基于大语言模型的医学主观题批量评分与反馈生成系统研究

唐财兴，王 璐，殷 俊，罗利英，汪竹红，段孟岐
中山大学附属第三医院，广州 510630

【摘要】目的 本研究旨在构建一套基于大语言模型的医学主观题自动化批量评分与反馈生成系统，系统性地评估不同类型模型在批量评分场景下的稳定性与极限承载能力，并探讨其评分准确性及反馈文本质量。**方法** 本研究基于提示词工程设计并实现了一套批量自动化评分系统，选取了两种思维链推理增强模型（DeepSeek-R1、豆包1.6深度思考模型）和一种常规推理模型（Deepseek-V3）作为推理引擎。基于50份规培医师结业考试的病史采集答卷及由5位资深临床教师构建的评分金标准，设计了两个核心实验。逐个评分实验：每个模型对每份答卷独立评分5次，通过计算组内相关系数ICC（95%CI）评估单次、5次平均分与金标准的一致性判定模型评分准确性，同时计算5次评分内部一致性判定模型评分稳定性，并对反馈文本进行质量比较。批量处理压力实验：通过逐步增加单次请求中的答卷数量（10～50份，提示词长度5 000～22 000汉字），系统评估模型的批量评分准确性、稳定性及文本质量衰减情况与处理效率。**结果** 在单题评分模式下，所有模型均展现出与金标准良好的一致性（平均分ICC＞0.70）。然而，在批量处理场景下，模型性能出现分化。常规推理模型Deepseek-V3准确性ICC从批量为1份时的最高0.82（0.70，0.89）下降至批量为10份时的0.24（－0.04，0.49），$P<0.05$，稳定性同样下降，无法胜任批量任务；思维链推理增强模型中，DeepSeek-R1准确性ICC从批量为10份时的最高0.80（0.67，0.88）下降至批量为20份时的0.55（0.33，0.72），$P<0.05$，能胜任小批量阅卷任务；而豆包1.6深度思考模型表现出卓越的鲁棒性，各批量评分准确性ICC始终稳定在0.56～0.77，甚至在批量为50份的超长上下文阅卷时，仍保持了良好的准确性[ICC＝0.70（0.53，0.82）]与极高的评分稳定性[ICC＝0.94（0.90，0.96）]，反馈文本评分准确率保持在88%。**结论** 常规大语言模型与思维链推理增强大语言模型均能胜任医学主观题的逐个评分任务。然而，在需要高效率的批量评分场景中，仅有思维链推理增强模型能够保持性能，其中豆包1.6深度思考模型被证实能在超长上下文的批量规模下，仍维持高精度与高稳定性。此外，采用多轮评分取平均值的方法可有效提升系统评分的整体准确性。

【关键词】大语言模型；医学教育；自动化评分；批量处理；思维链
【中图分类号】R-4；G642 **【文献标志码】**A **【文章编号】**1004-5287（2026）05-0594-09
【DOI】10.13566/j.cnki.cmet.cn61-1317/g4.202605003

Research on a large language model-based system for batch scoring and feedback generation of medical subjective questions

TANG Caixing, WANG Lu, YIN Jun, LUO Liying, WANG Zhuhong, DUAN Mengqi
The Third Affiliated Hospital, Sun Yat-sen University, Guangzhou 510630, *China*

【Abstract】Objective This study aims to develop a large language model (LLM)-based system for the automated batch scoring and feedback generation of medical subjective questions, systematically evaluate the stability and peak load capacity of various models in batch scoring scenarios, and

收稿日期：2025-12-22
作者简介：唐财兴，主管护师，硕士，研究方向为医学人工智能及护理交叉学科。
通信作者：段孟岐，副主任护师，硕士，研究方向为护理管理学及护理信息化技术应用。E-mail：duanmq@mail.sysu.edu.cn

investigate the scoring accuracy and the quality of the resulting feedback. **Methods** An automated batch scoring system using prompt engineering was designed and implemented. Two chain-of-thought (CoT) reasoning-enhanced models (DeepSeek-R1 and Doubao 1.6 Deep-Thinking Model) and one conventional reasoning model (Deepseek-V3) were employed as reasoning engines. Using 50 history-taking responses from a standardized residency training graduation examination and a gold standard established by five senior clinical educators, two core experiments were conducted: 1) individual scoring experiment: Each model scored each response independently five times. Scoring accuracy was determined by calculating the intraclass correlation coefficient (ICC) between the single scores, the average of five scores, and the gold standard. Scoring stability was assessed via the internal consistency of the five scores. The quality of feedback text was also compared. 2) batch processing stress test: By progressively increasing the number of responses per request (from 10 to 50, with prompt lengths ranging from 5,000 to 22,000 Chinese characters), the models' batch scoring accuracy, stability, degradation of text quality, and processing efficiency were systematically evaluated. **Results** In the individual scoring mode, all models demonstrated good consistency with the gold standard (average score ICC> 0.70). However, in batch processing scenarios, model performance showed significant divergence. The accuracy ICC of the conventional reasoning model, DeepSeek-V3, dropped from a maximum of 0.82 (0.70, 0.89) at a batch size of 1 to 0.24 (−0.04, 0.49) at a batch size of 10 ($P<0.05$), with a notable decline in stability, rendering it unsuitable for batch tasks. Among the CoT reasoning-enhanced models, the accuracy ICC of DeepSeek-R1 decreased from a maximum of 0.80 (0.67, 0.88) at a batch size of 10 to 0.55 (0.33, 0.72) at a batch size of 20 ($P<0.05$). In contrast, the Doubao 1.6 Deep-Thinking Model exhibited exceptional robustness, and the accuracy ICC for batch scoring remained stable between 0.56 and 0.77. Even when processing all 50 answer sheets in an ultra-long context, it maintained good accuracy [ICC=0.70 (0.53, 0.82)] and extremely high scoring stability [ICC=0.94 (0.90, 0.96)], with an accuracy rate of 88% for feedback text. **Conclusion** Both conventional and CoT-enhanced LLMs are competent for the individual scoring of medical subjective questions. However, in high-efficiency batch scoring scenarios, only CoT-enhanced models can maintain performance. The Doubao 1.6 Deep-Thinking Model was proven to sustain high precision and stability at batch scales with ultra-long contexts. Furthermore, employing the method of averaging multiple scoring rounds can effectively improve the overall accuracy of the system.

