·数智医学教育·

基于放射学历史题库数据分析的人工智能试题生成及考核效果评价

史玉书[1]　秦潇潇[1]　廖金生[1]　黄强[1]　楼海燕[1]　陈韶华[2]　肖文波[1]
[1]浙江大学医学院附属第一医院放射科，杭州 310003；[2]浙江大学医学院附属第一医院教学部，杭州 310003
通信作者：肖文波，Email：xiaowb.111@163.com

【摘要】 目的 探讨依托放射科历史题库数据的大语言模型在医学见习阶段放射影像试题智能生成中的可行性，评估其对试题难度与区分度等核心测量学参数的预测效能及实测一致性，为智能化医学命题提供实证依据。**方法** 采用定量实验设计，从放射科历史题库中按学科标准抽取50道经典试题作为提示基准，将其输入通义千问（Qwen-Plus）大语言模型生成新试题，并预测其难度与区分度参数。组织56名临床医学四年级见习生进行实测，采用非参数检验、Spearman秩相关及组内相关系数（intraclass correlation coefficient，ICC）等进行统计学分析。**结果** AI生成试题难度预测总体误差较小[平均绝对误差（mean absolute error，MAE）=0.052，均方根误差（root mean square error，RMSE）=0.062]，区分度预测误差略大（MAE=0.105，RMSE=0.131）。AI预测值与实测值呈高度正相关（难度ρ=0.896，区分度ρ=0.791，均P<0.001）。一致性检验显示，难度参数总体ICC=0.908（优秀），区分度参数总体ICC=0.801（良好）。分层分析表明，AI对容易题难度存在轻微高估，对中档题的区分度预测略偏乐观，但实测参数整体成功维持了预设的难度梯度与鉴别力分布特征，干扰项临床贴合度良好。**结论** 基于大语言模型的AI试题生成技术对放射影像见习试题的难度系数和区分度预测具有良好的一致性。依托历史题库实测参数进行交叉验证，AI辅助命题具备高度的可行性与教学适用性。建议构建“AI生成—历史数据校准—专家复核”的人机协同命题新模式，以推动医学影像学考核向标准化、精准化与智能化转型。

【关键词】 人工智能；　放射影像；　试题优化；　难度系数；　区分度
【中图分类号】 R-0；R-4；R445.4
基金项目：浙江大学医学院教育改革项目（jgzx2025010）
DOI：10.3760/cma.j.cn116021-20260129-02321

Artificial intelligence-based generation of radiology test items and evaluation of assessment effectiveness through analysis of historical test item bank data

Shi Yushu[1], Qin Xiaoxiao[1], Liao Jinsheng[1], Huang Qiang[1], Lou Haiyan[1], Chen Shaohua[2], Xiao Wenbo[1]
[1]Department of Radiology, The First Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou 310003, China; [2]Education Administration Office, The First Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou 310003, China
Corresponding author: Xiao Wenbo, Email: xiaowb.111@163.com

【Abstract】 Objective To evaluate the feasibility of a large language model (LLM) based on historical radiology test item bank data for the intelligent generation of radiological imaging test items for the medical clerkship stage, assess its predictive performance and agreement between predicted and empirical data for core psychometric parameters such as difficulty and discrimination, and provide empirical evidence for intelligent medical test item development. **Methods** A quantitative experimental design was adopted. Fifty classic items were extracted from a historical radiology test item bank according to subject-specific standards as baseline prompts, and were input into the Qwen-Plus LLM to generate new items and predict their difficulty and discrimination parameters. Fifty-six fourth-year clinical medicine clerkship students completed the generated test. Statistical analyses included non-parametric tests, Spearman's rank correlation,

and intraclass correlation coefficient (ICC). **Results** The predicted difficulty of artificial intelligence (AI)-generated test items showed a relatively small overall error (mean absolute error=0.052, root mean square error=0.062), while the predicted discrimination showed a relatively large error (mean absolute error=0.105, root mean square error=0.131). AI-predicted values were highly positively correlated with empirical values (difficulty ρ=0.896, discrimination ρ=0.791, both P<0.001). ICC analysis indicated excellent agreement for difficulty (ICC=0.908) and good agreement for discrimination (ICC=0.801). Stratified analysis revealed a slight overestimation of difficulty for easy items and a slightly optimistic prediction of discrimination for medium-difficulty items. However, the empirical parameters overall successfully maintained the preset difficulty gradient and discrimination distribution characteristics, with distractors demonstrating high clinical plausibility. **Conclusions** LLM-based item generation demonstrates good agreement between predicted and empirical values for the difficulty coefficient and discrimination of radiology clerkship test items. In cross-validation against the empirical data from a historical radiology test item bank, AI-assisted test item development is highly feasible and pedagogically applicable. It is recommended to construct a new human-AI collaborative test item development model of "AI generation, historical data calibration, and expert review" to advance the transformation of medical imaging assessment toward standardization, precision, and intelligence.

