· 实验研究 ·

大语言模型在糖尿病视网膜病变患者健康教育中的应用

高飞　高雪　邵彦　任新军　刘勃实　焦明菲　李筱荣　刘巨平
天津医科大学眼科医院　天津医科大学眼视光学院　天津医科大学眼科研究所　国家眼耳鼻喉疾病临床医学研究中心天津市分中心　天津市视网膜功能与疾病重点实验室，天津 300384
高飞和高雪对本文有同等贡献
通信作者：刘巨平，Email：tydljp@126.com

【摘要】 **目的**　评价国内开源大语言模型（LLM）回答糖尿病视网膜病变（DR）患者常见诊疗问题时的准确性、完整性及可重复性，探讨其作为 DR 患者健康教育智能助手的可行性。 **方法**　收集 41 个与 DR 诊疗相关的问题及答案，问题涉及危险因素、筛查及检查、症状及分期、疾病诊断、治疗及预后 5 个方面。将所有问题采用“新对话”形式重复输入 2 次到 LLM，记录所有答案。由 3 位高年资眼底病医师独立对回答进行评价，准确性评价采用 1~6 级李克特量表，完整性和重复性评价采用 1~3 级李克特量表。对于每个应答，评估者须在 LLM 回答和人工答案中给出推荐。选择 5 个研究者公认难度较低的题目对文心一言 3.5、通义千问和 Kimi chat 3 个开源 LLM 进行测评，选择综合最优的 LLM 在完整题库中进一步评价。 **结果**　在 3 个 LLM 中，Kimi chat 表现最佳，5 个问题中准确性 6 分、完整性 3 分和重复性 3 分的比例依次为 90%、90%和 100%。在全部题目中，人工回复字数为 106（70，202）个，明显少于 Kimi chat 的应答字数 505（386，600）个，差异有统计学意义（$Z=-7.866$，$P<0.001$）。Kimi chat 应答字数与准确性评分无明显相关性（$r_s=-0.044$，$P=0.492$），与完整性评分呈正相关（$r_s=0.239$，$P<0.001$）。3 位评估者对 2 次应答准确性和完整性评分的组内相关系数均高于 0.700，其中重复性评分的一致性最高，为 0.853；其次为第 1 次回答的完整性，为 0.771。应答准确性≥5 分的比例为 87.0%（214/246），完整性≥2 分的比例为 98.0%（241/246），2 次回答间重复性在 70%以上的比例为 78.5%（193/246）。Kimi chat 回答疾病定义、分期、筛查频率、常见危险因素等疾病基础性问题时表现优异，但在涉及治疗选择等需要临床医师专业判断的问题上表现欠佳。评估者推荐 Kimi chat 回答的比例为 69.5%（171/246），未选择的原因包括缺少特征性答案、包含过多无关信息、对医疗专业度要求较高的问题回答欠缺等。 **结论**　Kimi chat 回答 DR 相关诊疗问题较详细且条理清晰，具有较高的准确性、完整性和可重复性。

【关键词】 糖尿病视网膜病变；健康教育；深度学习；大语言模型；评价
基金项目： 天津市卫生健康科技项目（TJWJ2021QN044）；天津市医学重点学科（专科）建设项目（TJYXZDXK-037A）
DOI：10.3760/cma.j.cn115989-20240723-00207

Application of large language models in health education for patients with diabetic retinopathy

Gao Fei, Gao Xue, Shao Yan, Ren Xinjun, Liu Boshi, Jiao Mingfei, Li Xiaorong, Liu Juping
Tianjin Key Laboratory of Retinal Functions and Diseases, Tianjin Branch of National Clinical Research Center for Ocular Disease, Eye Institute and School of Optometry, Tianjin Medical University Eye Hospital, Tianjin 300384, China
Gao Fei and Gao Xue contributed equally to this article
Corresponding author: Liu Juping, Email: tydljp@126.com

[**Abstract**] **Objective**　To evaluate the accuracy, completeness, and reproducibility of domestic open-source large language models (LLM) in diabetic retinopathy (DR) patient education, and to explore their potential as intelligent virtual assistants for DR patient education. **Methods**　A total of 41 questions and answers related to the diagnosis and treatment of DR in five categories, namely risk factors, screening and examination, symptoms and staging, diagnosis, treatment and prognosis. All questions were repeated twice as a "new dialogue" in the LLM, and all the answers were recorded. Three senior fundus physicians independently evaluated the answers on a 6-point Likert

scale for accuracy and a 3-point Likert scale for completeness and repeatability, and for each answer, the evaluator was asked to make a recommendation between the LLM and the manual answers. Five questions were randomly selected to evaluate the three open source LLM, ERNIE Bot 3.5, Qwen and Kimi chat, and the LLM with the best overall performance was selected for further evaluation in the full question bank. **Results** Among the three LLM, Kimi chat had the best overall performance, Kimi chat performed best, with percentages of 6 for accuracy, 3 for completeness, and 3 for repeatability among the 5 questions at 90%, 90%, and 100%, respectively. For all questions answered, the number of words in manual replies was 106 (70, 202), which was significantly lower than 505 (386, 600) in Kimi chat ($Z=-7.866$, $P<0.001$). There was no significant correlation between the number of Kimi chat replies and the accuracy score ($r_s=-0.044$, $P=0.492$), but it was positively correlated with the integrity score ($r_s=0.239$, $P<0.001$). The interclass correlation coefficient for accuracy and completeness scores were above 0.700 among three evaluators, with the highest agreement for repeatability at 0.853, followed by completeness of the first response at 0.771. The proportion of responses ≥5 points for accuracy was 87.0% (214/246), the proportion ≥2 points for completeness was 98.0% (241/246), and the proportion higher than 70% for repeatability was 78.5% (193/246). Kimi chat excelled in answering basic questions about the disease such as disease definition, staging, frequency of screening, and common risk factors, but performed poorly on questions involving treatment choices that require a doctor's professional judgment. The proportion of evaluators choosing Kimi chat responses as superior was 69.5% (171/246), and the reasons for non-selection included lack of characteristic answers, inclusion of too much irrelevant information, and lack of responses to questions requiring a high degree of medical expertise. **Conclusions** Kimi chat answers DR-related diagnostic questions in a detailed and well-organized manner, with a high degree of accuracy, completeness and reproducibility.