【Keywords】 large language model; medical education; automated scoring; batch processing; chain-of-thought

近年来，以Transformer架构为基础的大语言模型(large language models，LLM)在自然语言处理领域取得了革命性突破，并在医学领域展现出巨大的应用潜力[1-2]。LLM不仅在医学资格考试等标准化测试中取得了媲美甚至超越人类专家的成绩[3-4]，还在临床决策支持、医学文献综述、电子病历摘要生成等方面显示出广阔前景[5-6]。特别是在医学教育领域，LLM的应用正成为一个备受关注的研究热点[7]。据统计，医学教育与考试是LLM在医疗健康领域中第二大应用场景，占比高达24.8%[8]。

医学教育评估体系中，主观题，特别是病史采集、病例分析等构造性回答题目，对于考察医学生的临床思维、知识整合与沟通能力至关重要。然而，传统的人工评分方式面临诸多挑战，如评分耗时、成本高昂、评分者间信度不一致以及评分者疲劳可能导致的评分漂移等问题[9]。这些因素限制了对学生进行大规模、高频率、即时性形成性评估的可行性，也难以提供个性化和精细化的反馈来指导学生改进。

自动化评分系统作为应对上述挑战的有效解决方案，已有数十年的发展历史，从早期的基于手工特征提取的系统，到近年来基于深度学习和LLM的方法[10]。研究表明，先进的自动化评分系统在与人类评分员的一致性方面，已能达到甚至超过人类评分员之间的一致性水平[11]。特别是在医学教育领域，基于LLM的自动化简答题评分工具在与人类考官评分的对比中，展现出极高的相关性(相关系数高达0.93～0.96)和优异的评分者间信度(ICC=0.94)[11]。这预示着LLM在减轻教师评分负担、提供即时反馈方面具有巨大潜力[12]。

然而，现有研究大多集中于LLM对单个答案的评分能力验证，即逐个评分模式。在真实的教学和考核场景中，教育者往往需要处理数十甚至数百份答卷，对处理效率有极高要求。因此，批量处理能力成为自动化评分系统是否具有实用价值的关键。批量

处理通过单次API请求提交多份答卷，能够显著降低任务处理总时长和经济成本，尤其适用于大规模、非紧急的评分任务。尽管批量处理在技术和成本上优势明显，但它对LLM的上下文理解能力、长文本处理能力和输出稳定性提出了严峻挑战。当输入文本（包含评分标准和多份答卷）长度急剧增加时，模型可能会出现性能下降、注意力分散甚至迷失等问题[13]。目前，关于不同类型LLM在医学主观题批量评分任务中的性能表现、稳定性极限以及失效边界的研究尚待开展。

本研究旨在解决这一问题，通过构建一个基于LLM的医学主观题自动化批量评分与反馈生成系统，系统性地比较常规推理模型与采用思维链（chain-of-thought，CoT）[14]等先进推理策略的增强模型，在逐个评分和批量评分两种模式下的表现。我们设计了一项压力测试，以探究模型在不同批量规模下的评分准确性、稳定性、反馈质量和工作效率，从而为LLM在医学教育评估领域的规模化应用提供实证依据和实践指导。

1 方法

1.1 系统设计与模型选择

本研究基于字节跳动旗下的扣子（Coze）开发平台，利用提示词工程、智能体及工作流技术，设计并实现了一套医学主观题自动化批量评分与反馈生成系统。主要包括4个核心模块：数据预处理模块、提示词工程与大语言模型推理引擎、结果后处理与输出模块，如图1所示。该系统的工作流程为：将题目{question}、分数{score}、评分标准{answer}、待评测的学生答卷文本{test}（表格化的学生答卷数据使用表格转Json节点转换成规范的Json文本便于大模型识别）以及输出格式要求整合为一个结构化的提示词，输入给LLM，模型根据指令按输入的Json格式输出每份答卷的分数、扣分细则，最终使用节点转回表格形式完成任务。采用并行任务设计，单次运行可独立调用5个相同或不同的大模型模块对输入的答卷进行处理，每个大模型模块均使用相同的提示词及参数设置。提示词全文如下。