【**Key words**】 Artificial intelligence; Radiological imaging; Test item optimization; Difficulty coefficient; Discrimination

Fund program: Education Reform Project of Zhejiang University School of Medicine (jgzx2025010)

DOI: 10.3760/cma.j.cn116021-20260129-02321

医学影像学是临床医学教育的核心组成部分，其考核试题的科学性直接决定影像人才临床思维与图像判读能力的培养质量[1-2]。放射影像学见习课程作为衔接理论与临床的关键环节，核心是培养学生对影像图像的精准判读、病灶识别及临床综合分析能力，其试题需深度融合影像图像与临床病例，兼具专业性、实践性与时效性。但传统命题模式高度依赖教师个人临床经验与教学积累，不仅存在主观性强、题目更新滞后、难度与区分度难以精准把控的固有弊端，还无法适配放射影像学快速发展的学科需求，难以真实反映学生的临床核心能力，严重制约考核标准化、精准化发展，成为影像人才高质量培养的瓶颈[3-4]。在此背景下，历经多年教学沉淀的放射科历史题库已成为破解上述困境的重要基础。该题库由资深临床带教教师与教学专家联合编撰，历经多届医学生实战考核验证，其内容安全性、临床贴合度与命题质量均已得到充分检验。更为关键的是，历史题库中每道试题均附有经大样本实测校准的难度系数、区分度等经典心理测量学参数，为试题质量的科学评价提供了客观标尺。这些高质量、带标注的命题数据不仅是临床教学考核的“金标准”，也为智能化命题研究奠定了不可替代的数据基石。

近年来，人工智能（artificial intelligence，AI）技术在医学教育评估领域的应用已展现出不可替代的优势，其可通过大数据建模、机器学习算法，实现试题难度、区分度等核心测量学指标的精准预测与优化，大幅提升命题效率与质量，有效弥补传统命题的诸多不足[5]。国家医学考试中心也将AI智能命题与影像题库建设纳入重点研究方向，凸显其在医学教育转型中的重要地位[6]。已有研究证实，AI生成的医学试题在难度、区分度上可与人类专家命题相媲美，甚至能通过“图灵测试”，让医学生与临床医生难以分辨其来源[7]。然而，目前AI在医学试题生成中的应用多聚焦于基础医学学科，而放射影像见习试题因需兼顾图像解析与临床思维，生成难度远高于基础医学试题，对模型的学科适配性要求极高，相关研究仍较为匮乏，导致见习课程深陷传统命题困境，无法实现精准考核与教学反馈[8]。在此背景下，探索面向放射影像见习课程的AI试题生成方法的重要性与必要性尤为突出，其既是响应国家医学考试智能化发展号召、填补学科研究空白的关键举措[6]，也是破解传统命题弊端、实现考核精准化的必由之路。本研究依托放射影像见习历史题库，构建适配放射影像学科特性的智能模型，优化试题生成质量，通过对试题核心指标进行预测试并开展

系统性对比分析，全面评估AI生成试题的可行性与实用价值。此举不仅能大幅提升命题效率、规范考核标准，更能精准对接学生学习需求，为构建科学高效的智能命题体系提供理论与实践支撑，推动医学影像学教学评估向智能化转型，助力高素质影像人才培养[3]。

1　资料与方法

1.1　研究对象

本研究选用的AI工具为服务器端部署的智能语言模型通义千问（Qwen-Plus-2025-12-01版本）。同时，选取浙江大学医学院附属第一医院见习学生（医学5年制本科，临床专业，第4年，医院见习阶段）56人作为试题考试结果的评估对象，对AI生成的试题进行测试，并将测试结果纳入研究。