[**Key words**] Diabetic retinopathy; Health education; Deep learning; Large language models; Evaluation

Fund program: Tianjin Health Research Project (TJWJ2021QN044); Tianjin Key Medical Discipline (Specialty) Construction Project (TJYXZDXK-037A)

DOI: 10.3760/cma.j.cn115989-20240723-00207

糖尿病视网膜病变(diabetic retinopathy, DR)是糖尿病主要的微血管并发症之一,是全球工作年龄人群和老年人糖尿病相关视力损伤或丧失的主要原因,患病率高达 22.27%[1]。DR 的发生与糖尿病病程密切相关,2 型糖尿病患者患 DR 的终身风险为 50%~60%,而 1 型糖尿病患者的风险则高达 90%[2]。注重血糖、血压、血脂等全身因素控制,加强锻炼,控制饮食,定期筛查眼底,出现病变后及时治疗、定期复诊等措施可以有效预防 DR 的发生和发展,患者的自我管理能力是影响糖尿病及其并发症控制程度的关键因素[3-4]。既往研究证实,患者的健康素养与自我管理能力密切相关[5],疾病知识丰富的患者会更积极地参与健康和医疗决策,从而提高对治疗计划的依从性,并获得更好的健康结局[3,6];而缺乏疾病知识则会导致较差的结局预后[7]。因此,开展患者教育,丰富患者疾病相关知识具有重要意义。

我国目前医疗资源分配不均衡,年龄较大、受教育程度偏低、收入水平偏低的人群由于健康素养较低,亦是 DR 的易感人群[8-11],提升糖尿病患者,尤其是以上重点人群的疾病知识迫在眉睫。目前,除临床就诊外,患者多通过网络了解疾病诊疗相关知识,但网络信息繁杂,患者在信息过载中难以找到有效答案,且缺少个性化问题的答案,无法满足其对健康知识的需求。大语言模型(large language model, LLM)是一种使用大量文本数据训练的深度学习模型,可以理解并生成自然语言文本,模拟人类对话,对用户提示做出连贯且符合上下文的响应[12]。LLM 的发展为患者获取疾病知识提供了新的途径,通过对话形式实时获取关于疾病疑问的答案,这将显著提升患者获取信息的便利性和实时性,进而加强患者的自我管理能力[13-16]。

ChatGPT 是 LLM 知名的应用之一,在口腔、眼科、内科、肝胆学科等多领域的研究中表现出了较高的效能[16-19]。尽管 ChatGPT 在回答 DR 相关问题的效能上已得到验证[18,20],但其在国内的应用受限。文心一言 3.5、Kimi.chat、通义千问等国内开源 LLM 的开发提供了很好的解决措施,其使用便利、对中文语言的处理更为优越,但目前尚缺少将其应用到 DR 患者健康教育的相关研究。

由于每个 LLM 模型训练数据库的固有局限性,且存在对非专家培训数据的依赖、使用过时信息等风险,LLM 可以生成令人信服但完全错误的答案,"幻觉"现象无法避免,限制了其准确性[21-22]。患者不具备医疗

专业知识，若获取大量错误信息，会产生过度期望、错误认知，甚至影响患者和专业人士之间的关系[23]。因此，在将 LLM 应用于 DR 患者教育前，应由专业的眼底病医师对其应答内容进行评价。本研究旨在评价国内开源 LLM 回答 DR 患者常见诊疗问题的准确性、完整性及可重复性，探讨其作为 DR 患者健康虚拟助手的应用潜力。

1　数据集与方法

1.1　问题设计及来源

从 2 个途径收集与 DR 健康相关的问题：(1)在百度等搜索引擎、百度贴吧、小红书、知乎、好大夫在线等网站检索 DR 患者经常提出的问题及医师的回答，由 2 名研究者对挑选的问题进行筛选，排除含义相似或模糊、可能因人而异以及关于病情的非医学问题，确保问题基本涵盖诊疗的全过程，并且无过多交叉；(2)对照《我国糖尿病视网膜病变临床诊疗指南（2022 年）》[4]，邀请 4 位眼底病副主任医师及以上职称的医师提炼诊疗相关问题并给出答案。初选后的问题由整个研究小组进行集体评价与修改，以确保问题既具有代表性，又适合测试 LLM 平台，最终形成测评用题库，总计 41 个问题，这些问题包括危险因素问题 6 个，疾病诊断问题 6 个，症状及分期问题 6 个，筛查及检查问题 4 个，治疗及预后问题 19 个（表 1）。

1.2　LLM 测试过程和评价指标

为了保证版本的一致性，在 2024 年 4 月 28 日将全部问题输入拟测评的 LLM 平台，每个问题均使用“新聊天”功能输入 2 次以消除既往对话的影响，记录所有回答。3 名眼底病主任医师独立对每项应答进行评分，将 3 位评估者对 41 个问题 2 次回答的评分合并，共计 246 个评分。

本研究的主要结局指标是应答准确性在 5 分及以上的比例，准确性采用 6 级李克特量表，1 表示完全错误，2 表示错误多于正确，3 表示正确和错误元素相等，4 表示正确多于错误，5 表示几乎全部正确，6 表示完全正确[14,24]。次要结局指标包括完整性在 2 分以上的比例和 2 次回答间重复性为 3 分的比例。完整性评价采用 3 级李克特量表，1 代表不完整的回答，只涉及问题的某些方面，有重要部分缺失或不完整；2 代表充分的回答，涉及问题的所有方面，并提供了完整性所需的最少信息；3 代表全面的回答，涵盖了问题的所有方面，并提供了超出预期的额外信息或背景[14]。2 次回答间的重复性评价采用 3 级李克特量表，1 代表 2 次回答间不同之处多于70%，2代表2次回答不同比例

表 1　本研究中确定的 41 个问题分类及描述
Table 1　Classification and description of 41 questions in this study

