

生成式大语言模型在生殖与妇产领域药学实践的实用性评估



刘勇军, 覃燕妮, 谭小兰, 田 新, 苏小清, 张春花

广西壮族自治区生殖医院药学部 (南宁 530029)

【摘要】目的 评估生成式大语言模型 (LLMs) 在生殖与妇产领域药学实践中的实用性与可行性, 为药师合理应用人工智能工具提供参考。**方法** 采用分层抽样法抽取广西壮族自治区生殖医院 2024—2025 年门诊处方、《妊娠及哺乳期用药咨询标准化手册》中用药咨询问题, 并随机抽取广西壮族自治区生殖医院出院带药用药教育记录构建问题库。分别采用 DeepSeek、ChatGPT 和 Grok 等 LLMs 进行问答, 采用双人核对的方式对 LLMs 输出的答案进行系统评估和分析。**结果** 处方点评方面, ChatGPT 准确率最高 (83.3%), 其次为 Grok (76.7%) 和 DeepSeek (73.3%); 但 3 个模型精确率均偏低 (45.5% ~ 57.1%); DeepSeek 和 ChatGPT 召回率均为 100.0%, Grok 仅 45%; F1 值以 ChatGPT 最佳 (72.7%)。用药咨询方面, DeepSeek 表现最佳, 平均得分为 (2.75 ± 0.45)。用药教育方面, ChatGPT 表现较好, 平均得分为 (2.50 ± 0.53)。**结论** 生成式 LLMs 在生殖与妇产领域的药学实践中具有一定实用价值, 但当前模型尚不能替代药师的专业判断, 建议采用“人机协同”模式, 以保障生殖与妇产领域特殊人群的用药安全。

【关键词】 大语言模型; 药学实践; 生殖与妇产科学; 处方点评; 用药咨询; 用药教育

【中图分类号】 R969.3

【文献标识码】 A

Evaluation of the practical utility of generative large language models in pharmacy practice within the reproductive and obstetrics-gynecology field

LIU Yongjun, QIN Yanni, TAN Xiaolan, TIAN Xin, SU Xiaoqing, ZHANG Chunhua

Department of Pharmacy, Guangxi Zhuang Autonomous Region Reproductive Hospital, Nanning 530029, China

Corresponding author: ZHANG Chunhua, Email: 191557768@qq.com

【Abstract】Objective To evaluate the practicality and feasibility of generative large language models (LLMs) in pharmaceutical practice within the reproductive and obstetrics-gynecology field, and to provide a reference for pharmacists in the rational application of artificial intelligence tools. **Methods** Stratified sampling was adopted to extract outpatient prescriptions from the Guangxi Zhuang Autonomous Region Reproductive Hospital between 2024 and 2025, as well as medication consultation questions from the *Standardized Manual for Medication Consultation in Pregnancy and Lactation*. Additionally, discharge medication education records of the above-mentioned hospital were randomly sampled to construct a question bank. DeepSeek, ChatGPT and Grok were separately used to generate question-answering outputs. A dual-review

DOI: 10.12173/j.issn.2097-4922.202605049

通信作者: 张春花, 硕士, 副主任药师, Email: 191557768@qq.com

<https://yxqy.whuznhmedj.com>

process was performed to systematically evaluate and analyze the responses produced by large language models (LLMs). **Results** In prescription review, ChatGPT achieved the highest accuracy (83.3%), followed by Grok (76.7%) and DeepSeek (73.3%). All three models showed relatively low precision (45.5%-57.1%). DeepSeek and ChatGPT both achieved 100.0% recall, whereas Grok achieved only 45%. The F1 score was best for ChatGPT (72.7%). In medication consultation, DeepSeek performed best, with an average score of 2.75 ± 0.45 . In medication education, ChatGPT performed best, with an average score of 2.50 ± 0.53 . **Conclusion** Generative LLMs have demonstrated practical utility in pharmaceutical practice in the reproductive and obstetrics-gynecology domain. However, current models cannot yet replace the professional judgment of pharmacists. A "human-machine collaborative" model is recommended to ensure medication safety for special populations in this field.

