·数字化口腔医学·

大语言模型在口腔执业医师考前学习中的应用比较研究

庞潇[1] 刘畅[1] 范嘉豪[2] 黄艳[1] 孙亚楠[2] 刘济远[1]
1.口腔疾病防治全国重点实验室 国家口腔医学中心 口腔疾病国家临床医学研究中心 四川大学华西口腔医院信息管理部，成都 610041；
2.四川大学计算机学院，成都 610065

[摘要] **目的** 比较不同的大语言模型（LLM）在口腔执业医师考前学习中的应用表现，重点评估其在答题、解析和教学效果方面的差异，为LLM在口腔医学教育中的应用提供参考。**方法** 设计3个评估场景，包括选择正确答案、答案解析和对抗性测试。选择DeepSeek-R1、通义千问2.5-MAX、豆包1.5 Pro、星火Spark-X1、文心一言4.0 Turbo、GPT-4o和华西口腔智联大模型进行对比测试，评估的指标为准确率、净正确率和教学效果评分。**结果** 在选择正确答案场景方面，所有LLM均达到了及格线，华西口腔智联大模型的准确率最高（84%）。在答案解析场景方面，华西口腔智联大模型的解析净正确率最高（92%），其次为DeepSeek-R1（89%）。在教学效果评分中，华西口腔智联大模型在相关性、实用性和清晰度方面得分最高，GPT-4o在简洁性方面分数最高。在对抗性测试场景下，华西口腔智联大模型和DeepSeek-R1分别在准确率和净正确率方面下降幅度最小。**结论** 在口腔执业医师考前学习应用场景中，针对口腔医学进行知识增强的LLM（如华西口腔智联大模型）表现优于以中文语料库为核心预训练的推理LLM（如DeepSeek-R1），优于以英文语料库为核心预训练的LLM（如GPT-4o）。在对抗性测试中，所有模型的表现均有所下降。未来相关模型的研究可针对评估中表现不足之处进行优化，提升其在口腔执业医师考前学习中的应用效果。

[关键词] 口腔医学； 大语言模型； 口腔执业医师考试； 口腔医学教育
[中图分类号] R78 **[文献标志码]** A **[doi]** 10.7518/hxkq.2025308

本文链接 开放科学标识码

Comparative study on the application of large language models in pre-exam learning for the Chinese dental licensing examination

Pang Xiao[1], Liu Chang[1], Fan Jiahao[2], Huang Yan[1], Sun Ya'nan[2], Liu Jiyuan[1]
1. State Key Laboratory of Oral Diseases & National Center for Stomatology & National Clinical Research Center for Oral Diseases & Dept. of Information Management, West China Hospital of Stomatology, Sichuan University, Chengdu 610041, China; 2. College of Computer Science, Sichuan University, Chengdu 610041, China

Surpported by: Industry-University-Research Innovation Fund for Chinese Universities (2024XL004); Sichuan Science and Technology Program (2025ZNSFSC0459); Sichuan University General Project on AI-Empowered Comprehensive Reform of Innovative Practical Education (2024-80)

Correspondence: Liu Jiyuan, E-mail: wchsljy@scu.edu.cn

[Abstract] **Objective** This study aims to compare the performance of different large language models (LLMs) in pre-exam learning for the Chinese dental licensing examination, with a focus on evaluating their differences in answering, explanation, and teaching effectiveness, to provide a reference for the application of LLMs in dental education. **Methods** Three evaluation scenarios were designed: selecting correct answers, providing answer explanations, and adversarial testing. DeepSeek-R1, Qwen 2.5-MAX, Doubao 1.5 Pro, Xinghuo Spark-X1, ERNIE 4.0 Turbo, GPT-4o, and Huaxi Zhilian

[收稿日期] 2025-07-28
[基金项目] 中国高校产学研创新基金（2024XL004）；四川省科技计划资助（2025ZNSFSC0459）；四川大学人工智能赋能创新型实践教育综合改革研究专项（2024-80）
[第一作者] 庞潇，工程师，硕士，E-mail：xiaopang@scu.edu.cn
[通信作者] 刘济远，副主任医师，博士，E-mail：wchsljy@scu.edu.cn

were selected for comparative testing. Evaluation metrics included accuracy, net accuracy, and pedagogical effectiveness. **Results** In the scenario of selecting correct answers, all LLMs exceeded the passing threshold, with Huaxi Zhilian achieving the highest accuracy (84%). In the answer explanation scenario, Huaxi Zhilian demonstrated the highest net accuracy (92%), followed by DeepSeek-R1 (89%), among models. Regarding pedagogical effectiveness, Huaxi Zhilian ranked highest in relevance, practicality, and clarity, whereas GPT-4o led in conciseness, among the investigated LLMs. In adversarial testing, Huaxi Zhilian and DeepSeek-R1 exhibited the smallest declines in accuracy and net accuracy, respectively, among the tested models. **Conclusion** In pre-exam learning for the Chinese dental licensing examination, knowledge-enhanced LLMs specifically optimized for dentistry (e.g., Huaxi Zhilian) outperform reasoning LLMs pretrained on general Chinese corpora (e.g., DeepSeek-R1) and those primarily trained on English corpora (e.g., GPT-4o). However, the performance of all models declines under adversarial conditions. Future research should focus on addressing identified weaknesses to enhance the utility of LLMs in dental education further.

[Key words] stomatology; large language model; Chinese dental licensing examination; dental education

随着大型语言模型（large language models，LLM）近年来在自然语言处理（natural language processing，NLP）领域的突破和发展[1-2]，其在医学领域的实践愈加广泛，在医学教育中的应用潜力也日益凸显，比如帮助医学生学习、解答执业医师考试试题等[3]。口腔执业医师考试是国内口腔行业的资格准入考试[4]，近2年的平均通过率为50%，以百分制计分，平均分约为60分[5]，表明仍需投入更多教育资源来帮助考生通过考试。

在多个英语语境的医学类考试的答题和解析场景中，LLM已经有良好的表现[6-7]，许多医学生已经开始将LLM应用于考前学习。在非英语医学的考试中，LLM的表现不一[8-10]，经常出现输出不可靠或不准确的“幻觉”问题[11]。在口腔医学领域，韩彩玲等[12]通过比较不同LLM对口腔诊疗问题、口腔健康咨询、口腔执业医师考试试题的回答，评估了其在辅助诊疗和健康咨询中的应用表现，结果表明，LLM具备辅助获取医疗信息的应用可行性，但其输出仍存在一定的错误风险，在实际应用中需审慎判断。

本研究聚焦于口腔执业医师考前学习场景，将试题按学科和题型分类，比较不同LLM的应用表现。本研究设计了选择正确答案、答案解析、对抗性测试共3个评估场景，选择DeepSeek-R1、通义千问2.5-MAX、豆包1.5 Pro、星火Spark-X1、文心一言4.0 Turbo、GPT-4o和华西口腔智联大模型，作为代表性的LLM，采用准确率、净正确率和教学效果评分作为模型表现的评估指标。本研究旨在通过评估分析不同LLM在口腔执业医师考前学习中的表现，为LLM将来在口腔医学教育领域的深入应用提供理论支持和参考。

