·数智医学教育·

中文大语言模型在放射医学中级专业技术考试中的正确率和答案解析质量

钟京谕[1]　星月[1]　胡扬帆[1]　闵庆华[1]　朱才松[1]　施丹丹[2]　范晓彧[3]　褚敬申[4]　张欢[5]　姚伟武[1]

[1]上海交通大学医学院附属同仁医院影像科，上海　200336；[2]上海交通大学医学院附属同仁医院教学科，上海　200336；[3]上海交通大学医学院附属同仁医院院长办公室，上海　200336；[4]上海交通大学医学院附属瑞金医院《诊断学理论与实践》编辑部，上海　200025；[5]上海交通大学医学院附属瑞金医院放射科，上海　200025

通信作者：姚伟武，Email：yaoweiwuhuan@163.com

【摘要】 目的　对比中文大语言模型（large language model，LLM）和医师在放射医学中级专业技术考试中的正确率，评估中文LLM提供的答案解析质量。**方法**　本研究分层随机抽取100题精选习题组成测试用题，网页访问文心一言要求其给出正确答案及解析，邀请15名不同经验的放射科医师参与测试。比较中文LLM及医师的答题正确率，2名医师使用5分制的半定量方法评估答案解析质量。**结果**　文心一言的正确率为60.00%，低于医师的正确率[67.00%（64.00%，73.00%）]，两者之间的差异具有统计学意义（W=2.47，P=0.013）。文心一言提供的答案解析的字数为（196.44 ± 99.25）字，正确和错误的答案解析的字数分别为（211.03 ± 107.53）字和（174.55 ± 81.84）字，差异无统计学意义（t=1.82，P=0.072）。正确答案的解析在质量评估中获得1~4分的试题数量分别为3题、9题、12题和36题，无获得5分的答案解析。**结论**　中文LLM具备一定的医学知识和临床推理能力，能够辅助带教医师开展教学，但尚不能单独辅导住院医师学习，缺乏引导式、启发式教学的能力。

【关键词】 大语言模型；　住院医师规范化培训；　考核；　放射科

【中图分类号】 R445

基金项目：上海交通大学2024年“生成式人工智能+教育”专项基金（CTLD24A0021）；2024年秋上海交通大学“人工智能+教育”专项基金（CTLD24A0081）；上海交通大学“数智课程”建设项目[医学影像学（预防医学专用）]；上海交通大学医学院2024年本科核心课程建设项目[医学影像学（预防医学专业）]；上海交通大学医学院2023年度本科教材建设项目[医学影像学（供预防医学专业用）]；上海市同仁医院萤星计划（TRYXJH28）

DOI: 10.3760/cma.j.cn116021-20231111-01961

Accuracy and quality of answer reasoning of Chinese large language model in Chinese middle level professional qualification examination of radiology

Zhong Jingyu[1], Xing Yue[1], Hu Yangfan[1], Min Qinghua[1], Zhu Caisong[1], Shi Dandan[2], Fan Xiaoyu[3], Chu Jingshen[4], Zhang Huan[5], Yao Weiwu[1]

[1]Department of Imaging, Tongren Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai 200336, China; [2]Department of Education, Tongren Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai 200336, China; [3]Dean's Office, Tongren Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai 200336, China; [4]Editorial Office of Journal of Diagnostics Concepts & Practice, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai 200025, China; [5]Department of Radiology, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai 200025, China

Corresponding author: Yao Weiwu, Email: yaoweiwuhuan@163.com

【Abstract】 Objective　To compare the accuracy of a Chinese large language model (LLM) and radiologists in Chinese middle level professional qualification examination of radiology, and evaluate the quality of answer reasoning provided by the Chinese LLM. **Methods**　In this study, 100 high-quality questions were selected using stratified random sampling to form a test set. We asked the ERNIE Bot by dialogues on the website to provide the correct answers and answer reasoning for these questions. These questions were also answered by 15 radiologists with different levels of experience. The accuracy of Chinese LLM and that of radiologists were compared. Two radiologists evaluated the quality of answer reasoning using a 5-point semi-quantitative scale. **Results**　The accuracy of ERNIE Bot was 60.00%, which was lower than

the median (interquartile) accuracy of 67.00% (64.00%, 73.00%) for radiologists, and the difference was statistically significant (W=2.47, P=0.013). The word count of the reasoning provided by Ernie Bot was (196.44 ± 99.25) words, with no significant difference in word count between correct and incorrect answer reasoning, which were (211.03 ± 107.53) words and (174.55 ± 81.84) words, respectively (t=1.82, P=0.072). Among the correct answers, the quality of reasoning was scored as follows: 1 point for 3 questions, 2 points for 9 questions, 3 points for 12 questions, and 4 points for 36 questions. No reasoning received a score of 5. **Conclusions** Chinese LLM demonstrates a certain level of medical knowledge and clinical reasoning ability, which can assist clinical teachers in educational activities. However, it is not yet able to independently tutor residents and lacks the ability of invitational and heuristic teaching.