【Keywords】 Large language models; Pharmaceutical practice; Reproductive and obstetrics-gynecology medicine; Prescription review; Medication counseling; Medication education

随着生活节奏的加快、饮食结构改变、自然环境污染以及晚婚晚育和人流率上升,我国不孕不育与生殖健康问题日趋显著^[1]。据世界卫生组织预测,在21世纪,人类生殖健康问题将成为继肿瘤、心血管疾病之后的第三大顽疾,这将是人类社会面临的一个重要公共卫生健康问题^[1]。药学实践指药师运用专业的知识与技能为医疗保健人员、患者和公众开展药品选择、管理和监测等方面的药学服务,这些服务活动主要包括处方点评、药品调配、用药教育、用药咨询、疑难病例会诊、药品不良反应监测等^[2]。

近年来,人工智能技术正在迅猛发展,凭借其强大的内容创造与复杂数据推理能力,为医疗健康领域带来了重大变革。生成式大语言模型(large language models, LLMs)是基于先进的人工智能系统,在海量数据集上进行预训练,使其能够学习复杂的语言模式,并通过微调或提示技术适配多样化任务,进而处理和生成多种复杂的人类语言^[3]。药学实践作为保障医疗质量与患者安全的关键,正面临从“以药品为中心”向“以患者为中心”深度转型的挑战与机遇,在此背景下,探索生成式LLMs如何革新药学服务,提升其精准性、实用性与可及性,具有重要的理论价值与迫切的现实意义^[4]。生殖与妇产领域的药学服务具有特殊性。首先,服务对象涵盖备孕期、妊娠期、哺乳期等特殊生理阶段,药物选择需同时考虑对母体和胎儿/婴儿的双重影响,用药风险阈值低、决策窗口期敏感;其次,辅助生殖技术用药方案复杂且超说明书用药现象普遍存在;此外,当前生殖专科药学实践中仍面临人力资源不

足,药师知识储备不够的痛点^[5]。当前国内外主流的LLMs包括ChatGPT、DeepSeek、Grok等。ChatGPT是全球应用最广泛的LLMs,也是各项医学研究应用报道最多的LLMs;DeepSeek是国内LLMs的代表,在中文自然语言处理任务中表现优异;Grok是近期新兴主流LLMs,具备较强的实时信息检索能力。目前国内外针对LLMs用于药学实践的研究尚处于起步阶段,相关实用性与可行性研究较为匮乏,现有研究大多局限于单一模型或单一任务,缺少多模型、多任务场景下的对比探索。此外,既往研究极少关注生殖与妇产这类对用药安全标准要求极高的亚专科。本研究围绕生殖及妇产领域药学实践,抽取门诊处方、标准化用药咨询问题及出院用药教育记录为研究资料,分别采用DeepSeek、ChatGPT和Grok进行问答测试。通过混淆矩阵与精确率、召回率、F1值等指标评估处方点评一致性,采用评分法评估用药咨询与用药教育的回答质量,探究LLMs应用的实用性与可行性,以期为药师合理运用LLMs提供依据,助力人工智能工具落地医疗健康场景,进而提升药师工作效能,缓解药学从业人员专业能力不均衡等现实问题。

1 资料与方法

1.1 研究资料

1.1.1 处方点评

采用分层抽样方法,利用电子病历信息系统抽取广西壮族自治区生殖医院2024—2025年门诊处方为研究资料,由2名药师(高级职称1人,中级职称1人)根据《处方管理办法》^[6]、药品

说明书以及广西壮族自治区生殖医院超说明书用药备案目录等资料, 对方剂进行细致点评, 中级职称药师进行初审, 高级职称药师复核。点评结果包括合理处方和不合理处方。合理处方定义为处方书写规范、适应证适宜、无禁忌证、药品选择符合诊疗规范、用法用量正确、无重复用药和相互作用问题; 不合理处方包括不规范处方、不适宜处方和超常处方。本研究仅对脱敏后的处方数据进行回顾性分析, 不涉及患者身份信息的收集、使用和披露, 且不干预患者的诊疗方案, 研究经广西壮族自治区生殖医院医学伦理委员会审查批准, 并豁免患者知情同意 (批准号: KY-LW-2026-11)。

1.1.2 用药咨询与用药教育

用药咨询问题来源于北京市医院管理局组织编写的《妊娠及哺乳期用药咨询标准化手册》^[7]。该手册共收录 54 条用药咨询条目, 涵盖药物疗效、药理作用、妊娠期用药、哺乳期用药、特殊剂型用法、不良反应、用法用量及药物相互作用等方面。为保证抽取问题在内容维度上的代表性。采用分层抽样方法, 从上述问题类型中每类随机抽取 1~2 条, 获得用药咨询问题作为研究资料。用药教育记录来源于广西壮族自治区生殖医院 2025 年生殖医学病区、妇科病区和产科病区的出院带药患者, 采用随机数字表法随机抽取用药教育记录作为研究资料。