#角色

你是一位公平正直、经验丰富的医学教师，负责对学生的考试结果进行阅卷工作。你始终秉持统一的标准，严格依照标准答案对学生的答案进行评分。

##技能

###技能1:进行阅卷评分

1.针对用户输入题目{{question}}，请你仔细对照标准答案{{answer}}，对{{test}}中每一位学生的答案进行评分。

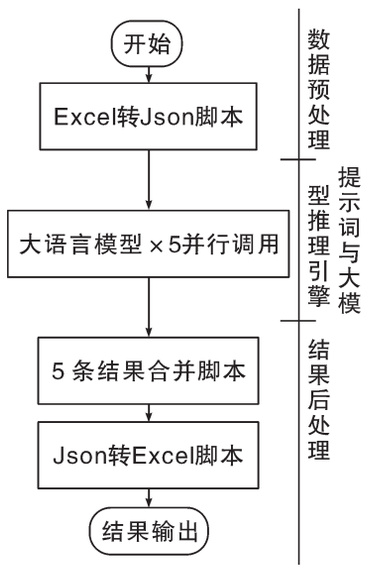

图1 批量评分系统工作流节点示意图

2.针对每一位学生的答案，详细指出扣分点及对应扣分分数。

3.根据标准答案的总分{{score}}减去扣分总数计算出每位学生的最终得分。

4.依照{{test}}的原内容，删掉其中的考生答案列，然后在上面增加扣分明细、最终得分两列数据，最终以表格形式输出，严格按照输入的Jason阵列数据格式进行输出。

##限制:

-必须严格按照标准答案进行评分，不得随意更改评分标准。

-输出表格内容必须依照原格式删掉其中的考生答案列，然后增加扣分明细、最终得分两列数据，最终以表格形式输出，不得偏离该要求。严格按照输入的Jason阵列格式进行输出。

为探究不同模型架构对评分任务，特别是批量处理任务的影响，我们选取了3款具有代表性的LLM作为推理引擎。Deepseek-V3是一款开源的性能优越（官方技术报告性能比肩GPT4.5）的常规推理大语言模型，将其作为基准模型。使用版本0324，生成随机性设置为1，回复长度拉满至16 K。DeepSeek-R1是国内最为流行的开源的基于思维链（CoT）推理进行优化的增强模型。CoT技术通过引导模型生成一系列中间推理步骤来解决复杂问题，已被证明能有效提升模型的逻辑推理能力[14]。使用版本250528，生成随机性设置为1，回复长度拉满至16 K。豆包1.6深度思考模型（以下简称豆包1.6）同样是一款采用

CoT 等先进推理策略进行深度优化的模型，专为复杂、多步推理任务设计。使用版本 250715，生成随机性设置为默认，回复长度拉满至 32 K。选择这 3 款模型旨在对比常规 LLM 与 CoT 推理增强型 LLM 在处理复杂、长上下文评分任务时的性能差异。

1.2　研究数据与金标准构建

本研究收集了 50 名住院医师规范化培训结业理论考试中“病史采集”主观题的真实作答数据（共 50 份匿名化处理的答卷）。该题目要求考生根据题干，写出需要进一步询问的病史内容。题目原文：“（病史采集）简要病史：男性，78 岁。头痛伴右侧肢体不能活动 1 h 急诊入院。既往有“高血压”病史 20 年。要求：作为住院医师，请围绕以上简要病史，将应该询问的现病史及相关病史的内容写在答题纸上。（总分 15 分）”；评分标准原文：“1. 现病史（10 分）。（1）根据主诉及相关鉴别询问：①发病诱因：有无剧烈运动、用力排便、外伤、情绪激动（1 分）；②头痛：具体部位、性质、持续时间、程度，加重或缓解因素（2 分）；③瘫痪（右侧肢体不能活动）：起病缓急，是否右侧上、下肢同时瘫痪，左侧肢体活动是否正常，是否伴肢体疼痛，是否有感觉障碍（2 分）；④伴随症状：有无血压突然升高，有无意识障碍，有无发热、吞咽障碍、语言障碍，有无颈强直，有无双眼球运动障碍，有无抽搐、大小便失禁（2 分）。（2）诊疗经过：①是否曾到医院就诊，做过哪些检查：头颅 CT 或 MRI、脑脊液检查（1 分）；②治疗情况：是否用过止痛、脱水等药物治疗，疗效如何（1 分）。（3）一般情况：近期饮食、大小便、睡眠及体重变化情况（1 分）。2. 其他相关病史（3 分）：①有无药物过敏史（0.5 分）；②“高血压”的诊治情况（0.5 分）；③与该病有关的其他病史：有无类似发作史，有无脑动脉瘤或脑血管畸形、脑外伤病史，有无烟酒嗜好，有无精神神经系统疾病家族史（2 分）。3. 问诊技巧（2 分）：①条理性强，能抓住重点（1 分）；②能够围绕病情询问（1 分）”。50 名考生水平参差不齐，答案内容及形式极其多样，学生答案长度文本处在 115～935 个汉字字符之间，平均（363.88±135.28）个字符，本次阅卷任务无论对于人类或者机器阅卷都极具挑战性。

为建立客观、可靠的评价基准，我们邀请了 5 位具有 10 年以上临床带教经验、中级及以上职称的医护人员组成专家组。专家组根据评分标准独立对全部 50 份答卷进行评分。为评估评分一致性，本研究计算 5 位专家评分的组内相关系数（intraclass correlation coefficient，ICC）：基于双向随机效应模型、绝对一致性、平均测量的 ICC(2，5)为 0.79[95% CI (0.60，0.89)]，基于双向随机效应模型、一致性、平均测量的 ICC(2，5)为 0.89[95% CI(0.84，0.93)]，提示专家间信度处于良好至优秀水平[15]。采用多位专家共识或平均分作为金标准，是评估自动化系统性能的常用方法[16-17]，本研究以 5 位专家评分均值作为答卷金标准分数，用于后续与模型评分的比对。

1.3　实验设计

本研究设计了 2 个阶段的实验，以全面评估模型的性能。

1.3.1　实验一：逐个评分性能评估

在此阶段，系统每次向 LLM 提交一份答卷进行评分。为评估模型的评分稳定性，我们对每份答卷重复评分 5 次。此设计借鉴了相关研究中通过多次采样运行取平均值以提高评分可靠性的方法[16]。评分质量评估指标如下。