1　材料和方法

1.1　不同LLM在口腔执业医师考前学习各场景中的表现比较

根据口腔执业医师考试的题型和学科分布，本研究收集了200道历年试题，分为4个学科：口腔颌面外科、口腔全科、口腔修复学和基础医学，每个学科包含50道试题。试题的类型包含单项选择题（A1题型，n=80）、病例摘要型选择题（A2题型，n=80）、病例组型/串型选择题（A3/A4题型，n=20）和标准配伍题（B1题型，n=20），题型和要求如表1所示。

在1名具有20年教学经验的资深口腔医学专家的指导下，笔者对每个试题的题干和选项都进行了细致修改，比如改写题干的语法、牙位、逻辑表述、选项顺序等，防止LLM的训练语料库包含了原试题，每个修改都经过2名临床经验在10年以上的口腔医学专家的验证。

根据LLM评估平台FlagEval[13]公布的榜单，笔者选择了其中性能表现良好的模型，以中文语料库为核心预训练的模型包括DeepSeek-R1[14]、通义千问2.5-MAX[15]、豆包1.5 Pro[16]、星火Spark-X1[17]、文心一言4.0 Turbo[18]，以英文语料库为核心预训练的GPT-4o（2025-03-27版本）[19]，另外还纳入了华西口腔智联大模型。华西口腔智联大模型融入了口腔医学领域知识，由于资源限制，尚未全面向公众开放，其团队正在逐步实施开放计划[20-21]。DeepSeek-R1、豆包1.5 Pro、星火Spark-X1、华西口腔智联大模型具备深度思考模式，在测试时均开启深度思考以获得最佳输出表现。

表 1 题型和要求

Tab 1 Question types and evaluation requirements

题型	样本量	回答要求	示例
A1	80	评估考生对基本原理的理解和记忆，要求考生从题目给出的选项中选择1个正确答案	上颌前磨牙区的麻醉方式是（） A.阻滞麻醉；B.浸润麻醉；C.局部麻醉；D.冷冻麻醉；E.全身麻醉
A2	80	评估考生理解并详细评估临床病例的能力，要求根据题干中给出的病历摘要，从选项中选择1个正确答案	患者，男性，40岁。诉左上颌后牙出现咬合痛7天求治。口腔检查：左上第一磨牙近远中向纵折，松动Ⅰ度，叩诊（+），牙龈稍红肿。拟局麻下拔除，采用的麻醉方法是（） A.上牙槽后神经阻滞麻醉+眶下神经阻滞麻醉；B.上牙槽后神经阻滞麻醉+腭前神经阻滞麻醉；C.上牙槽后神经阻滞麻醉+近中颊根局部浸润麻醉；D.上牙槽后神经阻滞麻醉+近中颊根局部浸润麻醉+眶下神经阻滞麻醉；E.上牙槽后神经阻滞麻醉+近中颊根局部浸润麻醉+腭前神经阻滞麻醉
A3/A4	20	评估考生全面的临床分析能力，题干中给出一个临床情景，要求考生进行综合分析和判断，回答一系列问题，每个问题都需要选择1个正确的答案。	一男性患者不慎被玻璃划伤面部软组织，查见右耳前区皮肤长4~5 cm纵行创口，创缘整齐，有活跃出血。 1）本病例应进一步检查的项目不包括（） A.面神经分支损伤；B.腮腺导管损伤；C.颌面骨骨折；D.腮腺腺体损伤；E.面前静脉及颌外静脉损伤 2）确切的诊断应是面部软组织的（） A.挫伤；B.刺伤；C.挫裂伤；D.切割伤；E.撕裂伤 3）正确的处理方法是（） A.清洁创面、止血、无菌敷料覆盖；B.尽早行清创缝合术；C.局部冷敷，加压包扎；D.暴露创口，全身应用抗生素；E.创面以油纱布覆盖
B1	20	评估考生对多个知识点之间关联性的理解。题目中提供备选答案，要求考生从每道试题备选答案中选择1个，每个备选答案可多次选用或不选用	A.现病史；B.既往史；C.个人史；D.月经及婚育史；E.家族史 1）预防接种史及具体日期记入（） 2）主诉疾病发生的诱因和发病情况记入（） 3）生活起居、生活条件和生活水平记入（） 4）全身健康情况，如有无高血压、肺结核等情况记入（） 5）药物不良反应及过敏史应记入（） 6）主诉疾病治疗史、方式及效果记入（）

本研究的工作流程如图1所示。

LLM对话任务的参数中，常用于控制输出的有温度、Top-p、Top-k和频率惩罚。温度通过调整概率分布的平滑度，控制输出的随机性；Top-p和Top-k均通过控制候选词生成范围以调节输出的多样性，其中Top-p能根据概率分布动态调整候选集规模，相比Top-k采样避免了固定截断带来的不连贯问题[22]，因此本文选择Top-p作为主要采样参数；频率惩罚用于防止词汇反复高频出现。在各场景的测试中，LLM的参数设置如表2所示，温度和Top-p均设为0.5，在保证对话输出质量的同时，也能兼顾输出的稳定性[23]，为了防止频率惩罚对结果产生的影响，频率惩罚均设置为0，不对词频做额外干预。

本研究在下述3个场景中输入每道试题对每个模型进行测试。

1.1.1 选择正确答案

给出题目和选项，让LLM从所有选项中选择1个正确答案。测试LLM基于本身知识，输出试题正确答案的能力。

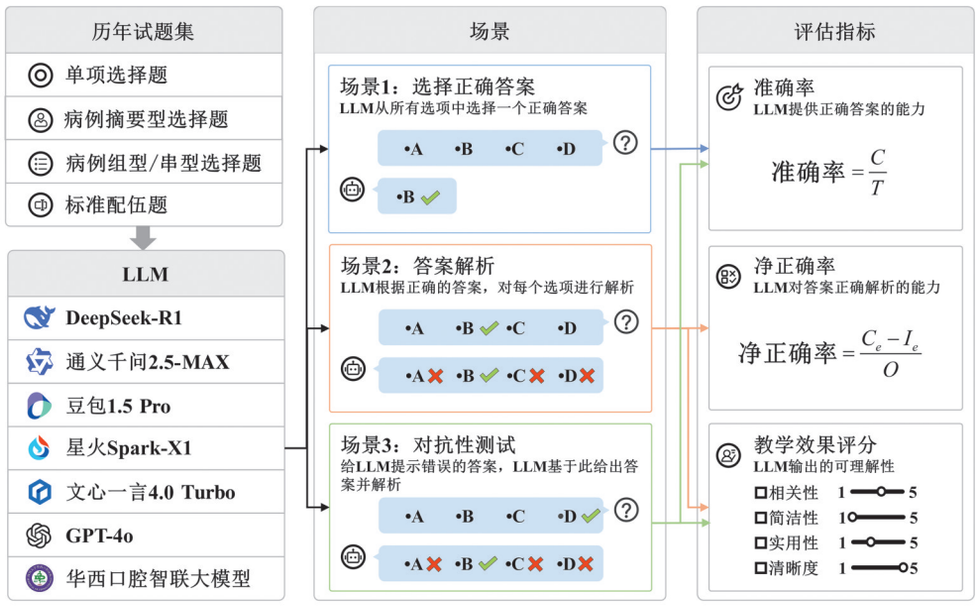
该测试场景旨在模拟考生在口腔执业医师考前学习中直接向LLM提问，希望LLM给出正确答案的情境。通过评估LLM在无外部提示下的答题准确性，可以反映其口腔医学知识储备水平及作为学习辅助工具的可靠性。

