·人工智能赋能检验医学专题·综述 DOI:10.3969/j.issn.1672-9455.2025.24.002

大语言模型在检验医学中的应用评测体系现状及进展*

刘 涛[1,2]综述,杨大干[3△]审校

1.浙江中医药大学医学技术与信息工程学院,浙江杭州 310053;2.永康市妇幼保健院检验科,浙江金华 321300;3.浙江大学医学院附属第一医院检验科,浙江杭州 310003

摘 要:大语言模型(LLM)是基于 Transformer 架构和海量数据训练的深度学习模型,具有对话、内容生成和推理能力。LLM 赋能智慧检验医学,在检验前、中、后及实验室管理等环节具有多种应用场景。但是,LLM 的应用伴随着幻觉、可解释性差等风险,其安全性和有效性亟待严格评估。应用评测体系用于衡量 LLM 在真实场景中的效果与价值,因此,构建一套科学、全面的应用评测体系至关重要。该文综述了 LLM 应用评测体系的构成要素,包括评测的维度、指标、评分、数据集、策略及方法,阐述 LLM 在检验医学领域的应用评测案例,发现评测数据集以公开及模拟数据为主,还面临着决策不透明、缺乏公认标准、隐私及数据安全等挑战。未来将聚焦于构建专用评测框架、采用真实世界数据集、健全应用监管体系及人机协同工作新范式等。探索 LLM 的应用评测体系,可为 LLM 在检验医学领域的安全、有效及合规应用提供理论框架与实践参考。

关键词:大语言模型; 检验医学; 人工智能; 应用评测体系; 幻觉

中图法分类号:R446;TP181 **文献标志码**:A **文章编号**:1672-9455(2025)24-3322-07

Current status and progress in the application evaluation system for applications of large language model in laboratory medicine*

LIU Tao[1,2],*YANG Dagan*[3△]

1. *College of Medical Technology and Information Engineering, Zhejiang Chinese Medical University, Hangzhou, Zhejiang* 310053, *China*; 2. *Department of Laboratory Medicine, Yongkang Maternal and Child Care Hospital, Jinhua, Zhejiang* 321300, *China*; 3. *Department of Laboratory Medicine, the First Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, Zhejiang* 310003, *China*

Abstract: Large language model (LLM) is deep learning models based on the Transformer architecture and trained on massive datasets, possessing capabilities for dialogue, content generation and reasoning. LLM empower intelligent laboratory medicine, presenting diverse application scenarios across pre-analytical, analytical and post-analytical phases, as well as laboratory management. However, the application of LLM is accompanied by risks such as hallucinations and poor interpretability; therefore, their safety and effectiveness need to be rigorously evaluated. Application evaluation systems are used to assess the effectiveness and value of LLM in real-world scenarios; therefore, establishing a scientific and comprehensive application evaluation system is crucial. This article reviews the constituent elements of LLM application evaluation systems, including evaluation dimensions, metrics, scoring schemes, datasets, strategies and methods, and elaborates on application evaluation cases of LLM in laboratory medicine. It is found that evaluation datasets are mainly based on public and simulated data, and that challenges such as opaque decision-making, lack of widely accepted standards, and issues of privacy and data security still remain. Future efforts will focus on constructing dedicated evaluation frameworks, adopting real-world datasets, improving application regulatory systems, and establishing new paradigms for human-machine collaboration. Exploration of application evaluation systems for LLM can provide a theoretical framework and practical reference for the safe, effective, and compliant application of LLM in laboratory medicine.

Key words: large language model; laboratory medicine; artificial intelligence; application evaluation system; hallucination

---

* **基金项目**:国家重点研发计划项目课题(2022YFC3602302)。

△ **通信作者**,E-mail:yangdagan@zju.edu.cn。

引用格式:刘涛,杨大干. 大语言模型在检验医学中的应用评测体系现状及进展[J]. 检验医学与临床,2025,22(24):3322-3327.