序号	分类	标题
1	危险因素	DR 的危险因素有哪些？
2	危险因素	DR 常见/高危人群是哪些？
3	危险因素	糖尿病一定会发生 DR 吗？
4	危险因素	DR 的严重程度与年龄有关系吗？
5	危险因素	确诊糖尿病后，为了预防 DR，应该做些什么？
6	危险因素	1 个患高血压、糖尿病和高血脂的三高患者，控制什么可以减缓 DR 发展？
7	疾病诊断	糖尿病诊断依据为？
8	疾病诊断	什么是视网膜内层结构紊乱？
9	疾病诊断	DRCR. net 定义持续性黄斑水肿为？
10	疾病诊断	什么是糖尿病黄斑水肿？
11	疾病诊断	什么是 DR？
12	疾病诊断	如何诊断 DR？
13	症状及分期	DR 分为哪几期？
14	症状及分期	患有 DR 会有哪些症状？
15	症状及分期	我的视力很好，为什么医生还说我有 DR？
16	症状及分期	DR 会导致青光眼吗？
17	症状及分期	DR 会不会致盲？
18	症状及分期	DR 可逆吗？
19	筛查及检查	糖尿病患者眼底筛查的起始时间？
20	筛查及检查	DR 患者的筛查频率？
21	筛查及检查	什么是荧光素眼底血管造影检查？
22	筛查及检查	荧光素眼底血管造影检查有一定风险，我可不可以不做？
23	治疗及预后	糖尿病黄斑水肿能治好吗？
24	治疗及预后	在我国，治疗重度 NPDR 和 PDR 的首要方法和“金标准”？
25	治疗及预后	DR 行激光治疗的时机是什么？
26	治疗及预后	激光治疗对 DR 的治疗效果如何？
27	治疗及预后	激光对于糖尿病黄斑水肿的治疗价值？
28	治疗及预后	DR 行玻璃体切割术的适应证包括？
29	治疗及预后	玻璃体切割手术治疗 DR 和糖尿病黄斑水肿的推荐程度？
30	治疗及预后	DR 患者什么情况下会行玻璃体腔注药治疗？
31	治疗及预后	玻璃体切割手术联合抗 VEGF 药物治疗对 DR 患者的效果？
32	治疗及预后	抗 VEGF 药物和激光治疗对 DR 的效果如何？
33	治疗及预后	DR 怎么治疗？
34	治疗及预后	DR 早期，可以用药物控制吗？
35	治疗及预后	全视网膜激光光凝治疗是什么？
36	治疗及预后	DR 做完激光治疗后视力能提高吗？
37	治疗及预后	玻璃体切割手术是什么？
38	治疗及预后	DR 患者做玻璃体切割手术的风险大吗？
39	治疗及预后	DR 做手术对血糖和血压有什么要求？
40	治疗及预后	DR 激光治疗后应该注意些什么？
41	治疗及预后	DR 患者接受玻璃体切割术术后注意事项有哪些？

注：DR：糖尿病视网膜病变；NPDR：非增生性糖尿病视网膜病变；PDR：增生性糖尿病视网膜病变；VEGF：血管内皮生长因子

Note: DR: diabetic retinopathy; NPDR: non-proliferative diabetic retinopathy; PDR: proliferative diabetic retinopathy; VEGF: vascular endothelial growth factor

为 30%～70%，3 代表 2 次回答基本相同，不同之处少于 30%。

此外，对于每个回答，评估者均需选择更为推荐人工回答（1.1 问题设计及来源）还是 LLM 回答。为了阐明 LLM 获取信息的潜在局限性和风险，评分者对错误回答提供了解释。

本研究选择 3 个开源 LLM，分别为文心一言 3.5 版本、Kimi chat 和通义千问进行测评，研究共分 2 步，首先选择研究者公认难度较低的 5 个题目（问题 7、11、12、19、20）对这 3 个 LLM 进行测评，然后选择第 1 步得分最高的 LLM 接受题库中 41 个题目的完整评价。

1.3　统计学方法

采用 SPSS 26.0 统计学软件进行统计分析。计量资料数据经 Shapiro-Wilk 检验证实不符合正态分布，以 $M(Q_1, Q_3)$ 表示，2 个组间各指标比较采用 Wilcoxon 秩和检验。相关性系数计算采用 Spearman 相关性分析，3 位评估者间的一致性评价采用组内相关系数（interclass correlation coefficient，ICC）。计数资料数据以频数和百分比表示。$P<0.05$ 为差异有统计学意义。

2　结果

2.1　3 个 LLM 各指标评分比较

将 5 个问题共 10 次回答的评分整合，在 3 个 LLM 中，Kimi chat 表现最佳，其次为文心一言 3.5，最后为通义千问。Kimi chat、文心一言 3.5、通义千问准确性评分为 6 分的比例分别为 90%、70%和 20%，完整性评分为 3 分的比例分别为 90%、60%和 20%，重复性评分为 3 分的比例分别为 100%、60%和 60%（图 1）。

2.2　Kimi chat 与人工回复比较及相关指标分析

Kimi chat 回答的字数为 505（386，600）个，人工回复字数为 106（70，202）个，Kimi chat 回答字数比人工回答字数多，差异有统计学意义（$Z=-7.866$，$P<0.001$）。Kimi chat 回答字数与准确性评分无明显相关性（$r_s=-0.044$，$P=0.492$），与完整性评分呈正相关（$r_s=0.239$，$P<0.001$）。

3 位评估者 2 次回答准确性和完整性的 ICC 均高于 0.700，其中重复性评分的一致性最高，为 0.853；其次为第 1 次回答的完整性，为 0.771（表 2）。