1.2　研究方法

从浙江大学医学院METESP（Medical Education Test & Evaluation Service Platform）考试系统中，根据设置条件限制性随机抽取历年影像学见习考试试题50题。选取标准如下：①按照人民卫生出版社《医学影像学》（第八版）中的病种进行选取，每题病种唯一，且覆盖4大系统（心胸、腹部、中枢、骨肌）。②试题的难度以答对该题目的考生比例来表示，难度值（p）的计算公式为$p=R/N$（p指试题难度，N为全体被测者人数，R为答对或通过该项目的人数）。难度值越大表示难度越低，其中容易题的难度系数为0.95~0.75，中档题为0.74~0.60，难题为0.59~0.20。选取试题的难度比例为3∶6∶1（容易题∶中档题∶难题，难度比例可根据学生知识掌握程度自行设定）。③区分度的计算公式为$D=(AH-AL)$/总分（D为区分度，AH为高分组成绩均值，AL为低分组成绩均值）。高分组定义为总分排名前27%的考生，低分组定义为总分排名后27%的考生。本研究选取区分度为0.3以上的试题。将抽取的历年试题文档作为样题和提示词一并输入AI模型进行分析（图1）。AI模型分析后总结出考核知识点，并分析和理解历史试题的难度和区分度；根据分析结果，自动生成试题并标注相应的难度系数区间和区分度。生成的试题经过研究者双人审核，确认题目无误后，作为研究对象纳入后续对见习学生的实际测试中。



图1　AI形成试题过程示例

1.3　统计学方法

采用SPSS 19.0和R软件(R 4.5.1)对数据进行统计分析。数据经Shapiro-Wilk检验呈非正态分布,计量资料以$M(Q_1,Q_3)$表示,组间比较采用Mann-Whitney U检验,计算U值、Z值、P值,检验水准α=0.05。为评估AI预测的精准性,计算以下指标。①预测误差:平均绝对误差(mean absolute error,MAE)和均方根误差(root mean square error,RMSE),分别反映预测值的平均绝对偏离程度和对较大误差的敏感程度。②相关性:采用Spearman相关系数分析AI预测值与实测值的单调关联强度,相关系数绝对值(ρ)≥0.8为高度相关。③一致性:采用组内相关系数(intraclass correlation coefficient,ICC),使用双向随机效应模型(绝对一致性)计算,置信区间基于F分布推导,以评价三者间的一致性;ICC>0.75表示一致性良好,ICC>0.90表示一致性优秀。

2　结果

2.1　一般资料

本研究共纳入50道试题,按预设认知梯度划分为容易题(1~15题,n=15)、中档题(16~45题,n=30)与难题(46~50题,n=5)。经正态性检验,数据呈非正态分布,故采用中位数及四分位距$M(Q_1,Q_3)$进行描述。3组试题在样题、AI预测(AI分析样题后自动生成的试题,包括题目、预测的难度和区分度)与AI实测(AI生成的试题经过见习学生测试后计算出试题的难度和区分度,用于评估AI生成试题的能力)3种条件下的难度系数与区分度统计结果见表1。

表1　各组难度系数与区分度的描述性统计[$M(Q_1,Q_3)$]

题目级别	指标类型	难度系数	区分度
容易题(n=15)	样题	0.83(0.79,0.92)	0.37(0.29,0.45)
	AI预测	0.89(0.86,0.93)	0.28(0.24,0.31)
	AI实测	0.86(0.79,0.91)	0.28(0.21,0.33)
中档题(n=30)	样题	0.68(0.64,0.72)	0.41(0.37,0.48)
	AI预测	0.67(0.64,0.69)	0.50(0.48,0.52)
	AI实测	0.70(0.68,0.71)	0.34(0.33,0.36)
难题(n=5)	样题	0.32(0.28,0.56)	0.55(0.49,0.57)
	AI预测	0.35(0.27,0.42)	0.42(0.39,0.43)
	AI实测	0.50(0.47,0.54)	0.35(0.34,0.35)

注:难度系数取值范围为0~1,值越高表示题目越容易;区分度值越高表示题目鉴别不同能力水平学习者的能力越强。

整体而言,难度系数随题目级别升高呈规律性递减,符合教育测量学中试题难度的梯度设计预期。区分度指标则呈现“中档题>难题≈容易题”的典型倒U型分布,与经典测量理论(classical test theory,CTT)的命题规律一致[9]。横向对比发现,AI生成试题在容易题难度上存在偏高倾向[M=0.89],在中档题区分度上呈现高估倾向[M=0.50]。此外,经实测发现,难度和区分度均介于样题与预测值之间,提示AI算法对试题属性的预估存在明显的难度层级特异性偏差。