1.2 方法

1.2.1 LLMs 的选择

为测试目前主流 LLMs 处理药理学领域常见问题的能力, 本研究在同一时间分别使用 DeepSeek R1、ChatGPT-5、Grok 4 对上述处方点评、用药咨询、用药教育问题依次进行问答并详细记录其回答情况。模型采用云端部署方式, 模型调用时间为 8:00~18:00, 模型主要参数统一设定为 temperature = 0.4, top_p = 0.9。

1.2.2 提示词构建

提示词旨在引导 LLMs 准确而全面地回答研究者提出的问题, 一般包含角色定义、任务说明、分析流程、注意事项、输出形式等^[8]。本研究所有提示词在正式提问前经过小样本预测试, 由 2 名研究者对模型输出质量进行初步评估, 根据预测试结果对提示词语句进行优化。

(1) 设计处方点评提示词框架。作为一名药

师, 请你点评以下处方是否合理。相关资料: 处方号; 患者信息 [性别、年龄、诊断]; 药品名称 [药品规格、药品数量、用法用量]; 我院药事伦理委员会超说明书用药审批情况 [超适应证用药、超用法用量用药、超人群用药等]; 任务要求: 点评处方, 判断是否合理, 并给出最终判断结论。

(2) 设计用药咨询提示词框架。咨询问题具体内容, 作为一名药师, 请你回答这个问题。

(3) 设计用药教育提示词框架。患者信息 [性别、年龄、诊断]; 药品名称 [药品规格、药品数量、用法用量]; 作为一名药师, 请你对该患者进行用药教育。

1.2.3 评价方法

处方点评评价指标: 本研究以药师处方点评结果作为参考标准, 分析药师点评与 LLMs 点评结果的一致性; 设定不合理处方为阳性, 合理处方为阴性; 将药师点评结果分别与 DeepSeek、ChatGPT、Grok 的点评结果整理为混淆矩阵, 包括真阳性 (true positive, TP)、假阳性 (false positive, FP)、真阴性 (true negative, TN)、假阴性 (false negative, FN) 4 个基本评价指标, 根据公式 1~4 计算准确率、精确率、召回率、F1 值。准确率表示模型判断正确的处方数量占所有处方数量的比例; 精确率表示模型和药师同时判断为不合理的处方数量占模型判断为不合理处方数量的比例; 召回率表示模型和药师同时判断为不合理的处方数量占实际不合理处方的比例; F1 值则表示模型的综合性能, 同时兼顾了精确率和召回率。

$$\text{准确率} = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (1)$$

$$\text{精确率} = \frac{TP}{TP + FP} \times 100\% \quad (2)$$

$$\text{召回率} = \frac{TP}{TP + FN} \times 100\% \quad (3)$$

$$F1 \text{ 值} = 2 \times \frac{\text{精确率} \times \text{召回率}}{\text{精确率} + \text{召回率}} \times 100\% \quad (4)$$

DeepSeek、ChatGPT、Grok 输出的用药咨询、用药教育结果由 2 名药师通过查阅药品说明书、临床诊疗指南、专家共识、高质量文献、专业数据库如 Uptodate (<https://www.uptodate.cn/contents/search>)、Drugs.com (<https://www.drugs.com/>)、LactMed (<https://www.ncbi.nlm.nih.gov/books/NBK501922/>)、丁香园用药助手 ([<https://yxqy.whuzhmedj.com>](https://drugs.</p></div><div data-bbox=)

dxy.cn/pc)、《妊娠及哺乳期用药咨询标准化手册》^[7] 标准答案等进行独立评分。如2名药师的

评估结果存在差异,提交至具有高级职称的第3名药师裁决。评分标准见表1。

表1 评分标准
Table 1. Scoring criteria

分值	回答质量	说明
3	完全符合	回答核心观点正确,信息准确且完整,经药师评估无偏差、遗漏或错误
2	基本符合	回答核心观点基本正确,但不完整,存在少量细节缺失或轻微偏差,经药师评估需要补充必要信息
1	部分符合	回答核心观点部分正确,部分错误,经药师评估需要大量修正
0	不符合	回答核心观点错误或完全偏离主题,经药师评估基本无可用信息

2 结果

2.1 一般情况

本研究共纳入90条门诊处方,其中药师评定合理处方70条、不合理处方20条;用药咨询问题12个;用药教育8个。

2.2 处方点评

处方点评混淆矩阵显示,DeepSeek、ChatGPT均可识别出全部不合理处方(FN=0),Grok漏判11条不合理处方;3种模型均存在不同程度FP误判,其中DeepSeek最多(24条),Grok最少(10条)。ChatGPT的准确率、精确率、F1值表现最优;DeepSeek与ChatGPT召回率可达100.0%,Grok的召回率、综合性能最低,3款模型精确率普遍偏低(45.5%~57.1%)。具体见表2。