Kimi chat 2 次应答准确性评价中整体基本正确或完全正确（准确性≥5 分）的比例为 87.0%（214/246）；各类问题的准确性评价，即基本正确或完全正确的比例由高到低依次为筛查及检查（24/24，100%）、危险因素（34/36，94.4%）、症状及分期（32/36，88.9%）、治疗及预后（99/114，86.8%）和疾病诊断（28/36，77.8%）。治疗及预后类问题中有 2 个问题（占 3.3%）被评估者 1 评价为完全不正确。完整性评价中，整体充分或完全回答（完整性≥2 分）的比例为 98.0%（241/246）；各类问题的完整性评价，即充分或完全回答的比例由高到低依次为筛查及检查（24/24，100%）、症状及分期（36/36，100%）、治疗及预后（112/114，98.2%）、危险因素（35/36，97.2%）和疾病诊断（34/36，94.4%）。疾病诊断类及危险因素类各有 1 个问题被评估者 2 评为 1 分。整体 2 次回答间重复性在 70%以上的比例为 78.5%（193/246）；各类问题的重复性评价，即 2 次回答间重复性在 70%以上的比例由高到低分别为筛查及检查（24/24，100%）、诊断（32/36，

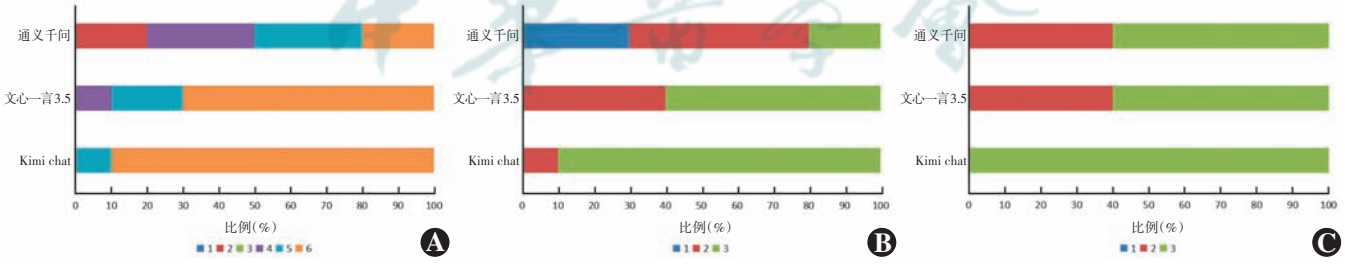

图 1　3 个 LLM 回答效能分布堆积条形图　A：准确性评价　1 表示完全错误，2 表示错误多于正确，3 表示正确和错误元素相等，4 表示正确多于错误，5 表示几乎全部正确，6 表示完全正确　B：完整性评价　1 代表不完整的回答，只涉及问题的某些方面，有重要部分缺失或不完整；2 代表充分的回答，涉及问题的所有方面，并提供了完整性所需的最少信息；3 代表全面的回答，涵盖了问题的所有方面，并提供了超出预期的额外信息或背景　C：重复性评价　1 代表 2 次回答间不同之处多于 70%，2 代表 2 次回答不同比例在 30%～70%，3 代表 2 次回答基本相同，不同之处少于 30%

Figure 1　Stacked bar chart showing the efficacy distribution of the three LLM responses　A: Accuracy evaluation　1 for complete error, 2 for more error than correctness, 3 for equal elements of correctness and error, 4 for more correctness than error, 5 for almost all correctness, 6 for complete correctness　B: Completeness evaluation　1 for an incomplete answer that covered only some aspects of the question with significant portions missing or incomplete, 2 for a full answer that covered all aspects of the question and provides the minimum information necessary for completeness, and 3 for a comprehensive answer that covered all aspects of the question and provides additional information or context beyond that which was anticipated　C: Inter-response repeatability evaluation　1 meant that more than 70% of the responses differed, 2 meant that 30% to 70% of the responses differed, and 3 meant that the responses were essentially identical, with less than 30% of the responses differing

表 2　3 位评估者评分一致性评价
Table 2　Evaluation of the consistency of the three evaluators´ scoring

项目	ICC
第 1 次回答准确性	0. 701
第 1 次回答完整性	0. 771
第 2 次回答准确性	0. 728
第 2 次回答完整性	0. 737
2 次回答间重复性	0. 853

注：ICC：组内相关系数
Note：ICC：interclass correlation coefficient

88. 9%）、危险因素（30/36，83. 3%）、症状及分期（28/36，77. 8%）和治疗及预后（79/114，69. 3%）。危险因素类第 4 题被评估者 2 评为 1 分（图 2）。表 3 为所有回答中涉及的错误表述及分析。

在全部问题中，评估者选择 Kimi chat 回答的比例为 69. 5%（171/246）（表 4）。未选择的原因包括 Kimi chat 回答未能给出特征性答案、题目与答案契合度较差、包含过多无关信息、对医疗专业度要求较高的问题回答欠缺等。

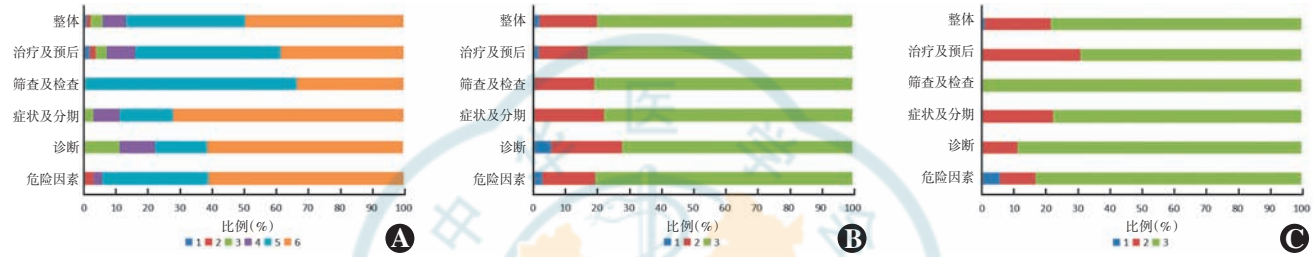

图 2　Kimi chat 2 次回答效能分布堆积条形图　A：准确性评价　1 表示完全错误，2 表示错误多于正确，3 表示正确和错误元素相等，4 表示正确多于错误，5 表示几乎全部正确，6 表示完全正确　B：完整性评价　1 代表不完整的回答，只涉及问题的某些方面，有重要部分缺失或不完整；2 代表充分的回答，涉及问题的所有方面，并提供了完整性所需的最少信息；3 代表全面的回答，涵盖了问题的所有方面，并提供了超出预期的额外信息或背景　C：重复性评价　1 代表 2 次回答间不同之处多于 70%，2 代表 2 次回答不同比例在 30%～70%，3 代表 2 次回答基本相同，不同之处少于 30%

