·数智医学教育·

AI与教师在医学呼吸系统试题解析撰写中的优劣势比较

姚一楠[1]　吕寒冰[2]　傅梦姣[1]　屈晶晶[1]　张培[1]　周建娅[1]
[1]浙江大学医学院附属第一医院呼吸与危重症医学科，杭州 310003；[2]浙江大学医学院，杭州 310058
通信作者：周建娅，Email：zhoujy@zju.edu.cn

【摘要】 目的 本研究旨在对比人工智能与教师在呼吸系统医学试题解析撰写中的优劣势，并分析不同考点类型下的表现差异。**方法** 采用问卷调查法，以浙江大学医学院临床医学本科生和浙江大学医学院附属医院的临床呼吸专业带教教师为调查对象。采用相对偏好评分法（5分制），并从专业准确严谨性、考点匹配精准度、结构逻辑清晰度、临床关联实用性及学习引导有效性这5个维度进行评价。数据采用SPSS 26.0软件进行统计分析。计量资料以均数±标准差表示，采用单样本t检验比较总体偏好与中立值（3）的差异，多维度比较采用配对样本t检验，均为双侧检验，检验水准α=0.05。**结果** 本研究共回收问卷60份，包含学生问卷38份，教师问卷22份。总体偏好均分为2.94±0.05，低于中立值3（t=−2.30，P=0.025），整体略倾向于AI解析。教师组（2.91±0.09）与学生组（2.95±0.05）均呈轻度倾向于AI，其中学生组与中立值的差异具有统计学意义（t=−2.05，P=0.047）。按不同考点进行亚组分析，诊断组均分为3.08±0.23，略倾向于教师解析；辅助检查组均分为2.87±0.23，略倾向于AI解析；治疗组均分为2.50±0.22，显著倾向于AI解析（t=−2.85，P=0.029）。多维度分析显示总体差异无统计学意义，但在治疗组中，AI解析在专业准确性方面的评分更高。**结论** AI解析在治疗类考点中表现较好，但在依赖临床经验与复杂情境的诊断考点解析上仍不及教师。因此，在实际教学中可根据不同考点特点探索“AI+教师”协同模式，在提升解析效率的同时，加强对临床思维的引导，从而为试题解析的优化提供参考。

【关键词】 医学试题；　人工智能；　AI解析；　教师解析
【中图分类号】 R-05
基金项目：浙大一院医学教育研究项目（zyjg202437、zyjg202518）
DOI: 10.3760/cma.j.cn116021-20260206-02298

Comparison of the advantages and disadvantages of artificial intelligence-and teacher-generated explanations for respiratory system medical examination questions

Yao Yinan[1], Lü Hanbing[2], Fu Mengjiao[1], Qu Jingjing[1], Zhang Pei[1], Zhou Jianya[1]
[1]Department of Pulmonary and Critical Care Medicine, The First Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou 310003, China; [2]Zhejiang University School of Medicine, Hangzhou 310058, China
Corresponding author: Zhou Jianya, Email: zhoujy@zju.edu.cn

【Abstract】 Objective To compare the advantages and disadvantages of artificial intelligence (AI) and teachers in generating explanations for respiratory system medical examination questions, and analyze performance differences across knowledge domains. **Methods** A questionnaire-based survey was conducted among undergraduate medical students from Zhejiang University School of Medicine and clinical instructors in respiratory medicine from the affiliated hospitals of Zhejiang University School of Medicine. A relative preference scoring method (5-point scale) was used, and evaluations were performed across five dimensions: professional accuracy and rigor, precision of alignment with key points, structural and logical clarity, clinical relevance and practicality, and effectiveness in learning guidance. Data were analyzed using SPSS 26.0. Continuous variables were expressed as mean ± standard deviation. A one-sample t-test was used to compare

overall preference scores with the neutral value of 3. A paired-sample *t*-test was used for multi-dimensional comparisons. All tests were two-tailed, with a significance level of α =0.05. **Results** A total of 60 questionnaires were collected, including 38 from students and 22 from teachers. The mean overall preference score was 2.94 ± 0.05, which was lower than the neutral value of 3 (t=−2.30, P=0.025), indicating a slight overall preference for AI-generated explanations. Both the teacher group (2.91 ± 0.09) and the student group (2.95 ± 0.05) showed a slight preference for AI, and the score of the student group was significantly different from the neutral value (t=−2.05, P=0.047). Subgroup analysis by knowledge domain showed that the diagnosis group scored 3.08 ± 0.23, indicating a slight preference for teacher explanations; the auxiliary examination group scored 2.87 ± 0.23, indicating a slight preference for AI; and the treatment group scored 2.50 ± 0.22, showing a significant preference for AI (t=−2.85, P=0.029). Multi-dimensional analysis revealed no significant overall differences. However, in the treatment group, AI-generated explanations scored higher in professional accuracy. **Conclusions** AI-generated explanations perform better in treatment-related knowledge domains but are inferior to those of teachers in diagnosis-related knowledge domains that rely on clinical experience and involve complex scenarios. Therefore, in teaching practice, an AI + teacher collaborative model can be explored based on the characteristics of different knowledge domains, so as to improve explanation efficiency and strengthen the guidance of clinical thinking, thereby providing a reference for optimizing medical examination question explanations.

