大语言模型在药物相互作用预测中的应用及研究进展

颜漫宇[1,2], 秦家安[2*], 鄢　丹[2*]

(1. 成都中医药大学药学院, 四川 成都 611137; 2. 首都医科大学附属北京友谊医院, 北京市临床药学研究所, 北京 100050)

摘要: 随着人工智能 (artificial intelligence, AI) 技术的飞速发展, 大语言模型 (large language model, LLM) 在生物医药和临床领域中的应用逐渐深化。药物相互作用 (drug-drug interaction, DDI) 预测是药物研发和应用中具有复杂特性的一种安全性研究, 传统方法依赖于实验验证或基于规则的计算模型, 具有耗时和更新滞后的缺点。大语言模型通过语义建模与整合多形态数据知识推理为药物相互作用预测提供了新范式。本文系统叙述了深度学习与大语言模型在DDI预测中的技术进展, 应用案例和技术难点。基于ChatGPT-4、DouBao、ERNIE Bot、Kimi Chat和DeepSeek-R1五种主流大语言模型在可验证的药物组合DDI分析任务中的表现, 探索大语言模型在药物相互作用预测中的应用价值并对未来发展趋势提出展望, 为系统开发、个体化预测优化以及标准化DDI预测框架构建提供参考, 以期推动LLM从实验室向临床的转化进程。

关键词: 大语言模型; 药物相互作用; 药物相互作用预测; 联合用药; 人工智能

中图分类号: R969　　**文献标识码**: A　　**文章编号**: 0513-4870(2025)07-2122-10

Performance evaluation and application value of large language models in the prediction of drug-drug interactions

YAN Man-yu[1,2], QIN Jia-an[2*], YAN Dan[2*]

(*1. College of Pharmacy, Chengdu University of Traditional Chinese Medicine, Chengdu 611137, China; 2. Beijing Institute of Clinical Pharmacy, Beijing Friendship Hospital, Capital Medical University, Beijing 100050, China*)

Abstract: With the rapid development of artificial intelligence technology, the application of large language model (LLM) in the field of biomedical and clinical applications has gradually deepened. Drug-drug interaction (DDI) prediction is a kind of safety study with complex characteristics in drug development and use, and traditional methods rely on experimental validation or rule-based computational models, which are time-consuming and lagging in updating. Big language models provide a new paradigm for drug interaction prediction through semantic modeling and integrating knowledge-based reasoning on polymorphic data. In this paper, we systematically review the technical progress, application cases and technical difficulties of deep learning and big language modeling in DDI prediction. Based on the performance of five mainstream big language models, namely ChatGPT-4, DouBao, ERNIE Bot, Kimi Chat, and DeepSeek-R1, in the verifiable drug combination DDI analysis task, we explore the value of the application of big language models in drug interaction prediction and provide an outlook on the future development trend, and system development, individualized prediction optimization, and standardized system development, individualized prediction optimization, and the construction of a standardized DDI prediction framework, with a view to promoting the translation process of LLM from the laboratory to the clinic.

收稿日期: 2025-05-25;　修回日期: 2025-06-16.
基金项目: 国家中医药管理局重点学科 (zyyzdxk-2023003); 青年北京学者 (2022-051).
*通讯作者 Tel: 86-10-63139318, E-mail: qinjiaan1@hotmail.com; yd277@126.com
DOI: 10.16438/j.0513-4870.2025-0590

Key words: large language model; drug-drug interaction; drug-drug interaction prediction; combination medication; artificial intelligence

药物相互作用 (drug-drug interaction, DDI) 是指两种或多种药物在同时或先后使用, 通过影响药物的理化性质、药代动力学和药效学所产生的药效变化或不良反应现象, 其临床发生率高达 30%, 尤其在老年患者和慢性病患者, 药物的多重使用使得这些风险更为突出[1,2]。DDI 的临床诊断多依赖于医生的临床经验, 具有一定的主观性和个体差异, 传统药学知识难以满足复杂的用药场景相互作用识别, 给临床用药带来安全隐患[3]。相比之下, 人工智能 (artificial intelligence, AI) 技术依靠其对大量、繁杂医学知识高效的分析与整合能力, 显著增进药物相互作用预测的客观性与可解释水平[4]。以 AI 技术实现 DDI 预测具有显著的临床意义, 但受高质量数据集获取、运用以及人工智能建模差异性的限制, 目前 AI 模型在临床中的实际应用效果仍需进一步验证[5]。本文首次提出对多款中文主流大语言模型 (large language model, LLM) 在 DDI 任务中的性能进行系统评估, 采用多维指标横向比较模型表现, 并针对传统的深度学习模型和 LLM 当前面临的技术难点和局限性进行介绍, 以可验证数据对主流的 LLM 进行 DDI 预测评估, 旨在为开发更可靠的 LLM 并优化临床决策方案提供建议与参考。

1 应用于 DDI 预测的 AI 模型分类

传统的深度学习模型和 LLM 在是两种进阶发展的 AI 技术, 深度学习是基础技术框架, 从结构化数据中学习复杂模式, LLM 是其扩展应用, 在 DDI 预测和研究中广泛应用, 显著提升了药物不良反应发现、临床决策和药物安全性评估的效率与准确性[6]。深度学习为 DDI 预测提供了强大的特征学习框架, LLM 可快速筛选海量文献, 辅助构建 DDI 知识库, 为深度学习模型提供高质量训练数据。

1.1 传统深度学习模型及其在 DDI 预测中的应用

深度学习通过模拟人脑神经元连接的多层神经网络架构, 自动从原始数据中学习多层次的特征表示, 并将简单的特征组合成更复杂的特征, 避免了人为的特征提取[7,8], 见图 1。在图像识别、语音识别、自然语言处理等多个领域取得了突出的应用效果[9]。在药物相互作用中深度学习在药物分子结构解析、生物通路、蛋白质相互作用等方面发挥重要作用, 通过对数据集信息的特征捕获, 实现药物相互作用的预测, 代表模型包括循环神经网络 (recurrent neural network, RNN)、图神经网络 (graph neural network, GNN)、卷积神经网络 (convolutional neural network, CNN) 等[10-12], 见表 1[13-24]。