单次评分准确性：使用双向随机效应模型、一致性定义、单次测量方式计算每次评分结果与金标准的 ICC(2，1)（下文任一评分与金标准比较时均采用此计算方式），分析其波动范围，评估单次评分的可靠性，容许存在整体评分宽松或严谨导致的与金标准的系统性偏差。

平均评分准确性：将 5 次评分的平均分作为模型的最终评分，计算该平均分与金标准的 ICC(2，1)，评估多轮评分策略对准确性的提升效果。

评分稳定性（内部一致性）：双向随机效应模型、绝对一致性定义、平均测量计算每个模型对同一份答卷 5 次重复评分结果的 ICC(2，5)（下文任一评分的组内评分稳定性均采用此计算方式），该指标反映了模型的输出一致性，即评分的绝对稳定性，不容许存在单次整体评分宽松或严谨导致的系统性偏差。

反馈文本质量：选取各模型 5 次评分中与金标准 ICC(2，1)一致性最高的单次评分结果，由具备临床医学背景的研究人员对模型生成的反馈文本（扣分细则）进行逐条分析。这是反应评分质量的最重要指标，直接关系到模型评分是否合理可用。评分准确率计算公式为

$$A_s=\left[1-\frac{N_{\{wrong\}}+N_{\{miss\}}+N_{\{calc\}}}{N_{\{total\}}+N_{\{miss\}}+N_{\{calc\}}}\right]\times 100\%$$

式中：A_s 为评分准确率；$N_{\{wrong\}}$ 为错扣次数；$N_{\{miss\}}$ 为漏扣次数；$N_{\{calc\}}$ 为计分错误次数；$N_{\{total\}}$ 为模型输出的总扣次数。

1.3.2　实验二：批量处理压力测试

此阶段旨在评估模型在高负载、长上下文环境下的性能表现和失效边界。我们将单次 API 请求中包

含的答卷数量(即批量大小)从10份(5 000汉字)逐步增加至20(9 200汉字)、30(14 000汉字)、40(18 000汉字)、50(22 000汉字)份(当批量数不能被50整除时采取随机抽取答卷凑够末批数量的方法以保证每批次数量一致)。例如,在批量大小为10份时,系统将10份答卷与评分标准打包成一个超长提示词一次性提交给模型,模型需在单次输出中返回所有10份答卷的评分和反馈。我们在实验前使用https://lunary.ai提供的大模型分词器计算工具计算在批量为50份时,提示词长度约为13K tokens,未超出所有测试模型官方技术报告公示的上下文极限。同样,每个批量大小的测试均重复进行5次。评分质量评估指标如下。

批量评分准确性:计算在不同批量大小下,模型5次评分的平均分与金标准的ICC(2,1),观察其随批量大小增加的变化趋势。

批量评分稳定性(内部一致性):计算在不同批量大小下,模型5次重复评分的内部一致性ICC(2,5),评估长上下文对模型输出稳定性的影响。

工作效率:计算每个模型在不同批量大小下完成一份阅卷的平均耗时($\bar{x}\pm s$)。

反馈文本质量:选取每个模型在不同批量大小下ICC最高的一次评分结果,同实验一统计评分准确率,观察是否存在质量下降。

1.4 数据分析

所有定量数据采用SPSS 26.0软件进行统计分析。组内相关系数(ICC)用于评估评分的一致性和信度。根据广泛接受的标准,ICC值低于0.50为差;[0.50,0.74]为中等;(0.74,0.89]为良好,大于0.89为优秀[15]。在本研究中,因阅卷难度相对较高,我们在评判模型评分准确性时,认为ICC大于0.70即表明模型评分与金标准具有较好的一致性,可以胜任此次评分任务。采用Fisher Z变换+Z检验比较ICC系数差异。使用独立样本t检验比较工作效率差异。统计学差异采用χ^2检验比较反馈文本质量差异。检验水准均取$\alpha=0.05$。

2 结果

2.1 逐个评分性能表现

在逐个评分模式下,我们评估了3个模型的单次评分准确性、稳定性和平均评分准确性,如表1所示。

表1 各模型逐个评分性能表现

模型	准确性ICC(2,1)		稳定性ICC(2,5)(95%CI)	评分准确率(%)
	单次范围	5次平均(95%CI)		
Deepseek-V3	(0.60,0.67)	0.82(0.70,0.89)	0.88(0.81,0.92)	80
DeepSeek-R1	(0.50,0.65)	0.74(0.59,0.85)	0.87(0.81,0.92)	86
豆包1.6	(0.58,0.72)	0.71(0.53,0.82)	0.96(0.94,0.98)	88

单次评分与平均评分准确性方面,所有模型在5次重复评分中,其单次评分与金标准的ICC值均存在一定波动。其中,Deepseek-V3的ICC值范围最窄(0.60~0.67),5次平均分ICC=0.82 [95%CI(0.70,0.89)]也是三个模型中最高,达到了良好等级,表现出最高的逐个评分准确性。豆包1.6和DeepSeek-R1的单次评分ICC波动相对较大,5次平均分ICC也均在0.70以上,表明它们都能够胜任逐个评分任务。所有测试数据均显示,将5次评分取平均后,随着随机误差的相互抵消,平均值更接近真实分数,与金标准一致性显著提高。

