

5种中文大语言模型对门诊处方中药物相互作用的审核能力比较^Δ

王紫灏^{1*}, 李晋宝¹, 张 研², 闫 森^{1#} (1. 黄河三门峡医院药学部, 河南 三门峡 472000; 2. 黄河三门峡医院眼科, 河南 三门峡 472000)

中图分类号 R95 文献标志码 A 文章编号 1001-0408(2026)19-2577-05

DOI 10.6039/j.issn.1001-0408.2026.19.16



摘要 **目的** 为审方软件的优化及医疗机构合理用药辅助审方工具的选择提供参考。**方法** 收集某三甲医院门诊2025年1月至2026年3月美康合理用药监测系统中符合纳入标准的处方559张。根据建立的复合参照标准(CRS)进行判定,其中阳性[存在药物相互作用(DDI)]处方206张、阴性(不存在DDI)处方353张。分别采用DeepSeek、豆包、文心一言、Kimi和通义千问对559张处方的DDI进行判定,并对其输出的DDI文本进行二值化判定。以CRS判定的结果为基准,构建混淆矩阵并计算模型的性能评价指标;比较任意2种大语言模型间敏感性的差异,计算二者的相对敏感性(RS)及其95%置信区间(CI),并以森林图展示。**结果** 文心一言的敏感性最高(74.8%),漏报风险最低;但其特异性仅为40.8%,且准确性最低(53.3%),存在较高的误报风险。通义千问的准确性最高(65.3%),F1分数最优(0.587)。DeepSeek的敏感性与特异性均居中。豆包与Kimi的敏感性均较低(分别为38.8%、45.1%),漏报风险均较高。两两比较结果显示,文心一言的敏感性显著高于DeepSeek[$\chi^2=4.84, P=0.028$; RS=1.124, 95%CI (1.018, 1.230)];通义千问的敏感性显著高于豆包[$\chi^2=4.77, P=0.029$; RS=1.239, 95%CI (1.035, 1.443)], DeepSeek的敏感性显著高于豆包[$\chi^2=37.65, P<0.001$; RS=1.715, 95%CI (1.491, 1.939)];Kimi与豆包、通义千问与Kimi的敏感性差异均无统计学意义($P>0.05$)。**结论** 文心一言适用于对漏报容忍度较低的初步筛查场景,且需联合人工复核;通义千问适合用于门诊辅助审核;DeepSeek可作为无审方软件医疗机构的过渡方案;豆包与Kimi不宜独立用于处方中DDI审核。

关键词 大语言模型; 药物相互作用; 处方审核; 敏感性; 辅助审方工具; 门诊处方

Comparative audit capabilities of five Chinese large language models in detecting drug-drug interaction in outpatient prescriptions

WANG Zihao¹, LI Jinbao¹, ZHANG Yan², YAN Sen¹ (1. Dept. of Pharmacy, Sanmenxia Hospital of the Yellow River, Henan Sanmenxia, 472000, China; 2. Dept. of Ophthalmology, Sanmenxia Hospital of the Yellow River, Henan Sanmenxia, 472000, China)

ABSTRACT **OBJECTIVE** To provide a reference for the optimization of prescription-review software and the selection of auxiliary prescription-review tools for medical institutions. **METHODS** A total of 559 eligible prescriptions were collected from the Meikang Prescription Automatic Screening System of a tertiary grade-A hospital from January 2025 to March 2026. Judged according to the composite reference standard (CRS) established in this study, there were 206 positive prescriptions [with drug-drug interaction (DDI)] and 353 negative prescriptions (without DDI). DeepSeek, Doubao, Wenxin yiyan, Kimi and Tongyi qianwen were separately adopted to identify DDI in the 559 prescriptions, and binary judgment was performed on their output texts related to DDI. Taking the results judged by CRS as the gold standard, the confusion matrix was constructed and model performance indicators were calculated. The differences in sensitivity between any two large language models were compared, the relative sensitivity (RS) and its 95% confidence interval (CI) were calculated, and the results were visualized by a forest plot. **RESULTS** Wenxin yiyan achieved the highest sensitivity (74.8%) with the lowest risk of missed reports, whereas its specificity was only 40.8%. It had the lowest accuracy (53.3%) and presented a high risk of false positive reports. Tongyi qianwen obtained the highest accuracy (65.3%) and the optimal F1-score (0.587). DeepSeek showed moderate performance in both sensitivity and specificity. Doubao and Kimi exhibited relatively low sensitivity (38.8% and 45.1%, respectively), accompanied by high risks of missed reports. Pairwise comparison revealed that the sensitivity of Wenxin yiyan was significantly higher than that of DeepSeek [$\chi^2=4.84, P=0.028$; RS=1.124, 95%CI (1.018, 1.230)]. The sensitivity of Tongyi qianwen was significantly higher than that of