RNN 是一种用于处理序列数据的神经网络架构, 在药物研究领域展现出多方面的重要应用[25]。RNN 能够有效处理药物的序列数据, 提取药物相互作用信息, 处理药物序列中的时间依赖性和长期依赖关系[26]。RNN 还常用于建模药物相互作用过程, 模拟药物在体内代谢、吸收等动态过程, Halder 等[13]提出的深度序贯学习架构 DDINet 模型, 基于排泄、吸收、代谢等机制预测和分类 DDI, RNN 对动态过程进行建模以提升预测准确性。

GNN 是一种用于处理图结构数据的深度学习模型, 在建模药物相互作用网络方面, 能够将药物及其相互作用关系以图结构呈现, 药物为节点, 相互作用为边, 进而有效捕捉药物间复杂关系[27]。Shen 等[28]利用 NPI-GNN 方法便借助 GNN 基于结构对 RNA-蛋白质相互作用预测, 根据网络信息和序列信息预测新的交互, 使药物结构特征可视化, 从而达到更好预测结果。

CNN 是一种特殊类型的人工神经网络, 主要用于分析视觉图像, 在药物研究领域展现出强大的应用潜

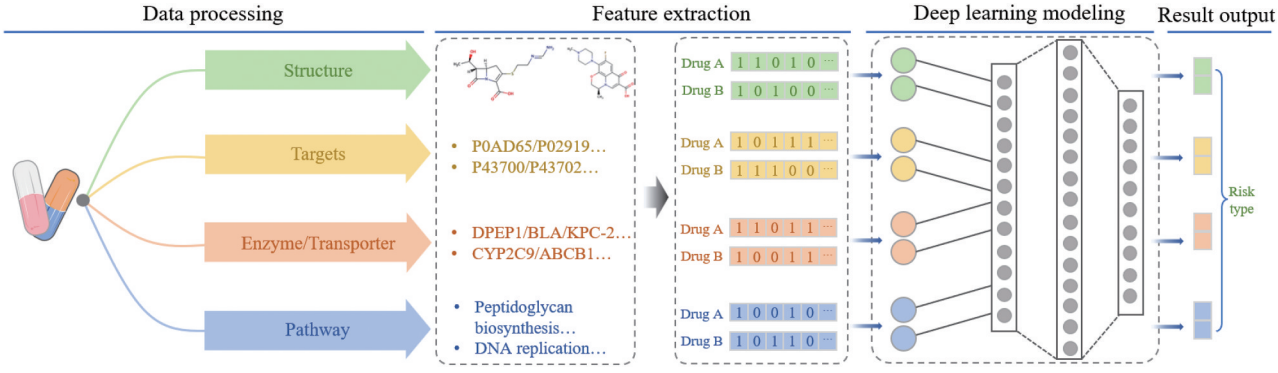


Figure 1 Application of deep learning model in drug interaction prediction

Table 1 Construction of deep learning models based on recurrent neural network (RNN), graph neural network (GNN), convolutional neural network (CNN) algorithms. GRU: Gated recurrent unit; LSTM: Long short term memory; BiLSTM: Bidirectional long short-term memory; GIN: Graph isomorphism network; MLP: Multilayer perceptron; GCL: Graph convolution layer; FCC: Fully connected layer; FW-RFE: Feature elimination based on feature weights; SVM: Support vector machines

Model architecture	Model	Composition	Dataset	Accuracy/%	Ref.
RNN	DDINet	RNN; GRU; LSTM	Kaggle; DrugBank	95	[13]
RNN	TA-RNN	RNN; LSTM; GRU	ADNI; NACC; MIMIC-III	73	[14]
RNN	RNN-VirSeeker	RNN; LSTM; NLP	CAMI; NCBI	92	[15]
RNN	RNN-BiLSTM-CRF	RNN; BiLSTM	BioVid	93	[16]
GNN	SS-GNN	GNN; GIN; MLP	PDBbindv2016; PDBbindv2013	85	[17]
GNN	LR-GNN	GNN; LR; GCN	LDA; MDA	95	[18]
GNN	BPI-GNN	GCN; ImageNet; MGC	ABIDE	91	[19]
GNN	DeepRank-GNN	GNN; GCL; FCC	CAPRI	82	[20]
CNN	SSF-DDI	CNN; Mix-attention; SGFE	DrugBank; Twosides	87	[21]
CNN	SGRU-CNN	GRU; CNN	DrugBank; MEDLINE	74	[22]
CNN	CNN-FWS	CNN; FW-RFE	PTB-XL; ECG	90	[23]
CNN	CNN-SVM	CNN; SVM	PhysioNet	99	[24]

力, 通过自动提取药物的化学结构特征, 借助多层卷积操作与激活函数等, 对药物化学结构等特征进行提取和转换, 用于DDI预测任务的药物特征表示[29-31]。CNN具备处理多模态数据的能力, SSF-DDI模型从药物的化学亚结构、药物序列和药物分子图结构等多模态数据中学习药物组合的特征表示, 更全面地捕捉药物间的相互作用信息[21]。

1.2 大语言模型的分类与技术特点 LLM是一种基于Transformer架构的深度学习模型, 专门用来处理自然语言任务的人工智能系统, 通过在大规模语料库中提取知识进行训练, 学习语言的统计规律、语义关系以及上下文逻辑, 专注于处理和生成自然语言[32,33], 见图2。作为自然语言处理领域的核心技术, LLM能够高效地处理自然语言数据, 捕捉语法结构的复杂性、词汇的细微差异以及语义的深层含义, 且对人工干预的需求低[34]。目前国内外常用的经典LLM主要分为ChatGPT-4、Kimi Chat、DouBao、ERNIE Bot和DeepSeek-R1等。依据训练方式的不同, 主要分为四种不同类型, 在语言推理和逻辑输出层面具有不同的优势, 适用于多种应用场景, 见表2。

ChatGPT借助自注意力机制, 突破了传统RNN、CNN模型远距离信息易丢失的难题, 能够高效捕捉长距离文本依赖关系[35]。同时ChatGPT通过自监督学习方式在大规模多源语料库上进行预训练, 与传统模型不同的是ChatGPT无需人工标注学习语言规律, 这极大节省了时间提高效率[36]。在技术优化上, OpenAI通过调整模型深度等架构参数增强表达能力, 并应用超参数调优和蒸馏模型等压缩技术来提升效率与推理速度[37]。引入强化学习的方法进一步优化了对话交互质量提升实时响应的表现[38]。Radha等[39]用ChatGPT-3.5对120名住院患者药物相互作用输出结果与药师进行比较, 在常见药对中模型输出结果较好, 对于有些药物