示例提示词：已知题目{输入1}，你需要根据题目要求选择正确的选项。你只需要给出你认为正确的选项即可。

1.1.2 答案解析

给出题目、选项和正确答案，让LLM对每个选项进行解析，说明正确答案的选择理由。测试LLM对题目进行正确解析的能力，以及输出解析的可理解性和教学效果。

该测试场景旨在模拟考生在考前学习时，详细对每个选项进行解析，以分析题目和考点的情境。通过对LLM解析选项的正确率和解析的教学效果进行评分，可评估其提供正确的题目解释、考点解析的能力和教学价值。



评估指标中，C表示正确回答的试题数，T表示总试题数，Ce表示模型正确解析的选项数量，Ie表示模型错误解析的选项数量，O表示选项的总数。

图 1　不同LLM在口腔执业医师考前学习中的应用比较研究工作流程图

Fig 1　Workflow of the comparative study on the application of different LLMs in pre-examination learning for the Chinese dental licensing examination

表 2　各LLM的参数配置

Tab 2　Parameter configurations of the evaluated LLMs

参数	DeepSeek-R1	通义千问2.5-MAX	豆包1.5 Pro	星火Spark-X1	文心一言4.0 Turbo	GPT-4o	华西口腔智联大模型
温度	0.5	0.5	0.5	0.5	0.5	0.5	0.5
Top-p	0.5	0.5	0.5	0.5	0.5	0.5	0.5
频率惩罚	0.0	0.0	0.0	0.0	0.0	0.0	0.0

示例提示词：已知题目{输入1}，其标准答案是{输入2}，你需要根据问题和正确的标准答案，解析每个选项，并解释选项正确或错误的理由。

1.1.3　对抗性测试

给出题目、选项和错误答案的提示，要求LLM在此干扰信息下重新作答并给出解析，以测试其在错误答案干扰下的纠错能力。

该测试场景旨在模拟考生在考前学习中向LLM询问时，可能因题目来源不准确或操作失误而提供错误答案的情况。通过观察LLM是否能够识别并纠正错误答案，并最终输出正确答案及合理解析，可评估不同模型在考前学习场景下的鲁棒性与安全性[24]。

示例提示词：已知题目{输入1}，其标准答案是{输入2}，你需要根据问题和正确的标准答案，解析每个选项，并解释选项正确或错误的理由。

所有的测试通过Python（版本3.10）调用各平台接口批量完成。测试结果通过1.2中的评估方法进行评估。

1.2　评估方法

本研究的评估指标为准确率、净正确率和教学效果评分。

1.2.1　准确率

准确率为正确回答的数量与总试题数量的比值，用于评估LLM准确响应问题，提供正确答案的能力，公式如下：准确率 $= \frac{C}{T}$，公式中C表示正确回答的试题数，T表示总试题数。

1.2.2　净正确率

净正确率为正确解析选项数与错误解析选项数的差值除以选项的总数，用于评估LLM对答案进行正确解析的能力。公式如下：净正确率 $= \frac{C_e - I_e}{O}$，公式中C_e表示模型正确解析的选项数量，I_e表示模型错误解析的选项数量，O表示选项的总数。

由上述2名口腔医学专家采用盲法独立判断选项的解析是否正确，若两者意见不一致，则由具有20年教学经验的资深口腔医学专家进行评判。

1.2.3　教学效果评分

教学效果评分基于5点李克特量表（1分表示强烈不同意，5分表示强烈同意），该量表包括4个不同的项目：相关性、简洁性、实用性和清晰度。由上述2名口腔医学专家采用盲法进行评估，最终得分为2名专家评分的平均值。如果2人评分差异过大，则由具备20年教学经验的资深口腔医学专家进行评判。

教学效果评分体现了专家从教学视角对LLM输出在教学价值方面的综合判断，可有效评估LLM的回答能否有效辅助考生理解与掌握知识。

1.2.4　统计分析

所有分析均使用SPSS（版本26.0）和Python（版本3.10）进行。所有场景均重复进行了5次独立测试，结果为5次测试的平均值。准确率和净正确率以百分比表示，记录标准差和95%置信区间（confidence interval，CI），教学效果评分以四分位数间距（interquartile range，IQR）和中位数表示。使用Cohen κ系数评估评分者之间的一致性，分类标准如下：0.01~0.20表示无或轻微一致性；0.21~0.40表示一般一致性；0.41~0.60表示中等一致性；0.61~0.80表示高一致性；0.81~1.00表示几乎完全一致。不同LLM以及各亚组的准确率、净正确率进行Wald χ^2检验，各组的中位数进行Kruskal-Wallis检验，检验结果是否存在统计学差异。进行多重比较时，使用了Bonferroni校正。双尾检验P<0.05被认为具有统计学差异。

2　结果

2.1　LLM答题的准确率

按照学科和题型分类统计的LLM答题的准确率结果如图2所示，图中每个学科或题型对应准确率坐标轴，将各模型的准确率用直线连接以对比其表现，红色虚线代表口腔执业医师考试的及格线。总体来看，所有LLM的总准确率为63%~84%，都达到了及格线，经Wald χ^2检验，各组间准确率的差异有统计学意义（χ^2=37.55，P<0.001）。华西口腔智联大模型的准确率最高（84%），文心一言4.0 Turbo的准确率最低（63%）。