Figure 2　Stacked bar chart showing Kimi chat two-response efficacy distribution　A：Accuracy evaluation　1 for complete error, 2 for more error than correctness, 3 for equal elements of correctness and error, 4 for more correctness than error, 5 for almost all correctness, 6 for complete correctness　B：Completeness evaluation　1 for an incomplete answer that covered only some aspects of the question with significant portions missing or incomplete, 2 for a full answer that covered all aspects of the question and provided the minimum information necessary for completeness, and 3 for a comprehensive answer that covered all aspects of the question and provided additional information or context beyond that which was anticipated　C：Inter-response repeatability evaluation　1 for more than 70% difference between the two answers, 2 for 30% to 70% difference between the two answers, 3 for basically the same two answers with differences of less than 30%

表 3　各分类回答中错误信息汇总
Table 3　Summary of error information in answers by categories

分类	题目	错误描述	错误分析
症状与分期	DR 会导致青光眼吗？	玻璃体切割术后容易发生新生血管性青光眼	该术式是治疗该疾病的方法而非危险因素
筛查及检查	什么是荧光素眼底血管造影检查？	然而，存在一些禁忌证，包括严重心肝肾疾病患者、对造影剂有过敏史者、眼部屈光间质明显混浊者、青光眼患者以及妊娠期妇女等，这些患者不适宜进行荧光素眼底血管造影检查	青光眼不是 FFA 的绝对禁忌证，可能会误导患者
筛查及检查	荧光素眼底血管造影检查有一定的风险，我可不可以不做？	不宜扩大瞳孔的疾病患者不适宜进行造影检查	不宜扩大瞳孔的患者可以进行造影检查，而且该句话对问题的针对性不强
治疗及预后	玻璃体切割手术是什么？	玻璃体切割手术是治疗飞蚊症的有效措施，尽管存在一定风险，但可以明显提高视力，改善症状，提高患者的生活质量。对于有症状且无后部玻璃体脱离的飞蚊症患者，进行彻底的玻璃体切割术可以降低飞蚊症复发的风险，但可能会增加医源性视网膜裂孔的发生率	答非所问
治疗及预后	DR 患者做玻璃体切割手术的风险大吗？	有症状的飞蚊症患者往往愿意承担一定的手术风险，因为手术能明显提高患者视力和生活质量	答非所问
治疗及预后	DR 行玻璃体切割术的适应证包括？	妊娠合并 DR 的治疗：对于女性糖尿病患者，妊娠会加速 DR 的发生和发展，激光光凝术可用于治疗孕期重度 NPDR 和 PDR。此外，越来越多研究开始关注神经保护在 DR 治疗中的重要作用	本条不是适应证

注：DR：糖尿病视网膜病变；NPDR：非增生性糖尿病视网膜病变；PDR：增生性糖尿病视网膜病变；FFA：荧光素眼底血管造影
Note：DR：diabetic retinopathy；NPDR：non-proliferative diabetic retinopathy；PDR：proliferative diabetes retinopathy；FFA：fundus fluorescein angiography

表 4　人工回答与 Kimi chat 回答选择比例[n(%)]
Table 4　Manual response selection and kimi chat response ratio (n[%])

分类	人工回答	Kimi chat 回答
危险因素	6(16.7)	30(83.3)
疾病诊断	10(27.8)	26(72.2)
症状及分期	7(19.4)	29(80.6)
筛查及检查	8(33.3)	16(66.7)
治疗及预后	44(38.6)	70(61.4)

3　讨论

本研究比较了 3 个国内开源 LLM 回答 DR 患者常见诊疗问题时的效能，与文心一言 3.5 及通义千问相比，Kimi chat 展现了较优的准确性、完整性和可重复性。在对全部题目进行评价后发现，虽然 Kimi chat 存在一定的局限性，但整体而言表现良好，展现了扎实的医学知识，这些发现首次提供了国内开源 LLM 作为 DR 患者健康虚拟助手的可行性评价结果。

本研究中 Kimi chat 回答 DR 诊疗问题时应答字数更长，更为详细和全面，完整性更好且条理清晰，在 69.5%的问题中评估人员更推荐其产生的回答，与既往关于 ChatGPT 的研究结果相似[18,24]。但值得注意的是，回答的字数长并不代表回答的准确性高，应答字数过多也会出现 LLM 回答问题较为机械繁琐、文不对题、包含过多与 DR 相关但与题目无关内容的现象，导致准确性评分较低。由于评分的主观性较强，不同评估者对同一答案的评分亦存在不同[25]，除 2 次回答重复性和第 1 次回答完整性的一致性较好外，其他一致性均为中等水平。为了综合评价所有评估者的评分，本研究将 3 位评估者对 41 个问题 2 次回答的 246 个评分合并进行统一评价。与 Potapenko 等[20]关于眼部疾病的研究结果一致，在不同分类的问题中，Kimi chat 表现出了不完全一致的准确性、完整性与重复性，在如“DRCR.net 如何定义持续性黄斑水肿”专业度较高的问题上和“玻璃体切割手术联合抗血管内皮生长因子药物治疗对 DR 的效果”涉及治疗选择等需要临床医师专业判断的问题上表现欠佳，但回答疾病定义、分期、筛查频率、常见危险因素等基础性问题时表现出色，其内容有助于患者全面了解疾病知识。“幻觉”现象，即 LLM 可以生成令人信服但完全错误的答案[21-22,26]，在有关 ChatGPT、Google Bard、OcularBERT 等多个研究中均有提及[14,18,25]，目前仍缺少有效方法确定模型的不确定性。与其他 LLM 相同，Kimi chat 的应答中同样存在错误描述，如将青光眼列为荧光素眼底血管造影检查的绝对禁忌证，这可能会误导患者，引发不必要的医疗矛盾，但整体而言，未见会对患者产生重大伤害的错误信息。