【**Key words**】 Large language model; Standardized residency training; Assessment; Radiology

Fund program: 2024 "Generative Artificial Intelligence + Education" Fund from Shanghai Jiao Tong University (CTLD24A0021); 2024 Autumn "Artificial Intelligence + Education" Fund from Shanghai Jiao Tong University (CTLD24A0081); "Digital Intelligence Courses" Construction Project from Shanghai Jiao Tong University [Medical Imaging (For Preventive Medicine)]; 2024 Undergraduate Core Courses Construction Project from Shanghai Jiao Tong University School of Medicine [Medical Imaging (For Preventive Medicine)]; 2023 Undergraduate Textbook Construction Project from Shanghai Jiao Tong University School of Medicine [Medical Imaging (For Preventive Medicine)]; Yingxing Project from Tongren Hospital, Shanghai Jiao Tong University School of Medicine (TRYXJH28)

DOI: 10.3760/cma.j.cn116021-20231111-01961

聊天生成式预训练模型(chat generative pre-trained transformer, ChatGPT)等大语言模型(large language model, LLM)是经过海量数据预训练的深度学习算法,能够理解语言的内在规律并将知识迁移到自然语言处理任务中,目前已经可以完成识别、总结、翻译和生成文本等任务[1]。在医学领域,LLM也已经显示出处理医疗文书、辅助临床决策、赋能医学教育的应用潜力[2]。LLM已经达到通过多种英文医学专业考试的水平[3-6],但在非英语医学专业考试中的表现不一[7-9],比如ChatGPT在中国执业医师资格考试中的表现就不如临床医学五年制专业毕业生[9]。在将LLM引入医疗领域之前,充分验证其在医学知识和临床推理上的能力至关重要。英文LLM已经能通过美国放射科医师资格认证考试[6],但目前尚无研究探索过中文LLM在医学专业考试中的表现。因此,本研究将以放射医学中级专业技术考试为例,对比中文LLM和医师在高阶医学专业考试中的正确率,并评估中文LLM提供的答案解析质量,为LLM在住培教学中的深入应用提供参考。

1 对象与方法

1.1 研究对象

本研究不涉及医学研究伦理,无须知情同意。本研究使用个案研究方法[10],使用可以公开购得的放射医学中级专业技术考试精选习题作为试题来源[11]。这一考试分为4个部分,包括基础知识、专业相关知识、专业知识和专业实践能力,每一部分的考试时间均为90 min,题量均为100题。其中专业实践能力部分的试题因涉及读图被排除在外。本研究认为精选习题中的试题分布近似地模拟了真实考试的情况,因此采用分层随机抽样的方法构建了能反映试题分布比例、题量为100题的测试用题,随机数法(https://www.random.org/integers/)抽样。具体的试题抽样分布情况见表1,测试用题的字数为(83.51 ± 30.21)字。

1.2 大语言模型测试

本研究采用网页对话的方式,使用简体中文提示词测试了文心一言(文心大模型3.5版,百度)[12]。2名医师设计了提示词并进行模拟测试,2名医师分别有5年和6年的放射诊断经验以及1个月和3个月的提示词构建经验[13]。本研究用于测试的提示词结构如下:“你是一名放射科住院医师,正在准备放射科的考试,请回答下面的问题:”&[题干]&[选项A]&[选项B]&[选项C]&[选项D]&[选项E]&“请给出正确选项,并给出具体解析”。提示词和回复的示例见图1。计入提示词后,每次对话测试者提问的字数为(133.51 ± 30.21)字。