大语言模型（LLM）是基于 Transformer 架构和海量数据训练的深度学习模型，具有对话、内容生成和推理能力，给智慧检验医学带来了新的机遇和挑战[1]。主流闭源的 LLM 有 ChatGPT、Gemini、Claude、Grok 等；免费开源的 LLM 有 LLaMA、DeepSeek、Qwen 等，不同版本的 DeepSeek 已在许多医疗机构本地化部署，用于疾病诊断、病历书写、报告解读、智能问答、科教赋能等。但 LLM 的输出可能伴有幻觉，存在可解释性差、结果不一致等隐患[1-2]，建立一套系统化、可复现的 LLM 应用评测体系（AES），用来衡量 LLM 在真实场景中的效果与价值，显得至关重要。AES 是一个多维度、多层次，兼顾安全和性能等的综合性框架。为提高 LLM 研究的透明度和可复现性，TRIPOD-LLM 等规范要求研究者详细报告研究设计、数据处理、模型调优和评价标准等信息[3]。因此，全面、科学地介绍 AES 现状和未来发展，可为 LLM 在检验医学领域的安全、有效及合规应用提供理论框架与实践参考。

1　LLM 在检验医学领域的应用场景

LLM 作为智能基座，涵盖智慧检验医学的全场景、全流程和全要素，可应用于实验室信息生态系统的各个环节，推动系统优化与智能升级。

1.1　LLM 在检验前的应用场景　LLM 凭借其自然语言处理和知识整合能力，在检验咨询、医嘱推荐、样本采集、流程监控等场景中具有应用潜力。ChatGPT-4o 用于检验医嘱的推荐，可提升医疗决策效率[4]。虽然 ChatGPT-4o 在检验前问题的智能回答方面表现最优，但仍需专业人员监管[5]。ChatGPT 在微生物样本采集要求、采集方法和运送等方面可进行智能交互，但需谨慎评估和正确使用[6]。

1.2　LLM 在检验中的应用场景　随着 LLM 及智能体的发展，其在检验中的典型应用场景具有中间件智能化、智能细胞识别、智能试剂管理、智能室内质控、仪器智能监控、检验流程无人化等特征。DeepSeek 可辅助将参考区间按特殊人群和特殊项目进行精准管理[7]。LLM 可在罕见病原体培养基的制备、分析性能数据协助质量控制（质控）和基准评估工作等应用方面提供支持[6]。4 种聊天机器人（ChatGPT-4o、ChatGPT-3.5、WriteSonic 和 Copy AI）在室内质控变异系数超出范围和室间质评不合格时，可提供智能解答，但仍需训练和监管[5]。凭借多模态能力，ChatGPT-4o 可识别正常血细胞，但在异常细胞分类方面表现不佳[8]。

1.3　LLM 在检验后的应用场景　LLM 在检验后可助力智能审核、报告解读、结果互认、患者沟通、辅助诊断等众多应用场景。有研究报道，ChatGPT 可协助报告解读、自动审核并提供决策支持[9]；一项研究在 36 例临床案例中使用 ChatGPT-4o 和 Gemini 1.5 对临床信息和检验结果进行解读，发现以上模型提升了医疗决策支持能力[10]。ChatGPT-4o 能准确识别 24 h 尿液分析中的异常值并推荐个性化饮食方案[11]。ChatGPT（4o 和 3.5 版本）在预测血红蛋白病方面表现良好，可作为辅助工具而非替代品[12]。

1.4　LLM 在实验室管理的应用场景　LLM 可优化实验室资源配置、提升风险管理能力，在即时检验、质量问答、检验考试、知识管理、医检协同、安全合规、科教赋能等场景中具有应用潜力。LLM 可用于质量管理的方案制订、实施监测、性能数据分析、风险评估等[7]。在实验室管理相关问题的智能回答方面，ChatGPT 展现出最佳性能[13]。ChatGPT-3.5 结合标准化病例和图谱库，可提高教学质量[14]。一项针对 363 种检验项目的参考区间判断和解释的研究发现，ChaGPT-4o 表现最佳[15]。

综上所述，LLM 在检验医学领域的智能问答、质控、流程优化、报告解读、决策支持等场景应用，可进一步提升质量和效率，降低人为错误，优化资源配置，实现更智能化的检验管理，成为实验室变革的重要工具。但是，LLM 存在报告单解释过于笼统、垂直领域知识不足、仅在结构化任务（如多项选择题）中表现较好，而在开放性问题（如血细胞形态识别）中表现较差（因为后者需要图像处理和特定领域知识）等问题。同时，LLM 性能高度依赖输入信息的质量（如提示词是否提供参考区间直接影响准确率），且对算力要求高、存在幻觉现象等。因此，当前 LLM 只能作为有用的智能辅助工具，需在专业人员的监督下谨慎使用。

2　LLM 的 AES

AES 包括评测维度、评测指标、评测评分、评测数据集、评测策略和评估方法。LLM 应用评测可分为以下 2 类[16]：（1）理解能力评测，评估检验报告解读、数据分析和知识检索等能力。（2）生成能力评测，评估报告生成、解释说明和与患者沟通等能力。AES 是 LLM 在检验医学领域应用中进行准入评估、性能对比和持续优化的核心依据。