2.3 用药咨询与用药教育

用药咨询问题类型涵盖药物疗效、药理作用、妊娠期用药、哺乳期用药、特殊剂型用法、

表2 LLMs的处方点评性能评价(%)
Table 2. Evaluation of prescription review performance of LLMs (%)

指标	DeepSeek	ChatGPT	Grok
准确率	73.3	83.3	76.7
精确率	45.5	57.1	47.4
召回率	100.0	100.0	45.0
F1值	62.5	72.7	46.2

不良反应、用法用量、药物相互作用等方面。药师对3个LLMs在用药咨询方面回答进行评分,DeepSeek的表现最优,平均得分为(2.75±0.45)分,75%(9/12)的问题获得满分,Grok表现次之,ChatGPT相对较弱。具体见表3。

用药教育方面,ChatGPT表现最佳,平均得分为(2.50±0.53)分,DeepSeek和Grok的平均得分分别为(2.00±0.93)分和(1.63±0.52)分;相较于药师的专业用药教育,LLMs的回答存在一些局限性,需要人工纠正或补充用药建议。具体见表4。

表3 LLMs在用药咨询、用药教育方面的评分情况
Table 3. Scores of LLMs in medication consultation and medication education

LLMs	用药咨询(n=12)			用药教育(n=8)		
	$\bar{x} \pm SD$	最低分(频次)	最高分(频次)	$\bar{x} \pm SD$	最低分(频次)	最高分(频次)
DeepSeek	2.75 ± 0.45	2 (3)	3 (9)	2.00 ± 0.93	1 (3)	3 (3)
ChatGPT	2.42 ± 0.67	1 (1)	3 (6)	2.50 ± 0.53	2 (4)	3 (4)
Grok	2.58 ± 0.67	1 (1)	3 (8)	1.63 ± 0.52	1 (3)	2 (5)

表4 LLMs回答用药教育问题的局限性
Table 4. Limitations of LLMs in answering medication education questions

涉及主要药物	DeepSeek	ChatGPT	Grok
黄体酮胶囊、磺达肝癸钠注射液、黄体酮注射液	缺乏常见不良反应、磺达肝癸钠注射液不需要2~8℃冷藏、磺达肝癸钠注射液开封后不可留存、未提及黄体酮胶囊建议远离进餐时间服用	未提及黄体酮胶囊建议远离进餐时间服用	磺达肝癸钠注射液不需要2~8℃冷藏

续表 4

涉及主要药物	DeepSeek	ChatGPT	Grok
阴道用乳杆菌活菌胶囊、甲硝唑片、头孢呋辛酯片	未提及阴道用乳杆菌活菌胶囊需 2~8℃贮藏	未提及阴道用乳杆菌活菌胶囊建议与抗菌药物间隔时间使用	阴道用乳杆菌活菌胶囊与抗菌药物建议间隔时间使用，而不是立即使用
生血宝合剂、多糖铁复合物胶囊	—	生血宝合剂成分表述错误	生血宝合剂成分表述错误、疗程计算错误
磺达肝癸钠注射液、左甲状腺素钠片、谷胱甘肽片、地屈孕酮片、黄体酮软胶囊	缺乏常见不良反应、磺达肝癸钠注射液不需要 2~8℃冷藏、左甲状腺素钠片与铁剂、钙剂等药物合用时模型提及的间隔时间不足、黄体酮软胶囊用法用量错误	—	磺达肝癸钠注射液不需要 2~8℃冷藏
戊酸雌二醇片、地屈孕酮片、左甲状腺素钠片、达肝素钠注射液、黄体酮注射液	缺乏常见不良反应、左甲状腺素钠片与铁剂、钙剂等药物合用时模型提及的间隔时间不足	—	达肝素钠注射液不需要 2~8℃冷藏、药理机制解释存在偏差
滋肾育胎丸、地屈孕酮片、间苯三酚注射液、黄体酮注射液	缺乏常见不良反应、间苯三酚注射液用法用量错误	未提及滋肾育胎丸说明书建议用淡盐水或蜂蜜水送服、忌食萝卜、薏苡仁、绿豆芽	间苯三酚注射液药理机制、禁忌证解释错误且该药不需要 2~8℃冷藏
磺达肝癸钠注射液、硫酸羟氯喹片	—	—	—
那屈肝素注射液、硝苯地平缓释片	—	—	那屈肝素注射液不需要 2~8℃冷藏