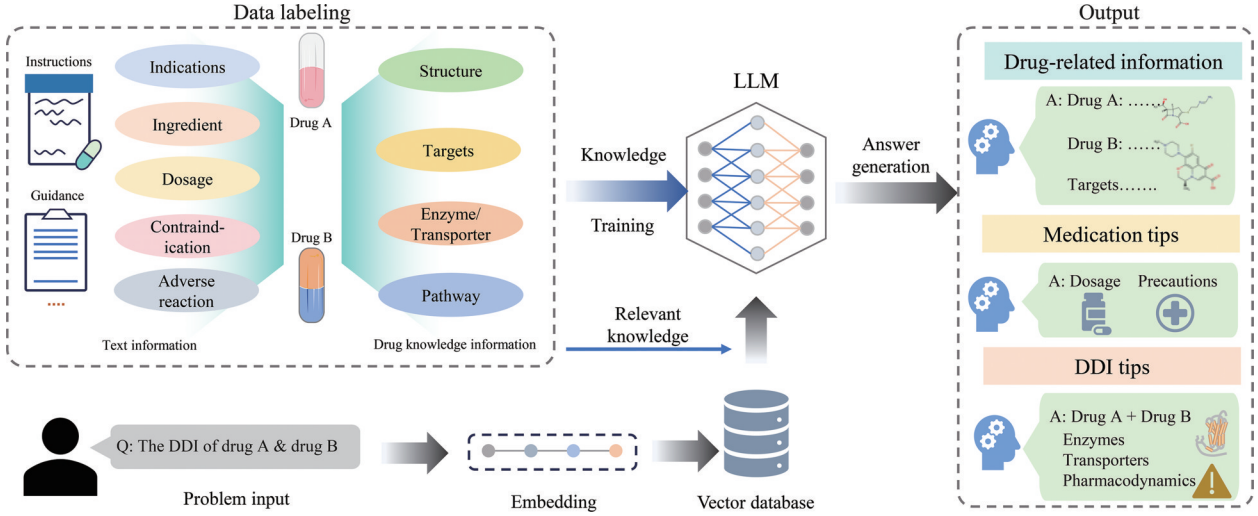


Figure 2 Application of large language model (LLM) in drug interaction prediction

Table 2 Common large language models and their applications in the field of drug interactions

Model	Training method	Feature	Clarification	Application scenario
ChatGPT-4	Supervised learning styles and intensive learning	Real-time response	Self-improvement without extensive manual labeling	Information for healthcare professionals, popularization of science for patients, prediction of potential interactions, assessment of drug interaction mechanisms
Kimi Chat	Sparse hybrid expert architecture and reinforcement learning	Long text analysis	Some experts are involved in the calculations to make optimal decisions	
DouBao	Mix expert models and fine-tuning	Optimization for lightweight training and recreational scenarios	Flexible selection or fusion of experts to introduce multiple modal information such as images and audio	
ERNIE Bot	Pre-training and fine-tuning	Outstanding Chinese comprehension skills	Pre-training on unsupervised corpus, small amount of labeled task-related data for fine-tuning	
DeepSeek-R1	Mix expert models and fine-tuning	Open source modeling and deep thinking	Simultaneous utilization of multiple expert models and tuning of expert models	

组合模型敏感度低, 不能准确识别相互作用。在一项临床情景的ChatGPT生成的鉴别诊断中ChatGPT-4正确诊断率超过80%远超Chat-GPT-3.5[40]。

ERNIE Bot通过预训练–微调方法集成的大语言模型, 模型的开发引入了互联网公开数据集, 显著提升了模型的泛化能力和专业问答准确性[41]。ERNIE Bot还深度融合百度知识图谱, 基于知识增强的预训练, 采用ERNIE架构, 针对中文语言特性进行深度优化, 提升中文理解和生成能力, 在中文处理领域比GPT系列更具优势[42]。在外科住院医师考试中的表现中ERNIE Bot优于住院医师表现[43]。在一项评估全球领域与中文领域LLM解决中国特定近视相关问题方面, 百度的ERNIE Bot和ERNIE 4.0大语言模型表现优秀, 无需格外提示就能输出预期的结果[44]。

DeepSeek-R1基于强化学习算法的混合专家模型架构, 提高模型效率的同时大大降低了训练和推理时间以及资源消耗[45]。通过多头潜在注意力技术降低计算复杂度, 提升模型对复杂语义和长距离上下文的理解能力, 大幅度提升训练效率[46]。DeepSeek利用含350万代码样本的CODEI/O数据集将代码等信息合成为自然语言思维链, 构建训练样本, 开创深度思考这一功能, 将模型思考过程输出便于用户理解的同时减少黑箱模式的不透明性[47]。在医疗任务和临床推理方面表现突出, DeepSeek提供的诊断推理步骤被认为比ChatGPT和Llama 3.1-405B提供的诊断推理步骤更准确[48]。DeepSeek-R1是开源模型, 允许用户查看模型的代码根据自己的需求进行微调, 以适应特定任务或领域, 同时构建性能更好的蒸馏版本, 适用于医学研究和临床实践[49,50]。

DouBao采用稀疏混合专家架构, 与DeepSeek-R1混合专家架构相比, DouBao只激活一部分“专家”模块, 通过稀疏混合专家架构设计, DouBao1.5 Pro实现了仅激活1/7参数就达到很好的使用效果, 显著降低了计算成本[51]。DouBao区别于其他LLM一大特点是多模态端到端训练和低延迟优化。DouBao提出Speech2端到端框架, 将语音与文本Token融合, 支持语音指令控制、低时延打断等功能, 实现真人级交互体验, 通过动态参数调整和算子级优化实现10 ms级低延迟响应, 远超平均大语言响应水平, 适合实时交互与资源受限场景[52,53]。豆包中国牙科执照考试中提供解释时与参考答案的一致性最高, 准确率能达到98%[54]。