2.1　LLM 的评测维度　评测维度用于描述 LLM 的应用场景和具体任务，是对 LLM 进行多方面、多维度效果评估的基础[16]。如 LLM 在检验报告解读中，需结合临床需求，对仪器报警信息分析、指标异常及程度识别、异常项目关联分析、初步实验室诊断方向、临床诊疗建议等维度进行评测。维度定义对于数据集构建、指标设定具有重要参考意义。

2.2　LLM 的评测指标　评测指标可分为客观指标和主观指标。客观指标是通过公式计算得到的指标，如准确率、召回率、精确率、Rouge-L 指标等[16]，客观指标较为成熟且稳定性好，但可能与人类主观评价不完全一致。主观指标根据具体任务要求进行定义，需明确指标含义和评分标准，但设计复杂，易受人类主观影响，常见主观指标[16-19]如下。（1）准确度：生成内容的事实符合程度，与医学标准相符，没有关键信息

错误或遗漏（如危急值等）。（2）相关度：生成内容紧扣问题或任务要求（如紧扣检验报告单内容），避免冗余无关信息。（3）完整度（全面性）：生成内容全面，无信息缺失或遗漏，回答覆盖所有关键的检验指标、潜在临床意义和诊断方向（包括罕见病等）。（4）有效性（实用性）：评估生成内容的实用程度，是否能辅助临床决策，提高效率或改善患者管理。（5）连贯性（清晰性）：生成内容简洁明了，逻辑连贯，专业术语使用恰当。（6）一致性：同一问题多次测试，生成内容应保持一致。（7）遵循性：符合问题要求、输出格式与约束条件等。（8）真实性：生成内容真实有效，无违反科学常识的虚假信息。（9）有害性（安全性）：生成内容没有偏见、隐私泄露、歧视等违反基本道德伦理和法律以及错误结论、误导性建议等。依据 LLM 应用场景，检验医学需重点关注的评测指标有准确度、完整度、有效性和安全性，以确保 LLM 性能，实现全过程监管及持续改进，减少不良事件发生。

2.3 LLM 的评测评分 评分标准需要定义单样本的评分标准，以及总体得分统计规则。对于客观指标，单样本评分和总体得分依据客观指标的计算公式确定。对于主观指标，单样本评分标准有：（1）二等级法，将单样本的结果进行判断，例如正确得 1 分，错误得 0 分；（2）多等级法，如李克特量表评分，3 或 5 分制，分值越高说明越好[9,20]。总体得分统计规则：通常采用将多个样本评分进行累加并进行归一化，将结果转换为百分制分数，作为最终评测评分。

2.4 LLM 评测数据集 质量高、覆盖全、具有真实数据特征的评测数据集是评估模型性能与可靠性的重要保障。数据集是测评的基石，但获取高质量的医疗数据集困难[1,21]。目前检验医学中用于 LLM 测评的数据集有：（1）公开数据集，如 MedMCQA 是印度医学考试的综合性选择题数据集[22]，MedQA-USMLE 是美国医学执业资格考试制定的多选题数据集[23]，含有解读检验结果的数据集。（2）合成数据集，由相关领域专家精心创建、编写大量虚拟但严格符合医学逻辑的案例，涵盖患者背景、检验结果及标准答案。（3）匿名真实世界数据集，对真实数据进行深度脱敏处理（去除标识符，数据调整）后形成的数据集，如检验报告、多学科诊疗等数据集。（4）多模态数据集，例如细胞形态学、疟疾细胞图像等数据集。

2.5 LLM 评估策略 评估策略的选择，取决于具体的评测需求，可分为打分评估（用于达标性判断）和对比评估（用于优劣性排序）[4,12,24]，见表 1。

表 1 打分评估和对比评估的差异

评价维度	打分评估	对比评估
评价参照系	参照系是金标准或历史数据	参照系是其他对象（对比模型、替代方案）
适用阶段	准入验证（能否发布？）、迭代验证（有改进吗？）	选型决策（用模型 A 还是模型 B？）、竞争优化（如何超越对手？）
数据要求	只需待评对象的数据和明确阈值，如≥90%	必须多对象同条件数据，在公开基准上对比多个 LLM
结论形式	二元化（通过/不通过），绝对分值（如 90/100 分），该 LLM 准确率为 90%，达标	排序（A＞B＞C＞D），相对差距（A 比 B 高 15%，比 D 高 10%），模型 A 准确率比模型 B 高 5%，排名第 1