注：—表示无相关内容。

3 讨论

随着自然语言处理技术及深度学习技术的快速发展，LLMs 已经渗入各行各业。LLMs 已广泛应用于解决医院药学实践问题，主要包括药品供应保障、药学服务、药事管理、药学研究、药学教育等方面，LLMs 凭借其强大的数据处理能力，可作为药师日常工作的辅助工具，提高药学服务的工作效率，改善医院药房的运营现状^[9]。本研究采用生成式 LLMs 回答生殖及妇产领域药学实践问题，初步评价了 DeepSeek、ChatGPT 和 Grok 3 个 LLMs 在生殖及妇产领域的处方点评、用药咨询、用药教育方面的实用性。研究结果表明，LLMs 可以辅助药师进行处方点评、用药咨询、用药教育等工作，通过输入文本数据进行学习训练，LLMs 能够理解用户需求检索相关数据并最后生成类似于人类语言的回答。3 个主流 LLMs 在处理不同类型的药学工作中具备不同的特点，首先，在处方点评方面，ChatGPT 的综合性能表现最佳；其次，在用药咨询方面，DeepSeek 的回答结果最为标准且全面；但对于用药教育方面，3 个 LLMs 的表现均不理想。DeepSeek、ChatGPT、Grok 等主流 LLMs 在预训练数据、模型架构与训练目标上存在显著差异；一方面，各模型所采用的自注意力机制与网络结构不同，使其在上下文

语义建模、长程依赖捕获及逻辑推理能力上表现各异；另一方面，由于训练数据源、领域覆盖与知识更新迭代节奏不同，模型所习得的知识分布与输出偏向也存在明显区别，最终导致不同模型在生成内容的准确性、专业性与风格上呈现差异化特征^[10-11]。

在生殖与妇产领域处方点评场景下，3 种 LLMs 均具备一定的分析处理能力。从准确率来看，ChatGPT 表现最佳（83.3%），其次为 Grok（76.7%）和 DeepSeek（73.3%），表明当前主流 LLMs 具备一定的处方合理性判断能力。然而，本研究中 3 个 LLMs 的精确率普遍偏低（45.5%~57.1%），反映出 LLMs 在判断“不合理处方”时存在较高的 FP 率。导致高 FP 率的原因一方面可能源于其对不合理处方的学习与训练不足或提示词设计不够精细，另一方面真实世界中处方的复杂性可能超出了 LLMs 的处理能力，导致其点评标准往往比药师更严格^[12]。例如，在辅助生殖超促排卵过程中，LLMs 认为仅开具 3 d 用量的注射用尿促卵泡素远未达到常规治疗疗程，判断处方不合理；而在真实世界中，患者可能仅仅开具本次就诊的药品用量，后续会根据超声和性激素等检查结果调整促排卵药品用量并重新开具处方。LLMs 难以对这种复杂多变的临床用药场景作出准确的判断，而药师可依托完整病历信息进

行综合评判。值得注意的是，DeepSeek 和 ChatGPT 在召回率上均达到 100%，表明这 2 个 LLMs 能够完全识别出所有不合理处方，这对于保障患者用药安全具有重要临床价值。然而，Grok 召回率仅为 45%，提示不同架构模型在敏感度方面存在显著差异。F1 值作为综合性能指标，ChatGPT (72.7%) 优于 DeepSeek (62.5%) 和 Grok (46.2%)，说明 ChatGPT 在精确率与召回率的平衡上表现更佳。从临床应用角度分析，LLMs 在处方点评中的优势在于能够快速处理大量处方数据，相较于设定静态规则的处方审核系统，LLMs 还能在检索大量文献资料的基础上进行深度思考并作出最终的判断，可有效减少药师人工审核的工作量。目前研究结果也提示 LLMs 尚不能完全替代药师的专业判断，特别是在生殖与妇产领域，涉及妊娠期、哺乳期等特殊人群的用药安全问题，任何误判都可能带来潜在的临床风险；因此，当前 LLMs 更适合作为药师辅助工具，采用“人机协同审核”模式是更为稳妥的实践方式^[13-14]。