【Key words】 Medical examination question; Artificial intelligence; Artificial intelligence-generated explanation; Teacher-generated explanation

Fund program: Medical Education Research Project of The First Affiliated Hospital, Zhejiang University School of Medicine (zyjg202437, zyjg202518)

DOI: 10.3760/cma.j.cn116021-20260206-02298

医学生备考正面临题量庞大与学习低效的现实困境。以执业医师资格考试为例，单场考试题量多达600题，现有的医学题库规模已超过十万级别。单纯依赖机械刷题难以有效提升临床思维能力。高质量规范化的解析不仅有助于明确正确答案，还能够揭示考点内涵、促进知识整合，并引导临床推理过程的形成，有助于提升医学生的备考效率和学习成效。在传统教学模式下，试题解析大多依赖教师完成，但受限于时间投入、经验差异及工作负荷，解析质量与产出效率之间往往难以平衡。近年来，人工智能（artificial intelligence，AI）在医学教育中的应用范围迅速扩大，涉及AI个性化学习路径、虚拟仿真技术、AI虚拟患者及协助教师进行智能评阅、构建数据驱动的教学质量评价体系等[1-2]。以ChatGPT和DeepSeek为代表的生成式AI的发展飞速提升了医学信息获取能力，使其在教学支持中的作用日益凸显[3-5]。一项基于700名医学生的调研发现，约80.3%的学生对AI持积极态度，66.6%的学生认为AI能提升学习质量[6]。而另一项包含17项随机对照试验的Meta分析显示，与传统教学相比，AI辅助教学能显著提升医学生的理论成绩、实践操作技能和教学满意度[7]。此外，众多医学AI工具已应用于诊疗辅助、解剖学习、课堂评估、临床考试等多个环节，为医学教育提供了多样化支持[8-9]。与教学应用相比，AI在医学教育评价领域的研究仍相对有限[10]。教育评价不仅涉及学习结果的判定，还包括对学习过程及思维能力的引导。既往研究提示，AI可通过自动化评分、学习行为分析及即时反馈等方式参与评价过程，在一定程度上提高评价效率并支持数据驱动的教学决策[11]。同时，AI与教师在教育实践中往往呈现互补关系，教师与AI合作可以呈现出5种概念化形式：交互型、情境型、操作型、实践型和协同型，进一步提升彼此的能力[12]。近期有研究基于人工智能辅助诊断平台构建了规培学员胜任力评价指标体系，将AI应用能力纳入评价框架，并验证了其良好的信效度及临床相关性，提示AI已逐步融入医学教育评价体系[13]。基于此，本研究以呼吸系统临床医学试题为例，对比AI与教师在试题解析撰写中的优劣势，并分析其在不同考点类型下的应用特点，以期为医学试题解析的优化策略提供一定的参考。

1　对象和方法

1.1　研究设计

本研究采用横断面调查设计与对比分析法，通过向教师和学生发放标准化问卷，收集2类人群对

AI与教师撰写的40道呼吸系统A2型选择题解析的评价数据，并进行组内及组间比较。接下来，本研究根据试题主要考查内容，将其分为诊断组、辅助检查组和治疗组，用于进一步分析不同考点类型下2类解析的表现差异。解析生成的方法如下：AI解析由DeepSeek生成，统一提示词为：梳理题干核心信息，给出推理过程并分析正确选项，逐一分析其余错误选项，最后总结核心考点和易错点，每题解析字数控制在200~300字。教师解析由1名具有3年以上命题经验的教师按照相同框架撰写，字数控制与AI解析一致。

1.2　研究对象

本研究的问卷调查对象为浙江大学医学院临床医学本科生及浙江大学医学院附属医院临床呼吸专业的带教教师。

1.3　研究工具与数据收集

研究工具为自编的《呼吸系统出科考试A2型选择题AI解析与教师解析对比评价问卷》。采用单盲设计，将问卷中的2种解析分别标记为“解析A”和“解析B”，其真实来源由系统后台随机设定，调查对象在作答时均不知晓其对应关系。