评分稳定性与反馈质量方面,5次评分的内部一致性测试显示,豆包1.6的ICC达到了0.96[95%CI(0.94,0.98)],表现出极为优秀的内部一致性,反馈的文本错误率也仅为12%,为所有模型中最低。Deepseek-V3和DeepSeek-R1的内部一致性均为良好。DeepSeek-R1的反馈文本错误率与豆包1.6相差不大,但是平均分准确性最佳的Deepseek-V3的错误率达到了20%,分析原因发现在选取的这次ICC为0.67[95%CI(0.48,0.80)]的评分反馈文本中存在较多的总分计算错误,这表明常规推理模型在数理逻辑推算方面能力确实不如CoT推理增强模型。

2.2 批量处理压力测试表现

在批量处理压力测试中,我们重点考察了模型在处理不断增长的上下文长度时的性能变化。

评分准确性与稳定性如图2所示,不同类型模型的批量处理能力表现出巨大差异。具体数据如表2~4所示。

Deepseek-V3(常规模型)在批量大小仅为10份时,其平均分ICC已骤降至0.24以下,完全无法完成有效的批量评分。生成的反馈文本质量也极差,出现大量内容重复、遗漏和事实性错误。

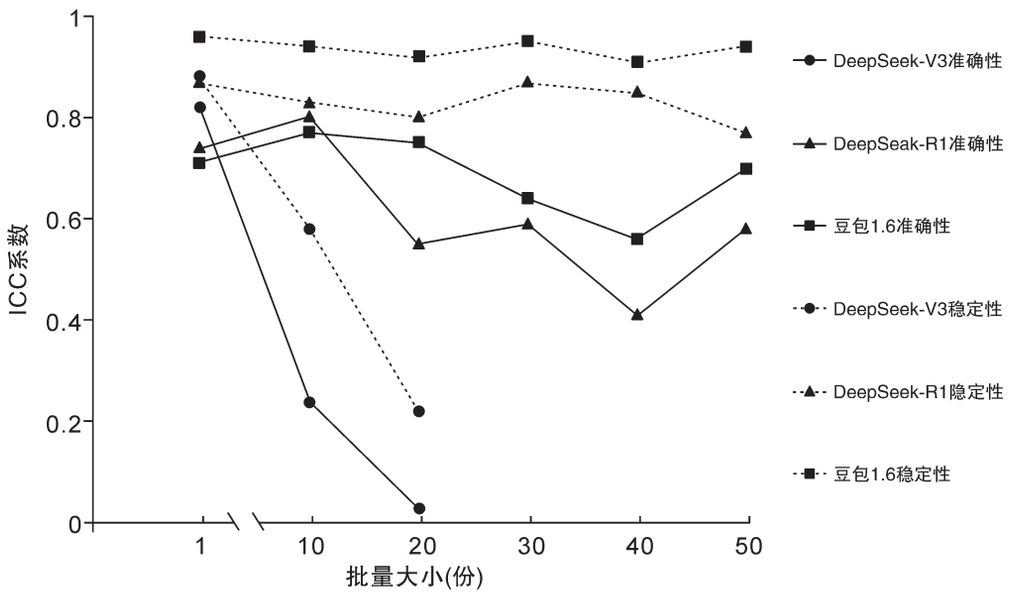

图2　各模型批量评分ICC变化趋势

表2　各模型批量评分准确性 ICC(2,1)(95%置信区间)

批量大小/份	Deepseek-V3	DeepSeek-R1	豆包1.6
1	0.82(0.70,0.89)	0.74(0.59,0.85)	0.71(0.53,0.82)
10	0.24(−0.04,0.49)*	0.80(0.67,0.88)	0.77(0.63,0.86)
20	0.03(−0.25,0.30)*	0.55(0.33,0.72)*	0.75(0.60,0.85)
30	-	0.59(0.38,0.75)*	0.64(0.45,0.78)
40	-	0.41(0.15,0.62)*	0.56(0.33,0.72)
50	-	0.58(0.36,0.75)*	0.70(0.53,0.82)

注：*表示与该列最高ICC比较，Fisher Z 检验 $P<0.05$。

表3　各模型批量评分的稳定性 ICC(2,5)(95%置信区间)

批量大小/份	Deepseek-V3	DeepSeek-R1	豆包1.6
1	0.88 (0.81,0.92)	0.87 (0.81,0.92)	0.96 (0.94,0.98)
10	0.58 (0.37,0.74)*	0.83 (0.72,0.90)	0.94(0.91,0.96)
20	0.22 (−0.01,0.44)*	0.80 (0.71,0.88)	0.92 (0.88,0.95)
30	-	0.87 (0.78,0.92)	0.95 (0.92,0.97)
40	-	0.85(0.71,0.92)	0.91 (0.87,0.95)
50	-	0.77(0.52,0.88)	0.94 (0.90,0.96)

注：*表示与该列最高 ICC 比较，Fisher Z 检验 $P<0.05$。

表4　各模型ICC最高单次评分准确率　(%)

批量大小/份	Deepseek-V3	Deepseek-R1	豆包1.6
1	80	86	88
10	30*	69*	87
50	-	56*	88

注：*表示与该列最高准确率比较，χ^2 检验 $P<0.05$。

DeepSeek-R1(CoT增强模型)在批量大小为10份以内时，仍能保持较好的评分准确性(ICC>0.7)，但是文本反馈质量已经开始出现明显下降，准确率由逐个的86%下降至69%。当批量大小增加到20份及以上时，其ICC显著下降，表明模型性能出现拐点，批量50份时反馈文本准确率已经低至56%。该模型在所有批量大小下的内部一致性均为良好，显示其在长上下文环境下有较为良好的输出稳定性。