Doubao [$\chi^2=4.77, P=0.029$; RS=1.239, 95%CI (1.035, 1.443)]. DeepSeek had significantly higher sensitivity than Doubao [$\chi^2=37.65, P<0.001$; RS=1.715, 95%CI (1.491, 1.939)]. No statistically significant differences in sensitivity were observed between Kimi versus Doubao and between

^Δ 基金项目 2024年度三门峡市科技计划项目(No.2024L02053)

* 第一作者 主管药师, 硕士。研究方向: 新药临床药理学研究、数据挖掘。E-mail: 3188085160@qq.com

通信作者 副主任药师, 硕士。研究方向: 医院药学管理、临床合理用药决策支持。E-mail: 149841646@qq.com

Tongyi qianwen versus Kimi ($P>0.05$). **CONCLUSIONS** Wenxin yiyan is applicable for preliminary-screening scenarios with low tolerance for missed reports and requires combined manual review. Tongyi qianwen is suitable for auxiliary outpatient prescription review. DeepSeek can act as a transitional solution for medical institutions without dedicated prescription-review software. Doubao and Kimi are not recommended for independent application in prescription DDI audit.

KEYWORDS large language models; drug-drug interaction; prescription review; sensitivity; auxiliary prescription-review tools; outpatient prescription

药物不良事件是临床用药安全面临的重要问题,药物相互作用(drug-drug interaction, DDI)是药物不良事件的重要成因^[1]。据 WHO 估算,全球约 1/10 的患者在医疗卫生服务中受到伤害,其中 50% 可预防伤害与药物有关^[2]。国家药品不良反应监测中心 2025 年度报告显示,全国共收到药品不良反应/事件报告 272.8 万份^[3]。另据文献报道,住院患者临床相关 DDI 发生率为 42.6%^[4]。处方审核是识别 DDI 的关键环节。国家卫生健康委办公厅 2025 年发布的《药事管理专业医疗质量控制指标(2025 年版)》,将药品收费前处方审核纳入质控,并明确处方审核方式包括人工审核、合理用药软件初审结合人工复核^[5],为“人机协作”的审方模式提供了政策基础。

美康合理用药监测系统(Prescription Automatic Screening System, PASS)是国内医疗机构部署最多、应用最成熟的合理用药信息化工具。美康 PASS 通过内置结构化规则库,可高效识别明确的不良相互作用^[6],是大型医疗机构保障用药安全的核心工具。然而,该系统存在一定局限性,如知识库更新依赖人工维护,对罕见相互作用、新上市药物风险、中西药联用的配伍禁忌覆盖不足;同时,该系统的采购与运维成本较高,难以在基层医疗机构普及。

大语言模型基于海量文本语料训练,具备语义理解、知识整合与自然语言生成能力,已广泛应用于金融、医疗健康等多个领域^[7]。在医疗健康领域,大语言模型初步展现了在医学问答、临床决策支持等方面的应用潜力^[8]。近年来,国内多家机构相继推出中文大语言模型^[9],如 DeepSeek、豆包、文心一言、Kimi 和通义千问等。已有研究初步探索了大语言模型在药学领域的应用,发现其在 DDI 审核领域具有潜在应用价值,但仍存在敏感性与特异性之间的平衡需进一步优化,缺乏对国内主流中文大语言模型的横向对比(主要以 1 个或 2 个大语言模型为研究对象),未系统计算敏感性、特异性及 F1 分数等大语言模型判别效能指标,以及缺乏对不同大语言模型间效能差异的统计学检验分析等问题^[10-11]。