问卷包含2部分内容：第1部分为基本信息，包括调查对象身份及一般资料（性别、年龄、教师职称、教师临床工作年限和临床带教年限）；第2部分为主体评价内容，包括总体偏好评价和多维度评价。总体偏好评价采用相对偏好评分法（5分制），用于反映受试者对2种解析的相对偏好，评分标准为：1分=解析A明显更好，2分=解析A稍好，3分=两者相近，4分=解析B稍好，5分=解析B明显更好。多维度评价中，受试者分别对解析A和解析B在专业准确严谨性、考点匹配精准度、结构逻辑清晰度、临床关联实用性和学习引导有效性5个维度进行评分，采用5分制评价法：1分=非常不足，2分=不足，3分=一般，4分=良好，5分=优秀。

问卷通过线上问卷星平台发放。所有参与者均知悉本研究仅用于学术分析，并在匿名、自愿的原则下完成填写。正式调查前，邀请3名医学教育及临床教学相关领域专家对条目设置及表述进行评议，并据此修订完善问卷；随后开展预调查，评估问卷的可理解性与可操作性。基于预调查数据，对多维度评价部分进行信度分析，Cronbach's α=0.80，提示问卷具有较好的内部一致性。

1.4　统计学方法

采用SPSS 26.0软件进行统计分析。计量资料以均数 ± 标准差表示，并对教师组与学生组一般资料进行描述性分析。根据后台对应关系进一步换算为AI解析或教师解析倾向。总体偏好评价以3分为中立值，采用单样本t检验分析其是否偏离中立水平，并根据评分方向判断对AI或教师解析的倾向；2类解析在多维度上的比较采用配对样本t检验。所有统计检验均为双侧检验，检验水准α=0.05。

2　结果

本研究共回收38份学生问卷和22份教师问卷，分析结果如下。

2.1　调查对象的一般资料

38名学生均为临床医学本科生，22名教师均为呼吸专业带教教师。所有带教教师均具有主治医师及以上职称，其中77.27%具有3年以上临床教学经验。详细信息见表1。

表1　调查对象人口学特征

一般资料		学生（n=38）	带教教师（n=22）
性别	男	17	7
	女	21	15
年龄（岁，$\bar{x} \pm s$）	–	23.40 ± 0.87	34.95 ± 2.44
职称	主治医师	–	12
	副主任医师	–	8
	主任医师	–	2
临床工作年限（年）	<5	–	3
	5~10	–	17
	>10	–	2
临床教学年限（年）	<3	–	5
	3~5	–	13
	>5	–	4

注：“–”表示无数据。

2.2　整体偏好分析

调查对象对40题解析的总体偏好平均分为2.94，略低于中间值3。从图1A可见，分数主要集中在2.60~3.20区间，频率最高的区间为2.90~3.10，但分布呈现轻微左偏特征。结合均值与分布，总体人群略有偏向AI解析的趋势（t=–2.30，P=0.025）。接下来，对教师组和学生组分别进行整体偏好分析。教师组的40题解析偏好平均分为2.91，低于总体均值。如图1B所示，分数分布的众数仍在2.90~3.10区间，其中2.60~2.90区间的频率较高，且2.00~2.40低分区间有一定分布，导致均值明显小于中位数，左偏特征较总体人群更为明显，但差异无统计学意义（t=–1.41，P=0.174）。学生组的40题解析偏

好平均分为2.95，非常接近中间值3，但仍略倾向于AI解析（t=–2.05，P=0.047）。如图1C所示，分数分布较为集中，主要在2.80~3.20区间，频率最高的区间为3.00~3.21。

2.3　具体偏好分析

本研究通过偏好分析散点图（图2）在40道试题解析中识别出2类分布特征：偏好明确的试题解析集中分布于图表两侧且标准差较低，如显著偏向AI解析的第9题（对应考点治疗）与第13题（对应考点辅助检查），以及显著偏向教师解析的第1题与第25题（对应考点诊断）；高度聚集于图表中心区域（均分2.80~3.20）且标准差显著偏高，如第7、30、33题（考点诊断/治疗），其均分接近中立值但在调查对象内部的评价分歧较大，这类试题解析被认为最具争议性。

进一步地，本研究基于试题考点类别进行了亚组分析，包含诊断组、辅助检查组和治疗组。诊断组的整体偏好均分为3.08，稍倾向于教师解析，但差异无统计学意义（t=0.91，P=0.374）；辅助检查组和治疗组的整体偏好均分分别为2.87和2.50，其中辅助检查组差异无统计学意义（t=–0.83，P=0.429），但治疗组明显倾向于AI解析（t=–2.85，P=0.029）。图2显示，诊断组和辅助检查组分布于中立值两侧，并且诊断组试题解析标准差>1.1的比例达到73.68%，表明2组在“倾向于AI生成”或“倾向于教师撰写”的判断上存在较大的意见分歧；相反，治疗组则多分布于左侧，并且意见相对一致。