在生殖与妇产领域用药咨询场景下，药师可以根据患者提出的咨询问题，向 LLMs 输入标准化的提问词，LLMs 即可快速在数据库及全网检索药物相关信息并经过整理给出具体建议，模型不仅具备快捷、准确、高效的检索能力，还可以针对不同患者给出个性化的建议，便于患者实际执行，极大提高了药师的工作效率^[15]。在本研究中，DeepSeek 表现最佳，对 12 个生殖及妇产领域用药问题的回答平均得分为 (2.75 ± 0.45) 分，分别获得 9 次 3 分和 3 次 2 分，显著优于 ChatGPT 和 Grok。ChatGPT 和 Grok 在回答用药咨询问题时存在较多不足。例如，在回答“哺乳期使用莫西沙星后需停药多久方可恢复哺乳”这一问题时，ChatGPT 认为，对于足月婴儿，停药后即可立即恢复哺乳并且未提及软骨损害的风险；而药师根据莫西沙星的药动学特征指出，该药物可少量分泌至母乳中，且喹诺酮类药物对婴幼儿骨骼发育及肠道菌群存在潜在不良影响，临床通常建议在服用莫西沙星后，至少经过 5 个半衰期再恢复哺乳更为稳妥，故药师最终点评该项回答仅 1 分。Grok 在回答“聚维酮碘凝胶引起阴道瘙痒”问题时提出的“短期内可试用高锰酸钾坐浴缓解症状”这一建议，需要指出，高锰酸钾是一种强氧

化剂，对阴道黏膜具有一定刺激性。而患者目前的瘙痒很可能是由聚维酮碘引起的接触性刺激或过敏反应，此时再使用高锰酸钾，可能会进一步损伤黏膜，导致灼痛、干燥或瘙痒加重。因此，药师认为 Grok 的回答存在部分问题，不建议患者自行使用高锰酸钾坐浴，应及时就诊。有研究者对 8 种主流 LLMs 在临床药学场景进行比较，结果 DeepSeek-R1 在用药咨询任务中取得 (9.4 ± 1.0) 分的最高综合评分，显著优于其他模型^[14]，与本研究结果一致。然而，尽管 DeepSeek 在本研究中表现突出，但 LLMs 回答的“正确性”仍主要基于标准化手册和专业数据库的比对，尚未能完全模拟药师在面对患者个体差异、合并用药、既往病史等复杂背景下的综合判断能力，因此，药师在实际应用中仍需保持谨慎。Krichevsky 等^[16]研究表明，LLMs 在真实世界药物治疗咨询中的回答质量显著劣于专业医生，出现错误信息的频率显著高于医生。

在生殖与妇产领域用药教育场景下，LLMs 回答的准确性一般，DeepSeek、ChatGPT 和 Grok 的回答平均得分分别为 2.00、2.50、1.63 分，回答稳定性较好的是 ChatGPT (标准差 0.53) 和 Grok (标准差 0.52)，DeepSeek 在回答稳定性方面欠佳 (标准差 0.93)。已有研究同样发现，LLMs 在用药教育任务中回答的正确性和全面性均不及专业药师的回答，往往给出“看起来正确”但信息可能错误或不全的回答，存在一定的安全隐患^[17]。深入分析模型回答的局限性，可将错误类型大致归纳为以下 3 类：第一，储存条件错误，多个模型对不需要冷藏的药品（如磺达肝癔钠、那屈肝素、间苯三酚等）作出了需要冷藏的错误判断；第二，用法用量建议不完整或存在错误，如缺乏与食物或其他存在相互作用药物的合理间隔时间、用药顺序建议不当等；第三，关键信息遗漏，包括常见不良反应、药品成分、药理机制解释错误等。LLMs 在用药教育场景中表现欠佳的原因具有多重性，用药教育本质上是一项高度个体化、情境化的药学服务。当前主流 LLMs 虽然能够检索并整合海量医学文献，却缺乏对患者具体临床背景的深度理解能力，也难以在多个相互制约的用药建议之间进行权衡取舍^[18]。此外，LLMs 在生成出院用药指导时还存在“幻觉”现象，即生成看似合理但实际错误的信息。有研究

报道,在100份出院总结中,约18%的回答存在潜在安全隐患,其中包括6%的幻觉和3%的新增药物相关错误^[19]。因此,在生殖与妇产这一涉及妊娠期、哺乳期等特殊人群的高风险用药领域,LLMs的回答质量尚不能替代药师的专业判断,现阶段更稳妥的实践路径是采用“药师主导、LLMs辅助”的协同工作模式,由LLMs提供初版用药教育素材,再由药师进行系统性审校、修正和个性化补充,以此保障患者用药安全。