本研究以某三甲医院真实门诊处方为测试集,以专家复核的 DDI 结果为参照标准,采用双人盲法,对 5 种国内主流中文大语言模型(DeepSeek、豆包、文心一言、Kimi 和通义千问)进行二值化判定,系统计算其敏感性、特异性、准确性及 F1 分数,并通过配对 McNemar 检验分析大语言模型间敏感性的统计学差异,旨在为审方软件的优化及医疗机构合理用药辅助审方工具的选择提供实证依据。

1 资料与方法

1.1 数据收集与复合参照标准建立

收集河南省某三甲医院门诊 2025 年 1 月至 2026 年 3 月美康 PASS 前置审方模块的处方信息。纳入标准:(1)仅包含 2 种药物的处方;(2)处方信息完整,包括医嘱内容、单次剂量、给药单位、给药频次;(3)美康 PASS 已输出 DDI 判定结果及问题描述的处方。根据设置的纳入标准,最终纳入 559 张处方,其中美康 PASS 判定为存在 DDI 的阳性处方有 201 张、不存在 DDI 的阴性处方有 358 张。处方涵盖心血管系统药物、降糖药物、抗感染药物、中枢神经系统药物、中成药等常见类别药物,具有较好的代表性。原始数据包含患者身份信息及用药信息,由信息技术人员在医院内网服务环境下进行脱敏处理,删除患者门诊号、姓名、性别、年龄等个人身份信息及诊断等敏感字段,仅保留处方号、医嘱内容及完整的用药信息,导出处理后的数据用于后续分析。本研究仅对脱敏处理后的处方数据进行回顾性分析,不涉及患者身份识别和隐私泄露,符合免除伦理审查的条件。

由 2 名具有副主任药师及以上职称的临床药师,独立对 559 张处方进行 DDI 判定,判定依据为《中国国家处方集》及 DrugBank 数据库;若二者意见不一致,则由第 3 位药师(具有主任药师职称)判定。最终以药师的共识作为复合参照标准(composite reference standard, CRS)。根据 CRS,在 559 张处方中,存在 DDI 的阳性处方有 206 张、不存在 DDI 的阴性处方有 353 张。

1.2 大语言模型的访问获取及提示词设置

5 种大语言模型的相关信息见表 1。5 种大语言模型均通过官方网站网页端访问,确保所有大语言模型在同一时间窗口内运行。各大语言模型的版本相对稳定,测试期间无版本更新。为保证输入一致性,5 种大语言模型均采用标准化提示词,其核心指令为:“根据《中国国家处方集》识别处方中 2 种药物可能存在的 DDI”。不同大语言模型对输出格式的要求不同,其提示词在格式上存在细微差异;但所有大语言模型的核心任务一致,均为判断是否存在 DDI。

表 1 5 种大语言模型的基本信息

大语言模型名称	开发公司	测试版本号	版本发布时间
DeepSeek	深度求索(DeepSeek)	DeepSeek-V3.2	2025 年 12 月
豆包	字节跳动(ByteDance)	豆包-2.0	2026 年 2 月
文心一言	百度(Baidu)	文心-5.0	2026 年 1 月
Kimi	月之暗面(Moonshot AI)	Kimi-K2.5	2026 年 1 月
通义千问	阿里云(Alibaba Cloud)	Qwen-3.5	2026 年 2 月

1.3 大语言模型DDI输出文本的二值化判定

将美康PASS判定的DDI结果及5种大语言模型的DDI输出文本分别整理成结构化表格。其中,大语言模型输出文本为原始文本,未作任何修改或重写。2名主管药师独立录入文本并交叉核对。为将大语言模型输出的DDI自然语言文本转变为可与CRS判定结果对比的二分类结果,由2名副主任药师独立对大语言模型DDI输出文本进行判读,按照统一规则将文本描述转化为“有相互作用”或“无相互作用”的二值化结果^[12]:(1)阳性判定规则:大语言模型DDI输出文本明确提及“相互作用”“影响”“风险增加”“需监测或调整剂量”“不宜合用”“慎用”“禁忌”等关键词,或给出明确风险描述或剂量调整建议。(2)阴性判定规则:大语言模型DDI输出文本明确表示“无明显相互作用”“联用安全”“无禁忌”“无明确不良相互作用”等关键词,或仅陈述药物用途而未提及DDI。若2名副主任药师对大语言模型输出文本的判读结果不一致,则由第3位主任药师裁决。最终获得5种大语言模型对206张阳性处方和353张阴性处方的二值化判定结果。