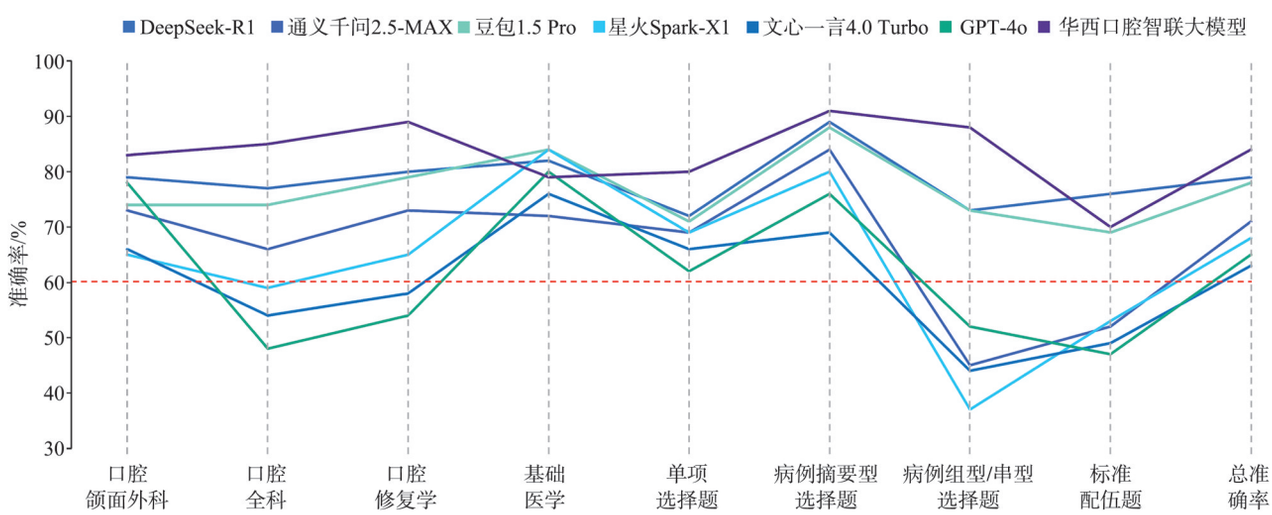


图 2　LLM回答不同学科和不同题型试题的准确率比较平行坐标图

Fig 2　Parallel coordinate plot of accuracy comparison among LLMs across different subjects and question types

LLM回答不同学科试题的准确率见表3、图3。华西口腔智联大模型在口腔颌面外科（83%）、口腔全科（85%）、口腔修复学（89%）中准确率最高，豆包1.5 Pro和星火Spark-X1在基础医学中准确率最高（84%）。

LLM回答不同题型试题的准确率见表4、图3。华西口腔智联大模型在单项选择题（80%）、病例摘要型选择题（91%）、病例组型/串型选择题（88%）中准确率最高；DeepSeek-R1在标准配伍题中准确率最高（76%）。

2.2　LLM解析试题的净正确率

LLM解析不同学科和题型试题的净正确率如图4所示。净正确率分布在73%~92%之间，Wald χ^2检验表明，各组之间的差异具有统计学意义（χ^2=39.71，P<0.001）。总体来看，华西口腔智联大模型的净正确率最高（92%），文心一言4.0 Turbo的净正确率最低（73%）。对于不同的学科，华西口腔智联大模型在口腔颌面外科（96%）、口腔全科（92%）、口腔修复学（91%）中准确率最高；星火Spark-X1和DeepSeek-R1在基础医学中净正确率最

高（96%）。对于不同的题型，华西口腔智联大模型在单项选择题（93%）、病例组型/串型选择题（88%）、标准配伍题（81%）中取得了最高的准确率；DeepSeek-R1在病例摘要型选择题中取得了最高的准确率（97%）。

表 3　LLM回答不同学科试题的准确率比较

Tab 3　Comparison of accuracy among LLMs in answering questions across different subjects

学科	DeepSeek-R1	通义千问2.5-MAX	豆包1.5 Pro	星火Spark-X1	文心一言4.0 Turbo	GPT-4o	华西口腔智联大模型	*P*值
口腔颌面外科	0.79±0.04	0.73±0.01	0.74±0.05	0.65±0.03	0.66±0.05	0.78±0.05	0.83±0.03	0.216
口腔全科	0.77±0.02	0.66±0.04	0.74±0.04	0.59±0.05	0.54±0.04	0.48±0.05	0.85±0.03	0.001
口腔修复学	0.80±0.04	0.73±0.03	0.79±0.03	0.65±0.04	0.58±0.04	0.54±0.03	0.89±0.02	<0.001
基础医学	0.82±0.04	0.72±0.05	0.84±0.04	0.84±0.04	0.76±0.04	0.80±0.04	0.79±0.03	0.735
合计	0.79±0.02	0.71±0.01	0.78±0.01	0.68±0.01	0.63±0.03	0.65±0.02	0.84±0.02	<0.001
（95%CI）	（0.77~0.82）	（0.70~0.72）	（0.77~0.79）	（0.67~0.70）	（0.60~0.67）	（0.62~0.67）	（0.82~0.86）	

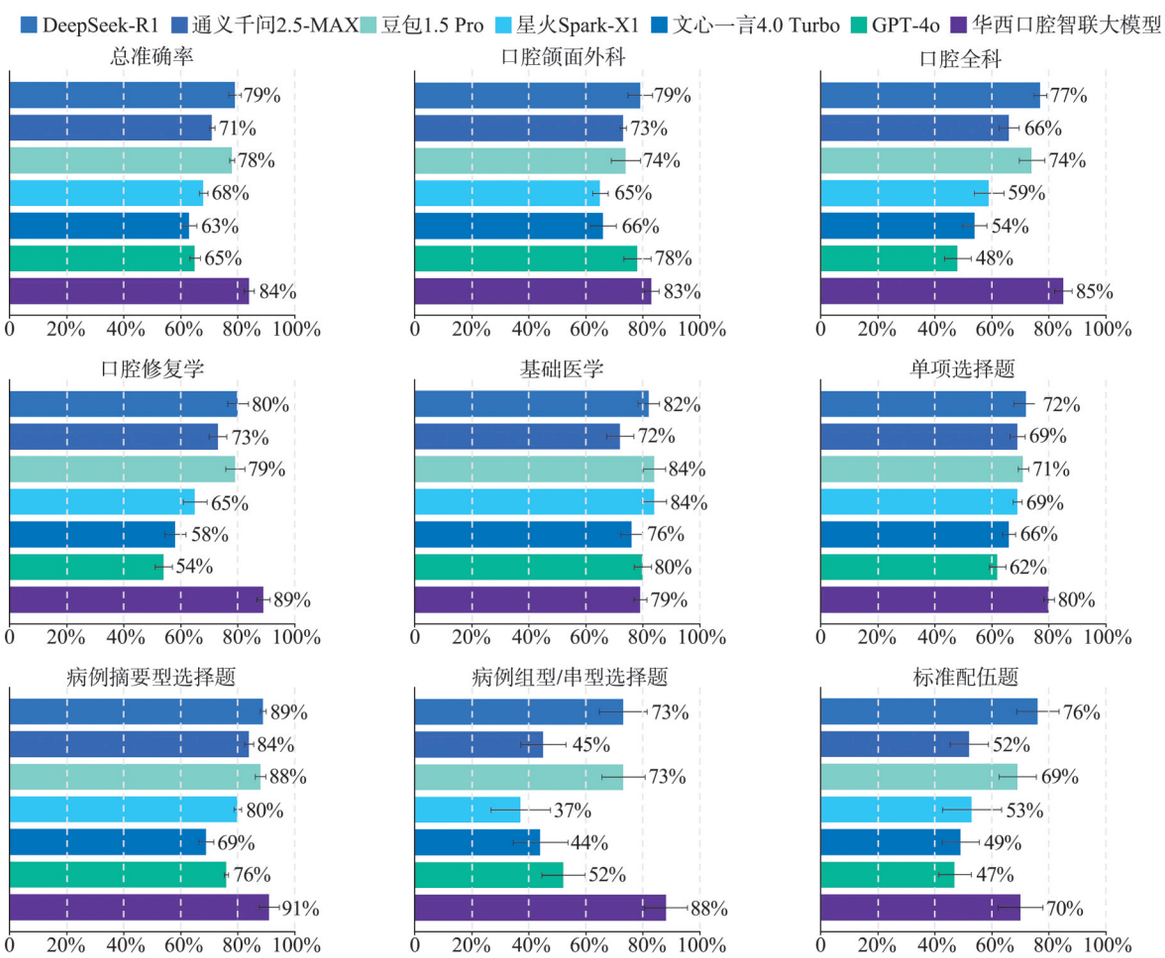


图 3　LLM回答不同学科和不同题型试题的准确率比较条形图

Fig 3　Bar chart of accuracy comparison among LLMs across different subjects and question types