Kimi Chat与DouBao相似都采用稀疏混合专家架构和强化学习技术, Kimi Chat仅需激活2.8B参数就能实现高效推理工作, 针对长链式思维任务, Kimi Chat通过长上下文扩展技术将推理窗口扩展至128K, 支持复杂任务的长链推理。结合部分回溯技术优化训练效率, 提升资源利用率[55,56]。Kimi Chat的核心优势在于能够支持200万字无损上下文输入, 远超ChatGPT-4等LLM[57]。目前Kimi Chat已广泛应用于护理诊断、医疗诊断等领域, 成为国产AI模型中的标杆产品。Cao等[58]基于神经外科患者的临床病例, 构建了一个提示框架, Kimi Chat在生成诊断和临床方面具有潜力。

1.3 大语言模型在药物相互作用中的应用 随着ChatGPT-4o、DeepSeek-R1等LLM的发布, LLM在药物相互作用方面的发挥出越来越多的潜力[59]。LLM在DDI方面的应用涵盖了临床分析、个性化医药以及安全性评估等多个领域, 很大程度上提升了药物研究和临床实践的效率[60]。ChatGPT作为生成式AI模型, 能够通过分析患者诊断信息和处方药物数据, 预测潜在的药物相互作用[61]。LLM还在患者用药教育中发挥作用, 尤其是在多药联用普遍的患者中, 生成通俗易懂的药物相互作用说明。对ChatGPT-4模型分别用通用提示和定向提示法, 分析FDA药品标签数据评估药物对肝、心、肾的毒性, 结果显示LLM的评估结果接近专家水平[62,63]。

DeepSeek-R1和ChatGPTo3模型在药物诱导的骨毒性预测中分别实现了87%和88%的准确率[64]。LLM能够快速处理大量多模态数据, 挖掘药物组合信息并预测药物不良反应, 在药物相互作用领域展现出巨大潜力。

1.4 深度学习和大语言模型的技术难点 深度学习在药物相互作用预测中的优势在于能够通过多层神经网络自动提取药物分子的化学结构、分子指纹等特征, 更偏重于结构化数据的分析和特征提取[65]。基于图神经网络等深度学习分析药物−药物相互作用或药物靶点预测得到很好的应用。但深度学习对高质量标注数据的依赖性高, 而药物相互作用数据通常稀缺且昂贵, 这限制了在药物领域的广泛应用[66]。同时深度学习“黑箱”模型缺乏可解释性难以直观展示药物相互作用的具体机制, 这在药物研发和临床应用中可能带来信任问题[67]。

基于深度学习开发的LLM, 通过从海量文本中提取语义特征, 在药物相互作用领域展现出强大的迁移能力和跨任务泛化能力, 可以快速适配多种任务, 从文献中挖掘药物组合信息或预测药物不良反应, 并通过模型微调实现特定领域的高效应用。LLM对训练数据的需求极大, 而药物研发和源于真实世界临床数据集相对有限, 导致在特定任务上的表现受限[68]。LLM复杂结构导致解释成本更高, 需兼顾真实医疗环境的复杂性, 小样本的数据稀疏问题, 易导致模型难以准确预测[69]。

2 大型语言模型在药物相互作用中应用价值评估

2.1 验证方法 本研究选取了ChatGPT-4、DouBao、ERNIE Bot、Kimi Chat、深度求索DeepSeek-R1这五种热门LLM。这些模型在自然语言处理任务中表现出色, 有望在药物相互作用预测中发挥重要作用。为了验证这些模型在药物相互作用预测中的预测效果, 选取了DDInter2.0和DrugBank数据库中已验证的西药−西药药物组合13种药物30组药物组合, 以及实验室前期验证过的中药−西药药物组合数据选取4组, 涵盖了转运体、代谢酶、理化作用和药效4个维度的信息, 为模型的验证提供基准数据信息[70]。

在模型输入环节, 对每个LLM注册全新账号并固定提示词, 确保模型输出一致性。提示词如下“对两种药物 (A药和B药) 进行以下分析: 结合PubMed、DrugBank、DDInter2.0数据库, 在转运体、代谢酶、理化性质、药效四个方面分析药物具体相互作用, 并输出参考文献名称。”记录输出的结果, 模型的结果输出经两位临床药师进行确认, 两名药师独立进行DDI严重程度评分, 为避免相互干扰,采用单盲设计, 最终分歧通过第三方资深药师仲裁解决。与验证数据做对比, 将0到2的分数分配给LLM回答, 分别表示不正确 (0)、部分正确 (1)、完全正确 (2)。根据LLM输出结果在代谢酶、转运体、理化性质、药效和文献支持度五个方面的准确率对模型的预测效果进行评估。

2.2 LLM输出结果与分析 五种LLM在药物相互作用预测与参考数据对比中, DeepSeek-R1模型综合准确率86%, 其他按照准确率高低依次为ChatGPT-4模型准确率84%、DouBao模型准确率79%、Kimi Chat模型准确率78%、ERNIE Bot模型准确率76% (预测结果最差)。加入中药−西药组合后, 增加了预测复杂性, 所有LLM与验证数据准确率略有下降, 但排名未发生变化, 其中DeepSeek-R1模型准确率84%在五个LLM中药物相互作用预测效果最佳, ChatGPT-4模型准确率82%、DouBao模型准确率78%、Kimi Chat模型准确率77%、ERNIE Bot模型准确率75%。整体来看DeepSeek-R1和ChatGPT-4模型在药物相互作用预测中表现最为稳定, 准确率都在80%以上, ERNIE Bot的表现相对较弱, 评价结果见图3。

在代谢酶方面, DouBao表现最优, 准确率为90%, 加入中药后为降低到79%。这五种LLM对常见药物的代谢机制理解较佳, 但是在涉及中药成分时, 不准确率的比例明显增加。在苯巴比妥和普萘洛尔组合中, DDInter2.0数据库中是没有代谢酶相关信息的, 但这五种LLM均输出相应代谢酶CYP1A2, 猜测是因为DDInter2.0数据库未更新相关数据。这说明LLM在某种程度上能够弥补传统数据库的不足, 尤其是在数据更新滞后或信息缺失的情况下, LLM通过其训练数据和知识整合能力, 推测出可能涉及的CYP1A2代谢酶。对于中药−西药药物组合Kimi Chat在代谢酶方面预测结果与验证数据一致, 预测准确率较好。西药−西药药物组合中, ERNIE Bot出现三组与参考数据出现完全不一致情况, 而DeepSeek-R1、ChatGPT-4LLM主要的问题是结果只涵盖部分代谢酶主要是CYP3A4, 未能全面考虑所有相关的代谢酶, 从而导致输出的信息不够完整。猜测可能是LLM在训练过程中对高频出现的CYP3A4代谢酶赋予更高的权重, 而对低频或特定场景下的CYP2D6等代谢酶关注不足。对五个LLM输出结果的细节描述分析, 发现LLM更侧重描述代谢酶在药物相互作用中的整体作用和临床意义, 但是在基因多态性对药物代谢和相互作用的影响上输出内容较少涉及。