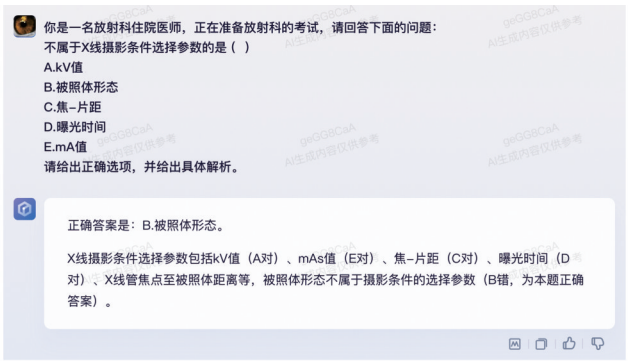
正式测试中,一名医师负责测试,在macOS 10.14.6版系统中使用Safari浏览器14.0.3版网页访问文心一言开展对话,依次给出测试用题第1~100题的提示词。另一名医师负责监督并计时,测试结束后记录了全部对话中的答案和解析。本研究于2023年9月27日20:00:00至20:20:47进行测试,总耗时20 min 47 s,期间网络状态良好,LLM未报告更新,生成过程未见明显卡顿。

1.3 放射科医师测试

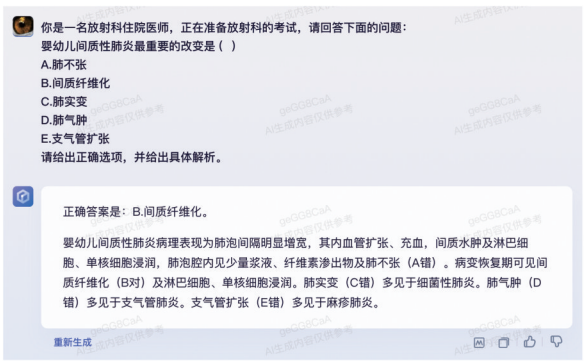
本研究邀请了15名放射科医师参与测试,要求在90 min内独立回答100道测试用题,其间不允许查阅资料。参与医师的年龄为33.00(28.00,44.00)岁;男性和女性分别为8人和7人;从事放射诊断的年

表 1　试题抽样结果分布[例(%)]

试题部分	试题亚专业	原题量	可用题量	抽样题量
基础知识	X线成像基础	63(3.39)	63(3.42)	3(3.00)
	CT成像基础	80(4.31)	80(4.35)	4(4.00)
	MRI成像基础	122(6.57)	122(6.63)	7(7.00)
	介入放射学	139(7.48)	139(7.55)	7(7.00)
	影像解剖	198(10.66)	198(10.76)	11(11.00)
	小计	602(32.40)	602(32.70)	32(32.00)
专业相关知识	神经头颈	161(8.67)	161(8.75)	9(9.00)
	胸部	90(4.84)	90(4.89)	5(5.00)
	消化	90(4.84)	90(4.89)	5(5.00)
	泌尿生殖	99(5.33)	99(5.38)	5(5.00)
	骨关节	97(5.22)	97(5.27)	5(5.00)
	小计	537(28.90)	537(29.17)	29(29.00)
专业知识	神经头颈	122(6.57)	105(5.7)	7(7.00)
	胸部	183(9.85)	183(9.94)	10(10.00)
	消化	134(7.21)	134(7.28)	7(7.00)
	泌尿生殖	156(8.4)	156(8.47)	8(8.00)
	骨关节	124(6.67)	124(6.74)	7(7.00)
	小计	719(38.70)	702(38.13)	39(39.00)
总计		1 858(100.00)	1 841(100.00)	100(100.00)



A. 第2题测试结果（本题正确答案为B，LLM回复正确答案B，记录其解析留待后续分析）



B. 第74题测试结果（本题正确答案为D，LLM回复错误答案B，并给出了错误的解析）

图 1　网页对话访问LLM测试示例

限为7.00(4.00,17.00)年;学士、硕士和博士分别为7人、7人和1人;住院医师、主治医师和副主任医师分别为5人、7人和3人;暂无亚专业者4人,神经头颈、胸部、消化和骨关节亚专业者分别为1人、4人、2人和4人。所有医师均遵守考试纪律并在规定时间内提交了答案。

1.4　解析质量评估

本研究采用半定量方法评估了正确答案的解析质量,评估标准如下:1分,解析内容存在知识、计算或逻辑错误;2分,解析内容简单重复试题内容;3分,解析内容仅解释了选择正确选项的原因;4分,解析内容解释了选择正确选项的原因以及错误选项的问题;5分,解析内容解释了试题的所有选项,还提供了引导式、启发式的教学帮助深入学习[10,14-15]。2名医师独立评估了答案解析的质量,2名医师分别有5年和6年的放射诊断经验以及1年和2年的教学经验。如有不一致,则协商达成一致;如协商不成,则由1名有21年放射诊断经验以及10年教学经验的医师判断。