综上所述,LLMs在药学服务方面具有提升药师工作效率的潜力,但现阶段模型尚不能独立处理生殖与妇产领域复杂药学问题,专业药师的审核把关仍然是保障用药质量与安全的核心环节。LLMs应定位为药师的辅助工具,采用“人机协同”模式,提高药学服务的效率。目前,处方审核仍面临处方量大、效率低,药师知识水平及注意力有限等问题,人工难以实时全面监控和分析用药风险,而当前处方前置审核系统多依赖静态规则,难以覆盖复杂的用药场景,往往会出现“过度拦截”和“漏网之鱼”的现象^[20]。未来可考虑对LLMs进行本地部署,LLMs可通过不断学习并动态标记疑似不合理处方和预警处方,药师可重点关注可疑处方,从而建立静态规则拦截—动态AI初审与智能监测预警—药师终审的工作模式,提高处方审核效率。经过训练的LLMs在逻辑清晰以及知识全面性方面可能考虑得比人工更佳,对于个性化较强的用药咨询和用药教育部分,LLMs可用于提供素材和结构化框架,药师进行审校后发出。本研究存在一定局限性:作为单中心探索性研究,样本量偏小,可能引入选择偏倚,结论外推性有限;此外,LLMs版本更新迅速,结论仅反映特定时间点的模型性能。

参考文献

- 1 孙文希,胡凌娟.国内外不孕不育现状及其干预策略探讨[J].人口与健康,2019,(12):19-23.[Sun WX, Hu LJ. Current situation of infertility at home and abroad and discussion on intervention strategies in China[J]. Population and Health, 2019, (12): 19-23.] https://kns.cnki.net/kcms2/article/abstract?v=iPTW3hFh4gfj2Q91RkdLq0FkKRKycPW3hzNLwM2-ce4dWgSB_KmMnW4e0j8TmQMwfjSyKHiHYRERqYqFItPVQbfYFEHfkkjQKv_qConDICMSaMYf4xi1LFuIXMxtcOWCgQXbjjiLnExVL04FTDlbvOnxSiKTWSpBVnwZz3kcwOIYICuduJVkQ==&uniplatform=NZKPT&language=CHS.
- 2 Paudyal V, Okuyan B, Henman MC, et al. Scope, content and quality of clinical pharmacy practice guidelines: a systematic review[J]. Int J Clin Pharm, 2024, 46(1): 56-69. DOI: 10.1007/s11096-023-01658-x.
- 3 Shool S, Adimi S, Saboori Amlashi R, et al. A systematic review of large language model (LLM) evaluations in clinical medicine[J]. BMC Med Inform Decis Mak, 2025, 25(1): 117. DOI: 10.1186/s12911-025-02954-4.
- 4 张微微,唐婧,冀召帅,等.人工智能在临床药学中的应用进展[J].药学报,2025,60(10):3045-3059.[Zhang WW, Tang J, Ji ZS, et al. Advances in the application of artificial intelligence in clinical pharmacy [J]. Acta Pharmaceutica Sinica, 2025, 60(10): 3045-3059.] DOI: 10.16438/j.0513-4870.2025-0565.
- 5 费姚,马晶晶,钱晨月.开展生殖-产科序贯药学门诊的SWOT分析及发展策略[J].中国临床药学杂志,2025,34(2):138-141.[Fei Y, Ma JJ, Qian CY. SWOT analysis and development strategies on carrying out sequential pharmaceutical clinics for reproduction and obstetrics[J]. Chinese Journal of Clinical Pharmacy, 2025, 34(2): 138-141.] DOI: 10.19577/j.1007-4406.2025.02.011.
- 6 国家卫生健康委员会.处方管理办法[S/OL].(2007-02-14)[2026-05-01]. https://www.gov.cn/zhengce/2007-02/14/content_5713779.htm.
- 7 冯欣.妊娠及哺乳期用药咨询标准化手册[M].北京:人民卫生出版社,2016:2-70.
- 8 张志玲,李强,闫盈盈,等.大语言模型为药品不良反应报告者生成反馈信息的探究[J].临床药物治疗杂志,2024,22(10):38-42.[Zhang ZL, Li Q, Yan YY, et al. Applying large language models to generate feedback for reporters of spontaneous adverse drug reactions [J]. Clinical Medication Journal, 2024, 22(10): 38-42.] DOI: 10.3969/j.issn.1672-3384.2024.10.008.
- 9 牛露露,张宏亮.大型语言模型在医院药学中的应用——探索、总结与思考[J].广西医学,2025,47(8):1106-1114.[Niu LL, Zhang HL. Application of large language model in hospital pharmacy: a study on exploration, summary and reflection[J]. Guangxi Medical Journal, 2025, 47(8): 1106-1114.] DOI: 10.11675/j.issn.0253-4304.2025.08.06.
- 10 张浩严,吕文涛,余润泽,等.大语言模型研究现状[J].无线电工程,2025,55(1):163-174.[Zhang HY, Lyu WT, Yu RZ, et al. Research status of large language models[J]. Radio Engineering, 2025, 55(1): 163-174.] DOI: 10.3969/j.issn.1003-3106.2025.01.020.
- 11 Wang X, Xiong Z, Zou K, et al. Reasoning-driven large language models in medicine: opportunities, challenges, and the road ahead[J]. Lancet Digit Health, 2026, 8(1): 100931. DOI: 10.1016/j.landig.2025.100931.
- 12 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.[Ma HY, Zhou X, Zhang BY, et al. Practicability of DeepSeek large language model-assisted prescription rationality evaluation[J]. Herald of Medicine, 2025, 44(12): 2057-2061.] DOI: 10.3870/j.issn.1004-0781.2025.12.026.
- 13 Ong JCL, Jin L, Elangovan K, et al. Large language model as clinical decision support system augments medication safety in