在转运体方面, Kimi Chat在转运体预测中表现最佳, 尤其是在西药−西药药物组合场景下的预测准确率最高, 分析输出结果发现Kimi Chat、ERNIE Bot等更侧重于描述转运体在药物吸收、分布、排泄等过程中

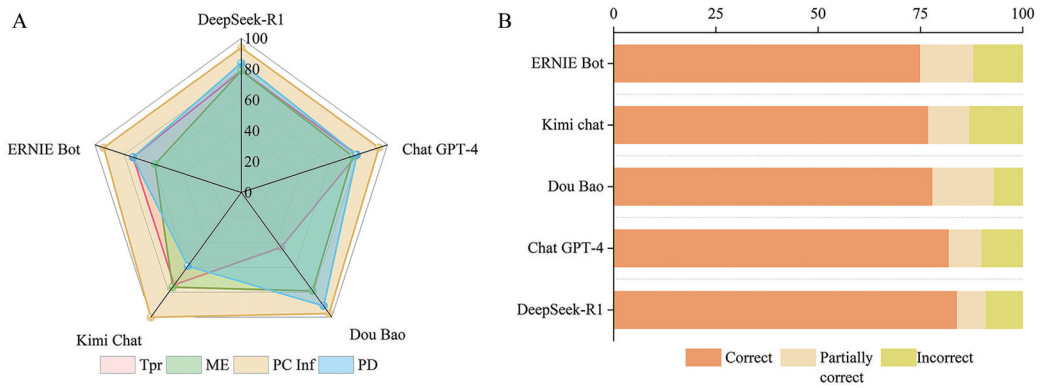


Figure 3 Performance of large language model in drug-drug interaction (DDI) prediction and response quality assessment. A: Accuracy analysis of large language models based on DDI prediction indicators; B: Classification distribution of large language model responses by correctness (correct, partially correct, incorrect); Tpr: Transporter; ME: Metabolic enzyme; PC Inf: Physical-chemical information; PD: Pharmacodynamics

的宏观作用, 输出结果仅为笼统的P-gp并未输出具体的转运体名称。单从结果准确率来看DouBao表现最差, 完全正确率仅为44%, 但通过分析输出结果发现DouBao能够识别到OATP、OCTs等其他转运体, 相较于其他四个LLM转运体的识别, DouBao在这一方面表现较为突出。总体来看, 这五种LLM在转运体预测方面并不理想, 均不能够准确地预测两药连用时可能发生的具体转运体名称。

从理化性质预测结果来看, 五种LLM表现都非常出色, 尤其在西药–西药药物组合预测中实现了完全一致。虽然加入中药–西药药物组合后部分模型的准确率略有下降, 但整体表现仍然非常稳定。这五种LLM都能够考虑药物的理化性质对药物相互作用的影响, 总结药物的溶解度、稳定性、分子量等理化性质分析药物在体内的吸收、分布和代谢过程。但不同模型在对理化性质的综合分析和应用能力上有所不同。其中DeepSeek-R1能够结合文献, 具体分析药物的理化性质与转运体和代谢酶的相互作用关系, 进而影响药物相互作用的发生和程度。Kimi Chat、DouBao、ERNIE Bot、ChatGPT-4只是简单提及理化性质对药物相互作用的潜在影响, 缺乏具体的分析和应用。

在药效方面, DeepSeek-R1同样表现最佳, 准确率为80%, 而Kimi Chat的准确率最低60%, 可能与其推理能力的局限性以及影响药效因素复杂性有关。这五种LLM均能够理解药物相互作用对药效的影响, 输出结果能提示药物相互作用导致药效是增强还是减弱, 同时会提示两药同用发生的不良反应。能够分析药物相互作用对药物疗效的影响机制, 具体说明如何通过影响药物的血药浓度、作用靶点等来改变药效。但是DeepSeek-R1、ChatGPT-4能够结合药物的转运体、代谢酶和理化性质等因素, 综合分析更准确地预测药物相互作用对药效的影响, DouBao、ERNIE Bot等只是进行简单的描述和分析, 缺乏深入的机制探讨和准确的预测能力。

从总体来看, DeepSeek-R1在预测药物相互作用的整体性能上表现最为稳定, 其次是ChatGPT-4, 而ERNIE Bot、Kimi Chat和DouBao在特定领域存在一定短板。在推理链条方面DeepSeek-R1和ChatGPT-4推理思维逻辑更缜密, DouBao与Kimi Chat响应速度更快。通过西药–西药药物组合与加入中药–西药的药物组合对比分析发现, 中药–西药的预测结果并不理想, 不一致的比例明显增加, 可能与中药成分复杂性和多靶点作用机制有关。各模型在不同任务中的表现差异反映了其在药物相关领域的知识覆盖深度和推理能力的差异, 未来可通过增加专业领域数据和优化建模方法进一步提升预测准确性。

本文还验证了LLM输出结果中的参考文献在PubMed检索中的准确率, 结果显示不同的LLM表现出较大差异。其中KimiChat输出的文献真实度最高, 对输出文献真实度进行验证, 发现KimiChat在西药–西药药物组合输出的文献在PubMed数据库中都可以检索到真实度100%, 但是在关于中药–西药药物组合文献检索真实度并不理想。DouBao输出的文献真实度低于50%, 而DeepSeek-R1、ChatGPT-4、ERNIEBot输出文献名称全为模型编撰, 这一类错误称之为模型幻觉。推测可能与模型的训练数据有关, KimiChat、DouBao训练数据库中可能包含与DDI相关的高质量生物医学文献, 在PubMed检索时较为准确识别出相关的文献。但是在中药方面输出的结果还是有所欠缺, 在检索中药–西药药物组合时, KimiChat会在输出时明确提醒用户“由于A药和B药之间的相互作用研究较少, 以下是一些可能相关的文献, 但请注意这些文