2.3　LLM解析试题的教学效果评分

LLM解析试题的教学效果在相关性、简洁性、实用性和清晰度方面的评分结果如图5所示。Cohen κ系数结果表示评分者的评分之间具有高一致性（κ=0.68）。Kruskal-Wallis检验显示，各组之间的得分差异具有统计学意义（H=63.26，P<0.001）。相关性方面，华西口腔智联大模型取得了最高的分数（中位数3.8，IQR 3.1~4.3）。简洁性方面，GPT-4o得分最高（中位数3.6，IQR 3.1~4.2）。实用性方面，华西口腔智联大模型获得最高得分（中位数3.8，IQR 2.1~4.4），DeepSeek-R1与之表现接近，得分整体更为稳定（中位数3.1，IQR 2.3~4.2）。清晰度方面，华西口腔智联大模型取得了最高得分（中位数3.4，IQR 2.5~4.6）。

表 4　LLM 回答不同题型试题的准确率比较

Tab 4　Comparison of accuracy among LLMs in answering different types of questions

题型	DeepSeek-R1	通义千问 2.5-MAX	豆包 1.5 Pro	星火 Spark-X1	文心一言 4.0 Turbo	GPT-4o	华西口腔智联大模型	*P* 值
单项选择题	0.72±0.04	0.69±0.03	0.71±0.02	0.69±0.02	0.66±0.02	0.62±0.03	0.80±0.02	0.330
病例摘要型选择题	0.89±0.01	0.84±0.02	0.88±0.02	0.80±0.01	0.69±0.03	0.76±0.01	0.91±0.04	0.002
病例组型/串型选择题	0.73±0.08	0.45±0.08	0.73±0.08	0.37±0.09	0.44±0.09	0.52±0.08	0.88±0.08	0.002
标准配伍题	0.76±0.07	0.52±0.07	0.69±0.07	0.53±0.09	0.49±0.06	0.47±0.06	0.70±0.08	0.300
合计	0.79±0.02	0.71±0.01	0.78±0.01	0.68±0.01	0.63±0.03	0.65±0.02	0.84±0.02	<0.001
（95%CI）	（0.77~0.82）	（0.70~0.72）	（0.77~0.79）	（0.67~0.70）	（0.60~0.67）	（0.62~0.67）	（0.82~0.86）	