1.5　统计学方法

本研究采用R语言4.1.3版(https://www.r-project.org)在Rstudio软件1.4.1106版(https://posit.co)中进行统计分析,计数资料采用频数及百分比表示;计量资料根据是否正态分布分别使用(均数±标准差)或中位数(四分位数)表示,比较使用t检验或Wilcoxon符号秩和检验。检验水准α=0.05。

2　结果

2.1　答题正确率

图2展示了LLM和医师的答题结果。LLM的正确率为60.00%(60/100),基础知识、专业相关知识和专业知识部分的正确率分别为59.38%(19/32)、

55.17%(16/29)和61.54%(24/39)。15名医师的正确率为67.00%(64.00%,73.00%),基础知识、专业相关知识和专业知识部分的正确率分别为62.50%(57.81%, 55.17%)、62.07%(55.17%, 72.41%)和74.36%(67.95%,78.21%)。不同职称医师的答题正确率见表2。医师和LLM的总体和专业知识部分答题正确率差异有统计学意义(W=2.47,P=0.013;W=2.79,P=0.005),但基础知识和专业相关知识部分的正确率差异无统计学意义(W=1.25,P=0.221;W=1.59,P=0.111)。

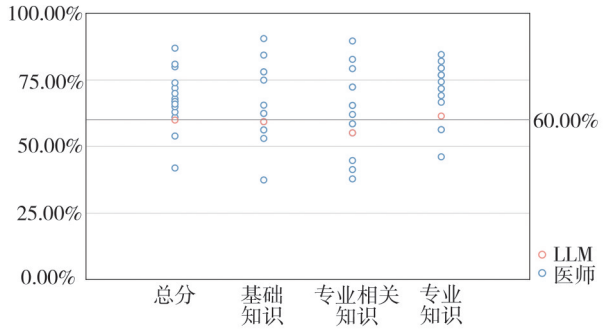


图2　LLM和医师的正确率对比

2.2　解析质量评估

答案解析的字数为(196.44 ± 99.25)字,正确和错误答案解析的字数分别为(211.03 ± 107.53)字和(174.55 ± 81.84)字,两者字数差异无统计学意义(t=1.82,P=0.072)。图3展示了LLM为正确答案提供的解析质量评估结果。评估中获得1~4分的试题数量分别为3题、9题、12题和36题,无获得5分的答案解析。

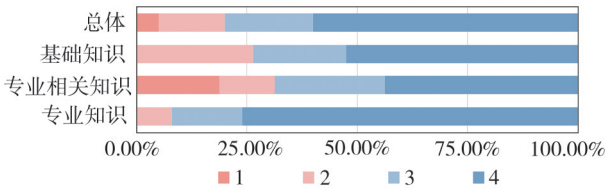


图3　LLM提供的解析质量评估结果

3　讨论

3.1　LLM具备辅助教学的潜力

人工智能已经进入医学影像学课程建设[16],但LLM在医学影像学中的作用尚未得到充分探索。本研究发现中文LLM可以通过放射医学中级专业技术考试,但水平仍不如放射科医师。中文LLM不论答案是否正确均能进行答案解析,其中正确答案的相应解析具备辅助教学的潜力,但仍缺乏引导式、启发式教学能力,尚不能单独辅导住院医师。先前的研究测试了ChatGPT-3.5在中国执业医师资格考试中的表现,结果不如临床医学五年制专业毕业生[11]。这可能与语言模型训练时所使用的数据语种有关,其在英文医学专业考试[3-6]的表现就较非英文医学专业考试[7-9]优秀且稳定。此次研究结果为开发基于中文数据的LLM提供了参考,证实了中文LLM同样具备理解海量知识并高效生成文字的能力。本次基于中文LLM的结果与英文LLM相近,均能通过中美两国相应语种的放射科医师资格考试[6]。先前的研究认为LLM能较好理解有直接关联的知识点,但在系统性理解上的表现不尽如人意[10]。然而,本研究认为中文LLM在回答和解析专业知识问题时表现良好,能在解读病例资料的基础上给出较为合理的回答,表现出了对医学知识的系统性理解和临床推理能力。这一差异可能与2次研究使用的LLM不同有关,同时本研究精心构建的提示词也可能更好地激发了LLM的智能潜力。不过,中文LLM所提供的解析目前仅能一定程度上帮助住院医师学习,其生成的内容仍需带教医师审核。