献可能不直接涉及两者的相互作用。”LLM中的模型幻觉是指模型在生成文本时，有时会生成与输入内容不相关、无意义或偏离输入的输出[71]。在本次验证过程中ChatGPT-4等大语言模型输出的结果会提供具体的文献名称、作者姓名以及杂志名称，但通过对文献检索验证，结果显示文献并不是真实存在的，模型输出的结果极具迷惑性，在使用LLM时需要格外注意这个问题。

通过对于上述药物组合预测的综合分析，这五种LLM各有优缺，DeepSeek-R1、ChatGPT-4在药物相互作用预测总体性能上比其他几个LLM表现更优，但是在文献支持上KimiChat表现更加出色。DouBao在各个方面表现中等，ERNIEBot综合表现最差。总体而言，大多数LLM在回答药物相互作用问题时表现出了广泛的知识基础和推理能力，能够帮助专业药学工作者和医生掌握药物基础知识，在预测药物相互作用方面的潜力是巨大的。在药物相互作用方面LLM主要是对药物机制理解不够，对生物医学领域的专业知识建模不足，尤其是涉及到中药药物预测时，可能的原因包括模型参数、提示设置和训练数据。这体现出有必要进一步改进，增加训练数据量、调整模型超参数配置、改进训练手段、进行更细致的评估以及引入领域专业知识以增强模型表现。在一项不同参数大语言模型对比中参数规模从7B提升至33B时，模型准确率从21%显著提升至68%[72]。本次仅验证了34组药物组合，样本量有限可能无法完全反映模型在不同类型DDI上的应用能力。选取的药物组合只针对两两用药，未涉及到多药联用复杂临床情境，这些因素可能影响结论的普适性。但根据本次验证结果，也能看出不同LLM的优点与不足，结果具有参考意义。未来期望能引入更多药物机制研究及相关实验验证数据，以提升模型对复杂药物任务的预测能力，针对LLM输出文献时的模型幻觉，有望借助增加对DDI相关文献的标注力度，加大标注数据的质量与数量，同时把DrugBank、PharmGK等权威数据库以及药学领域的专业知识与背景信息纳入模型训练，以呈现最新的科研成果和药物互作的知识内容，增进模型的综合性。

3 讨论与展望

3.1 大语言模型现有挑战 LLM在药物相互作用中的应用仍存在一定的局限性。基于上述测试结果显示几款LLM在药物相互作用预测方面表现较好，但其针对不同数据信息的理解仍然存在偏差，预测结果存在不确定性，需结合使用者自身经验选择参考回答结果。模型的局限性主要体现在以下几个方面，首先算法合理性问题突出，通用模型因缺乏对医学领域复杂性的适配性，易产生错误判断，需结合专业算法进行修正。其次结果可验证性缺陷制约着临床信任，黑箱决策机制与模拟场景评估的局限性导致真实临床环境表现差异显著[73]。伦理与安全风险同样不容忽视，错误信息传播可能引发误诊风险，训练数据中的性别、种族差异等社会偏见及隐私泄露问题亟待解决[74]。LLM需要海量数据支撑，但医疗数据存在标注成本高、隐私限制等问题不能直接用于临床。而通过Lab2clean等算法清洗工具，虽然在一定程度上标准化格式和提升数据质量，但也不能保证结果真实性，仍存在噪声[75]。当前LLM主要为辅助工具使用并不能给出独立解决方案，其适用场景主要限于信息整合或初步筛查，最终决策仍需专业人员审核，而在高风险领域需额外验证。解释黑箱模型也是未来LLM面对的一大难题[76]。通用的LLM未针对临床用途进行优化，容易生成虚假信息，在使用中要注意仔细辨别。

3.2 大语言模型发展脉络 跨学科多模态数据的引入是未来LLM预测性能限制突破的有效手段。以AlphaFold 3为代表的分子结构图像跟文本联合建模技术，依靠构建多模态基础模型在图像、视频等领域实现的突破，表明跨模态语义理解将成为技术演进的关键要点[77]。在医疗范畴中，多模态模型在医学影像分析、DDI预测等场景已呈现出显著优势，其具备的跨模态推理能力将助力医疗AI向更精准的临床辅助方向前进[78]。随着医疗、药学等专业范畴的深度需求，LLM正从通用语言模型向领域专用模型过渡，对细分领域的EAST-GPT、ACS-GPT等专业模型予以开发，代表着模型架构经由专家知识注入与领域数据微调实现精准契合[79]。临床决策支持系统得建立严格的验证机制去应对数据偏见，并采用专家协作框架去平衡生成能力与伦理上的风险[80-82]。应对模型生成内容潜在的虚假性风险，要拟定数据治理的伦理守则，医疗AI领域已开始开拓“专家−开发者”协同验证机制的探索之路，采用建立数据溯源体系与临床验证流程的方式，修订ICMJE/COPE指南，保障模型输出合乎药物安全标准[83-85]。该治理模式将带动形成可检验可追溯的AI伦理框架。针对传统大语言模型因参数量庞大引发的算力瓶颈，混合专家架构、模型压缩技术加上监督学习协同策略成为研究热门方向[86-88]。采用动态稀疏激活机制以及知识蒸馏技术，模型推理效率能实现多个数量级的增长，让其在移动设备与边缘计算的场景中具备可行性[89,90]。

LLM在药学领域的技术潜力与临床挑战并存，DDI预测的结果需要经过严格的临床验证，确保在实际应用中的安全性和有效性。现今LLM的预测结果

多数是在研究阶段, 未进行充分的临床验证, 没办法直接应用于临床实践, 未来需借助专业化模型开发工作、标准化评估体系搭建和伦理规范健全, 推动其安全开展, 同时警示过度依赖产生的隐患, 此过程不仅仅要求技术创新, 更需要医疗、技术、法律等多领域的携手合作, 以实现人工智能跟临床实践的深度融合。

作者贡献: 颜漫宇负责文献调研和文章撰写; 秦家安负责文章修改; 鄢丹负责文章框架设计和修改。

利益冲突: 本文所有作者声明不存在利益冲突关系。

References

[1] Zazzara MB, Palmer K, Vetrano DL, et al. Adverse drug reactions in older adults: a narrative review of the literature [J]. Eur Geriatr Med, 2021, 12: 463-473.

