·智能眼科·

# 不同大型语言模型与不同水平医学专业人士回答眼科问题的对比研究

黄 慧[1,2*],胡瑾瑜[1,2*],王晓宇[1,2],叶书苑[2],吴世楠[3],陈 程[1,2],何良琪[1,2],曾艳梅[1,2],魏 红[1,2],邵 毅[1,2]

**引用**:黄慧,胡瑾瑜,王晓宇,等. 不同大型语言模型与不同水平医学专业人士回答眼科问题的对比研究. 国际眼科杂志, 2024, 24(3):458-462.

**作者单位**:[1](330006)中国江西省南昌市,南昌大学第一附属医院眼科;[2](200126)中国上海市,复旦大学附属眼耳鼻喉科医院;[3](361104)中国福建省厦门市,厦门大学眼科研究所
*:黄慧和胡瑾瑜对本文贡献一致。
**作者简介**:黄慧,女,在读硕士研究生,研究方向:角膜病、人工智能;胡瑾瑜,女,在读硕士研究生,研究方向:角膜病、人工智能。
**通讯作者**:邵毅,博士,主任医师,博士研究生导师,国际转化医学会副会长兼眼科专委会主任委员,中国医药教育协会眼科影像与智能医疗分会主任委员,研究方向:角膜病、智能医疗. freebee99@163.com

收稿日期:2023-11-02 修回日期:2024-01-26

**摘要**

**目的**:评估 3 种不同的大型语言模型(LLM,包括 GPT-3.5、GPT-4 和 PaLM2)在回答眼科专业问题中的表现并与 3 种不同水平的专业人群(医学本科生、医学硕士、主治医师)进行比较。

**方法**:分别对三种不同的 LLM 和 3 种不同水平的专业人群(包括了本科生 9 名,专业型研究生 6 名,主治医师 3 名)进行一项由 100 道眼科单项选择题组成的测试,问题涵盖了眼科基础知识、临床知识、眼科检查诊断方法以及眼病相关治疗手段。从平均得分、答题稳定性和答题自信心等方面综合评估 LLM 的性能并与人类组进行比较。

**结果**:在平均测试得分中,每个 LLM 都在总体上优于本科生(GPT-4:56 分,GPT-3.5:42 分,PaLM2:47 分,本科生:40 分),其中 GPT-3.5、PaLM2 略低于硕士水平(硕士:51 分),而 GPT-4 则表现出与主治医师相当的水平(主治医师:62 分)。另外,GPT-4 表现出明显高于 GPT-3.5 和 PaLM2 的答题稳定性和答题自信心。

**结论**:以 GPT-4 为代表的 LLM 在眼科领域表现的较为出色,LLM 模型可为临床医生和医学教育进行临床决策及教学辅助。

**关键词**:大型语言模型(LLM);自然语言处理;眼科问题
DOI:10.3980/j.issn.1672-5123.2024.3.24

## Comparative study of different large language models and medical professionals of different levels responding to ophthalmology questions

**Huang Hui[1,2*], Hu Jinyu[1,2*], Wang Xiaoyu[1,2], Ye Shuyuan[2], Wu Shinan[3], Chen Cheng[1,2], He Liangqi[1,2], Zeng Yanmei[1,2], Wei Hong[1,2], Shao Yi[1,2]**

[1]Department of Ophthalmology, the First Affiliated Hospital of Nanchang University, Nanchang 330006, Jiangxi Province, China; [2]Eye & ENT Hospital of Fudan University, Shanghai 200126, China; [3]Eye Institute of Xiamen University, Xiamen 361104, Fujian Province, China
[*]Co-first authors: Huang Hui and Hu Jinyu
**Correspondence to**: Shao Yi. Department of Ophthalmology, the First Affiliated Hospital of Nanchang University, Nanchang 330006, Jiangxi Province, China; Eye & ENT Hospital of Fudan University, Shanghai 200126, China. freebee99@163.com
Received:2023-11-02 Accepted:2024-01-26

## Abstract

• AIM: To evaluate the performance of three distinct large language models (LLM), including GPT-3.5, GPT-4, and PaLM2, in responding to queries within the field of ophthalmology, and to compare their performance with three different levels of medical professionals: medical undergraduates, master of medicine, and attending physicians.