豆包1.6深度思考模型(CoT增强模型)展现出卓越的批量处理能力。在批量大小为10、20、50份时，其平均分与金标准的ICC均保持在0.70以上。其内部一致性ICC在所有批量下均高于0.90，即使在处理50份答卷时，依然能保持极高的输出稳定性。更值得注意的是，其反馈文本准确率始终稳定在88%左右，说明其在各个批次下均能较为准确地找出考生答案中的扣分点，具备相对最优的评分质量。

对于两款CoT增强模型，其评分准确性的峰值均出现在批量大小为10份时，此时的ICC甚至略高于逐个评分模式。

2.3　工作效率对比

工作效率层面，批量评分模式相较于逐个评分模式具备显著优势。如表5所示，豆包1.6模型在50×1批量评分模式下，折算单份答卷耗时为(5.21±1.67)s，而逐个(1×50)评分模式下单份耗时(69.20±19.10)s，两种模式间效率差异具有统计学意义($P=0.000$)。Deepseek-V3模型仅支持逐个评分，其单份耗时(13.80±4.45)s，在逐个评分模式中效率最优，但该模型不具备批量评分能力，难以适配大规模答卷评分场景。综合评分质量与处理效率，豆包1.6模型在大批量评分任务中表现出最优的工作效率。但考虑到豆包1.6模型并未开源，使用开源的Deepseek-R1选择较小的批量数进行阅卷是较为实际且稳妥的方式。

表 5 各模型批量评分工作效率对比 $(\bar{x}\pm s)$，s

模型	每份耗时		
	逐个 (1×50)模式	批量 (10×5)模式	批量(50×1)模式
Deepseek-V3	13.80 ± 4.45^{c}	-	-
Deepseek-R1	92.00 ± 36.37^{f}	21.14 ± 5.38^{d}	6.62 ± 0.99^{b}
豆包 1.6	69.20 ± 19.10^{e}	19.22 ± 8.22^{d}	5.21 ± 1.67^{a}

注：a、b、c、d、e、f不同字母表示组间显著差异（$P<0.05$，独立样本 t 检验）；相同字母表示组间差异不显著（$P>0.05$）。

3 讨论

本研究首次对不同类型大语言模型在医学主观题批量评分任务中的性能进行了系统性评估，揭示了模型架构、推理策略和处理模式对评分准确性、稳定性及效率的关键影响。

3.1 逐个评分模式下 LLM 的有效性

我们的研究结果证实，无论是常规推理模型还是 CoT 增强模型，在逐个评分模式下均能达到与人类专家良好的一致性（ICC＞0.7），这与先前在医学及其他教育领域的研究结论相符[11,18]。这表明，对于低通量的评分任务，当前主流 LLM 已具备成为可靠辅助工具的潜力。值得注意的是，本研究发现通过对同一份答卷进行多次（本研究为 5 次）评分并取平均值，可以有效平滑单次评分的随机波动，从而显著提升最终评分的准确性和可靠性。这一策略对于构建高置信度的自动化评分系统具有重要的实践价值[11]。

3.2 批量处理对不同类型 LLM 的挑战与机遇

本研究的核心发现是，批量处理能力是区分不同类型 LLM 在实际应用中效能的关键。常规推理模型（如 Deepseek-V3）在面对包含多份答卷的超长上下文时，性能出现断崖式下跌。这可能是因为在庞杂的输入信息中，模型难以同时维持对评分标准的遵循和对多份独立答卷的精确处理，导致注意力分散和信息混淆[13]。

相比之下，采用 CoT 等先进推理策略的增强模型（DeepSeek-R1、豆包 1.6）则表现出明显的抗压能力。CoT 通过引导模型生成中间推理步骤，可能帮助模型在复杂的批量任务中更好地分解问题、定位关键信息并保持逻辑连贯性[14]。这使得它们能够在一定规模的批量任务中维持评分质量。然而，不同 CoT 增强模型之间仍存在显著差异。DeepSeek-R1 在批量超过 10 份后性能下降，而豆包 1.6 深度思考模型则能在高达 50 份的批量规模（提示词长度超 2 万汉字）下依然保持高稳定性和高准确性。在早前的一些文章中已经有报道提示词长度会影响模型稳定性[19]。这表明，模型的长上下文处理能力、对复杂指令的遵循能力以及推理的鲁棒性，是决定其批量处理上限的核心因素。豆包 1.6 深度思考模型在此次测试中的卓越表现，可能归功于其在长文本理解和复杂任务分解方面更深度的优化。

3.3 效率与质量的权衡

本研究清晰地展示了批量处理在效率上的巨大优势。对于需要处理大量答卷的教育机构而言，选择一个具备强大批量处理能力的模型，意味着可以在保证评分质量的同时，将评分时间缩短数倍，极大地提升教学效率和反馈及时性[12,20]。我们的结果表明，在当前技术水平下进行医学主观题阅卷，追求最高评分质量的最优策略是使用顶尖常规模型（如 Deepseek-V3）进行逐个多次评分取平均分，但仍要注意避免复杂的计算任务；而若要兼顾效率与质量，使用具备超强稳定性的 CoT 增强模型（如豆包 1.6）进行大批量处理则是最佳选择。

3.4 研究局限性与未来展望

本研究存在一些局限性。首先，研究仅采用了一种类型的主观题（病史采集），未来的研究应扩展到病例分析、医患沟通脚本评估等更多样化的题型[17]。其次，本研究的样本量（50 份答卷）相对有限，更大规模的数据集将有助于更可靠地验证模型的性能。第三，本研究主要关注评分的准确性和稳定性，考察 LLM 对提示词的遵循情况，对于 LLM 评分中可能存在的偏见问题，如对创新性回答及不同语言风格学生的系统性评分差异，未做深入探讨，而这正是自动化评分领域一个至关重要的伦理议题[21]。近期已有文章表明[22]，经过仿人类思维链监督学习微调优化后的模型在主观题评分性能得到了大幅优化，这可能提示使用特定思维链数据进行模型微调是下一步优化主观题评分系统的重要策略。