2.2　试题难度和区分度

2.2.1　难度系数的组间差异分析

采用Mann-Whitney U检验对3种条件下的难度系数进行两两比较,结果见表2。

表2　3种题组的难度系数组间差异检验

题目类型	对比组	U值	Z值	P值
容易题(n=15)	样题 vs. AI预测	60	−2.64	0.008
	样题 vs. AI实测	110	−0.04	0.972
	AI预测 vs. AI实测	53	−2.71	0.007
中档题(n=30)	样题 vs. AI预测	389	−0.69	0.491
	样题 vs.AI实测	241	−2.97	0.003
	AI预测 vs. AI实测	213	−3.44	0.001
难题(n=5)	样题 vs. AI预测	8	−0.81	0.421
	样题 vs. AI实测	6	−1.01	0.310
	AI预测 vs. AI实测	3	−1.92	0.056

在容易题中,样题与AI预测难度之间的差异具有统计学意义(Z=−2.64,P=0.008),而样题与AI实测难度之间的差异无统计学意义(P=0.972),且AI预测与AI实测难度之间的差异有统计学意义(Z=−2.71,P=0.007)。这表明AI对低难度试题的理论预测值显著高于原始样题,但经见习学生实测后难度系数回落至与样题无显著差异的水平,提示AI生成容易题时可能存在“预设过易”的倾向。

在中档题中,样题与AI预测的差异无统计学意义(P=0.491),但样题与AI实测(Z=−2.97,P=0.003)及AI预测与AI实测(Z=−3.44,P=0.001)的差异均有统计学意义。这说明AI对中档题难度的理论设定较为合理,但在实际施测中受题干信息复杂度或知识迁移要求的影响,实测难度发生了正向偏移。

在难题中,3组两两比较的P值均大于0.05(P=0.310~0.421),差异无统计学意义。受限于样本量较小(n=5),统计检验效能可能不足,但描述性趋势

显示，AI 实测难度（*M*=0.50）明显高于样题与预测值，提示高难度试题在真实测试情境中易出现“难度衰减”现象。

2.2.2　3 种题组的区分度组间差异分析

本研究对 3 种题组的区分度进行组间差异分析，结果见表 3。

表 3　3 种题组的区分度组间差异分析

题目类型	对比组	*U* 值	*Z* 值	*P* 值
容易题（*n*=15）	样题 vs. AI 预测	68	−2.21	0.027
	样题 vs. AI 实测	64	−2.43	0.014
	AI 预测 vs. AI 实测	106	−0.35	0.724
中档题（*n*=30）	样题 vs. AI 预测	92	−5.11	<0.001
	样题 vs.AI 实测	92	−5.11	<0.001
	AI 预测 vs. AI 实测	96	−5.23	<0.001
难题（*n*=5）	样题 vs. AI 预测	3	−1.92	0.056
	样题 vs. AI 实测	0	−2.40	0.016
	AI 预测 vs. AI 实测	3	−1.92	0.056

统计检验表明，AI 介入对试题区分度的影响模式与难度系数存在明显分化。在容易题中，样题与 AI 预测（*Z*=−2.21，*P*=0.027）及样题与 AI 实测（*Z*=−2.43，*P*=0.014）的差异有统计学意义，且 AI 预测与 AI 实测差异无统计学意义（*P*=0.724）。结合表 1 数据可知，AI 生成的容易题区分度[*M*≈0.28]显著低于原始样题[*M*=0.37]，表明低难度题目经 AI 处理后，其鉴别高低水平学习者的功能有所削弱，可能源于 AI 在生成选项时未能有效设置具有认知梯度的干扰项。在中档题中，样题与 AI 预测区分度差异有统计学意义（*Z*=−5.11，*P*<0.001），AI 预测值[*M*=0.50]显著高于样题。然而，AI 预测与实测值之间存在较大落差，提示算法对中档题区分度的理论预估较为乐观，实际作答反馈受考生认知负荷与答题策略调节后区分效能出现衰减。难题描述性数据显示 AI 预测与实测区分度均低于样题（0.42 vs. 0.55；0.35 vs. 0.55）。该趋势符合经典测量理论中“过高难度易引发地板效应，进而压低区分度”的规律，提示 AI 在生成高难度试题时，需进一步强化对高阶思维路径的差异化设计。综合难度与区分度指标，发现 AI 自动生成试题在理论预测阶段表现出较强的参数可控性，但实测结果揭示其在不同难度层级上存在系统性偏差（容易题鉴别力不足，中档题实测难度与区分度偏离预测值，难题参数易受实测情境影响发生漂移）。该结果为后续优化 AI 命题算法、建立“预测—实测”动态校准机制提供了实证依据。