- 16 clinical specialties[J]. *Cell Rep Med*, 2025, 6(10): 102323. DOI: [10.1016/j.xcrm.2025.102323](https://doi.org/10.1016/j.xcrm.2025.102323).
- 14 Li L, Du P, Huang X, et al. Comparative analysis of generative artificial intelligence systems in solving clinical pharmacy problems: mixed methods study[J]. *JMIR Med Inform*, 2025, 13: e76128–e76128. DOI: [10.2196/76128](https://doi.org/10.2196/76128).
- 15 刘静, 孙静, 昂韦, 等. 人工智能大语言模型在用药咨询和用药教育中的多维度评估及与药师服务的比较研究[J]. *临床药物治疗杂志*, 2025, 23(6): 38–44, 50. [Liu J, Sun J, Ang W, et al. Multi-dimensional evaluation of artificial intelligence large language models in medication consultation and medication education and a comparative study with pharmacist services[J]. *Clinical Medication Journal*, 2025, 23(6): 38–44, 50.] DOI: [10.3969/j.issn.1672-3384.2025.06.007](https://doi.org/10.3969/j.issn.1672-3384.2025.06.007).
- 16 Krichevsky B, Engeli S, Bode-Böger SM, et al. Human vs. artificial intelligence: Physicians outperform ChatGPT in real-world pharmacotherapy counselling[J]. *Br J Clin Pharmacol*, 2026, 92(3): 891–902. DOI: [10.1002/bcp.70321](https://doi.org/10.1002/bcp.70321).
- 17 Abanmy NO, Al-Ghreimil N, Alsabhan JF, et al. Evaluating the accuracy of ChatGPT in delivering patient instructions for medications: an exploratory case study[J]. *Front Artif Intell*, 2025, 8: 1550591. DOI: [10.3389/frai.2025.1550591](https://doi.org/10.3389/frai.2025.1550591).
- 18 Aydin S, Karabacak M, Vlachos V, et al. Navigating the potential and pitfalls of large language models in patient-centered medication guidance and self-decision support[J]. *Front Med (Lausanne)*, 2025, 12: 1527864. DOI: [10.3389/fmed.2025.1527864](https://doi.org/10.3389/fmed.2025.1527864).
- 19 Stanceski K, Zhong S, Zhang X, et al. The quality and safety of using generative AI to produce patient-centred discharge instructions[J]. *NPJ Digit Med*, 2024, 7(1): 329. DOI: [10.1038/s41746-024-01336-w](https://doi.org/10.1038/s41746-024-01336-w).
- 20 雷珍珍, 梁艳, 张章, 等. 人工智能赋能我国药师开展药学服务模式探讨[J]. *中国药事*, 2026, 40(2): 235–239. [Lei ZZ, Liang Y, Zhang Z, et al. Exploring the application of artificial intelligence in empowering pharmacists to provide pharmaceutical services in China[J]. *Chinese Pharmaceutical Affairs*, 2026, 40(2): 235–239.] DOI: [10.16153/j.1002-7777.2025-08-0038](https://doi.org/10.16153/j.1002-7777.2025-08-0038).

收稿日期: 2026 年 05 月 13 日 修回日期: 2026 年 08 月 02 日
本文编辑: 李 阳 冼静怡