2.4　AI与教师撰写医学试题解析的多维度对比评价

本研究从专业准确严谨性、考点匹配精准度、

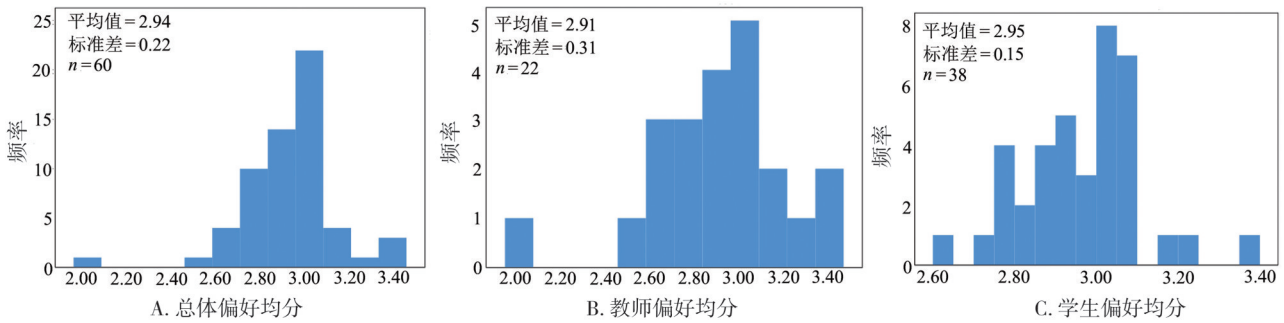


图1　总人群、教师组和学生组整体偏好均分直方图

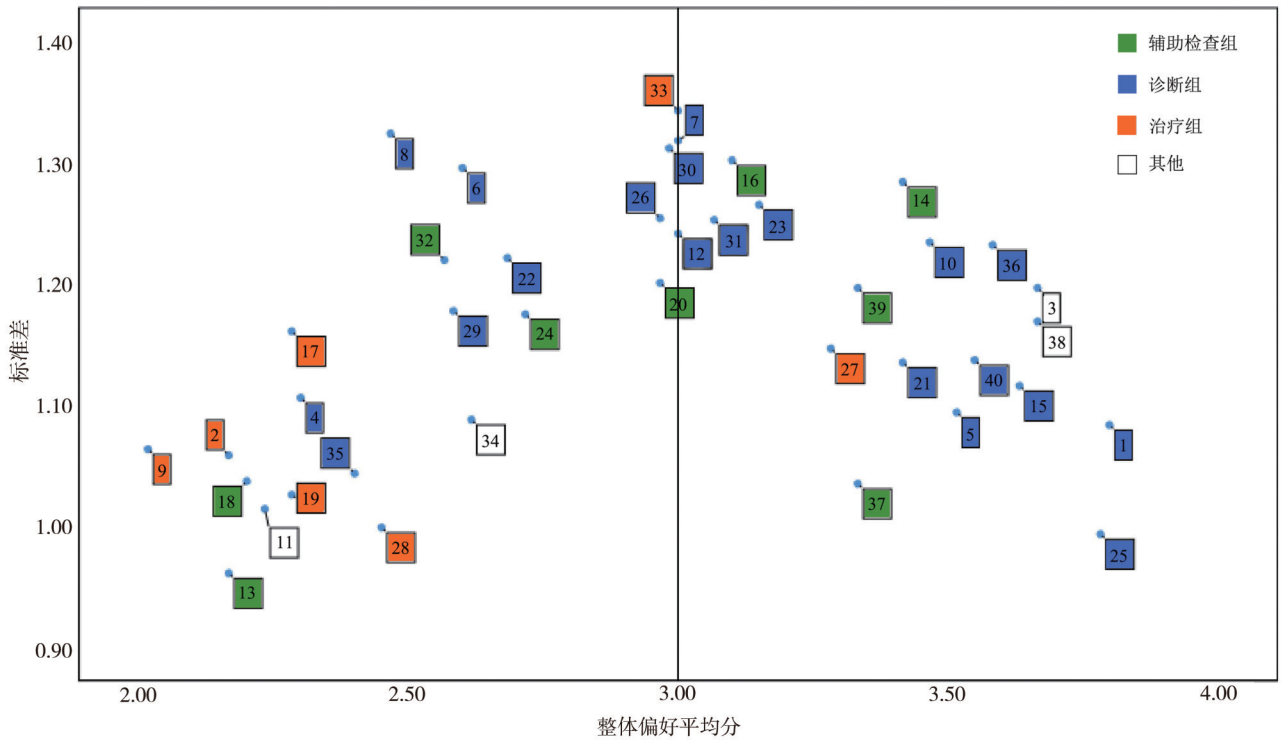


图2　整体偏好评分散点图

结构逻辑清晰度、临床关联实用性、学习引导有效性等5个维度对AI解析和教师解析进行评估。如表2所示，总体上，AI解析和教师解析在5个维度上差异无统计学意义。进一步针对不同考点进行对比发现，在辅助检查组和诊断组，两者解析差异无统计学意义。而在治疗组，总体人群均认为AI解析的专业准确严谨性优于教师解析。从学生角度看，AI解析在专业准确严谨性和考点匹配精准度上优于教师解析；从教师角度看，更是认为在5个维度上AI解析均更胜一筹。

表2　AI解析和教师解析多维度对比评价

分组		AI解析	教师解析	*t*值	*P*值
全部考题					
总人群	专业准确严谨性	3.98 ± 0.03	3.97 ± 0.03	0.22	0.830
	考点匹配精准度	4.00 ± 0.02	3.96 ± 0.02	1.03	0.311
	结构逻辑清晰度	4.02 ± 0.03	3.97 ± 0.03	0.88	0.385
	临床关联实用性	3.99 ± 0.02	3.96 ± 0.02	0.73	0.469
	学习引导有效性	3.98 ± 0.02	3.92 ± 0.03	1.25	0.218
教师	专业准确严谨性	4.04 ± 0.06	4.01 ± 0.05	0.40	0.690
	考点匹配精准度	4.06 ± 0.05	4.00 ± 0.05	1.07	0.290
	结构逻辑清晰度	4.05 ± 0.07	4.00 ± 0.05	0.64	0.524
	临床关联实用性	4.07 ± 0.05	4.01 ± 0.05	0.92	0.361
	学习引导有效性	4.07 ± 0.06	3.95 ± 0.06	1.62	0.112
学生	专业准确严谨性	3.94 ± 0.03	3.92 ± 0.03	0.22	0.824
	考点匹配精准度	3.97 ± 0.01	3.94 ± 0.02	0.73	0.468
	结构逻辑清晰度	3.99 ± 0.01	3.96 ± 0.03	0.84	0.405
	临床关联实用性	3.93 ± 0.01	3.93 ± 0.01	0.18	0.854
	学习引导有效性	3.91 ± 0.02	3.88 ± 0.02	0.79	0.432
辅助检查组					
总人群	专业准确严谨性	4.03 ± 0.02	3.97 ± 0.02	0.70	0.502
	考点匹配精准度	4.04 ± 0.03	3.89 ± 0.03	1.40	0.199
	结构逻辑清晰度	4.08 ± 0.04	3.85 ± 0.04	1.78	0.113
	临床关联实用性	4.01 ± 0.03	3.90 ± 0.01	1.20	0.264
	学习引导有效性	4.00 ± 0.04	3.85 ± 0.04	1.27	0.241
教师	专业准确严谨性	4.10 ± 0.06	4.00 ± 0.03	0.67	0.518
	考点匹配精准度	4.10 ± 0.05	3.93 ± 0.03	1.39	0.201
	结构逻辑清晰度	4.13 ± 0.06	3.95 ± 0.05	1.19	0.270
	临床关联实用性	4.10 ± 0.04	3.96 ± 0.06	1.08	0.311
	学习引导有效性	4.11 ± 0.05	3.91 ± 0.04	1.59	0.151