2.6 LLM 评估方法 LLM 评估方法有人工评估和自动化评估，或二者相结合。人工评估适用于产品调研及最小可行产品阶段。自动化评估适用于应用场景及能力固定、且有确切答案的任务，可针对一些维度和指标进行评估。人工评估需要设置评估标准、评估流程，并组织相关领域专家参与评估。在评估过程中，可分为：（1）采用分工协作，不同评估人员负责不同的评估集，即将评估任务分解为多个子任务，由不同成员分别负责，然后通过协作整合结果。这种方式可以减少个人偏见，并确保评估的全面性。（2）双盲评估，消除主观偏见和期望效应，减少评估者间的主观差异，保证结果的一致性，即不同的评估人员独立负责相同的评估集，最后对不同的打分结果进行合并[20,25]。

以 LLM 对检验报告解读为例，具体评估方法的流程如下：首先，评估前对所有评估人员进行统一培训和考核，培训内容包括相关领域知识、评测判断标准、典型案例等，选择有能力的人员进行预评估，预评估结果的一致性和可靠性达到要求后才能进行正式评估。然后，不同 LLM 回答的结果由专业评估人员采用分工协作或双盲评估方法，依据评测指标和打分标准进行评估打分，评估时对 LLM 的模型类型和版本设盲，并且模型的排列顺序不固定，采用 5 分制李克特量表进行打分。最后，将所有的评分按照模型和评测指标进行分析，得分和排名越高说明模型的评测性能越好。

总之，AES 已经形成评测维度、评测指标、评测评分、评测数据集、评估策略和评估方法等体系框架[16]，利用公开、合成及真实脱敏数据等高质量数据集，对 LLM 的理解与生成能力从多维度、多方面的主客观指标进行量化的效果评估，可为 LLM 的准入、选型及持续优化提供依据。在检验医学的应用评测中应充分考虑医疗领域对专业性、准确性和安全性低容忍度的特定需求，进行全面系统的评估。

3　LLM 在检验医学中的应用评测案例

LLM 在检验医学中的应用评测案例主要包括知识问答、检验结果解读、实验室管理、教学科研等方面。当前，AES 领域的研究数量有限，尚处于初始阶段，LLM 在准确性、可解释性、安全性及临床实用性等方面尚需更多案例进行评测。

3.1　知识问答准确性与推理能力的评测　评测 LLM 是否掌握正确的检验医学知识，并能进行临床逻辑推理，包括：(1)标准化知识问答。通过设计涵盖不同难度的基础理论和临床实践的多选题、判断题及问答题，评估模型的准确性。(2)案例推理分析。通过案例分析题，评价模型对结果的解读能力，判断结果是否异常，并解释其临床意义，列出可能诊断及下一步诊疗建议。该场景的 AES 案例见表 2。

表 2　知识问答准确性与推理能力的评测案例

案例	评测维度	评测数据集	评估方法	评估策略	评测指标	评测结果
DeepSeek R1 在检验医学中的辅助决策能力[20]	诊断、鉴别诊断和检查建议	100 个临床案例	资深医师人工测评	人工打分	准确性、完整性	DeepSeek R1 整体表现较好，但在鉴别诊断和检查建议方面存在局限。
ChatGPT 在检验医学基础知识与结果推理的回答能力[25]	问题回答和结果推理	医学知识、临床情境解读、标准操作规程等 65 道题目	人工测评回答质量	人工打分	准确性、相关性、一致性	ChatGPT 可回答医学问题，但准确度和相关性有待提高，存在知识过时现象。
ChatGPT 作为可靠的检验医学顾问[17-18]	智能回答质量	来自 Reddit 和 Quora 的患者提问及专业人员回答数据	实验室医学专家人工测评	多等级法人工打分	准确性、清晰性、一致性、稳定性	ChatGPT 回复具有更高质量和应用潜力，但存在准确性波动、回答不精确、缺乏透明度等问题。
ChatGPT-4o 能否作为临床实验室检查推荐工具[4]	检验医嘱推荐	15 个模拟临床案例	参照标准人工测评	人工打分	一致性、精确率、召回率	ChatGPT 在检查项目推荐上具有潜力，但召回率偏低。
ChatGPT-4.0 在血细胞识别与分类中的能力[8]	正常/异常血细胞分类	33 张血细胞数字图像	与专家分类结果进行人工测评	人工打分	分类准确率、推理合理性	在多模态视觉任务中，ChatGPT-4.0 在识别正常血细胞分类上略优于异常细胞，但其性能仍不足以单独用作诊断工具，也无法完全取代当前的血液学数字成像软件。
ChatGPT 能否预测血红蛋白病[12]	疾病预测	59 例确诊血红蛋白病和 59 例阴性患者数据	人工测评	依据准确率人工打分	准确率、稳定性(重复性)	GPT-4.0 总体准确率为 76%，但阴性病例准确率低；无参考区间时准确率下降。
LLM 回答乙型肝炎感染相关问题的表现[26]	专业问答	选取主观和客观共 64 个问题，涵盖风险因素、临床表现等	2 位专业医师进行人工评测	人工打分	准确性、一致性、实用性	ChatGPT-4o 在诊断方面表现更好，Gemini 在临床症状描述上表现出色。
评估 LLM 在 RhD 血型输血决策支持中的准确性[19]	输血相关问题的性能	15 道基于真实临床案例的多选题和判断题	人工测评	与专家结果对比评分	准确性	ChatGPT-4o 表现出最佳的整体性能，可能有助于 RhD 血制品输注决策。