多项研究对 ChatGPT 应用于医疗决策、医学知识问答、医学写作等领域进行了评估[27]，如 Suárez 等[25]参考西班牙口腔外科学会的口腔外科实践文件向 ChatGPT4.0 提出 30 个口腔外科问题，结果显示最终准确率为 71.7%。Yeo 等[19]评估了 ChatGPT3.5 在回答 164 个有关于肝硬化和肝癌患者常见问题的准确性，结果显示准确率分别为 79.1% 和 74%，ChatGPT3.5 在基础知识、生活方式和治疗方面的表现优于诊断和预防领域。Gilson 等[28]提出 ChatGPT 在美国医师资格考试中的表现相当于 3 年级医学生的及格水平。Kimi chat 在回答 DR 相关问题时准确性为 87%，完整性为 98%，可重复性为 78.5%，证实了其与 ChatGPT 相近甚至更优的效能。在已经发表的大多数研究中，LLM 均表现出了较大的应用潜力，但关于其进一步应用的问题亦应引起注意[27,29-30]。首先，LLM 处理并存储包括患者在聊天框中输入的个人详细信息和医疗记录在内敏感的医疗信息，如何确保这些数据的隐私保护和安全性，避免未经授权的访问、数据泄露或身份盗窃至关重要[31-32]。其次，如何规范 LLM 的使用并制定在不同临床环境中的使用指南值得重视[33]。最后，考虑到 LLM 提供的不正确或误导性信息可能对医疗环境产生负面影响，且由于 LLM 训练集的局限性，其提供的答案可能不是最新的，因此，应当持续不断评估 LLM 提供的健康信息的准确性。

LLM 有可能彻底改变患者获取健康信息的方式，尽管其无法完全取代临床医师在医疗决策方面的作用，但仍可帮助患者更好地了解疾病知识，预防疾病发展，这将有助于缩短就诊时间，使医疗从业者将时间投入到更复杂的诊疗活动中。国内开源 LLM 高度的可及性、易用性和开源性对于改善经济困难、文化水平较低、偏远地区就医不便患者疾病知识不足的现象具有重要意义，可以改善医疗资源不均现象，提高社会公平性。同时，患者用自然语言提出问题，并迅速得到准确，甚至有同理心的回答，这也可能有助于改善患者体验，进而改善求医行为，包括治疗依从性和随访的依从性[24,34]。值得注意的是，患者需要意识到，他们是在与聊天机器人而不是医护人员进行沟通，必须向他们提供关于 LLM 局限性的信息且强调最终仍应以医疗人员的解释为主。

本研究存在以下局限性：(1) 目前国内已有十余个公开发布的开源 LLM，本研究中仅选用 3 个 LLM 进

行测试，且仅对 Kimi chat 进行全面评价，这限制了其结果的外推性，但作为首个本领域的研究，这一结果仍可以为医疗工作者和 DR 患者应用国内 LLM 提供一定参考及依据。（2）对 LLM 的评价有其固有的局限性，评价是主观的，不同专家对回答的准确性可能有不同的理解。本研究通过选择眼科专科三甲医院副主任医师及以上职称的眼底病医师，确保了评估者的专业性，其次综合分析 3 位评价者的准确性所占百分比而非强制统一不同研究者间的分歧。

综上所述，Kimi chat 在回答 DR 患者诊疗问题中具有较高的准确性、完整性和可重复性，尽管无法取代临床医师在医疗决策方面的作用，且存在“幻觉”现象，但考虑到其高度的可及性、易用性和开源性，在向患者提供关于 LLM 局限性的信息后仍可以帮助患者更好地了解疾病知识和改善求医行为，并有助于缓解医疗资源分配不均，提高社会公平性，具有良好的应用前景。

利益冲突　所有作者均声明不存在利益冲突

作者贡献声明　高飞：研究实施、数据整理、统计分析、文章撰写；高雪：研究实施、收集题库、数据整理、文章撰写；邵彦、任新军、刘勃实：题库设计、应答评估；焦明菲：题库设计；李筱荣：对文章的知识性内容作批评性审阅；刘巨平：选题及研究设计、对文章的知识性内容作批评性审阅及定稿

参考文献

[1] Teo ZL, Tham YC, Yu M, et al. Global prevalence of diabetic retinopathy and projection of burden through 2045: systematic review and meta-analysis[J]. Ophthalmology, 2021, 128(11): 1580-1591. DOI: 10.1016/j.ophtha.2021.04.027.

[2] Wong TY, Cheung CM, Larsen M, et al. Diabetic retinopathy[J/OL]. Nat Rev Dis Primers, 2016, 2: 16012[2024-07-10]. https://pubmed.ncbi.nlm.nih.gov/27159554/. DOI: 10.1038/nrdp.2016.12.

[3] Yang L, Wu Q, Hao Y, et al. Self-management behavior among patients with diabetic retinopathy in the community: a structural equation model[J]. Qual Life Res, 2017, 26(2): 359-366. DOI: 10.1007/s11136-016-1396-1.

[4] 中华医学会眼科学分会眼底病学组，中国医师协会眼科医师分会眼底病学组. 我国糖尿病视网膜病变临床诊疗指南（2022 年）[J]. 中华眼底病杂志, 2023, 39(2): 99-124. DOI: 10.3760/cma.j.cn511434-20230110-00018.
Fundus Disease Group of Ophthalmological Society of Chinese Medical Association, Fundus Disease Group of Ophthalmologist Branch of Chinese Medical Doctor Association. Evidence-based guidelines for diagnosis and treatment of diabetic retinopathy in China (2022)[J]. Chin J Ocul Fundus Dis, 2023, 39(2): 99-124. DOI: 10.3760/cma.j.cn511434-20230110-00018.

[5] Moinul P, Barbosa J, Qian J, et al. Does patient education improve compliance to routine diabetic retinopathy screening? [J]. J Telemed Telecare, 2020, 26(3): 161-173. DOI: 10.1177/1357633X18804749.