学生	专业准确严谨性	3.99 ± 0.01	3.96 ± 0.01	0.48	0.641
	考点匹配精准度	3.99 ± 0.02	3.95 ± 0.02	0.48	0.641
	结构逻辑清晰度	4.02 ± 0.01	3.98 ± 0.03	0.55	0.601
	临床关联实用性	3.96 ± 0.01	3.90 ± 0.02	0.90	0.396
	学习引导有效性	3.99 ± 0.01	3.83 ± 0.02	2.30	0.051
诊断组					
总人群	专业准确严谨性	3.91 ± 0.03	4.02 ± 0.02	−1.59	0.127
	考点匹配精准度	3.98 ± 0.02	3.94 ± 0.02	0.62	0.541
	结构逻辑清晰度	4.00 ± 0.03	3.96 ± 0.02	0.61	0.552
	临床关联实用性	3.96 ± 0.01	3.94 ± 0.01	0.37	0.718
	学习引导有效性	3.96 ± 0.02	3.90 ± 0.02	1.01	0.327
教师	专业准确严谨性	3.95 ± 0.04	4.06 ± 0.04	−1.34	0.196
	考点匹配精准度	3.99 ± 0.04	4.04 ± 0.06	−0.59	0.565
	结构逻辑清晰度	3.95 ± 0.06	4.05 ± 0.05	−0.99	0.332
	临床关联实用性	4.00 ± 0.04	4.05 ± 0.03	−0.75	0.463
	学习引导有效性	4.00 ± 0.05	4.00 ± 0.07	0.02	0.981
学生	专业准确严谨性	3.88 ± 0.03	3.97 ± 0.02	−1.45	0.162
	考点匹配精准度	3.93 ± 0.01	3.98 ± 0.02	−1.18	0.253
	结构逻辑清晰度	3.95 ± 0.01	4.02 ± 0.02	−1.44	0.166
	临床关联实用性	3.90 ± 0.01	3.95 ± 0.01	−1.44	0.164
	学习引导有效性	3.86 ± 0.02	3.92 ± 0.02	−1.27	0.221
治疗组					
总人群	专业准确严谨性	4.12 ± 0.02	3.78 ± 0.02	3.07	0.022
	考点匹配精准度	4.05 ± 0.01	4.03 ± 0.02	0.35	0.740
	结构逻辑清晰度	4.02 ± 0.01	4.01 ± 0.02	−1.04	0.340
	临床关联实用性	4.03 ± 0.02	4.02 ± 0.03	0.11	0.917
	学习引导有效性	4.02 ± 0.02	4.04 ± 0.03	−0.15	0.882
教师	专业准确严谨性	4.26 ± 0.07	3.83 ± 0.02	2.84	0.030
	考点匹配精准度	4.24 ± 0.05	3.89 ± 0.03	2.54	0.044
	结构逻辑清晰度	4.26 ± 0.05	3.85 ± 0.03	2.78	0.032
	临床关联实用性	4.23 ± 0.05	3.86 ± 0.02	2.96	0.025
	学习引导有效性	4.28 ± 0.05	3.78 ± 0.03	3.47	0.013
学生	专业准确严谨性	4.05 ± 0.01	3.76 ± 0.02	3.09	0.021
	考点匹配精准度	4.04 ± 0.01	3.80 ± 0.01	3.27	0.017
	结构逻辑清晰度	4.10 ± 0.02	3.81 ± 0.05	2.41	0.052
	临床关联实用性	4.00 ± 0.01	3.85 ± 0.01	2.44	0.050
	学习引导有效性	3.97 ± 0.02	3.81 ± 0.02	2.04	0.088