3.2　报告生成与解释能力的评测　LLM 报告生成与解释能力是指其将原始检验数据转化为临床医师或患者易于理解且表述准确的能力。例如，给定原始数据(如血常规各指标数值)，LLM 应能生成结构化的初步报告，包括异常值提示及其可能原因概述，并提供患者易于理解的报告解读，以评估其解释的清晰度和准确性。一项研究利用 ChatGPT 解读 10 个模拟检验报告，然后实验室专家从相关性、正确性、有用性和安全性评估其回答结果，发现 ChatGPT 能识别检测项目并判断结果是否偏离参考区间，但解读较肤浅、不全面，很少提供后续诊断建议[9]。有研究评估 ChatGPT、Gemini 和 Le Chat 对 100 个在线健康论坛上患者全血细胞计数解读的能力，并与在线医师的回复进行对比，发现 LLM 的解读不如医师，常生成错误或过度概括的回答，还存在高估病情的情况[27]。有研究评估 LLM 思维链推理能力在检验案例解读中的表现，使用去标识化的 30 例真实会诊案例，采用 OpenAI o1、Gemini 1.5Pro 与 ChatGPT-4o 进行解读，由 3 位实验室主任从事实正确性、推理逻辑、回答完整性和建议实用性进行评估比较，结果显示 OpenAI o1 和 Gemini 1.5Pro 在推理能力、正确性、回答完整性、实用性和知识更新上表现优于 ChatGPT-4o[28]。

3.3 实验室管理应用场景的评测 LLM 在实验室管理中具有广阔应用前景，但其性能因模型和任务类型而异，需结合专家评估以确保可靠性。DeepSeek 可推动检验结果互认，在实验室质量管理的监测、方案制订、风险评估等方面有优势[7]。ChatGPT 在基础知识和技术方面的回答具有较高的准确性，但在技术操作和监管措施方面的回答错误率较高（约 31%）[25]。有研究评测 LLM 在全流程管理中对常见错误的识别能力，发现 ChatGPT-3.5 准确性显著低于 CopyAI 和 ChatGPT-4o，不同 LLM 在不同类型问题上表现有差异[5]。有研究将 109 个基于临床问题的测验题，采用零样本提示法输入各评估模型，由相关领域专家评估模型的回答情况，结果显示 ChatGPT-4o 整体准确率最高，能有效处理专业问题[13]。

综上所述，当前 LLM 在检验医学中的评测研究主要从智能问答、报告解读、疾病诊断、医嘱推荐、管理应用等评测维度出发，采用标准化试题、模拟及真实临床案例作为评测数据集，以实验室专家进行打分评价为主，评估其准确性、相关性、一致性、完整性和安全性等，评测结果普遍显示 LLM 在基础问答和报告生成上具有应用潜力，但在推理深度、复杂案例鉴别诊断及操作细节上的准确性仍不稳定，存在回答较为肤浅与过度解读等局限性。而且，目前的 AES 普遍存在以下共性问题：缺乏专门的检验数据集和评测指标，缺少在真实临床环境中的长期验证，以及不同 LLM 训练数据间存在差异。这些问题导致不同 LLM 在处理同一问题时，或同一 LLM 在面对不同问题时，其表现存在显著差异的情况。