• METHODS: A total of 100 ophthalmic multiple-choice tests, which covered ophthalmic basic knowledge, clinical knowledge, ophthalmic examination and diagnostic methods, and treatment for ocular disease, were conducted on three different kinds of LLM and three different levels of medical professionals (9 undergraduates, 6 postgraduates and 3 attending physicians), respectively. The performance of LLM was comprehensively evaluated from the aspects of mean scores, consistency and confidence of response, and it was compared with human.

• RESULTS: Notably, each LLM surpassed the average performance of undergraduate medical students (GPT-4: 56, GPT-3.5:42, PaLM2:47, undergraduate students:40).

Specifically, performance of GPT-3.5 and PaLM2 was slightly lower than those of master's students (51), while GPT-4 exhibited a performance comparable to attending physicians (62). Furthermore, GPT-4 showed significantly higher response consistency and self-confidence compared with GPT-3.5 and PaLM2.

• CONCLUSION: LLM represented by GPT-4 performs well in the field of ophthalmology, and the LLM model can provide clinical decision-making and teaching aids for clinicians and medical education.

• KEYWORDS: large language models (LLM); natural language processing; ophthalmology question

**Citation**: Huang H, Hu JY, Wang XY, et al. Comparative study of different large language models and medical professionals of different levels responding to ophthalmology questions. Guoji Yanke Zazhi(Int Eye Sci), 2024,24(3):458-462.

## 0 引言

大型语言模型(large language model,LLM)的兴起是自然语言处理(natural language processing,NLP)领域的一大飞跃[1]。近年热门的机器学习、深度学习等人工智能方法在眼科中的应用日益广泛,而其主要分析的为高质量的典范性数据样本和图像数据训练模型[2-3]。LLM是基于大规模纯文本数据训练的深度学习模型[4-5]。具体来说,LLM是建立在转换器架构之上的,通过对现有语料库(书本、报纸、网络等)的学习获得通用技能(语言理解、文本输出等),再通过指令调和对齐调对其能力做进一步调整,使其行为更加符合人类的价值观和偏好,从而更好的完成任务,生成高质量的输出文本。LLM与传统小规模语言模型的最大区别在于其参数容量更大,包含千亿甚至更多的参数,如此大的参数规模不仅使语言模型的性能得到显著提升,还表现出了一些小规模语言模型不具备的功能,比如上下文、指令遵循、推理等[6]。目前,由OpenAI公司发布的最新版本ChatGPT/GPT-4被认为是最强大的LLM之一。以GPT-4为代表的新一代LLM显示出强大的功能,已经在一般领域取得了成功。例如:通过对模型进行农业相关文本数据的预训练,LLM便能出色完成粮食和农业领域的NLP任务[7]。在美国注册会计师考试和美国律师考试中,ChatGPT也获得了相当可观的分数[8]。此外,也有数据表明LLM在像医学这样需要高级和复杂思维的领域也极具潜力。Gilson等[9]发现ChatGPT达到了通过美国医师执照考试(the US Medical Licensing Examination, USMLE)的标准。不仅如此,ChatGPT还可以帮助外科医生采集患者病史[10]、分析医学成像特征、精确诊断、优化手术计划、预测手术结果、提高手术效率和安全性、加强术后管理和康复[11-12]。甚至在某些方面,ChatGPT表现的比医生更出色,在一项评估ChatGPT回答患者问题能力的试验中,Ayers等[13]发现,ChatGPT在问题回复的质量和同理心上都优于医生。但这些测试都是基于基础或科普类知识,有太多的现存资料,甚至可以在现有语料库中轻松寻找答案,从而无法系统、全面地评价LLM的性能。然而,截止目前,在医学中使用LLM存在潜在风险,包括受限和过时的医学数据训练,导致不准确的医疗建议[13]。例如,一项涉及LLM在视网膜相关疾病训练上的研究显示,当涉及来自患有视网膜疾病的患者的信息来源时,其识别准确率仅为45%。这表明了在眼科临床环境中应用人工智能存在显著差距[14]。基于上述情况,想要科学、系统地对模型性能进行探索和分析,不仅要选择更加细微、专业壁垒更高的领域,还需确保测试内容不包括在训练数据中[15]。为此,在研究中,我们新创建了100道眼科单项选择题用于测试,旨在评估3种不同的LLM(GPT-3.5、GPT-4、PaLM2)在回答眼科问题中的表现,并将结果与3种不同水平的专业人群(医学本科生、医学硕士、主治医师)进行比较,从而匹配出不同模型对应的专业人群。此外,我们还探讨了不同模型在测试中的答题稳定性,并将继续探讨ChatGPT-4用于医学教育和临床决策的可靠性。