3　讨论

在医学教育中，试题解析的质量直接影响学习成效与临床思维培养。基于教学实践与研究，高质量的试题解析应系统满足以下5方面核心诉求：第一，专业准确严谨性。这是解析的基石，要求知识点准确无误、符合时效，医学术语与标准规范统一，从根源上杜绝知识误导。第二，考点匹配精准度。解析需要紧扣题目考查的核心知识与能力目标，针对性强，避免泛化或偏移，确保教学与评估目标一致。第三，结构逻辑清晰度。解析应呈现清晰的论述逻辑，如对正确选项的论证、错误选项的排除依据及题干与选项间的内在关系，以利于学生理解与吸收。第四，临床关联实用性。解析须注重将理论知识置于真实临床情境中阐释，通过构建“患者画像”、模拟诊疗流程，有效贯通理论与临床实践。第

五，学习引导有效性。超越答案本身，解析应致力于思维训练，引导学生建立并实践从症状到诊疗的完整临床逻辑链，并可根据学习者认知阶段（如见习、实习）进行分层设计[14]，促进其思维向高阶发展。这5个维度共同构成了评估与优化试题解析质量的系统性框架，旨在实现从知识传递到临床胜任力培养的教学转化。

3.1 AI与教师撰写试题解析的偏好分化及影响因素探讨

本研究采用问卷调查法，以临床医学本科生与临床带教教师为调查对象，通过设计相对偏好评分量表，对比评估二者对AI与教师撰写的医学试题解析总体和不同维度的接受度与偏好。整体结果显示，总体人群均略倾向于AI解析，但2组差异无统计学意义；然而，针对不同知识模块进行细分评估后，发现倾向性存在分化。其中，在治疗相关考点上，学生与教师一致倾向于认为AI解析更优。这可能归因于治疗原则通常基于明确的专业知识点和指南规范，逻辑结构清晰，AI在信息整合的完整性与表述的标准化方面具有优势，能高效满足学习者对准确性、规范性与效率的核心诉求。相比之下，在诊断考点上，评估者对解析来源的倾向出现明显分歧，且诊断组内部评分一致性较低。这一现象揭示了当前AI解析的局限所在。诊断思维的核心在于对复杂临床信息的整合、鉴别与决策，这远超越标准知识的陈述[15]。教师解析往往能融入临床经验与情境化判断，通过模拟诊疗决策过程有效引导思维；而在此类需要高阶推理与灵活关联的环节中，AI解析在临床思维深度与教学引导针对性方面尚有不足，因此导致不同评估者基于各自的评价标准和关注点给出了差异化的评价。