4 LLM 应用评测面临的问题及对策

LLM 在通用知识任务中表现出色，然而，当其作为独立工具应用于专业领域问答时面临着局限性问题，尤其是在对准确性和安全性要求极高的检验医学等医疗场景中[29]。目前主要面临的问题和对策如下：(1)LLM 在评测中暴露出严重缺陷、潜在风险或不符合伦理规范的行为时，尚缺乏系统地记录、披露并追踪修复的失效报告机制。LLM 的黑箱特性和幻觉现象导致其决策过程难以追溯，进而动摇用户对其的信任[30-31]。可采用思维链提示工程增强透明度，使用检索增强生成技术接入权威知识库，同时建立标准化的失效报告机制和权威评测基准，对模型输出进行持续验证，可有效缓解此问题[32-33]。(2)缺乏公认的 LLM 在检验医学领域中的应用评估标准，导致不同模型性能难以进行客观比较，且多数研究缺乏在真实临床环境中的长期有效性验证。针对评测标准与真实世界证据的缺失问题，应开发涵盖复杂场景的标准化数据集，并通过多中心临床研究系统评估其准确性、安全性及对临床结局的实际影响。(3)LLM 应用时，存在患者隐私泄露的情况，发生问题时法律责任不明，训练数据偏见可能加剧医疗不公等风险，必须在技术性能、社会信任与医疗领域的伦理考量之间取得平衡[34]。针对数据安全、责任归属与公平性问题[35-36]，可以通过制订严格的数据隐私规范和法律法规明确各方责任，并建立长效机制持续监测和纠正算法偏见，确保 LLM 应用的公平与安全。

5 LLM 评测的未来展望

未来，LLM 在检验医学领域的深度应用将引发工作范式变革。为应对挑战并充分释放其潜力，今后 LLM 的发展应重点关注以下内容。(1)构建透明、可溯源的评估体系：制订检验医学领域专用的评测框架，通过透明、可追溯的失效报告与反馈体系，全面反映其真实能力与潜在风险，而非仅仅依赖简单的“打分排名”。(2)建立评测数据集和标准：构建一个融合教科书、临床指南、匿名数据及形态学图像等多模态信息的检验医学专用、持续更新的真实世界数据集[37]，并在此基础上，形成开放、多中心、多层次且贴合实际的评测标准。(3)健全 LLM 应用监管体系：健全隐私、伦理准则与监管评测框架，建立生成内容的安全服务体系，明确开发者、医疗机构和使用者在 LLM 辅助诊断中的权利与责任[1]。通过主动防御、多层防护与人工协同，重点解决有害内容、幻觉、合规性及责任归属模糊[38]等安全问题，为 LLM 的合规落地提供坚实的制度保障。(4)人机协同工作新范式：探索人机协同的模式，在将 LLM 应用到检验医学实践中时，在人工智能技术专家和医疗专业人员的共同指导下，构建一个结构化的协同框架，使检验人员能够从烦琐工作中解放出来，而更加专注于复杂病例分析、质量管理及临床协作等高阶任务，从而提升检验医学的整体价值与效率，使其在精准医疗时代发挥更为关键的核心枢纽作用。如：LLM 可以作为“初审者”处理海量常规报告，标记异常和潜在问题，同时还能根据医学专业知识给出初步的解释和建议，再由检验医师进行“复核与决策”，以提高审核的效率。

6 小 结

LLM 为智慧检验医学带来了范式变革，包括检验前咨询与医嘱推荐、检验中智能识别与质控、检验后报告解读与辅助诊断以及实验室管理优化等场景。但是，目前 LLM 存在黑箱特性、幻觉现象及输出不一致等风险，亟须建立系统化、可复现、透明的 AES，包括评测维度、指标、数据集、策略与方法，通过主客观指标对 LLM 的知识问答、检验报告生成与解释、临床决策支持和实验室管理等进行人工或自动化评估。虽然目前 LLM 在检验医学部分领域中表现良好，展现出了较好的应用价值，但仍存在准确率不足、解释不够深入、结果波动等问题，以及面临可解释性和透明度差、真实数据集和评测标准缺失、数据安全与责任归属不完善等挑战。未来，需制订检验医学领域专

用的评测框架，构建多模态和持续演进的真实世界数据集及标准，完善安全与伦理治理体系，并推动人机协同机制，以促进 LLM 在检验医学领域安全、有效、合规地应用。

参考文献

[1] YU E L，CHU X H，ZHANG W W，et al. Large language models in medicine：applications，challenges，and future directions[J]. Int J Med Sci，2025，22(11)：2792-2801.

[2] AZAMFIREI R，KUDCHADKAR S R，FACKLER J. Large language models and the perils of their hallucinations[J]. Crit Care (Fullerton)，2023，27(1)：120.

[3] GALLIFANT J，AFSHAR M，AMEEN S，et al. The TRIPOD-LLM reporting guideline for studies using large language models[J]. Nat Med，2025，31(1)：60-69.