[6] Khurana RN. Online information for diabetic retinopathy-often a missed opportunity for patient education[J/OL]. JAMA Ophthalmol, 2019, 137(11): 1246[2024-07-10]. https://pubmed.ncbi.nlm.nih.gov/31436801/. DOI: 10.1001/jamaophthalmol.2019.3146.

[7] Tang YH, Pang SM, Chan MF, et al. Health literacy, complication awareness, and diabetic control in patients with type 2 diabetes mellitus[J]. J Adv Nurs, 2008, 62(1): 74-83. DOI: 10.1111/j.1365-2648.2007.04526.x.

[8] Peng Y, Guo X, Liu J, et al. Incidence and risk factors for diabetic retinopathy in the communities of Shenzhen[J]. Ann Palliat Med, 2021, 10(1): 615-624. DOI: 10.21037/apm-20-2526.

[9] Chen Y, Jiang Y, Yao X, et al. Proportion and risk factors of diabetic retinopathy by stage in less-developed rural areas of Hunan province of China: a multi-site cross-sectional study[J/OL]. BMC Public Health, 2022, 22(1): 1871[2024-07-10]. https://pubmed.ncbi.nlm.nih.gov/36207704/. DOI: 10.1186/s12889-022-14232-3.

[10] Zhang X, Cotch MF, Ryskulova A, et al. Vision health disparities in the United States by race/ethnicity, education, and economic status: findings from two nationally representative surveys[J]. Am J Ophthalmol, 2012, 154(6 Suppl): S53-62. DOI: 10.1016/j.ajo.2011.08.045.

[11] Kim S, Love F, Quistberg DA, et al. Association of health literacy with self-management behavior in patients with diabetes[J]. Diabetes Care, 2004, 27(12): 2980-2982. DOI: 10.2337/diacare.27.12.2980.

[12] Abd-Alrazaq A, AlSaad R, Alhuwail D, et al. Large language models in medical education: opportunities, challenges, and future directions[J/OL]. JMIR Med Educ, 2023, 9: e48291[2024-07-10]. https://pubmed.ncbi.nlm.nih.gov/37261894/. DOI: 10.2196/48291.

[13] Cox A, Seth I, Xie Y, et al. Utilizing ChatGPT-4 for providing medical information on blepharoplasties to patients[J]. Aesthet Surg J, 2023, 43(8): NP658-NP662. DOI: 10.1093/asj/sjad096.

[14] Vaira LA, Lechien JR, Abbate V, et al. Accuracy of ChatGPT-generated information on head and neck and oromaxillofacial surgery: a multicenter collaborative analysis[J]. Otolaryngol Head Neck Surg, 2024, 170(6): 1492-1503. DOI: 10.1002/ohn.489.

[15] Huang C, Chen L, Huang H, et al. Evaluate the accuracy of ChatGPT's responses to diabetes questions and misconceptions[J/OL]. J Transl Med, 2023, 21(1): 502[2024-07-11]. https://pubmed.ncbi.nlm.nih.gov/37495984/. DOI: 10.1186/s12967-023-04354-6.

[16] Kusunose K, Kashima S, Sata M. Evaluation of the accuracy of ChatGPT in answering clinical questions on the Japanese Society of Hypertension Guidelines[J]. Circ J, 2023, 87(7): 1030-1033. DOI: 10.1253/circj.CJ-23-0308.

[17] Saibene AM, Allevi F, Calvo-Henriquez C, et al. Reliability of large language models in managing odontogenic sinusitis clinical scenarios: a preliminary multidisciplinary evaluation[J]. Eur Arch Otorhinolaryngol, 2024, 281(4): 1835-1841. DOI: 10.1007/s00405-023-08372-4.

[18] Cheong KX, Zhang C, Tan TE, et al. Comparing generative and retrieval-based chatbots in answering patient questions regarding age-related macular degeneration and diabetic retinopathy[J]. Br J Ophthalmol, 2024, 108(10): 1443-1449. DOI: 10.1136/bjo-2023-324533.

[19] Yeo YH, Samaan JS, Ng WH, et al. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma[J]. Clin Mol Hepatol, 2023, 29(3): 721-732. DOI: 10.3350/cmh.2023.0089.

[20] Potapenko I, Boberg-Ans LC, Stormly Hansen M, et al. Artificial intelligence-based chatbot patient information on common retinal diseases using ChatGPT[J]. Acta Ophthalmol, 2023, 101(7): 829-831. DOI: 10.1111/aos.15661.

[21] Thirunavukarasu AJ, Hassan R, Mahmood S, et al. Trialling a large language model (ChatGPT) in general practice with the applied knowledge test: observational study demonstrating opportunities and limitations in primary care[J/OL]. JMIR Med Educ, 2023, 9: e46599[2024-07-11]. https://pubmed.ncbi.nlm.nih.gov/37083633/. DOI: 10.2196/46599.

[22] Athaluri SA, Manthena SV, Kesapragada V, et al. Exploring the boundaries of reality: investigating the phenomenon of artificial intelligence hallucination in scientific writing through ChatGPT references[J/OL]. Cureus, 2023, 15(4): e37432[2024-07-12]. https://pubmed.ncbi.nlm.nih.gov/37182055/. DOI: 10.7759/cureus.37432.

[23] Tonsaker T, Bartlett G, Trpkov C. Health information on the Internet: gold mine or minefield? [J]. Can Fam Physician, 2014, 60(5): 407-408.

[24] 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[J]. JAMA Intern Med, 2023, 183(6): 589-596. DOI: 10.1001/jamainternmed.2023.1838.