3.2　LLM的合理应用需要引导

LLM进入医学教育领域已是大势所趋,但如何在住培中利用这一工具需要师生等多方面的协作和合理引导[17]。对于住院医师而言,LLM出色的文字生成能力有很大的吸引力。这就需要引导住院医师规避对这一工具的不合理滥用,学会借助人机协同的契机,实现随时随地、个性化的学习[17]。同时,需要注意LLM可能存在的知识性错误。由于错误的医学知识可能导致致命的后果,因此目前LLM在医学教学中的应用仍需带教医师监督,从而及时纠正LLM对住院医师的误导[18]。对于带教医师而言,需要尽快接触、学习、使用LLM,尝试将其用于辅助教学,纠正其对住院医师的误导,构建全新的住培教学生态[10]。带教医师可以借助LLM强化医学课程教学和考核体系,从知道、理解的层次跃升至展示和实践的层次,最终达到培养住院医师岗位

表2　不同职称医师的答题正确率[M_d(P_{25},P_{75}),%]

职称	人数	基础知识(n=32)	专业相关知识(n=29)	专业知识(n=39)	总分(n=100)
住院医师	5	56.25(56.25,62.50)	44.83(41.38,50.00)	74.36(56.41,76.92)	65.00(54.00,67.00)
主治医师	7	62.50(60.94,76.56)	65.52(62.07,68.75)	71.79(70.51,76.92)	70.00(67.00,76.00)
副主任医师	3	65.63(59.38,78.13)	62.07(60.34,65.63)	79.49(67.95,82.05)	67.00(64.00,77.00)
总计	15	62.50(57.81,55.17)	62.07(55.17,72.41)	74.36(67.95,78.21)	67.00(64.00,73.00)

胜任力的目标[19]。对于住培基地而言，需要思考如何在日常教学中充分发挥LLM的潜力[17]，比如提供资源接入设备，开设应用培训课程以及引导相关教学研究等。同时，对于LLM的日常监管也必不可少，如何建立合理的伦理监督、管理机制和激励制度也是亟待解决的问题。

3.3 本研究的局限性

本研究存在一定的局限性：①由于医师执业资格考试及住培结业理论考核的真题难以获得，本研究使用了放射医学中级专业技术考试精选习题作为试题来源，仅能从一个侧面展示LLM在医学专业考试中的表现，LLM在医学其他专科考试中的表现仍需探索。②放射医学中级专业技术考试的难点之一为影像解读，但本研究的纸质版试题来源无法为研究提供高质量的图像。所以本研究排除了需要读图的试题，无法完全评估其医学知识和临床推理水平，未来需要应用更多高质量的图文试题评估LLM的图文综合理解和生成能力，挖掘其辅助教学的潜在价值。③本研究仅对一种中文LLM进行一次评估，不能反映所有LLM即时的最新能力水平，但本研究的结果仍具有较好的代表性。随着LLM的不断推陈出新，其能力有望不断增强。④本研究的提示词经过精心构建，尚未达到最优，也未针对LLM的解析进行深入提问。实际应用中可以针对特定问题多次给出提示词，开展多轮对话，以充分激发其潜在智能。⑤本研究提示LLM具有作为客观教学辅导手段的潜力，但远未达到人类教师的水平，仍有可能误导学生。如果LLM在未来具备接近人类教师的教学辅导能力，研究者们将设法对比这一手段与现有教学辅导手段的效果，或探索联合这一手段能否进一步提高教学效果。

志谢　感谢常军教授和桑刘睿老师对本文提出的建设性意见和指导；感谢上海交通大学医学院附属同仁医院影像科的诸位医师积极参与研究

利益冲突　所有作者声明无利益冲突

作者贡献声明　钟京谕：研究构思、资料收集、数据分析、论文撰写；星月、胡扬帆、闵庆华、朱才松：研究构思、资料收集、数据分析、论文修改；施丹丹、范晓彧、褚敬申：研究构思、论文修改；张欢、姚伟武：研究构思、论文审定、项目组织

参考文献

[1] Khurana D, Koli A, Khatter K, et al. Natural language processing: state of the art, current trends and challenges [J]. Multimed Tools Appl, 2023, 82(3): 3713-3744. DOI: 10.1007/s11042-022-13428-4.