[4] ZAYED A M，FRANS G，DELVAUX N. Evaluating large language models as clinical laboratory test recommenders in primary and emergency care：a crucial step in clinical decision making[J]. Clin Chem Lab Med，2025，63(11)：2186-2197.

[5] ABUSOGLU S，SERDAR M，UNLU A，et al. Comparison of three chatbots as an assistant for problem-solving in clinical laboratory[J]. Clin Chem Lab Med，2023，62(7)：1362-1366.

[6] EGLI A. GPT-4，and other large language models：the next revolution for clinical microbiology[J]. Clin Infect Dis，2023，77(9)：1322-1328.

[7] 李波，袁旭，李小强，等. 人工智能驱动下的检验医学创新探索与实践[J]. 国际检验医学杂志，2025，46(17)：2056-2061.

[8] NEGRINI D，PIGHI L，TOSI M，et al. Evaluating the accuracy of ChatGPT in classifying normal and abnormal blood cell morphology[J]. Clin Chem Lab Med，2025，63(6)：e143-e145.

[9] CADAMURO J，CABITZA F，DEBELJAK Z，et al. Potentials and pitfalls of ChatGPT and natural-language artificial intelligence models for the understanding of laboratory medicine test results：an assessment by the European Federation of Clinical Chemistry and Laboratory Medicine (EFLM) Working Group on Artificial Intelligence (WG-AI)[J]. Clin Chem Lab Med，2023，61(7)：1158-1166.

[10] FELDMAN M J，HOFFER E P，CONLEY J J，et al. Dedicated AI expert system vs generative AI with large language model for clinical diagnoses[J]. JAMA Netw Open，2025，8(5)：e2512994.

[11] KIRIAKEDIS S，DUTY B，CHASE T，et al. Using ChatGPT-4 to analyze 24-hour urine results and generate custom dietary recommendations for nephrolithiasis[J]. J Endourol，2024，38(8)：719-724.

[12] KURSTJENS S，SCHIPPER A，KRABBE J，et al. Predicting hemoglobinopathies using ChatGPT[J]. Clin Chem Lab Med，2024，62(3)：e59-e61.

[13] HEO W Y，PARK H D. Assessment of large language models in medical quizzes for clinical chemistry and laboratory management：implications and applications for healthcare artificial intelligence[J]. Scand J Clin Lab Invest，2025，85(2)：125-132.

[14] 谭婷婷，徐志晔，王森，等. 大语言模型在医学检验标准化病例和图谱库教学中的应用[J]. 江苏卫生事业管理，2025，36(3)：434-437.

[15] BHASURAN B，JIN Q，DEVILLE A，et al. LabQAR：a manually curated dataset for question answering on laboratory test reference ranges and interpretation[EB/OL]. medRxiv，2025 [2025-08-31]. https://www.medrxiv.org/content/10.1101/2025.06.03.25328882v1.

[16] 国家市场监督管理总局，国家标准化管理委员会. 人工智能大模型第 2 部分：评测指标与方法：GB/T 45288.2-2025[S]. 北京：中国标准出版社，2025.

[17] EL-KHOURY J M. ChatGPT：a reliable laboratory medicine consult[J]. Clin Chem，2024，70(9)：1089-1091.

[18] GIRTON M R，GREENE D N，MESSERLIAN G，et al. ChatGPT *vs*. medical professional：analyzing responses to laboratory medicine questions on social media[J]. Clin Chem，2024，70(9)：1122-1139.

[19] LEE J K，CHOI S，PARK S，et al. Evaluation of six large language models for clinical decision support：application in transfusion decision-making for RhD blood-type patients[J]. Ann Lab Med，2025，45(5)：520-529.

[20] LI Q，ZHAN L，CAI X. Assessing DeepSeek-R1 for clinical decision support in multidisciplinary laboratory medicine [J]. J Multidiscip Healthc，2025，18：4979-4988.

[21] WU C，LIN W，ZHANG X，et al. PMC-LLaMA：toward building open-source language models for medicine[J]. J Am Med Inform Assoc，2024，31(9)：1833-1843.

[22] YANG H，LI M C，ZHOU H X，et al. Large language model synergy for ensemble learning in medical question answering：design and evaluation study[J]. J Med Internet Res，2025，27：e70080.

[23] SINGHAL K，AZIZI S，TU T，et al. Large language models encode clinical knowledge [J]. Nature，2023，620(7972)：172-180.

[24] 陆小琴，佳薇，武宇翔，等. 大语言模型在检验医学领域的应用潜力与挑战评估[J]. 临床检验杂志，2024，42(8)：619-623.