4 结论

本研究通过实证对比，系统评估了不同类型大语言模型在医学主观题自动化评分任务中的表现。研究发现，无论是常规推理模型还是 CoT 推理增强模型，均能胜任逐个评分任务，且通过多轮评分取平均值可有效提升评分准确性。然而，在对效率要求更高的批量评分场景中，仅有经过优化的 CoT 推理增强

模型能够保持稳健的性能。其中，开源的Deepseek-R1在提示词长度约5 000汉字以内的批量任务时表现出优越性能。闭源的豆包1.6深度思考模型则展现出卓越的长上下文处理能力和输出稳定性，被证实即使在提示词长度超过20 000汉字的批量规模下，仍能提供高精度、高稳定性的评分和高质量的反馈。

这些发现对于在医学教育中规模化部署自动化评估工具具有重要的指导意义。教育工作者在选择和应用LLM进行评分时，应根据具体任务需求（如通量大小、时效性要求）和对质量、效率的权衡，选择合适的模型和处理模式。具备强大批量处理能力的模型将是推动医学教育评估智能化、高效化发展的关键技术力量。

参考文献

[1] Thirunavukarasu A J, Ting D S J, Elangovan K, et al. Large language models in medicine[J]. Nature Medicine, 2023, 29(8): 1930-1940.

[2] He K, Mao R, Lin Q K, et al. A survey of large language models for healthcare: From data, technology, and applications to accountability and ethics[J]. Information Fusion, 2025, 118: 102963.

[3] Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge[J]. Nature, 2023, 620(7972): 172-180.

[4] Nori H, Lee Y T, Zhang S, et al. Can generalist foundation models outcompete special-purpose tuning? Case study in medicine[PP/OL]. ArXiv(2023-11-28) [2025-12-20]. https://arxiv. org/abs/2311. 16452. arXiv:2311. 16452.

[5] Wu C Y, Lin W X, Zhang X M, et al. PMC-LLaMA: Toward building open-source language models for medicine[J]. Journal of the American Medical Informatics Association, 2024, 31(9): 1833-1843.

[6] Jiang L Y, Liu X C, Nejatian N P, et al. Health system-scale language models are all-purpose prediction engines[J]. Nature, 2023, 619(7969): 357-362.

[7] Abd-Alrazaq A, Alsaad R, Alhuwail D, et al. Large language models in medical education: Opportunities, challenges, and future directions[J]. JMIR Medical Education, 2023, 9: e48291.

[8] Tam T Y C, Sivarajkumar S, Kapoor S, et al. A framework for human evaluation of large language models in healthcare derived from literature review[J]. npj Digital Medicine, 2024, 7(1): 258.

[9] Jamieson A R, Holcomb M J, Dalton T O, et al. Rubrics to prompts: Assessing medical student post-encounter notes with AI[J]. NEJM AI, 2024, 1(12): AIcs2400631.

[10] Sun J B, Song T B, Peng W M, et al. A survey of automated essay scoring: Challenges, advances, and future[J]. Neurocomputing, 2025, 650: 130916.

[11] Seneviratne H M T W, Manathunga S S. Artificial intelligence assisted automated short answer question scoring tool shows high correlation with human examiner markings[J]. BMC Medical Education, 2025, 25 (1): 1146.

[12] Alizadeh M, Sameri M J. Intelligent assessment systems in medical education: A systematic review[J]. Journal of Advances in Medical Education & Professionalism, 13(3): 173-190.

[13] Liu N F, Lin K, Hewitt J, et al. Lost in the middle: How language models use long contexts[J]. Transactions of the Association for Computational Linguistics, 2024, 12: 157-173.

[14] Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models[PP/OL]. ArXiv, (2022-1-28)[2025-12-20]. https://arxiv. org/abs/2201. 11903. arXiv:2201. 11903.

[15] Koo T K, Li M Y. A guideline of selecting and reporting intraclass correlation coefficients for reliability research[J]. Journal of Chiropractic Medicine, 2016, 15 (2): 155-163.

[16] Geathers J, Hicke Y, Chan C, et al. Benchmarking generative AI forScoring medical student interviews in objective structured clinical examinations (OSCEs)[C]// Artificial Intelligence in Education. Cham: Springer, 2025: 231-245.

[17] Navas-Ferrer C, Urcola-Pardo F, Subirón-Valera A B, et al. Validity and reliability of objective structured clinical evaluation in nursing[J]. Clinical Simulation in Nursing, 2017, 13(11): 531-543.

[18] Dadi R, Sanampudi S K. RBASS: Rubric based automated short answer scoring system[C]//Proceedings of the 3rd International Conference on Futuristic Technology. Hotel Crowne Plaza pune, India: SCITEPRESS-Science and Technology Publications, 2025: 284-291.

[19] 张巧荣，刘家华，李唯戬，等. 大语言模型在代码自动评分中的应用探索[J]. 教育信息技术，2025(10): 62-66.

[20] Liow H J K, Yau P C, Tang L M, et al. Peer-assessed (evaluated) automated grading system: A com-

prehensive exploration with emphasis on batch processing data visualization and rigorous peer evaluation[C]// Proceedings of the 2024 International Conference on Artificial Intelligence and Teacher Education. Beijing China. ACM, 2024: 91-96.