## 1 对象和方法

**1.1 对象** 本研究中用于测试的100道眼科单项选择题由经验丰富的眼科专家创建。考试问题列于附录A部分。在测试中,将100道眼科单项选择题分别输入到每个LLM中。每次测试都以新线程或重置后的初始化提示开始。随后,LLM会以20个问题为一个批次出现提示,直到测试结束。在每一次测试中都向LLM说明必须返回绝对正确的答案。如果LLM无法一次性处理20个问题,则将批量处理问题个数改为10。如果LLM未返回该批次中所有问题的答案,则下一批测试将包括未回答问题和下一批中所有问题。将LLM测试结果与3种专业人群(医学本科生、医学硕士、主治医师)的结果进行比较。在南昌大学中随机抽取9名临床医学专业的本科生和6名眼科学专业的硕士为本科生组和硕士组,在南昌大学第一附属医院中随机抽取3名眼科主治医师为主治医师组。每位考生都参加时长为3 h的闭卷考试。在比较LLM和人类成绩时,评估了平均得分、答题稳定性和答题自信心。

**1.2 方法** 为了量化准确性,我们通过计算每次试验的平均分来表征每种LLM的得分。对于人类测试组,每组得分用各组均值来表示。为了量化不同LLM的答题稳定性,我们计算了试验之间的得分标准差和不同LLM在测试时给出答案与正确答案的平均相关性。相关性分析方法采用的是Pearson相关性分析。并且为了更加清楚地看到不同LLM在眼科测试中的差距,分别统计并计算了其得分和相关性均值与各自的方差。为了量化LLM的答题自信心,我们在所有测试中计算每个问题的正确答案数。例如,如果每个LLM正确回答同一问题5次,则所有5个答案都正确的问题的百分比会增加1%(因为有100个问题)。此外,将测试结果与考生随机猜测时可能出现的预期分布进行比较。随机猜测时,5次试验中的预期正确答案数平均约为$0.25\times5=1.25$,选择题均为4个选项。使用该值,可以根据所得的poison分布来估计每个问题的正确答案出现的次数,并根据LLM与各资历人员测试题目过程多次试验的回答正确率计算评分。最后,将ChatGPT(GPT-3.5和GPT-4)和PaLM2累计计算得到的分数与人类分数进行比较。

## 2 结果

**2.1 LLM分数与人类组分数的比较** 原始得分和平均测试得分分别见图1、2,其中LLM平均测试得分为5次测试

得分的平均值,人类组平均测试得分为组内每人得分的平均值。在 LLM 的原始得分中可见,不同的 LLM 在总分不确定性及问题回答正确次数方面表现不同,GPT-4 涵盖试题的正确答案范围更广(图 1)。在平均测试得分中可见,每个 LLM 都在总体上优于本科生(平均得分=40 分),其中 GPT-3.5(平均得分=42 分)、PaLM2(平均得分=47 分)略低于硕士水平(平均得分=51 分),而 GPT-4(平均得分=56 分)则表现出与主治医师(平均得分=62 分)相当的水平(图 2)。