1.4 评价指标

以CRS判定的结果为基准,为每种大语言模型构建2×2混淆矩阵,详见表2。在大语言模型性能评估中,敏感性(以召回率表示)反映大语言模型对存在DDI处方的识别能力,其值越高表明漏报率越低;特异性反映大语言模型对不存在DDI处方的正确排除能力,其值越高表示误报率越低;准确性反映大语言模型判定正确的比例,即真阳性与真阴性之和占总处方数的比例;F1分数为精确率(阳性预测值)与召回率的调和平均,用于综合反映大语言模型判定阳性处方的准确性与完整性^[13]。

表2 2×2混淆矩阵结构

大语言模型判定结果	CRS判定为阳性(存在DDI)	CRS判定为阴性(不存在DDI)
阳性(存在DDI)	真阳性	假阳性
阴性(不存在DDI)	假阴性	真阴性

1.5 统计学分析

1.5.1 敏感性差异分析

以206张CRS判定的阳性处方为研究对象,对任意2种大语言模型(分别记为模型A、模型B)进行配对比较,构建配对四格表:2种大语言模型均判定为阳性(存在DDI,下同)的处方数记为*a*,大语言模型A判定为阳性而大语言模型B判定为阴性(不存在DDI,下同)的处方数记为*b*,大语言模型A判定为阴性而大语言模型B判定为阳性的处方数记为*c*,2种大语言模型均判定为阴性的处方数为*d*。*b*与*c*为大语言模型判定不一致的配对数据,是检验2种大语言模型敏感性差异的核心依据^[14]。采用经连续性校正的配对McNemar检验比较2种大语言模型敏感性的差异。该检验适用于配对二分类数据的差异性比较,主要通过比较不一致的配对数据

b、*c*来判断2种大语言模型的敏感性是否具有统计学差异^[15]。检验水准 $\alpha=0.05$ 。

1.5.2 相对敏感性分析

采用相对敏感性(relative sensitivity, RS)作为效应量,计算不同大语言模型两两比较的RS及其95%置信区间(confidence interval, CI),并以森林图展示。由于RS的抽样分布呈偏态,直接构建其CI较为困难,故本研究采用Delta法近似估计RS对数尺度的标准误,在对数尺度上构建其95%CI,经指数变换后获得原始尺度下的RS及其95%CI^[16]。RS>1且95%CI不跨越1,表明大语言模型A的敏感性显著高于大语言模型B;RS<1且95%CI不跨越1,表明大语言模型A的敏感性显著低于大语言模型B;95%CI跨越1,表明2种大语言模型的敏感性差异无统计学意义^[17]。

2 结果

2.1 大语言模型的DDI判定结果

美康PASS判定的201张阳性处方中,5张经专家复核确认为假阳性;美康PASS判定的358张阴性处方中,10张经专家复核确认为假阴性。美康PASS的总体错判率为2.7%(15/559)。结果表明,以美康PASS作为参照标准存在一定的假阳性与假阴性偏倚。5种大语言模型判定的DDI结果见表3。其中,文心一言判定的阳性处方数最多(363张),提示其阳性判定阈值较低;其次为DeepSeek(310张),提示其阳性判定阈值亦较低,但比文心一言稍严格;通义千问和Kimi判定的阳性处方数分别为186、198张,提示二者阳性判定阈值适中;豆包判定的阳性处方数为169张,提示其可能存在漏报。

表3 5种大语言模型的DDI判定结果(张)

大语言模型	阳性处方数量	阴性处方数量
DeepSeek	310	249
豆包	169	390
文心一言	363	196
Kimi	198	361
通义千问	186	373