2.3　AI 预测精准性评估

为系统评估 AI 生成试题参数的预测效能，本研究从预测误差、秩相关性及绝对一致性 3 个维度进行量化分析。

2.3.1　预测误差分析

以实测值为基准，计算 AI 预测值与实测值之间的 MAE 与 RMSE，结果见表 4。

表 4　AI 生成试题的预测值与实测值的预测误差分析

组别	难度值		区分度值	
	MAE	RMSE	MAE	RMSE
容易题	0.045	0.057	0.046	0.064
中档题	0.042	0.048	0.154	0.161
难题	0.144	0.166	0.065	0.077
总体	0.052	0.062	0.105	0.131

总体而言，难度预测的 MAE 为 0.052，RMSE 为 0.062；区分度预测的 MAE 为 0.105，RMSE 为 0.131。难度指标的整体预测误差显著低于区分度指标。经过分层级观察发现：容易题的预测误差最小（难度 MAE=0.045，区分度 MAE=0.046）；中档题的区分度预测误差最大（MAE=0.154，RMSE=0.161）；难题的难度预测误差最为突出（MAE=0.144，RMSE=0.166）。该结果表明，AI 对基础难度层级的把控较为稳健，但在面对认知复杂度较高的中档题鉴别力预估，以及高难度试题的难度跃升时，预测偏差呈现层级特异性放大趋势。

2.3.2　相关性分析

采用 Spearman 等级相关系数（ρ）检验 AI 预测值与实测值之间的单调关联强度，结果见表 5。

表 5　Spearman 等级相关系数分析

指标	组别	样题 vs. AI 预测	样题 vs. 实测	AI 预测 vs. 实测
难度	容易题	0.978	0.768	0.846
	中档题	0.973	0.860	0.902
	难题	0.951	0.863	0.983
	总体	0.966	0.820	0.896[a]
区分度	容易题	0.864	0.639	0.748
	中档题	0.784	0.672	0.712
	难题	0.798	0.712	0.824
	总体	0.920	0.645	0.791[a]

注：所有相关系数检验 *P*<0.001；a 为 AI 预测效能核心指标。

整体上，难度预测值与实测值呈高度正相关（ρ=0.896），区分度预测值与实测值呈中度至高度正相关（ρ=0.791）。分层级看，难度参数在各组别中均保持强相关（ρ=0.846~0.983），表明AI能高度还原试题难度的相对排序。另外，区分度参数的相关性虽略低，但依然显著（ρ=0.712~0.824），提示AI对题目鉴别力的预测趋势与实测反馈基本一致，但其绝对量值的映射精度会因实测情境的调节而存在一定波动。

2.3.3　一致性检验

本研究基于双向随机效应模型，采用绝对一致性公式计算ICC，以评价样题参数、AI预测参数与实测参数三者之间的一致性程度，结果见表6。参考Koo和Li[10]的判定标准：差（ICC<0.50）、中等（0.50≤ICC<0.75）、良好（0.75≤ICC<0.90）、优秀（ICC≥0.90）。

表6　组内ICC分析

指标	组别	ICC值	95%*CI*	一致性水平
难度	容易题	0.892	[0.756,0.958]	良好
	中档题	0.927	[0.879,0.960]	优秀
	难题	0.891	[0.612,0.986]	良好
	总体	0.908	[0.867,0.938]	优秀
区分度	容易题	0.758	[0.527,0.893]	中等良好
	中档题	0.713	[0.553,0.830]	中等
	难题	0.755	[0.279,0.966]	良好
	总体	0.801	[0.715,0.868]	良好