**2.2 LLM 答题稳定性的比较**　各试验之间的标准差和平均相关性见图 3。3 种 LLM 都显示出高度一致的分数和答案,评分标准差极小(图 3A);GPT-4 表现出更高的平均相关性(平均值高达 0.83),与 GPT-3.5(0.59)和 PaLM2(0.61)表现出明显差异,PaLM2 显示出略高于 GPT-3.5 的平均相关性,但两者未表现出明显差异(图 3B)。在不同 LLM 下的答题平均相关性中,相关性 $r$ 值越大,代表其答题稳定性越高,GPT-4 与 GPT-3.5 和 PaLM2 之间存在显著差异,而 GPT-3.5 和 PaLM2 之间无显著差异。

**2.3 LLM 答题自信心的比较**　从图 4 显示的结果来看,3 种LLM 与随机猜测没有相似之处,这说明 LLM 往往要么表现为自信,要么表现为困惑,随机猜测答案的可能性很小。其中,GPT-3.5 回答正确率为 52%,错误率为 48%(图 4A);GPT-4 在测试中表现出高度自信,得出 64%的正确答案,但仍有 36%的概率存在混淆(图 4B);PaLM2 的水平则介于两者之间(图 4C)。

## 3 讨论

由于现代生活中视频显示终端设备的普及,眼病的发病率不仅呈现上升趋势,还表现出年轻化的趋势,这给眼科临床工作带来巨大压力和负担。作为眼科专业人员,必须具备扎实的眼科知识以确保为患者提供安全可靠的治疗。此外,由于眼科医生在不同地区,包括城市和农村地区的分布不平衡,医疗资源的获取存在显著差距[16]。在