3.2 多维度评估下AI与教师医学试题解析质量的知识点特异性差异

本研究进一步从5个核心维度评估了AI与教师医学试题解析的质量。总体而言，两者差异无统计学意义，但深入分析显示其表现仍具有知识点特异性。在治疗考点上，AI解析获得一致认可，教师群体尤为肯定AI解析在全维度的优势，这与前面相对偏好评分的结果一致。治疗决策紧密依赖持续更新的指南与循证证据，教师往往面临繁重的准确性核查与内容更新压力。AI能够精准、高效地整合最新权威知识，提供稳定规范的输出，这极大地保障了教学材料的基础质量与时效性，从而解放教学生产力。学生群体则主要认可AI在专业准确性与考点匹配上的优势。他们的核心诉求在于获得明确、紧扣题目的答案。AI解析多以较为规范的方式呈现治疗思路和基本原则，结构相对清晰、重点突出，在需要快速把握核心知识的学习情境中更具优势。而在辅助检查与诊断考点上，双方解析差异则无统计学意义。这提示，在依赖临床情境整合与经验性决策的复杂思维领域，当前AI解析尚未建立起如同在治疗考点上的相对优势。此前诊断类考点中评估者间评分差异较为明显，提示该类解析在给出结论的同时，更需要展开思维过程，以帮助理解其推理路径。

3.3 医学试题解析中AI与教师的优势互补及融合应用思考

基于本研究及临床教学实践，AI与教师撰写的医学试题解析展现出良好的互补性。AI凭借强大的信息整合能力，在基础知识准确率、术语规范性、生成速度、风格一致性、大体量处理等方面都有明显的优势，但它在复杂的临床情境中容易产生逻辑碎片化和诊断误差等问题。相比之下，教师解析往往依赖个人的临床经验和教学积累，能够较好地指出易错点并引导高阶思考，但不同教师之间，受经验深浅和投入精力的影响，其效率和表达的稳定性也会有所差异。基于此，建议构建“AI+教师”协同工作模式，具体可采取以下方法。①角色分工：在实施过程中，AI负责初步的题目解析、知识点标注、术语校正、多版本草稿生成，教师则对AI产出的内容进行临床复核、易错点精准提取及高阶思维引导的整合。②双向流程：在试题解析修订的过程中加入双向工作流，先由AI生成解析草稿及分层难度、不同题型和考点的变体，之后再经由教师审核、补充临床判断，教师的修改可以回流至模型中，作为调整的参照，从而不断优化AI的试题解析。③质量控制：常规采用教师抽样复核AI产出的关键题目解析，逐步形成统一且规范的评分模板，记录常见错误并更新错误库。④个性化教学：通过提取答题数据实现解析的个性化推荐，由AI生成不同难度层次或关注重点不同的解析，再由教师根据学生群体特点调整教学重点和反馈方式。⑤持续评价与改进：围绕试题解析建立周期性评价机制，结合解析准确性、术语使用规范性、知识点覆盖情况、教师复核耗时及师生反馈等指标进行综合判断。根据每次考试后的反馈对解析进行调整，在反复修订中逐步优化该模式，形成持续提升的闭环。⑥安全与责任界定：应对AI在试题解析中的使用范围进行界定，并

由教师负责最终审核。同时，还需关注数据隐私及结果的可追溯性，在实践中逐步加以规范。

两者的协同策略既可以把 AI 在高效规模化和一致性方面的优势发挥出来，又能保持教师在临床判断和教学价值传递上所起的核心作用，为构建可控、可评估的“AI+教师”医学试题解析体系提供了可操作的有效途径。

3.4 局限性

本研究仍存在一定局限。首先，本研究为单中心研究，样本量相对有限，且研究对象来源较为集中，因此研究结果的外推性仍需谨慎对待。其次，本研究主要基于调查对象的主观评价，而学生和教师在学习背景和认知水平上的差异，可能在一定程度上影响评分结果。再次，本研究采用自编问卷，虽然对其进行了内部一致性检验，但在结构效度方面仍需进一步完善。此外，本研究所选试题均为呼吸系统 A2 型选择题，题型和学科范围相对单一，相关结果在其他题型或学科中的适用性仍需进一步验证。后续研究可在扩大样本量的基础上，结合更客观的评价指标进行深入分析。