2.2 大语言模型的DDI识别性能评价结果

5种大语言模型DDI识别的性能评价指标见表4。由表4可见,文心一言的敏感性最高(74.8%),漏报风险最低(假阴性处方数最少);然而,其特异性仅为40.8%,提示在CRS判定的阴性处方中,约有59.2%的处方被该大语言模型误判为存在DDI,存在较高的假阳性率,这可能导致无临床意义的DDI被警示,增加人工复核负担。文心一言的准确性最低(53.3%),这可能是因为其敏感性较高,存在较高的误报风险(假阳性处方数量较多),反而导致其准确性降低。通义千问的准确性最高(65.3%),F1分数表现最优(0.587)。DeepSeek的敏感性与特异性均居中,整体表现均衡。豆包与Kimi的敏感性较低,漏报风险较高。5种大语言模型的敏感性和特异性存在一定差异。

表4 5种大语言模型DDI识别的性能评价指标比较

大语言模型	DDI判定结果/张				性能评价指标			
	真阳性	假阳性	假阴性	真阴性	敏感性/%	特异性/%	准确性/%	F1分数
DeepSeek	137	173	69	180	66.5	51.0	56.7	0.577
豆包	80	89	126	264	38.8	74.8	61.5	0.511
文心一言	154	209	52	144	74.8	40.8	53.3	0.528
Kimi	93	105	113	248	45.1	70.3	61.0	0.549
通义千问	99	87	107	266	48.1	75.4	65.3	0.587

2.3 大语言模型的敏感性分析结果

文心一言的敏感性显著高于DeepSeek($\chi^2=4.84$, $P=0.028$);二者的RS=1.124,95%CI(1.018,1.230),95%CI不跨越1。通义千问的敏感性显著高于豆包($\chi^2=4.77$, $P=0.029$);二者的RS=1.239,95%CI(1.035,1.443),95%CI不跨越1。DeepSeek的敏感性显著高于豆包($\chi^2=37.65$, $P<0.001$);二者的RS=1.715,95%CI(1.491,1.939),95%CI不跨越1。Kimi与豆包的敏感性差异无统计学意义($\chi^2=2.76$, $P=0.093$);二者的RS=1.163,95%CI(0.983,1.342),95%CI跨越1。通义千问与Kimi的敏感性差异亦无统计学意义($\chi^2=0.49$, $P=0.484$);二者的RS=1.065,95%CI(0.916,1.214),95%CI跨越1。结果见表5、图1。

表5 5种大语言模型的敏感性分析结果(张)

对比组	a	b	c	d	χ^2	P
文心一言 vs. DeepSeek	120	34	17	35	4.84	0.028
通义千问 vs. 豆包	55	44	25	82	4.77	0.029
DeepSeek vs. 豆包	68	69	12	57	37.65	<0.001
Kimi vs. 豆包	61	32	19	94	2.76	0.093
通义千问 vs. Kimi	72	27	21	86	0.49	0.484

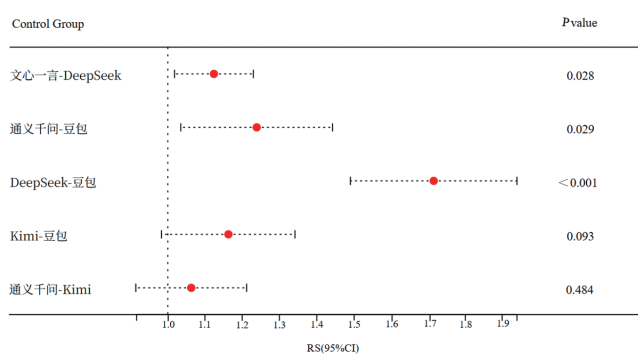


图1 5种大语言模型敏感性分析的森林图

3 讨论

3.1 5种大语言模型的性能差异

本研究通过构建CRS,对美康PASS作为参照标准的潜在偏倚进行了定量评估。结果显示,美康PASS的总体错判率约2.7%,提示以单一数据库作为参照标准存在系统性偏倚风险。在本研究建立的CRS基础上,5种大语言模型在DDI识别任务中的性能指标存在显著差异,各模型在敏感性与特异性之间的权衡取向不同。这一差异可能与不同大语言模型训练数据构成、架构设计、指令调优策略及安全对齐机制不同等相关^[18]。