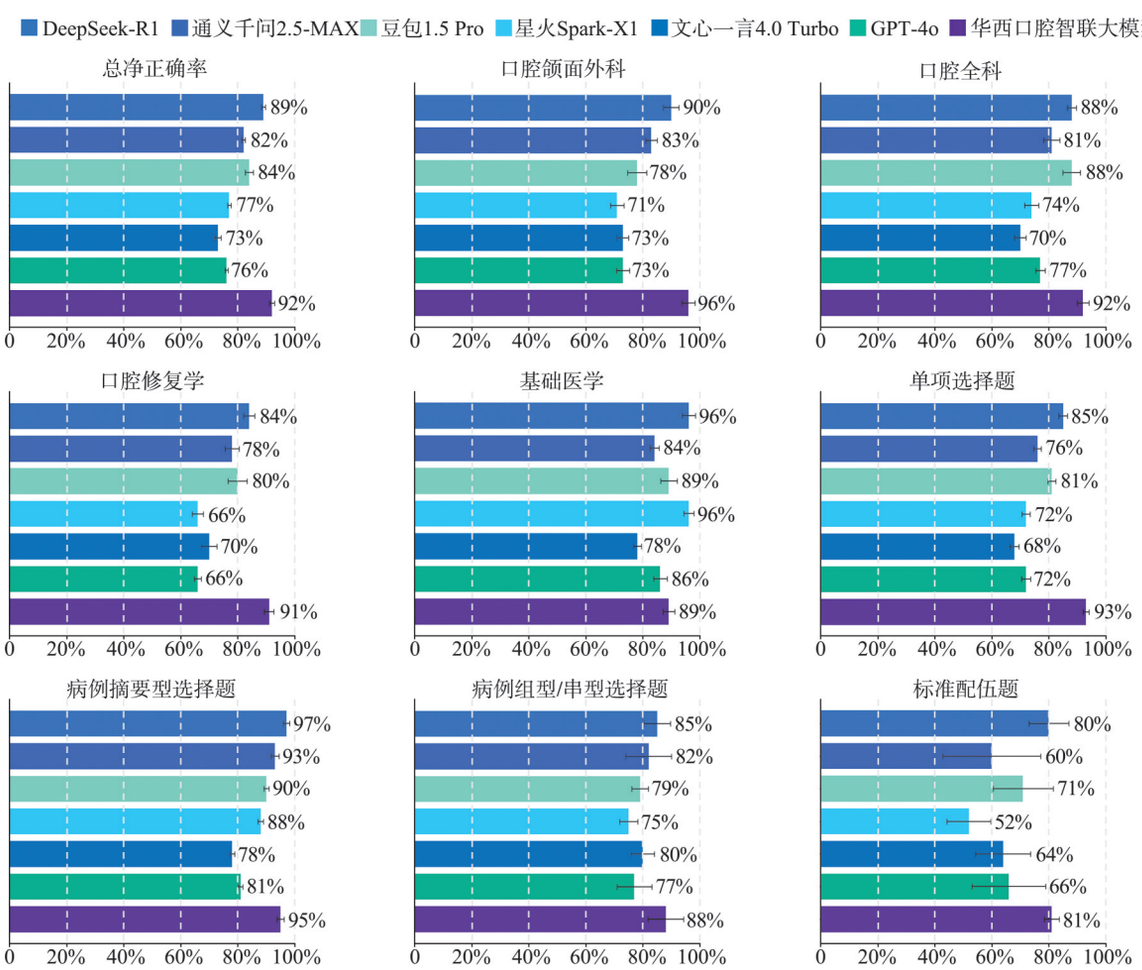


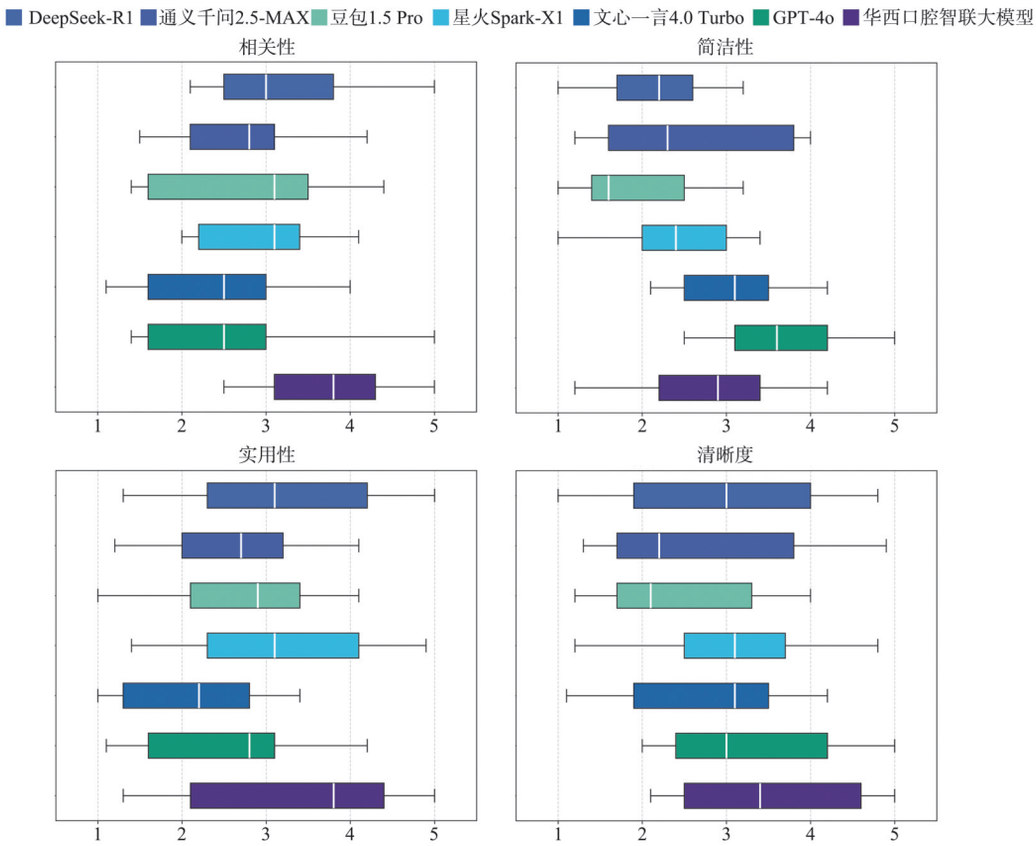
图 4　LLM 解析不同学科和不同题型试题的净正确率比较条形图

Fig 4　Bar chart of net accuracy comparison among LLMs in explaining different subjects and question types

2.4　对抗性测试中准确性和净正确率的变化

当受到错误答案提示的影响，LLM 的输出会发生显著变化，各 LLM 的对抗性测试结果详情如图 6 所示。该图表示了在对抗性测试前后，LLM 在回答和解析试题场景中，准确率和净正确率的变化，其中箭头上方的数值是由于对抗性测试而导致的性能下降幅度，右边的 *P* 值表示 Wald χ^2 检验的结果，用于评估观察到的下降的统计学意义，双尾检验 $P<0.05$ 被认为差异具有统计学意义。

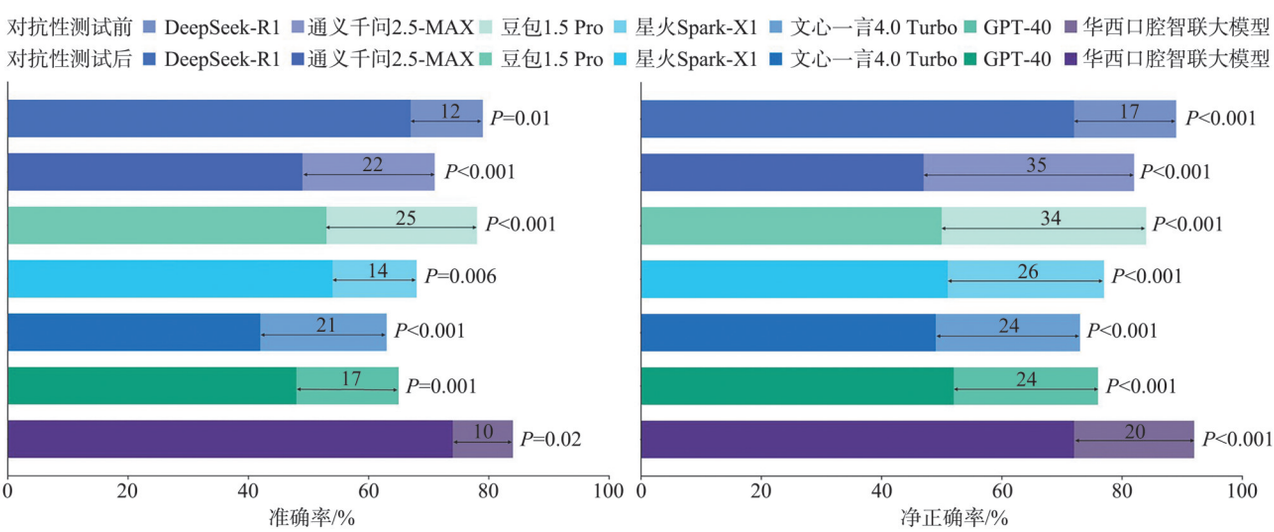
在对抗性测试场景中，各 LLM 与选择正确答案场景相比，准确率下降了 10%~25%；其中豆包 1.5 Pro 下降幅度最大（25%，$P<0.001$），华西口腔智联大模型下降幅度最小（10%，$P=0.02$），差异具有统计学意义。与答案解析场景相比，对抗性测试中净正确率变化范围为 17%~35%，其中通义千问 2.5-Max 下降幅度最大（35%，$P<0.001$），Deep-Seek-R1 的净正确率下降最少（17%，$P<0.01$）。



箱体代表不同LLM在此方面评分的IQR，箱体中的白线代表中位数，两侧的须线代表非异常的最大值和最小值（IQR的1.5倍范围内）。

图 5　LLM解析试题的教学效果评分比较箱形图

Fig 5　Box plot of pedagogical effectiveness scores for LLM-generated explanations



左：准确率变化；右：净正确率变化；箭头上方数值为性能下降幅度（%）。

图 6　LLM在对抗性测试中性能变化条形图

Fig 6　Bar chart of LLM performance variations in adversarial testing

3　讨论

本研究比较了代表性的LLM在口腔执业医师考试的各场景中答题和解析的表现。所选的LLM都成功达到了口腔执业医师考试的及格线（60%），其中华西口腔智联大模型取得了最高的准确率（84%）和净正确率（92%）。所有的LLM都表现出了一定的教学有效性，其中华西口腔智联大模型在相关性、实用性和清晰度方面得分最高，

DeepSeek-R1在实用性方面表现与之接近，且得分更稳定，GPT-4o则在简洁性方面分数最高。在对抗性测试中，所有LLM的性能都有所下降，华西口腔智联大模型的准确率下降幅度最小，DeepSeek-R1的净正确率下降幅度最小。