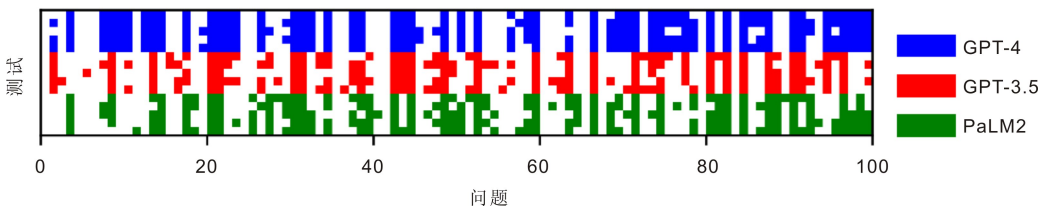


图 1　不同 LLM 在 100 道问题测试得分分布热图。

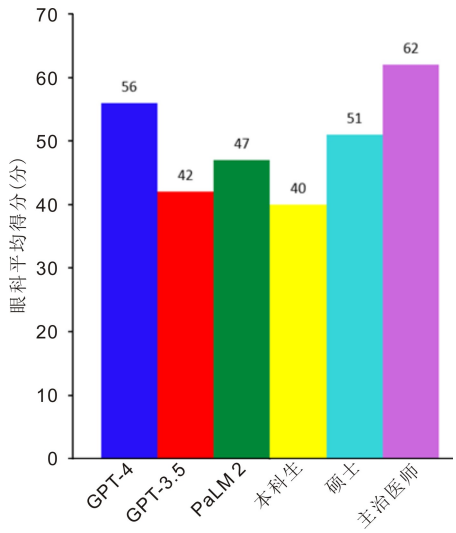


图 2　不同 LLM 与 3 种不同水平的专业人群在 100 道问题测试得分分布柱状图。

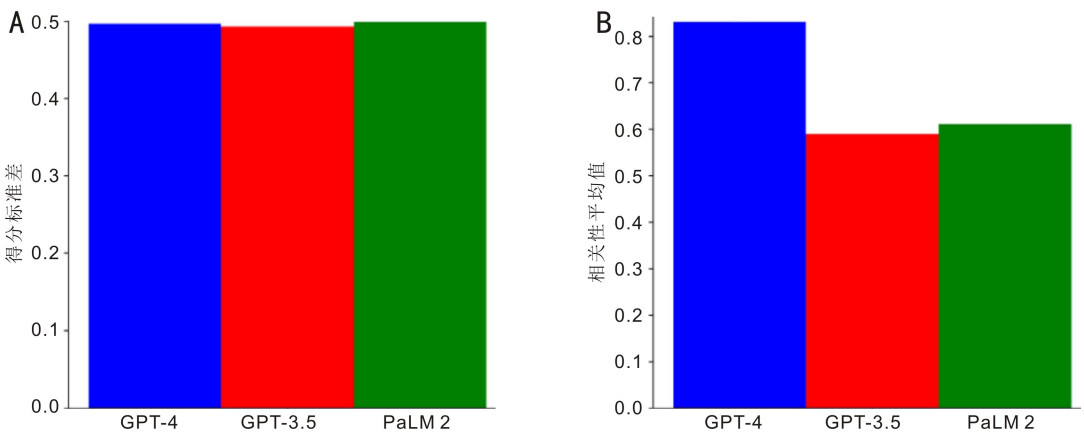


**图 3　不同大语言模型在眼科试题中答题稳定性的比较**　A:3 种 LLM 眼科试题得分标准差;B: 3 种 LLM 在测试时给出答案与正确答案的平均相关性。

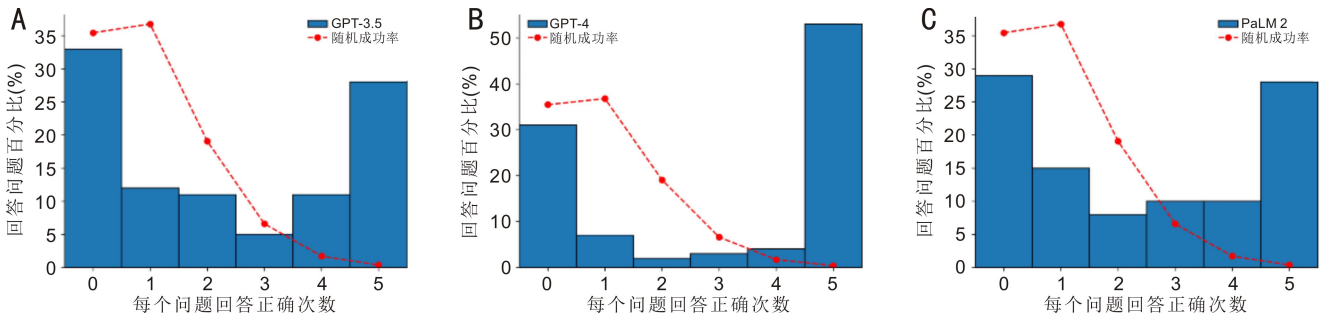


**图 4 不同 LLM 答题自信心比较** A:GPT-3.5 答题随机成功率;B:GPT-4 答题随机成功率;C:PaLM2 答题随机成功率。

基层水平有效提升眼科诊断和治疗能力可以极大地改善眼病患者的预后。先前有项研究[17]比较了一组眼科专家提供的治疗建议与人工智能生成的建议的准确性。专家组和人工智能组在治疗建议方面的一致性达到了 61%。另一方面,专家组和人工智能组在生成医疗处方信息方面无显著差异。且截至目前为止,已有相应的人工智能辅助诊断医疗器械在眼科领域成功定型产品,包括了智能辅助诊断产品、智能辅助治疗产品、智能监护与生命支持、智能康复理疗和智能中医诊疗产品[18]。这些研究结果表明,LLM 等人工智能具有为众多眼病患者提供适当治疗建议的广泛应用潜力。Google Research 利用 Fine-tuning 和指令提示开发了其用于医疗领域的 PaLM 大语言模型,从而产生了 Med-PaLM,在美国医学执照考试中取得了 85%的准确率[19]。这项研究进一步强调了 LLM 在医学领域的适用性。