[2] Jiang Y, Yang M, Wang S, et al. Emerging role of deep learning-based artificial intelligence in tumor pathology [J]. Cancer Commun (Lond), 2020, 40: 154-166.

[3] Wu HY, Shendre A, Zhang S, et al. Translational knowledge discovery between drug interactions and pharmacogenetics [J]. Clin Pharmacol Ther, 2020, 107: 886-902.

[4] Gupta R, Srivastava D, Sahu M, et al. Artificial intelligence to deep learning: machine intelligence approach for drug discovery [J]. Mol Divers, 2021, 25: 1315-1360.

[5] Sadeghi S, Bui A, Forooghi A, et al. Can large language models understand molecules? [J]. BMC Bioinformatics, 2024, 25: 225.

[6] Zhang T, Leng J, Liu Y. Deep learning for drug-drug interaction extraction from the literature: a review [J]. Brief Bioinform, 2020, 21: 1609-1627.

[7] Stahlschmidt SR, Ulfenborg B, Synnergren J. Multimodal deep learning for biomedical data fusion: a review [J]. Brief Bioinform, 2022, 23: 569.

[8] Wu S, Roberts K, Datta S, et al. Deep learning in clinical natural language processing: a methodical review [J]. J Am Med Inform Assoc, 2020, 27: 457-470.

[9] Wang H, Hu X. Deep learning in bioinformatics and biomedicine [J]. Methods, 2023, 209: 38-39.

[10] Gan Y, Liu W, Xu G, et al. DMFDDI: deep multimodal fusion for drug-drug interaction prediction [J]. Brief Bioinform, 2023, 24: 397.

[11] Zhang N, Wang XY, Han B, et al. Research progress of artificial intelligence technology in pharmacology of traditional Chinese medicine [J]. Acta Pharm Sin (药学学报), 2025, 60: 550-558.

[12] Sun Z, Lu X, Duan H, et al. Deep dynamic patient similarity analysis: model development and validation in ICU [J]. Comput Methods Programs Biomed, 2022, 225: 107033.

[13] Halder A, Saha B, Roy M, et al. A novel deep sequential learning architecture for drug drug interaction prediction using DDINet [J]. Sci Rep, 2025, 15: 9337.

[14] Al OM, Bozdag S. TA-RNN: an attention-based time-aware recurrent neural network architecture for electronic health records [J]. Bioinformatics, 2024, 40: 169-179.

[15] Liu F, Miao Y, Liu Y, et al. RNN-VirSeeker: a deep learning method for identification of short viral sequences from metagenomes [J]. IEEE/ACM Trans Comput Biol Bioinform, 2022, 19: 1840-1849.

[16] Pouromran F, Lin Y, Kamarthi S. Personalized deep Bi-LSTM RNN based model for pain intensity classification using EDA signal [J]. Sensors, 2022, 22: 8087.

[17] Zhang S, Jin Y, Liu T, et al. SS-GNN: a simple-structured graph neural network for affinity prediction [J]. ACS Omega, 2023, 8: 22496-22507.

[18] Kang C, Zhang H, Liu Z, et al. LR-GNN: a graph neural network based on link representation for predicting molecular associations [J]. Brief Bioinform, 2022, 23: 513.

[19] Zheng K, Yu S, Chen L, et al. BPI-GNN: Interpretable brain network-based psychiatric diagnosis and subtyping [J]. Neuroimage, 2024, 292: 120594.

[20] Réau M, Renaud N, Xue L C, et al. DeepRank-GNN: a graph neural network framework to learn patterns in protein-protein interfaces [J]. Bioinformatics, 2023, 39: 759.

[21] Zhu J, Che C, Jiang H, et al. SSF-DDI: a deep learning method utilizing drug sequence and substructure features for drug-drug interaction prediction [J]. BMC Bioinformatics, 2024, 25: 39.

[22] Wu H, Xing Y, Ge W, et al. Drug-drug interaction extraction *via* hybrid neural networks on biomedical literature [J]. J Biomed Inform, 2020, 106: 103432.

[23] Zhu J, Lv J, Kong D. CNN-FWS: a model for the diagnosis of normal and abnormal ECG with feature adaptive [J]. Entropy, 2022, 24: 471.

[24] Ozaltin O, Yeniay O. A novel proposed CNN-SVM architecture for ECG scalograms classification [J]. Soft Comput, 2023, 27: 4639-4658.

[25] Shi Y, Ding C, Li S, et al. New RNN algorithms for different time-variant matrix inequalities solving under discrete-time framework [J]. IEEE Trans Neural Netw Learn Syst, 2025, 36: 5244-5257.

[26] Zou J, Zhao L, Shi S. Generation of focused drug molecule library using recurrent neural network [J]. J Mol Model, 2023, 29: 361.

[27] Chen Y, You J, He J, et al. SP-GNN: learning structure and position information from graphs [J]. Neural Netw, 2023, 161: 505-514.

[28] Shen ZA, Luo T, Zhou YK, et al. NPI-GNN: predicting ncRNA – protein interactions with deep graph neural networks [J]. Brief Bioinform, 2021, 22: 51.

[29] Young SI, Zhe W, Taubman D, et al. Transform quantization for CNN compression [J]. IEEE Trans Pattern Anal Mach Intell, 2022, 44: 5700-5714.

[30] Son CS, Kang WS. Multivariate CNN model for human locomotion activity recognition with a wearable exoskeleton robot [J]. Bioengineering, 2023, 10: 1082.

[31] Herbreteau S, Kervrann C. DCT2net: an interpretable shallow CNN for image denoising [J]. IEEE Trans Image Process, 2022, 31: 4292-4305.

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

[33] Huang H, Zheng O, Wang D, et al. ChatGPT for shaping the future of dentistry: the potential of multi-modal large language model [J]. Int J Oral Sci, 2023, 15: 29.

[34] Gibney E. China's cheap, open AI model DeepSeek thrills scientists [J]. Nature, 2025, 638: 13-14.