在其他医学考试中，LLM的准确率表现参差不齐，如西班牙医学住院医师考试（86.81%）[25]、印度国家资格暨入学考试（42.9%）[26]、全国临床医学硕士研究生招生考试（51%）[27]，但都有接近或超过人类平均表现水平的趋势，这与本研究中的表现一致。此外，本研究显示针对口腔医学进行知识增强的LLM（如华西口腔智联大模型）表现优于中文语料库为核心预训练的LLM（如DeepSeek-R1），而多数以中文语料库为核心预训练的LLM准确率都超过了以英文语料库为核心预训练的GPT-4o，表明GPT-4o虽然是一个功能强大的模型，但在需要对中文及其文化背景进行深入理解的任务中，以中文语料库为核心预训练的LLM表现更加出色[28]，因此，LLM的选择应以具体应用场景和语言需求为导向。

在选择正确答案的场景中，LLM在不同学科的准确率结果差异不明显，说明试题的难度在不同学科间的分布较为平均。而在不同题型分类比较时，可以明显看出LLM在不同题型之间的表现有较大差异，与基础知识相关的单项选择题和病例摘要型选择题难度最小，在回答病例组型/串型选择题和标准配伍题时，LLM的准确度显著下降。因此有必要将这两类题型的试题输入LLM进行针对性学习，以提高对该类题型的解答能力[29]。

在答案解析的场景中，LLM根据正确答案，对每个选项进行解析，解析的净正确率是对试题中所有选项的解析中，正确解析所占的比例。净正确率为73%~92%，其中华西口腔智联大模型的净正确率最高，可能是由于其在预训练模型的基础上，用口腔医学领域的数据进行了微调[20-21]，增强了口腔医学知识方面的答题和解析能力。

在对抗性测试的场景中，华西口腔智联大模型在对抗性测试中准确率下降最小，表明其有足够的口腔领域知识以不被错误的提示所干扰。DeepSeek-R1的净正确率下降幅度最小，可能由于其采用了强化学习训练方法，能够生成长链思考[30]，对错误或干扰性输入具有较强的鲁棒性。

目前已有许多评价方法用于评估LLM的输出质量，其中大多数是基于李克特量表构建的，包括中文综合医学基准[24]、LLM评分框架[31]。教学效果评分也是基于李克特量表的指标，旨在评估LLM输出结果的可理解性和教学价值。评估的结果显示在相关性、实用性和清晰度方面，大多数LLM的中位数不小于2.5分，表明当前主流LLM的输出通常能够达到一定的教学效果[32]，然而在简洁性方面，以DeepSeek-R1为代表的推理模型需要进一步优化以输出更加简洁的答案解析。

LLM的幻觉是指进行无效输出，或编造偏离原始材料的内容[33-34]。本研究中观察到3种类型的幻觉。第1种是在选择正确答案的场景中，LLM输出了错误的答案。第2种是在答案解析的场景中，LLM的解析不正确，给出了错误的知识点。第3种是在对抗性测试的场景中，当受到错误选项的干扰时，LLM输出了错误的答案并给出了错误的解析。另外本研究还观察到，由于LLM本身具有的不确定性，其幻觉可能在任意情况出现，甚至在相同条件下的不同对话中都可能随机出现。遗憾的是，由于缺乏持续的监督学习，LLM的幻觉问题在当前发展阶段仍然难以避免[11]。

在某些特定的NLP任务中，传统的NLP模型[如双向编码器表示模型（bidirectional encoder representations from transformers，BERT）]仍具有明显的优势，例如情感分析[35]、命名实体识别[36]和文本分类[37]等。然而，这些NLP模型在回答问题和多轮对话的任务表现出明显的局限性，尤其是需要进行高级推理和深刻理解专业领域知识的场景[38]。随着LLM领域的快速进步，能够整合各类包括医学在内的专业知识，综合上下文进行逻辑推理，快速精确地完成上述任务[39]，这使得LLM在回答和解析问题的任务中优于传统的NLP模型，因此本文选取代表性的生成式LLM进行比较研究。

本研究仍然存在一些局限性。第一，由于时间和资源的限制，只选择了当前具有代表性的LLM进行对比，无法评估所有LLM在口腔执业医师考前学习中的表现。第二，目前还没有标准的评价方法评估LLM对开放式问题的输出质量（如试题解析的教学效果），本研究采用李克特量表进行评估，最大程度地减少主观评价带来的潜在偏倚。此外，LLM产生幻觉的根本原因仍不明确，因此有必要进一步研究相关模型的可解释性。

综上所述，本文通过选择正确答案、答案解析、对抗性测试3个测试场景，以准确率、净正确率和教学效果评分为指标，评估了不同LLM在口腔医学执业医师考前学习应用中的表现，结果显示：针对口腔医学进行知识增强的LLM（如华西

口腔智联大模型）表现优于以中文语料库为核心预训练的推理LLM（如DeepSeek-R1），优于以英文语料库为核心预训练的LLM（如GPT-4o）；在对抗性测试中，所有模型的表现均有所下降。未来的研究可以基于评估结果，对各模型进行针对性优化，比如提高对错误答案干扰的鲁棒性、回答的简洁性等，以充分发挥LLM在口腔执业医师考试和医学教育中应用的潜力。

利益冲突声明：作者声明本文无利益冲突。

[参考文献]

[1] Wu TY, He SZ, Liu JP, et al. A brief overview of ChatGPT: the history, status quo and potential future development[J]. IEEE/CAA J Autom Sin, 2023, 10(5): 1122-1136.

[2] Li JN, Dada AM, Puladi B, et al. ChatGPT in healthcare: a taxonomy and systematic review[J]. Comput Methods Programs Biomed, 2024, 245: 108013.

[3] Thirunavukarasu AJ, Ting DSJ, Elangovan K, et al. Large language models in medicine[J]. Nat Med, 2023, 29(8): 1930-1940.

[4] 全国人民代表大会常务委员会. 中华人民共和国执业医师法[M]. 北京: 中国法制出版社, 1998.
Standing Committee of the National People's Congress of the People's Republic of China. Law of the People's Republic of China on medical practitioners[M]. Beijing: China Legal Publishing House, 1998.

[5] 国家医学考试中心. 国家医学考试网[EB/OL]. (2024-12-02) [2025-03-27]. https://www.nmec.org.cn/.
China National Medical Examination Center. China National medical examination network[EB/OL]. (2024-12-02)[2025-03-27]. https://www.nmec.org.cn/.

[6] 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]. JMIR Med Educ, 2023, 9: e45312.

[7] 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.

[8] 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.

[9] Wang XY, Gong ZY, Wang GX, et al. ChatGPT performs on the Chinese national medical licensing examination[J]. J Med Syst, 2023, 47(1): 86.

[10] Chen YK, Huang XJ, Yang FJ, et al. Performance of ChatGPT and Bard on the medical licensing examinations varies across different cultures: a comparison study [J]. BMC Med Educ, 2024, 24(1): 1372.

[11] Farquhar S, Kossen J, Kuhn L, et al. Detecting hallucinations in large language models using semantic entropy [J]. Nature, 2024, 630(8017): 625-630.