在5种大语言模型中,文心一言的敏感性最高、特异

性最低,提示文心一言在判定决策中倾向于将模糊或信息不足的样本判为阳性,该判定倾向在需要高敏感性的临床场景中具有优势,但可能带来较高的假阳性率;通义千问的特异性最高、F1分数最优,但敏感性较低,提示通义千问判定阈值可能较保守,倾向于将模糊或信息不足的样本判定为阴性,可能增加漏报风险;DeepSeek的敏感性与特异性均居中,整体表现相对均衡,这可能与其训练数据覆盖广泛及调优策略有关;豆包与Kimi的敏感性均较低,提示其漏报风险较高。5种大语言模型均未同时兼顾高敏感性与高特异性,临床用其辅助审核处方中DDI时需权衡漏报与误报风险^[19]。

3.2 基于DDI审核需求的大语言模型选取建议

本研究在同一数据集上对5种中文主流大语言模型进行横向比较,并基于各大语言模型在DDI识别任务中的性能差异,结合审核场景提出以下选取建议:(1)以DDI漏报风险最小化为目标,推荐采用文心一言。该模型的敏感性达74.8%,在5种大语言模型中最高,且漏报风险最低,适用于对DDI漏报容忍度较低、可接受较高误报率的临床场景,如重症监护室、肿瘤化疗病房等^[20]。但在具体使用时,建议联合人工复核,如由药师对假阳性警示进行复核。(2)以减少DDI误报、提升审核效率为目标,推荐采用通义千问。该模型的特异性达75.4%,且误报率最低,可减轻药师对不存在DDI处方的无效复核负担,适合作为日常门诊处方中DDI的辅助审核工具。在实际应用中,药师可优先复核通义千问判定为阳性的处方,对判定为阴性的处方可减少人工复核。(3)DeepSeek可作为过渡选择。该模型的敏感性为66.5%、特异性为51.0%,整体表现均衡,适用于尚无专业审方系统、需兼顾DDI识别敏感性与特异性的基层医疗机构。豆包与Kimi的漏报率较高,不宜独立用于处方DDI审核。

3.3 研究局限性与展望

本研究存在如下不足:(1)处方样本量和药物组合类型有限,不足以代表各级医疗机构的处方特征,结论的外推性有待多中心、大样本验证;此外,本研究中敏感性差异检验未涵盖5种大语言模型的全部配对比较,结论的全面性有待后续研究验证。(2)本研究仅评价了不同大语言模型的二分类判别能力,未对其输出处理建议的质量(如具体监测指标、剂量调整建议)进行定量评分,而后者对临床实际应用价值的影响更直接。未来研究应扩大数据来源,并纳入住院医师中的DDI数据进行分析;构建结构化评分量表,系统评估大语言模型输出处理建议的临床实用性;探索检索增强生成等知识增强方法,在统一知识库条件下重新比较各大语言模型性能,以减少大语言模型训练数据不同而带来的比较偏倚^[21],但需要评估此方法在基层医疗机构中的可及性与部署成本^[22]。

4 结语

各大语言模型用于DDI识别的敏感性与特异性存在差异:文心一言的敏感性最高,特异性最低,适用于对DDI漏报容忍度较低、可接受较高误报率的初步筛选场景,但建议联合人工复核;通义千问特异性、准确性及F1分数均优于其他大语言模型,误报最少,适合作为日常处方DDI辅助审核工具;DeepSeek的敏感性与特异性相对均衡,可作为暂无专业审方系统医疗机构的过渡方案;豆包与Kimi的敏感性最低,漏报风险高,不宜独立用于处方DDI审核。本研究为医疗机构选择DDI辅助审核工具及专业审方软件优化提供了初步的实证依据,并为后续大语言模型在DDI领域的专项微调与测评框架构建提供了参考。但本研究仅基于单中心门诊处方数据展开,样本量和药物组合类型有限,上述选型建议仍需结合多中心、大样本数据及更全面的模型配对比较研究进一步验证。

利益冲突 本文所有作者声明不存在利益冲突

参考文献

[1] Aksoy N, Ozturk N. A meta-analysis assessing the prevalence of drug-drug interactions among hospitalized patients[J]. *Pharmacoepidemiol Drug Saf*, 2023, 32 (12): 1319-1330.

[2] World Health Organization. Patient safety: fact sheet[EB/OL]. (2023-09-11)[2026-09-23]. <https://www.who.int/news-room/fact-sheets/detail/patient-safety>.