表中数据说明，难度参数总体一致性达优秀水平（ICC=0.908）。其中，中档题一致性最高（ICC=0.927，95%*CI* [0.879,0.960]），容易题与难题均达良好水平；而区分度参数总体一致性达良好水平（ICC=0.801），但各组别均处于中等良好水平（0.713~0.758）。这些结果表明，AI在难度维度的绝对一致性表现优异，能够稳定复现原始试题的难度结构；而区分度参数虽具有一致性基础，但受实测群体能力分布、作答策略及干扰项有效性等多因素交互影响，AI预测值与实测值之间的绝对吻合度仍有优化空间。

3　讨论

3.1　AI命题的难度可复现，区分度难迁移

本研究通过系统对比放射学历史试题、AI预测试题与实测试题的测量学参数，首次量化评估了大语言模型在放射影像见习试题生成中的预测效能与实测偏差。结果显示，AI预测试题难度与实测难度呈现优异的一致性（ICC=0.908），该结果与近期AI辅助医学教育评估的研究趋势相一致[11]，表明依托历史题库结构化提示的生成策略能有效复现经典试题的难度梯度。历史题库中经多届学生实测校准的难度系数及“知识点—临床情境”映射关系，为AI提供了高信度的先验知识锚点，使其在文本语义解析与认知层次匹配方面具备较强的泛化能力与参数可控性。相比之下，区分度预测一致性为良好（ICC=0.801），但各组别均处于中等良好水平（0.713~0.758），提示AI在模拟试题鉴别能力时仍面临学科特异性挑战。

3.2　层级偏差揭示纯文本命题的认知断层

本研究对生成试题的难度层级分析发现，AI生成试题在不同难度层级存在系统性偏差：容易题实测区分度偏低，中档题参数出现预测漂移，难题组则受“地板效应”压制导致鉴别力下降。这一现象符合经典测量理论规律，也揭示了当前纯文本大语言模型在放射命题中的内在局限。区分度高度依赖干扰项的隐蔽性、临床决策路径的复杂性及考生群体的能力离散度。AI虽能依据历史题库规律生成结构合规的题干与选项，但在动态构建“高迷惑性干扰项”与“多维度临床推理陷阱”时，仍缺乏对见习生认知盲区及影像征象误判路径的深度建模。此外，放射见习试题需同步调用“图像特征提取—病理机制关联—临床决策推导”的多模态认知负荷，而纯文本提示难以完全还原该认知链条，导致实测区分度与预测值产生交互性偏离[12-14]。

3.3　AI命题需以实测为锚反向校准生成偏差

针对上述偏差，历史题库的核心价值不仅在于提供生成语料，更在于构建“预测—实测”动态校准的基准参照系。本研究证实，AI命题并非“一键生成即用”，而需依托历史题库的实测参数进行迭代优化。未来智能命题系统应建立“AI初稿生成→历史题库参数锚定→小样本实测校准→项目反应理论（item response theory，IRT）等值→正式发布”的闭环流程[15]。同时，可将历史题库中高频易错知识点、典型干扰项模式及学生答题轨迹向量化，嵌入提示词工程与强化学习反馈回路，以强化AI对放射学科特异性认知路径的模拟能力，从而有效弥合预测与

实测之间的差距，提升中高阶试题的鉴别效能。

3.4 研究局限性

本研究存在若干局限，需在后续工作中完善。首先，样本量有限（n=50，难题组仅5题），可能导致统计检验效能不足，尤其是难题组参数估计的置信区间较宽，未来需开展多中心、大样本的跨院校验证。其次，本研究仅采用CTT指标，未引入IRT或认知诊断模型（cognitive diagnosis models，CDM）进行参数等值与能力剖面分析，可能掩盖试题在不同能力层级考生中的功能差异。另外，AI模型仅选用单一文本大模型（Qwen-Plus），未比较集成图像理解模块的多模态模型在影像判读题生成中的效果，后续研究可探索“图文双模态输入+临床病例结构化提示”的生成范式。最后，本研究聚焦测量学指标，未系统评估AI试题的内容效度、表面效度及学生认知负荷体验，未来应结合专家Delphi法与答题轨迹分析，构建更立体的试题质量评价体系。