[25] MUNOZ-ZULUAGA C，ZHAO Z，WANG F，et al. Assessing the accuracy and clinical utility of ChatGPT in laboratory medicine[J]. Clin Chem，2023，69(8)：939-940.

[26] LI Y，HUANG C K，HU Y，et al. Exploring the performance of large language models on hepatitis B infection-related questions：a comparative study[J]. World J Gastroenterol，2025，31(3)：101092.

（下转第 3334 页）

·人工智能赋能检验医学专题·综述 DOI:10.3969/j.issn.1672-9455.2025.24.003

基于血常规参数的人工智能疾病模型临床应用进展*

陈晓玲[1]综述,曹 科[2],崔胜金[1],刘永棠[1],罗小娟[2△]审校
1.香港大学深圳医院检验医学部,广东深圳 518000;2.广东省深圳市儿童医院检验科,广东深圳 518038

摘 要:血常规参数直接反映了机体血液系统的关键生理和病理状态,具有检测简便、高效、蕴含信息量大等特点。随着人工智能(AI)技术的迅猛发展,机器学习和深度学习在医学数据分析及部分影像识别中的应用为深度挖掘血常规参数的临床价值开辟了新途径。该文综述了基于血常规参数的 AI 诊断模型构建与临床应用研究进展,涵盖线性模型、树模型/集成学习、支持向量机和神经网络等方法,通过整合血常规参数和其他临床数据,可构建用于疾病筛查、诊断与风险预测的高效模型,在恶性肿瘤、感染性疾病及良性非感染性疾病的早期识别、预后与转归判断方面展现出应用价值。然而,AI 模型的构建和应用仍面临数据隐私和安全性、模型泛化能力及临床整合等挑战。未来需进一步推动多中心前瞻性研究、完善方法学规范,并加强与临床信息系统的深度融合,以促进 AI 技术在医疗领域的规模化与规范化应用。

关键词:血常规参数; 人工智能; 疾病诊断; 风险预测; 深度学习
中图法分类号:R446.11;TP181 **文献标志码**:A **文章编号**:1672-9455(2025)24-3328-07

Advances in the clinical applications of artificial-intelligence disease models based on complete blood count parameters*

CHEN Xiaoling[1],*CAO Ke*[2],*CUI Shengjin*[1],*LIU Yongtang*[1],*LUO Xiaojuan*[2△]
1.*Department of Laboratory Medicine,the University of Hong Kong-Shenzhen Hospital,Shenzhen,Guangdong* 518000,*China*;2.*Department of Laboratory Medicine,Shenzhen Children's Hospital,Shenzhen,Guangdong* 518038,*China*

Abstract:Complete blood count parameters directly reflect key physiological and pathological states of the hematologic system and are characterized by simplicity,high efficiency and rich information content. The rapid advancement of artificial intelligence (AI),particularly in machine learning and deep learning as applied to medical data analysis and image recognition,has opened new avenues for the in-depth exploration of the clinical value of complete blood count parameters. This article reviews recent advances in the development and clinical application of AI diagnostic models based on complete blood count parameters. Methods including linear models,tree-based/ensemble learning,support vector machines and neural networks are covered. By integrating complete blood count parameters with other clinical data,efficient models can be developed for disease screening,diagnosis and risk prediction. These models have demonstrated application value in early detection,prognosis assessment,and outcome prediction for malignancies,infectious diseases and benign non-infectious conditions. However,the development and application of AI models are still confronted with challenges related to data privacy and security,model generalizability and clinical integration. Future efforts should prioritize multicenter prospective studies,refined methodological standards,and deeper integration with clinical information systems to facilitate the scalable and standardized application of AI in healthcare.

Key words:complete blood count parameter; artificial intelligence; disease diagnosis; risk prediction; deep learning

血常规(CBC)检测是临床诊断的重要工具,其参数反映机体功能状态,具有广泛的临床应用价值。CBC 是常见的实验室检测项目之一,可提供患者健康状况信息并揭示潜在疾病风险。其主要参数包括红细胞计数(RBC)、白细胞计数(WBC)、血小板计数(PLT)等,这些指标对疾病的诊断和预后评估具有重要意义。RBC 和血红蛋白(Hb)可评估贫血情况,WBC 提示感染或炎症情况,PLT 及其形态学参数与

* **基金项目**:广东省基础与应用基础研究基金企业联合基金(2023A1515220156,2024A1515220079)。
△ **通信作者**,E-mail:luoxiaojuan1983@126.com。
引用格式:陈晓玲,曹科,崔胜金,等.基于血常规参数的人工智能疾病模型临床应用进展[J].检验医学与临床,2025,22(24):3328-3334.