4 结语

AI 在医学试题解析中的整体认可度与教师相近，其表现在不同知识领域间存在一定差异。在结构化程度较高、依赖临床指南和规范的治疗类考点中，AI 在信息整合及表达效率方面具有一定优势，能够较好地满足学习需求；而在诊断等需要构建临床思维的内容中，AI 虽可提供基本解析，但在情境理解及思维引导方面仍不及教师。从教学应用角度来看，AI 可在一定程度上作为教师的辅助工具，在提升解析效率及规范表达方面发挥作用；而教师则更适合在临床思维引导及复杂情境解释中发挥优势。因此，在未来的医学教育中，可根据不同内容特点对两者进行合理分工与配合，使其在各自擅长的环节中发挥作用。相关应用仍有待在更多教学情境中进一步探索和完善。

利益冲突　所有作者声明不存在利益冲突

作者贡献声明　姚一楠：研究设计、实施研究、采集数据、分析/解释数据、撰写文章；吕寒冰：修改和审阅文章；傅梦姣、屈晶晶、张培：采集数据；周建娅：研究设计、审阅文章

参考文献

[1] Ahsan Z. Integrating artificial intelligence into medical education: a narrative systematic review of current applications, challenges, and future directions [J]. BMC Med Educ, 2025, 25(1): 1187. DOI: 10.1186/s12909-025-07744-0.

[2] 袁靳闲，李金政. 人工智能应用于医学教育的现状、挑战与前景[J]. 中华医学教育探索杂志，2025，24(6)：777-783. DOI: 10.3760/cma.j.cn116021-20240711-02049.

[3] Cooper A, Rodman A. AI and medical education: a 21st-century Pandora's box [J]. N Engl J Med, 2023, 389(5): 385-387. DOI: 10.1056/nejmp2304993.

[4] Boscardin CK, Gin B, Golde PB, et al. ChatGPT and generative artificial intelligence for medical education: potential impact and opportunity [J]. Acad Med, 2024, 99(1): 22-27. DOI: 10.1097/ACM.0000000000005439.

[5] Guo D, Yang D, Zhang H, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning [J]. Nature, 2025, 645(8081): 633-638. DOI: 10.1038/s41586-025-09422-z.

[6] Sami A, Tanveer F, Sajwani K, et al. Medical students' attitudes toward AI in education: perception, effectiveness, and its credibility [J]. BMC Med Educ, 2025, 25(1): 82. DOI: 10.1186/s12909-025-06704-y.

[7] 魏嵬，罗高萌，刘润铭，等. 人工智能辅助教学在医学教育中应用效果的 Meta 分析[J]. 医学新知，2025，35(10)：1242-1248. DOI: 10.12173/j.issn.1004-5511.202505084.

[8] Zhou J, Zeng Z, Li L. A meta-analysis of Watson for oncology in clinical application [J]. Sci Rep, 2021, 11: 5792. DOI: 10.1038/s41598-021-84973-5.

[9] Balaya V, Guimiot F, Bruzzi M, et al. Feasibility of a fetal anatomy 3D atlas by computer-assisted anatomic dissection [J]. J Gynecol Obstet Hum Reprod, 2020, 49(9): 101880. DOI: 10.1016/j.jogoh.2020.101880.

[10] Holmes W, Bialik M, Fadel C. Artificial intelligence in education: promises and implications for teaching and learning [M]. Boston: Center for Curriculum Redesign, 2019.

[11] Zawacki-Richter O, Marín VI, Bond M, et al. Systematic review of research on artificial intelligence applications in higher education: where are the educators? [J]. Int J Educ Technol High Educ, 2019, 16(1): 39. DOI: 10.1186/s41239-019-0171-0.

[12] Cukurova M, Suraworachet W, Zhou Q, et al. Towards synergistic teacher-AI interactions with generative artificial intelligence[EB/OL]. (2025-11-24)[2026-02-01]. https://arxiv.org/abs/2511.19580.

[13] 杨士凤，孙博实，刘旭云，等. 基于人工智能辅助诊断平台的结直肠肿瘤规培学员胜任力评价指标体系构建[J]. 中华医学教育探索杂志，2025，24(12)：1592-1599. DOI: 10.3760/cma.j.cn116021-20250122-02148.

[14] Anderson LW, Krathwohl DR, Bloom BS. A taxonomy for learning, teaching, and assessing: a revision of Bloom's taxonomy of educational objectives [M]. New York: Longman, 2001.

[15] Shimizu T. DECLARE: a comprehensive, multifaceted cognitive forcing strategy to confront complex cases [J]. Int J Gen Med, 2023, 16: 1505-1511. DOI: 10.2147/ijgm.s406165.

（收稿日期：2026-02-06）

（本文编辑：邬小梅）