本研究旨在通过新建 100 道专门为测试设计的单项选择题来评估 LLM 在眼科这样高度专业化的主题上的表现,上述测试题涵盖了眼科的基础知识、临床知识、眼科检查以及眼病诊断和治疗各种类型。同时将其结果与不同水平的眼科专业人群进行对比,以探究 LLM 用于医学教育和临床决策的可靠性。用于测试的 100 道题列于附录 A 部分。测试结果表明,各个 LLM 在总体性能上表现各不相同,3 种 LLM 在总体上都优于医学本科生,医学硕士的水平略高于 GPT-3.5 和 PaLM2,最令人吃惊的是,GPT-4的水平几乎与主治医师持平。医学本科生在医学领域狩猎范围较广,接触的知识也较为基础和科普,在眼科专业领域相当于未入门级别,而 LLM 的训练数据几乎涵盖了大部分本科生所学知识,因此 LLM 轻而易举地超越了医学本科生的答题水平。对于医学硕士,在眼科知识储备和临床经验上虽不及主治医师但也经历了系统化地培训,在专业上已经有一定水平的突破,同时,其具备的逻辑推理能力让他们在不知道正确答案的情况下也能通过推理和猜测得到不太可能出错的答案,从而表现出高于 GPT-3.5 和 PaLM2 的答题水平。而对于已经具有一定临床工作经验的主治医师来说,由于工作性质的要求其必然具备扎实的专业知识和临床技能。GPT-4 表现出与其相当的答题水平,这足以证明 GPT-4 在眼科这样高度专业化的主题上性能也非常出色。此外,在美国眼科学会(AAO)基础与临床科学课程自我评估计划的模拟考试成绩评估中,尽管 GPT-4 仅接受了一般性认知培训,没有经过特定医学领域的数据培训,但在基于文本的练习题中,其准确性明显高于眼科住院医生和执业眼科医生[20]。这与我们的研究结果一致,表明一般人工智能在眼科领域具有潜在有价值的应用前景。

虽然 LLM 在总体上都优于医学本科生,但 GPT-4 表现出比另两种 LLM 更高的答题相关性,这表明 GPT-4 在不同的测试中对于同一题更为倾向地选择同一个答案。与人类相比,GPT-4 在答题相关性上也占据优势,因为如果可以以测试 LLM 相同的方式重复测试同一个人,他也可能会在某一些问题上感到困惑而无法每次选择出相同的答案并表现出一定程度的混乱。在答题自信心方面,LLM 与随机猜测不同,它们总是要么表现为自信要么表现为困惑。此时,GPT-4 又在三者之间脱颖而出,虽然 GPT-4 也显示出一定程度的困惑,有 36%的概率出现混淆,但在答题正确率上以 64%完胜 GPT-3.5 和 PaLM2。

上述结果显示 GPT-4 在各项性能中都表现优秀,能做到与主治医师相似的答题水平,能在不同的测试中对于同一题总是选出同一个答案,表现出高度的答题自信心,但这也不能成为其可以完全取代主治医师的理由,原因有以下几点:(1)GPT-4 在回答问题时,倘若回答正确那么绝对是自信的,但在回答错误时它也总是自信地选择同一个错误答案并自信地认为它是正确的。作为人类的主治医师,他在高度专业化的主题上会根据自己的经验而思考,懂得何时应该猜测并知道如何去明智地猜测,即使在某些问题上拿不准正确答案,也能通过推理得出不太可能错的答案。(2)主治医师就算教育经历、专业背景一致,个人能力和知识储备也会出现很大差异,在某些特定领域有更为深入的研究。(3)也是极为重要的一点,回答问题并不能完全等同于临床医生的日常工作,临床工作是复杂而繁琐的,扎实的专业知识只是其中一个要素。由此可见,单用 LLM 的做题能力来评估其在临床工作中的表现可能会产生一定的偏差。