[3] 国家药品不良反应监测中心. 国家药品不良反应监测年度报告: 2025 年[EB/OL]. (2026-04-27)[2026-09-23]. https://www.cdr-adr.org.cn/center_news/202604/t20260-427_51496.html.

[4] Li L, Baker J, Quirk R, et al. Drug-drug interactions and actual harm to hospitalized patients: a multicentre study examining the prevalence pre- and post-electronic medication system implementation[J]. *Drug Saf*, 2024, 47 (6): 557-569.

[5] 国家卫生健康委办公厅. 药事管理专业医疗质量控制指标: 2025 年版[EB/OL]. (2025-09-08)[2026-09-18]. https://www.gov.cn/zhengce/zhengceku/202509/content_7039760.htm.

[6] 夏铮铮, 李剑屏, 颜海兰, 等. 合理用药监测系统在肿瘤专科医院的应用实践[J]. *肿瘤药学*, 2022, 12 (5): 649-655.

[7] Ray P P. ChatGPT: a comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope[J]. *Internet Things Cyber Phys Syst*, 2023, 3: 121-154.

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

[9] 籍欣萌, 咎红英, 崔婷婷, 等. 中文医疗大模型综述: 进展、评估与挑战[J]. *中文信息学报*, 2024, 38(11): 1-12.

[10] 牛露露, 刘滔滔, 黄天敏, 等. 评价 DeepSeek 与 ChatGPT 作为咨询工具在真实世界中回答药物-药物相互作用和超说明书用药的能力[J]. *中国现代应用药学*, 2025, 42 (17): 2929-2935.

[11] 马红燕, 周晓, 张碧瑶, 等. DeepSeek 大语言模型辅助处方合理性评估的实用性[J]. *医药导报*, 2025, 44(12): 2057-2061.

[12] Mohammed M, Khan M A, Chaudhary V, et al. Comparative performance of large language models and Drugs.com versus Lexicomp for antiepileptic medication drug-drug interactions: a cross-sectional study with iterative prompting analysis[J]. *Explor Res Clin Soc Pharm*, 2026, 22: 100733.

[13] Domián B M, Ashraf A R, Fittler A T, et al. Comparative evaluation of artificial intelligence platforms and drug interaction screening databases using real-world patient data[J]. *Explor Res Clin Soc Pharm*, 2025, 20: 100655.

[14] Wu Y G. Weighted McNemar's test for the comparison of two screening tests in the presence of verification bias[J]. *Stat Med*, 2022, 41(16): 3149-3163.

[15] Hou B, Mukherjee P, Batheja V, et al. One year on: assessing progress of multimodal large language model performance on RSNA 2024 case of the day questions[J]. *Radiology*, 2025, 316(2): e250617.

[16] Oehlert G W. A note on the delta method[J]. *Am Stat*, 1992, 46(1): 27-29.

[17] Sarkar S, Baidya D K. Meta-analysis-interpretation of forest plots: a wood for the trees[J]. *Indian J Anaesth*, 2025, 69(1): 147-152.

[18] Bhati D, Neha F, Bandaru D S, et al. Mapping the LLM landscape: a cross-family survey of architectures, alignment methods, and benchmark performance[J]. *AI*, 2026, 7 (4): 142.

[19] Sicard J, Montastruc F, Achalme C, et al. Can large language models detect drug-drug interactions leading to adverse drug reactions? [J]. *Ther Adv Drug Saf*, 2025, 16: 20420986251339358.

[20] Bakker T, Klopotoska J E, Dongelmans D A, et al. The effect of computerised decision support alerts tailored to intensive care on the administration of high-risk drug combinations, and their monitoring: a cluster randomised stepped-wedge trial[J]. *Lancet*, 2024, 403 (10425): 439-449.

[21] Hu H, Wang L, Chen L Q, et al. Benchmarking large language models for drug combination alerts: achieving expert-level reliability via knowledge grounding and contextual reasoning[J]. *J Med Chem*, 2026, 69(3): 3477-3489.

[22] Xu K X, Xu Y R, Shi Q. Harnessing computational power for intelligent oncology in the age of large models: status, challenges, and prospects[J]. *Intell Oncol*, 2026, 2 (1): 100038.

(收稿日期: 2026-05-25 修回日期: 2026-09-26)

(本文责编: 林 静)