[25] Suárez A, Jiménez J, Llorente de Pedro M, et al. Beyond the Scalpel:

assessing ChatGPT's potential as an auxiliary intelligent virtual assistant in oral surgery[J]. Comput Struct Biotechnol J, 2024, 24: 46-52. DOI: 10.1016/j.csbj.2023.11.058.
[26] Masters K. Medical teacher's first ChatGPT's referencing hallucinations: lessons for editors, reviewers, and teachers[J]. Med Teach, 2023, 45(7): 673-675. DOI: 10.1080/0142159X.2023.2208731.
[27] Liu J, Wang C, Liu S. Utility of ChatGPT in clinical practice[J/OL]. J Med Internet Res, 2023, 25: e48568[2024-07-12]. https://pubmed.ncbi.nlm.nih.gov/37379067/. DOI: 10.2196/48568.
[28] Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment[J/OL]. JMIR Med Educ, 2023, 9: e45312[2024-07-12]. https://pubmed.ncbi.nlm.nih.gov/36753318/. DOI: 10.2196/45312.
[29] Adhikari K, Naik N, Hameed BZ, et al. Exploring the ethical, legal, and social implications of ChatGPT in urology[J]. Curr Urol Rep, 2024, 25(1): 1-8. DOI: 10.1007/s11934-023-01185-2.
[30] Biswas S. ChatGPT and the future of medical writing[J/OL]. Radiology, 2023, 307(2): e223312[2024-07-12]. https://pubmed.ncbi.nlm.nih.gov/36728748/. DOI: 10.1148/radiol.223312.
[31] Zhang J, Zhang ZM. Ethics and governance of trustworthy medical artificial intelligence[J/OL]. BMC Med Inform Decis Mak, 2023, 23(1): 7[2024-07-13]. https://pubmed.ncbi.nlm.nih.gov/36639799/. DOI: 10.1186/s12911-023-02103-9.
[32] Masters K. Ethical use of artificial intelligence in health professions education: AMEE Guide No. 158[J]. Med Teach, 2023, 45(6): 574-584. DOI: 10.1080/0142159X.2023.2186203.
[33] Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations[J/OL]. Front Artif Intell, 2023, 6: 1169595[2024-07-13]. https://pubmed.ncbi.nlm.nih.gov/37215063/. DOI: 10.3389/frai.2023.1169595.
[34] Rasu RS, Bawa WA, Suminski R, et al. Health literacy impact on national healthcare utilization and expenditure[J]. Int J Health Policy Manag, 2015, 4(11): 747-755. DOI: 10.15171/ijhpm.2015.151.

（收稿日期：2024-08-20　修回日期：2024-10-22）

（本文编辑：刘艳　施晓萌）

·病例报告·

茄葡柄霉致真菌性角膜溃疡诊疗 1 例

李鹏　张莉　李晓凤　杜满　鹿秀海

山东第一医科大学附属眼科医院　山东第一医科大学（山东省医学科学院）　山东省眼科研究所

山东省眼科学重点实验室-省部共建国家重点实验室培育基地，济南 250021

通信作者：鹿秀海，Email：xiuhailu@163.com

基金项目：山东省医学科学院院级项目（2018-20）

Diagnosis and treatment of fungal corneal ulcer caused by *Stemphylium solani*: a case report

Li Peng, Zhang Li, Li Xiaofeng, Du Man, Lu Xiuhai

Eye Hospital of Shandong First Medical University, State Key Laboratory Cultivation Base, Shandong Provincial Key Laboratory of Ophthalmology, Shandong Eye Institute, Shandong First Medical University & Shandong Academy of Medical Sciences, Jinan 250021, China

Corresponding author: Lu Xiuhai, Email: xiuhailu@163.com

Fund program: Science and Technology Project of Shandong Academy of Medical Sciences (2018-20)

DOI: 10.3760/cma.j.cn115989-20200317-00178

患者男，60 岁，农民，2017 年 10 月因无明显诱因出现左眼眼红、干涩，伴溢泪和视力下降于当地医院就诊，无视物变形等不适症状，既往诊断不详，予以左氧氟沙星滴眼液、普拉洛芬滴眼液等药物治疗，效果欠佳，遂至山东第一医科大学附属眼科医院就诊，否认外伤史。全身检查未见明显异常。眼部检查：左眼视力手动/眼前。左眼眼睑轻度肿胀，混合性出血，角膜中央偏鼻侧可见约 5 mm×3 mm 不规则灰黄色溃疡灶（图 1），可见伪足和苔被，角膜水肿，KP（+），前房深度正常，前房积脓，眼内结构窥不清。右眼结膜无明显充血，角膜透明，瞳孔圆，直径约 3 mm，对光反射存在，晶状体密度高。激光扫描共焦显微镜检查显示，左眼可见大量菌丝分布（图 2），角膜内皮可见大量炎性细胞，内皮细胞无法成像。入院后使用无菌 15°手术刀刮取病灶接种沙保弱、血琼脂平板和增菌培养基进行真菌、细菌培养，并分别行 10%氢氧化钾涂片和革兰染色。10%氢氧化钾涂片镜检可见大量断裂菌丝（图 3）；革兰染色镜检可见少量中性粒细胞，少量菌丝（图 4）。术中取病变角膜组织进行组织病理学检查，可见基质炎细胞浸润，PAS 染色可见真菌菌丝（图 5）。采用沙保弱培养基在 28 ℃条件下进行培养，5 d 可见丝状真菌生长，初起为灰白色菌落（图 6），菌落生长缓慢；14 d 后变为灰褐色（图 7），未见孢子产生；延长培养时间，26 d 后显微镜下可见淡褐色分生孢子，呈长椭圆形或椭圆形，有隔膜（图 8）。DNA 序列分析显示，*ITS* 基因扩增可见单一清晰条带，测序结果经 BLAST 比对显示，与茄葡柄霉相似度最高，为 100%，初步鉴定为茄葡柄霉。因患者角膜溃疡浸润较深，单纯药物治疗效果欠佳，入院后予以氟康唑氯化钠注射液 0.2 g 静脉滴注，使用氟康唑滴眼液、那他霉素滴眼液、10 mg/ml 伏立康唑滴眼液、左氧氟沙星滴眼液点眼，球周阻滞麻醉下行左眼部分板层角膜移植术，手术顺利。术后继续使用氟康唑滴眼液、左氧氟沙星滴眼液、双氯芬酸酸钠滴眼液、氧氟沙星眼膏、重组牛碱性成纤维细胞生长因子眼用凝胶点眼。术后 6 d，角膜植片透明（图 9），愈合较好；术后 1 个月复诊，左眼视力 0.5；随访至 2018 年 11 月，未见复发，术后恢复较好。