我们的研究显示,LLM 特别是 GPT-4 在眼科领域有很大的应用前景,但在医学领域还有许多高度专业化的领域有待进行评估。随着 LLM 的不断升级,使用 LLM 完成更加高端复杂的任务是其趋势,因此人类应该把 LLM 的重心从获取简单基础知识转移到更加高度专业化的领域。此外,基于 LLM 的系统并不是设计来替代眼科医生的。相反,它们可以在一定程度上辅助眼科医生的工作。例如,患有眼疾的患者可以在前往眼科诊所之前咨询 LLM 系统,获取有关常见眼部疾病的初步信息。咨询后,模型可以进一步总结并提供具体个性化的就诊摘要和后续建议。这种方法可以实现个性化患者教育,同时节省眼科医生宝贵时间,并可应用于更复杂和具有挑战性的病例。

当然,不可否认,在眼科这样高度专业化的领域,以 GPT-4 为代表的 LLM 表现的较为出色。我们的研究表明,LLM 可为临床医生和医学教育进行临床决策和教学提供有利辅助。LLM 显示出的强大功能使其广泛应用于眼科,目前已有将 ChatGPT 用于眼科方面的测试。Bernstein 等[17]进行了一项评估 LLM 聊天机器人生成的眼科建议与眼科医生撰写的建议相比的质量研究,发现聊天机器人生成的答案在指导性、安全性、可靠性方面与人类答案没有差异。由此可见,LLM 可以作为一个平台,在患者出现眼部问题时跨越语言障碍提供有用的意见,满足患者的咨询需求[21]。同时,ChatGPT 还能用于编写手术记录、出院小结[22-23],在适当的提示下将手术步骤[24]、特定药物[25]、随访说明等信息记录在文书的恰当位置,医生只需对其进行微调便能生成可以归档的医疗文书[26],这无疑可以大大提高医生效率。此外,ChatGPT 也日益成为辅助科研设计和论文写作的有力工具[27-28]。

此外,GPT 在生成对用户输入的响应时的具体分析方法和学习过程是未知的,类似于机器学习方法,它们缺乏清晰的可解释性,这使得临床专业人员和用户在完全接受它们过程中存在一定挑战。我们需要进一步增强这些“黑匣子模型”的可解释性。最后,从法律和监管考虑的角度来看,像 ChatGPT 这样的人工智能产品,如果用于指导临床实践,必须遵守国家药品监督管理局(NMPA)设定的相关法规,以确保医疗保健中人工智能技术的安全、可靠和可控发展,并解决医疗人工智能责任和问责问题。

我们本次研究也存在一定的局限性:(1)LLM 的测试题中未包括眼科学影像相关内容,而眼科图像在基础疾病诊断中发挥重要作用,我们将在后续的研究中进一步扩大多模态 LLM 的测试效能。(2)测试眼科专业试题的准确率并不能完全模拟真实临床应用场景,而患者在临床咨询过程中问题的咨询应更普惠化,LLM 对该类问题的准确充分回答在后续临床应用中发挥重要作用。(3)LLM 在眼科学中的应用仍需要更多的实际临床验证,获得更多真实病例数据,以确保我们的模型在实际医疗环境中的鲁棒性和可靠性。

我们的研究表明,以 GPT-4 为代表的 LLM 在眼科这样高度专业化的领域表现的较为出色,具有较高的答题正确率、稳定性和自信心。随着对 LLM 的不断完善,LLM 将在医学教育和临床决策中发挥重要作用。