[2] Thirunavukarasu AJ, Ting DSJ, Elangovan K, et al. Large language models in medicine [J]. Nat Med, 2023, 29(8): 1930-1940. DOI: 10.1038/s41591-023-02448-8.

[3] Gilson A, Safranek CW, Huang T, et al. Correction: 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]. JMIR Med Educ, 2024, 10: e57594. DOI: 10.2196/57594.

[4] 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. DOI: 10.1371/journal.pdig.0000198.

[5] St Huang R, Lu KJQ, Meaney C, et al. Assessment of resident and AI chatbot performance on the university of Toronto family medicine residency progress test: comparative study [J]. JMIR Med Educ, 2023, 9: e50514. DOI: 10.2196/50514.

[6] Bhayana R, Krishna S, Bleakney RR. Performance of ChatGPT on a radiology board-style examination: insights into current strengths and limitations [J]. Radiology, 2023, 307(5): e230582. DOI: 10.1148/radiol.230582.

[7] Huh S. Are ChatGPT's knowledge and interpretation ability comparable to those of medical students in Korea for taking a parasitology examination? A descriptive study [J]. J Educ Eval Health Prof, 2023, 20: 1. DOI: 10.3352/jeehp.2023.20.1.

[8] Takagi S, Watari T, Erabi A, et al. Performance of GPT-3.5 and GPT-4 on the Japanese medical licensing examination: comparison study [J]. JMIR Med Educ, 2023, 9: e48002. DOI: 10.2196/48002.

[9] Wang X, Gong Z, Wang G, et al. ChatGPT performs on the Chinese National Medical Licensing Examination [EB/OL]. (2023-02-16) [2023-10-01]. https://www.researchsquare.com/article/rs-2584079/v1.

[10] 贺樑，应振宇，王英英，等. 教育中的ChatGPT：教学能力诊断研究[J]. 华东师范大学学报（教育科学版），2023，41（7）：162-176. DOI: 10.16382/j.cnki.1000-5560.2023.07.015.

[11] 刘斯润，张水兴. 2021全国卫生专业资格技术考试题集丛书：放射医学精选习题集[M]. 北京：人民卫生出版社，2020.

[12] 百度. 文心一言[EB/OL]. (2023-08-31) [2023-09-27]. https://yiyan.baidu.com.

[13] 张华平，李林翰，李春锦. ChatGPT中文性能测评与风险应对[J]. 数据分析与知识发现，2023，7（3）：16-25. DOI: 10.11925/infotech.2096-3467.2023.0214.

[14] Lyu Q, Tan J, Zapadka ME, et al. Translating radiology reports into plain language using ChatGPT and GPT-4 with prompt learning: results, limitations, and potential [J]. Vis Comput Ind Biomed Art, 2023, 6(1): 9. DOI: 10.1186/s42492-023-00136-5.

[15] Stoehr F, Kämpgen B, Müller L, et al. Natural language processing for automatic evaluation of free-text answers-a feasibility study based on the European Diploma in radiology examination [J]. Insights Imaging, 2023, 14(1): 150. DOI: 10.1186/s13244-023-01507-5.

[16] 陈勇，张欢. 线上自适应教程在医学影像学专业型硕士研究生教学实践中的应用[J]. 诊断学理论与实践，2021，20（6）：596-598. DOI: 10.16150/j.1671-2870.2021.06.017.

[17] 倪闽景. 从学习进化的视角看ChatGPT/生成式人工智能对学习的影响[J]. 华东师范大学学报（教育科学版），2023，41（7）：151-161. DOI: 10.16382/j.cnki.1000-5560.2023.07.014.

[18] 李袁爽，张子航，王运武. 基于ChatGPT的人机共教、共育和共创模式[J]. 中国医学教育技术，2023，37（4）：383-389，396. DOI: 10.13566/j.cnki.cmet.cn61-1317/g4.202304002.

[19] 袁欢欢，全祉悦，林祎，等. 当医学课程考核遇上生成式人工智能：价值、挑战与未来展望[J]. 中国医学教育技术，2023，37（3）：254-259. DOI: 10.13566/j.cnki.cmet.cn61-1317/g4.202303002.

（收稿日期：2023-11-11）
（本文编辑：唐宗顺）