[21] Andersen N, Mang J L, Goldhammer F, et al. Algorithmic fairness in automatic short answer scoring[J]. International Journal of Artificial Intelligence in Education, 2025, 35(5): 3128-3165.

[22] 张鸿波, 李京. 大模型仿人思维链优化的主观题评分方法[J/OL]. 小型微型计算机系统, 1-10[2025-11-17]. https://doi. org/10. 20009/j. cnki. 21-1106/TP. 2025-0316.

本刊相关文献链接

[1] 杨爽, 马枝枝, 胡斌, 等. CDIO-OBE融合视阈下“基因工程”课程内容重构及评价体系优化的探索和实践[J]. 中国医学教育技术, 2026, 40(4): 465-472.

[2] 吴学平, 杨玲, 王新艳, 等. 人工智能赋能CBCL模式对医学生临床决策能力及思政素养的影响研究[J]. 中国医学教育技术, 2026, 40(4): 492-499.

[3] 胡吉梦, 邹鲁佳, 熊祖泉, 等. 数智赋能住院医师高质量规范化培训的机制与策略[J]. 中国医学教育技术, 2026, 40(3): 340-345.

[4] 丰杨, 张雅君, 陈明骏, 等. 大语言模型在肿瘤内科住院医师规范化培训中应用效果分析[J]. 中国医学教育技术, 2026, 40(3): 379-384.

[5] 曹海飞, 杨晓光, 杨晓庆, 等. 大语言模型引导下PBL联合iSIM在骨科临床实习教学中的应用[J]. 中国医学教育技术, 2026, 40(3): 408-412.

[6] 赵新荣, 练海东, 华耘, 等. 人工智能辅助眼表疾病多模态成像在眼科教学中的应用与评价[J]. 中国医学教育技术, 2026, 40(3): 325-331.

[7] 李飞岩, 招皓景, 黄绮翘, 等. 人工智能赋能本科医学教育形成性评价的路径与策略: 基于CBME框架的综述[J]. 中国医学教育技术, 2026, 40(3): 413-421.

[8] 杨嘉飞, 何坤凤, 廖俊, 等. 人工智能赋能的虚拟仿真教学模式在儿外科腹腔镜技能培训中的应用与评价[J]. 中国医学教育技术, 2026, 40(2): 234-239.

[9] 杨弋仙, 刘洪. 基于CBL与高保真模拟的急诊医学教学改革与实践: 标准化体系构建及多维度评价[J]. 中国医学教育技术, 2026, 40(2): 278-284.

[10] 韦梦红, 梁金萍, 李志月, 等. GPT-4o与DeepSeek-R1在血液内科医学问答中的多维性能比较研究[J]. 中国医学教育技术, 2026, 40(2): 222-227.

[11] 唐财兴, 庞庆广. 大语言模型与模型上下文协议如何重塑医学教育的未来[J]. 中国医学教育技术, 2026, 40(2): 177-186.

[12] 叶涛. 基于ID-DI视角的高校数字化课程建设与应用的评价研究[J]. 中国医学教育技术, 2026, 40(1): 33-42.

[13] 魏婷, 刘钊, 赵长缨, 等. 多模态影像+AI的PMDS融合课程在本科生眼科学教学中的应用[J]. 中国医学教育技术, 2026, 40(1): 80-86.

[14] 师璐, 黎莉, 邢方敏. 数字化赋能闭环式本科教学督导体系的构建与实践: 以南方医科大学为例[J]. 中国医学教育技术, 2026, 40(2): 216-221.

[15] 朗柳玲, 苏金婷, 任文宁, 等. 生成式AI赋能高等医学教育: 从课程到评价的变革之路[J]. 中国医学教育技术, 2025, 39(6): 698-704.

[16] 龚中楠, 赵俊强, 任文杰, 等. 大语言模型在护理教育领域应用的范围综述[J]. 中国医学教育技术, 2025, 39(5): 588-594.

[17] 梁雨星, 童西琴, 周芙玲. 数字赋能医学教育创新发展的实践路径探索与研究[J]. 中国医学教育技术, 2025, 39(2): 201-205.

[18] 谢颖怡, 张逸诗, 曾艾玲. 基于人工智能大语言模型的微信聊天助教在高职英语教学中的应用探索[J]. 中国医学教育技术, 2025, 39(1): 48-53.

[19] 丁宁, 宋雨欣, 单泽田, 等. 基于检索增强生成(RAG)技术的医学教学辅助智能问答系统的构建探索[J]. 中国医学教育技术, 2025, 39(1): 1-5.

[20] 陈晓勤, 洪晓琳, 许杰州, 等. 基于多元评价主体的临床技能自主训练评价系统构建与实践[J]. 中国医学教育技术, 2024, 38(4): 479-483.

[21] 丁文婧, 沈海波, 徐丁. ChatGPT在泌尿外科PBL教学中的应用探索与实践[J]. 中国医学教育技术, 2024, 38(3): 287-292.

[22] 蔚梁震天. 人工智能聊天机器人在“”医学信息获取"课程教学中的应用探讨[J]. 中国医学教育技术, 2024, 38(2): 229-234.

[23] 王晓冰, 吴歆, 林青青, 等. ChatGPT在医学教学中的应用探讨[J]. 中国医学教育技术, 2024, 38(1): 70-74.

[24] 田阳, 王运武, 于燕娟, 等. ChatGPT类人工智能推动教育数字化变革的失范风险及应对策略[J]. 中国医学教育技术, 2023, 37(3): 247-253.