**参考文献**

[1] Zhao L, Zhang L, Wu ZH, et al. When brain-inspired AI meets AGI. Meta-Radiology, 2023,1(1):100005.
[2] Wiedeman P, 惠延年. 眼科学人工智能. 国际眼科杂志, 2023,23(9):1417-1420.
[3] 马晓宇, 张力, 毕燕龙. 人工智能在糖尿病视网膜病变领域的研究进展. 国际眼科杂志, 2022,22(11):1818-1821.
[4] Thirunavukarasu AJ, Ting DSJ, Elangovan K, et al. Large language models in medicine. Nat Med, 2023,29(8):1930-1940.
[5] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature, 2015,521(7553):436-444.
[6] Zhao WX, Zhou K, Li J, et al. A survey of large language models. arXiv,2023.
[7] Rezayi S, Liu ZL, Wu ZH, et al. AgriBERT: Knowledge-Infused Agricultural Language Models for Matching Food and Nutrition. Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, 2022.
[8] Bommarito J, Bommarito M, Katz DM, et al. Gpt as knowledge worker: A zero-shot evaluation of (ai) cpa capabilities. arXiv, 2023.
[9] 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. JMIR Medical Education, 2023,9(1):e45312.
[10] Waisberg E, Ong J, Masalkhi M, et al. Text-to-image artificial intelligence to aid clinicians in perceiving unique neuro-ophthalmic visual phenomena. Ir J Med Sci, 2023,192(6):3139-3142.
[11] Cheng K, Sun Z, He Y, et al. The potential impact of ChatGPT/GPT-4 on surgery: will it topple the profession of surgeons? Int J Surg, 2023,109(5):1545-1547.
[12] Patel SB, Lam K. ChatGPT: the future of discharge summaries? The Lancet Digital Health, 2023,5(3):e107-e108.
[13] Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med, 2023,183(6):589-596.
[14] Potapenko I, Boberg-Ans LC, Stormly Hansen M, et al. Artificial intelligence-based chatbot patient information on common retinal diseases using ChatGPT. Acta Ophthalmol, 2023,101(7):829-831.
[15] Kalyan KS, Rajasekharan A, Sangeetha S. AMMUS: a survey of transformer-based pretrained models in natural language processing. arXiv, 2021.
[16] Teo ZL, Ting DSW. AI telemedicine screening in ophthalmology: health economic considerations. Lancet Glob Health, 2023,11(3):e318-e320.
[17] Bernstein IA, Zhang YV, Govil D, et al. Comparison of ophthalmologist and large language model chatbot responses to online patient eye care questions. JAMA Netw Open, 2023,6(8):e2330320.
[18] 张乔, 黄思芮, 张楠, 等. 国内人工智能医疗器械发展现状分析. 医疗卫生装备, 2023,44(5):64-68.
[19] Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. arXiv, 2022.
[20] Lin JC, Younessi DN, Kurapati SS, et al. Comparison of GPT-3.5, GPT-4, and human user performance on a practice ophthalmology written examination. Eye, 2023,37(17):3694-3695.
[21] Will ChatGPT transform healthcare? Nat Med, 2023,29(3):505-506.
[22] Waisberg E, Ong J, Masalkhi M, et al. GPT-4 and ophthalmology operative notes. Ann Biomed Eng, 2023,51(11):2353-2355.
[23] Singh S, Djalilian A, Ali MJ. ChatGPT and ophthalmology: exploring its potential with discharge summaries and operative notes. Semin Ophthalmol, 2023,38(5):503-507.
[24] Bhattacharya K, Bhattacharya AS, Bhattacharya N, et al. ChatGPT in surgical practice—a new kid on the block. Indian J Surg, 2023,85(6):1346-1349.
[25] Chakraborty C, Bhattacharya M, Lee SS. Artificial intelligence enabled ChatGPT and large language models in drug target discovery, drug discovery, and development. Mol Ther Nucleic Acids, 2023,33:866-868.
[26] Waisberg E, Ong J, Masalkhi M, et al. GPT-4: a new era of artificial intelligence in medicine. Ir J Med Sci 1971, 2023,192(6):3197-3200.
[27] Salimi A, Saheb H. Large language models in ophthalmology scientific writing: ethical considerations blurred lines or not at all? Am J Ophthalmol, 2023,254:177-181.
[28] 惠延年. 人工智能聊天机器人助力眼科和科学论文写作. 国际眼科杂志, 2024,24(1):1-4.