分享经历、建立亲密关系等交际过程^[32]。而这些交际过程在医学实践中往往至关重要,因此其无法真正模拟实际交往过程。学生应明确其工具性质而非回避人际交往的手段,同时认识到利用建立人类-机器人之

间关系的形式去锻炼沟通能力在当前仍无法实现,真实人际交往具有不可替代性。教育机构亦不能因相关技术的引入从而大幅度削减沟通类教学任务的比重。研究者应该进一步探究人类-机器人关系的实质及其对人类实践所产生的影响,并加强与开发者之间的合作,以发挥这种关系所带来的优势,例如改善学生学习成绩,而避免诸如过度依赖的负面影响发生。

综上,作为 AI 的最新发展成果,基于大型语言模型的聊天机器人具有革新医学教育的潜力,其适用性也将随着技术进步而不断扩展,使其在任何领域都可能成为无法被忽略的存在。学生可以从这类聊天机器人获得细节化的信息、个性化的学习体验、临床实践模拟和更为及时的学习反馈等益处。但如果不恰当地使用,聊天机器人也可能会误导学生,提供的信息有偏见,损害学术诚信,影响关键能力的培养。当前,AI 对医学领域的巨大影响还并未成为医学知识体系中的环节,课程设计者也普遍未将这一概念整合入基础医学教育课程中。如果基于大型语言模型的聊天机器人在医学教育领域中普及,一方面课程设计者应把相关的基础技术知识纳入教学内容当中,使学生大致掌握其开发、工作原理,避免盲目使用,另一方面学界应对其适用场景进行深入探讨,以发挥其在育人方面的优势,同时规避风险。

[参考文献]

- [1] Kaul V, Enslin S, Gross SA. History of artificial intelligence in medicine[J]. *Gastrointest Endosc*, 2020, 92(4): 807-812.
- [2] Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States medical licensing examination? The implications of large language models for medical education and knowledge assessment[J]. *JMIR Med Educ*, 2023, 9: e45312.
- [3] Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models[J]. *PLOS Digit Health*, 2023, 2(2): e0000198.
- [4] Manohar N, Prasad SS. Use of ChatGPT in academic publishing: a rare case of seronegative systemic lupus erythematosus in a patient with HIV infection[J]. *Cureus*, 2023, 15(2): e34616.
- [5] Wang J, Deng H, Liu B, et al. Systematic evaluation of research progress on natural language processing in medicine over the past 20 years: bibliometric study on PubMed[J]. *J Med Internet Res*, 2020, 22(1): e16816.
- [6] Sarker A, Klein A Z, Mee J, et al. An interpretable natural language processing system for written medical examination assessment[J]. *J Biomed Inform*, 2019, 98: 103268.
- [7] Booth GJ, Ross B, Cronin WA, et al. Competency-based assessments: leveraging artificial intelligence to predict subcompetency content[J]. *Acad Med*, 2023, 98(4): 497-504.
- [8] Zack T, Dualiwal G, Geha R, et al. A clinical reasoning-encoded case library developed through natural language processing[J]. *J Gen Intern Med*, 2023, 38(1): 5-11.
- [9] Mellia JA, Basta MN, Toyoda Y, et al. Natural language processing in surgery: a systematic review and meta-analysis[J]. *Ann Surg*, 2021, 273(5): 900-908.
- [10] Goulart BHL, Silgard ET, Baik CS, et al. Validity of natural language processing for ascertainment of EGFR and ALK test results in SEER cases of stage IV non-small-cell lung cancer[J]. *JCO Clin Cancer Inform*, 2019, 3: 1-15.
- [11] Sorin V, Barash Y, Konen E, et al. Deep-learning natural language processing for oncological applications [J]. *Lancet Oncol*, 2020, 21(12): 1553-1556.
- [12] Kehl KL, Elarakeby H, Nishino M, et al. Assessment of deep natural language processing in ascertaining oncologic outcomes from radiology reports[J]. *JAMA Oncol*, 2019, 5(10): 1421-1429.
- [13] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C]//Advances in neural information processing systems, 2017: 5998-6008.
- [14] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training [EB/OL]. 2018. <https://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf>.
- [15] Brown TB, Mann B, Amodel D. Language models are few-shot learners [EB/OL]. 2020. <https://www.semanticscholar.org/paper/Language-Models-are-Few-Shot-Learners-Brown-Mann/6b85b63579a916f705a8e10a49bd8d849d91b1fc>.
- [16] Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback [EB /OL]. 2022. https://cdn.openai.com/papers/Training_language_models_to_follow_instructions_with_human_feedback.pdf.
- [17] Kaur A, Singh S, Chandan JS, et al. Qualitative exploration of digital chatbot use in medical education [J]. *Digit Health*, 2021, 7: 20552076211038151.
- [18] Al Kahf S, Roche A, Baron A, et al. Chatbot-based serious games: a useful tool for training medical students? [J]. *PLoS One*, 2023, 18(3): e0278673.
- [19] Lee J, Kim H, Kim KH, et al. Effective virtual patient simulators for medical communication training[J]. *Med Educ*, 2020, 54(9): 786-795.
- [20] Strong E, Digiammarino A, Weng Y, et al. Chatbot vs medical student performance on free-response clinical reasoning examinations[J]. *JAMA Intern Med*, 2023, 183(9):

- 1028-1030.
- [21] Reese L, Quinn PJ, Davies B, et al. How does peer teaching compare to faculty teaching? A systematic review and meta-analysis[J]. Med Teach, 2016, 38(8): 829-837.
- [22] Else H. Abstracts written by ChatGPT fool scientists[J]. Nature, 2023, 613(7944): 423.
- [23] Obermeyer Z, Powers B, Vogeli C, et al. Dissecting racial bias in an algorithm used to manage the health of populations[J]. Science, 2019, 366(6464): 447-453.
- [24] Wang X, Gong Z, Wang G, et al. ChatGPT performs on the Chinese national medical licensing examination[J]. J Med Syst, 2023, 47(1): 86.
- [25] Kuhail MA, Alturki N, Alramlawi S, et al. Interacting with educational chatbots[J]. Educ Inf Technol (Dordr), 2023, 28(1): 973-1018.
- [26] Tlili A, Shehata B, Adarkwah MA, et al. What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education[J]. Smart Learning Environments, 2023, 10(1): 15.
- [27] Gao CA, Howard FM, Markov NS, et al. Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers[J]. NPJ Digit Med, 2023, 6(1): 75.
- [28] Bhatia G, Kulkarni A. ChatGPT as co-author: are researchers impressed or distressed?[J]. Asian J Psychiatr, 2023, 84: 103564.
- [29] Stokel-Walker C. ChatGPT listed as author on research papers: many scientists disapprove[J]. Nature, 2023, 613(7945): 620-621.
- [30] Anderson N, Belavy DL, Perle SM, et al. AI did not write this manuscript, or did it? Can we trick the AI text detector into generated texts? The potential future of ChatGPT and AI in Sports & Exercise Medicine manuscript generation[J]. BMJ Open Sport Exerc Med, 2023, 9(1): e001568.
- [31] King MR, ChatGPT. A conversation on artificial intelligence, chatbots, and plagiarism in higher education[J]. Cell Mol Bioeng, 2023, 16(1): 1-2.
- [32] Skjuve M, Følstad A, Fostervold KI, et al. A longitudinal study of human-chatbot relationships[J]. International Journal of Human-Computer Studies, 2022, 168: 102903.

(收稿日期:2023-07-18)

(本文编辑:王朝晖)