[12] 韩彩玲, 白石柱, 张婷民, 等. 口腔辅助诊疗和健康咨询领域5种大语言模型应用初探[J]. 中华口腔医学杂志, 2025, 60(8): 871-878.
Han CL, Bai SZ, Zhang TM, et al. Preliminary application of 5 large language models in the field of dental auxiliary diagnosis and health consultation[J]. Chin J Stomatol, 2025, 60(8): 871-878.

[13] 北京智源人工智能研究院. FlagEval[EB/OL]. (2025-04-06) [2025-04-06]. https://flageval.baai.ac.cn/.
Beijing Academy of Artificial Intelligence. FlagEval[EB/OL]. (2025-04-06) [2025-04-06]. https://flageval.baai.ac.cn/.

[14] 杭州深度求索人工智能基础技术研究有限公司. DeepSeek–深度求索[EB/OL]. (2025-03-24) [2025-04-12]. https://www.deepseek.com/.
Hangzhou DeepSeek Artificial Intelligence Co., Ltd. DeepSeek[EB/OL]. (2025-03-24) [2025-04-12]. https://www.deepseek.com/.

[15] 阿里云计算有限公司. 通义–你的超级个人助理[EB/OL]. (2025-01-29) [2025-04-12]. https://www. tongyi.com/.
Alibaba Cloud Computing Co., Ltd. Tongyi[EB/OL]. (2025-01-29) [2025-04-12]. https://www.tongyi.com/.

[16] 北京抖音信息服务有限公司. 豆包–字节跳动旗下AI助手[EB/OL]. (2025-03-17) [2025-04-06]. https://www.doubao.com/chat/.
Beijing Douyin Information Service Co., Ltd. Doubao [EB/OL]. (2025-03-17) [2025-04-06]. https://www.doubao.com/chat/.

[17] 科大讯飞股份有限公司. 讯飞星火–懂我的AI助手[EB/OL]. (2025-04-20) [2025-04-22]. https://xinghuo.xfyun.cn/.
iFLYTEK Co., Ltd. Spark Desk[EB/OL]. (2025-04-20) [2025-04-22]. https://xinghuo.xfyun.cn/.

[18] 百度在线网络技术(北京)有限公司 . 文心一言[EB/OL]. (2024-11-05) [2025-04-07]. https://yiyan.baidu.com. Baidu Online Network Technology (Beijing) Co., Ltd. ERNIE Bot[EB/OL]. (2024-11-05) [2025-04-07]. https://yiyan.baidu.com/.

[19] OpenAI. Chatgpt[EB/OL]. (2025-03-27) [2025-04-12]. https://chatgpt.com.

[20] 刘济远, 宋孝天, 孙亚楠, 等 . 基于小样本学习的执医考试辅导大模型精调方法及系统: CN119151737B[P]. 2024-08-20.
Liu JY, Song XT, Sun YN, et al. Fine-tuning method and system for medical licensing examination tutoring large model based on few-shot learning: CN119151737B[P]. 2024-08-20.

[21] 孙亚楠, 宋孝天, 刘济远, 等 . 用于构建口腔医学大模型的问答对话数据生成系统及方法, CN118520082B[P]. 2024-05-07.
Sun YN, Song XT, Liu JY, et al. Q&A dialogue data generation system and method for constructing stomatology large model: CN118520082B[P]. 2024-05-07.

[22] Holtzman A, Buys J, Forbes M, et al. The curious case of neural text degeneration[EB/OL]. (2020-02-14)[2025-06-27]. https://arxiv.org/abs/1904.09751.

[23] Du WH, Yang YM, Welleck S. Optimizing temperature for language models with multi-sample inference[EB/OL]. (2025-06-16) [2025-06-27]. https://arxiv. org/abs/2502.05234.

[24] Yuan XY, He P, Zhu QL, et al. Adversarial examples: attacks and defenses for deep learning[J]. IEEE Trans Neural Netw Learn Syst, 2019, 30(9): 2805-2824.

[25] Guillen-Grima F, Guillen-Aguinaga S, Guillen-Aguinaga L, et al. Evaluating the efficacy of ChatGPT in navigating the Spanish medical residency entrance examination (MIR): promising horizons for AI in clinical medicine[J]. Clin Pract, 2023, 13(6): 1460-1487.

[26] Farhat F, Chaudhry BM, Nadeem M, et al. Evaluating large language models for the national premedical exam in India: comparative analysis of GPT-3.5, GPT-4, and bard[J]. JMIR Med Educ, 2024, 10: e51523.

[27] Yu P, Fang CC, Liu XL, et al. Performance of ChatGPT on the Chinese postgraduate examination for clinical medicine: survey study[J]. JMIR Med Educ, 2024, 10: e48514.

[28] Glm T, Zeng A, Xu B, et al. ChatGLM: a family of large language models from GLM-130B to GLM-4 all tools [EB/OL]. (2024-07-30) [2025-06-05]. https://arxiv. org/abs/2406.12793.

[29] Zhang MR, Press O, Merrill W, et al. How language model hallucinations can snowball[EB/OL]. (2023-05-22)[2025-06-05]. https://arxiv.org/abs/2305.13534.

[30] Wang XZ, Wei J, Schuurmans D, et al. Self-consistency improves chain of thought reasoning in language models [EB/OL]. (2023-05-07) [2025-06-17]. https://arxiv. org/abs/2203.11171.

[31] Tan TF, Elangovan K, Ong J, et al. A proposed S.C.O.R.E. evaluation framework for large language models: Safety, consensus, objectivity, reproducibility and explainability[EB/OL]. (2024-07-10) [2025-06-18]. https://arxiv.org/abs/2407.07666.

[32] Rafailov R, Sharma A, Mitchell E, et al. Direct preference optimization: your language model is secretly a reward model[J]. Adv Neur Informat Proc Sys, 2023, 36: 53728-53741.

[33] Nadeem M, Sohail SS, Cambria E, et al. Negation blindness in large language models: Unveiling the NO syndrome in image generation[EB/OL]. (2024-09-04)[2025-07-03]. https://arxiv.org/abs/2409.00105.

[34] Ji ZW, Lee N, Frieske R, et al. Survey of hallucination in natural language generation[EB/OL]. (2024-07-14) [2025-07-03]. https://arxiv.org/abs/2202.03629.

[35] Anas M, Saiyeda A, Sohail SS, et al. Can generative AI models extract deeper sentiments as compared to traditional deep learning algorithms[J]. IEEE Intell Syst, 2024, 39(2): 5-10.

[36] Chandra C, Ojima Y, Bendarkar MV, et al. Aviation-BERT-NER: named entity recognition for aviation safety reports[J]. Aerospace, 2024, 11(11): 890.

[37] Eang C, Lee S. Improving the accuracy and effectiveness of text classification based on the integration of the bert model and a recurrent neural network (RNN_Bert_Based)[J]. Appl Sci, 2024, 14(18): 8388.

[38] Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining[J]. Bioinformatics, 2020, 36(4): 1234-1240.

[39] Meskó B. Prompt engineering as an important emerging skill for medical professionals: tutorial[J]. J Med Internet Res, 2023, 25: e50638.

（本文编辑　吴爱华）