4 结语

综上所述，基于放射科历史题库的AI试题生成模型在难度预测上表现出优异的一致性，但在区分度模拟与高阶认知鉴别能力上仍存在系统性偏差。历史题库不仅是AI学习的先验数据源，更是实现“预测—实测”参数校准与命题质量可控性的核心标尺。未来应通过多模态模型迭代、IRT参数等值及动态反馈机制，推动AI命题从“文本生成”向“认知适配”跃迁，为放射影像学考核的标准化、精准化与智能化转型提供可验证、可推广的实践范式。

利益冲突　所有作者声明不存在利益冲突

作者贡献声明　史玉书：项目设计、撰写论文；秦潇潇、廖金生：项目实施；黄强、楼海燕：方案优化、统计分析；陈韶华、肖文波：项目指导、凝练主要观点

参考文献

[1] 陈欣，戴社教，赵暹，等. 不同类型试题对医学影像学考试评价的效能分析[J]. 中华医学教育探索杂志，2012，11(11)：1160-1163. DOI: 10.3760/cma.j.issn.2095-1485.2012.11.022.

[2] 刘朋，张洁，杨大为，等. 临床医学专业本科生影像学教学现状与改革[J]. 医学教育管理，2023，9(增刊1)：96-99. DOI: 10.3969/j.issn.2096-045X.2023.S1.023.

[3] 曲波. 推动AI与教育深度融合构建精准学习新路径[J]. 中华医学教育探索杂志，2025，24(10)：1297-1300. DOI: 10.3760/cma.j.cn116021-20250326-02133.

[4] 徐德林，史军华，姚秋丽，等. 医学院校试卷命题的问题分析与解决策略[J]. 遵义师范学院学报，2023，25(3)：121-123.

[5] Aster A, Laupichler MC, Rockwell-Kollmann T, et al. ChatGPT and other large language models in medical education: scoping literature review [J]. Med Sci Educ, 2025, 35(1): 555-567. DOI: 10.1007/s40670-024-02206-6.

[6] 国家医学考试中心. 国家医学考试中心创新性研究目录（2025—2026年度）[R]. 北京：国家医学考试中心，2025.

[7] 魏千，郝瑞. 基于AI互评的中国眼科住院医师规范化培训大语言模型应用现状分析[J]. 中华医学教育探索杂志，2025，24(12)：1585-1591. DOI: 10.3760/cma.j.cn116021-20250710-02166.

[8] 王斌，田士来，王小成，等. 以ChatGPT为代表的人工智能技术在医学教育中的机遇与挑战[J]. 医学教育研究与实践，2023，31(4)：402-405. DOI: 10.13555/j.cnki.c.m.e.2023.04.002.

[9] 平措玉珍，兰亚佳，马景璇，等. 基于经典测量理论和项目反应理论的工作内容量表（JCQ-22）评价与优化[J]. 现代预防医学，2025，52(17)：3220-3226. DOI: 10.20043/j.cnki.MPM.202503385.

[10] Koo TK , Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research [J]. J Chiropr Med, 2016, 15(2): 155-163. DOI: 10.1016/j.jcm.2016.02.012.

[11] 黄广仕，周梦强，韩春梅，等. 人工智能深度学习技术在医学考试试题难度预估中的应用研究[J]. 中华医学教育杂志，2021，41(10)：932-935. DOI: 10.3760/cma.j.cn115259-20201203-01668.

[12] Staziaki PV, Sarangi R, Parikh UN, et al. An objective structured clinical examination for medical student radiology clerkships: reproducibility study [J]. JMIR Med Educ, 2020, 6(1): e15444. DOI: 10.2196/15444.

[13] Rezaei-Zadeh M, Cerbin-Koczorowska M. A research roadmap for AI opportunities in student assessment for medical education [J]. BMC Med Educ, 2025, 25(1): 1653. DOI: 10.1186/s12909-025-08078-7.

[14] Zhang JX, Zuo CS. GRPO-LEAD: a difficulty-aware reinforcement learning approach for concise mathematical reasoning in language models [EB/OL]. [2026-01-05]. https://arxiv.org/abs/2504.09696.

[15] Pereira DSM, Mourão F, Ribeiro JC, et al. ChatGPT as an item calibration tool: psychometric insights in a high-stakes examination [J]. Med Teach, 2025, 47(4): 677-683. DOI: 10.1080/0142159X.2024.2376205.

（收稿日期：2026-01-29）

（本文编辑：邬小梅）