[35] Dergaa I, Chamari K, Zmijewski P, et al. From human writing to artificial intelligence generated text: examining the prospects and potential threats of ChatGPT in academic writing [J]. Biol Sport, 2023, 40: 615-622.

[36] Wang R, Liu X, Ren P, et al. What are the public's concerns about ChatGPT? A novel self-supervised neural topic model tells you [J]. Mathematics, 2025, 13: 183.

[37] Tan S, Xin X, Wu D. ChatGPT in medicine: prospects and challenges: a review article [J]. Int J Surg, 2024, 110: 3701-3706.

[38] Byrne MD. Generative artificial intelligence and ChatGPT [J]. J Perianesth Nurs, 2023, 38: 519-522.

[39] Radha RP, Hung EH, Ashford M, et al. Evaluating the capability of ChatGPT in predicting drug-drug interactions: real-world evidence using hospitalized patient data [J]. Br J Clin Pharmacol, 2024, 90: 3361-3366.

[40] Hirosawa T, Kawamura R, Harada Y, et al. ChatGPT-generated differential diagnosis lists for complex case-derived clinical vignettes: diagnostic accuracy evaluation [J]. JMIR Med Inform, 2023, 11: 48808.

[41] Sahin MF, Atea H, Keles A, et al. Responses of five different artificial intelligence chatbots to the top searched queries about erectile dysfunction: a comparative analysis [J]. J Med Syst, 2024, 48: 38.

[42] Yuan XT, Shao CY, Zhang ZZ, et al. Comparing the performance of ChatGPT and ERNIE Bot in answering questions regarding liver cancer interventional radiology in Chinese and English contexts: a comparative study [J]. Digit Health, 2025, 11: 20552076251315511.

[43] Huang L, Hu J, Cai Q, et al. The performance evaluation of artificial intelligence ERNIE bot in Chinese National Medical Licensing Examination [J]. Postgrad Med J, 2024, 100: 952-953.

[44] Guo S, Li G, Du W, et al. The performance of ChatGPT and ERNIE Bot in surgical resident examinations [J]. Int J Med Inform, 2025, 200: 105906.

[45] Deng Z, Ma W, Han Q L, et al. Exploring DeepSeek: a survey on advances, applications, challenges and future directions [J]. IEEE/CAA J Autom Sin, 2025, 12: 872-893.

[46] Temsah A, Alhhasan K, Altamimi I, et al. DeepSeek in healthcare: revealing opportunities and steering challenges of a new open-source artificial intelligence frontier [J]. Cureus, 2025, 17: 79221.

[47] Ullah E, Parwani A, Baig MM, et al. Challenges and barriers of using large language models (LLM) such as ChatGPT for diagnostic medicine with a focus on digital pathology - a recent scoping review [J]. Diagn Pathol, 2024, 19: 43.

[48] Rasool A, Shahzad MI, Aslam H, et al. Emotion-aware embedding fusion in large language models (Flan-T5, Llama 2, DeepSeek-R1, and ChatGPT 4) for intelligent response generation [J]. AI, 2025, 6: 56.

[49] Liu Q, Huang C, Zhan G, et al. The need for expansion of global collaborations on AI in oncology [J]. Lancet, 2025, 405: 1339.

[50] Okaiyeto SA, Bai J, Wang J, et al. Success of DeepSeek and potential benefits of free access to AI for global-scale use [J]. Int J Agric Biol Eng, 2025, 18: 304-306.

[51] He J, Dai H, Sui R, et al. EvilPromptFuzzer: generating inappropriate content based on text-to-image models [J]. Cybersecurity, 2024, 7: 70.

[52] Wang J, Zhu Z, Liu C, et al. LLM-enhanced multimodal detection of fake news [J]. PLoS One, 2024, 19: 0312240.

[53] Yu S, Hou Y, Li H. Suitability of Chinese GenAI platforms for early childhood education: a multifaceted evaluation [J]. AI Brain Child, 2025, 1: 1-14.

[54] Xiong YT, Zhan ZZ, Zhong CL, et al. Evaluating the performance of large language models (LLMs) in answering and analysing the Chinese Dental Licensing Examination [J]. Eur J Dent Educ, 2025, 29: 332-340.

[55] Hori K, Suzuki K, Ogata T. Enhancement of long-horizon task planning *via* active and passive modification in large language models [J]. Sci Rep, 2025, 15: 7113.

[56] Lai H, Liu J, Bai C, et al. Language models for data extraction and risk of bias assessment in complementary medicine [J]. NPJ Digit Med, 2025, 8: 74.

[57] Tarecha RI, Choirina P. Compressing large language models (LLMs) using knowledge distillation for optimizing inference time and model size [J]. G-Tech J Technol Terapan, 2025, 9: 1056-1062.

[58] Cao Y, Hu L, Cao X, et al. Can large language models facilitate the effective implementation of nursing processes in clinical settings? [J]. BMC Nurs, 2025, 24: 394.

[59] Shin E, Hartman M, Ramanathan M. Performance of the ChatGPT large language model for decision support in community pharmacy [J]. Br J Clin Pharmacol, 2024, 90: 3320-3333.

[60] Zhou H, Li M, Xiao Y, et al. LEAP: LLM instruction-example adaptive prompting framework for biomedical relation extraction [J]. J Am Med Inform Assoc, 2024, 31: 2010-2018.

[61] Cavazzotto TG, Dantas DB, Queiroga MR. ChatGPT and exercise prescription: human *vs*. machine or human plus machine? [J]. J

Sport Health Sci, 2023, 13: 661-662.

[62] Connor S, Wu L, Roberts RA, et al. Is ChatGPT ready for public use in organ-specific drug toxicity research? [J]. Drug Discov Today, 2025, 30: 104297.

[63] Deng L, Wang T, Zhai Z, et al. Evaluation of large language models in breast cancer clinical scenarios: a comparative analysis based on ChatGPT-3.5, ChatGPT-4.0, and Claude2 [J]. Int J Surg, 2024, 110: 1941-1950.

[64] Chen YQ, Yu T, Song ZQ, et al. Application of large language models in drug-induced osteotoxicity prediction [J]. J Chem Inf Model, 2025, 65: 3370-3379.

[65] Krittanawong C, Johnson KW, Rosenson RS, et al. Deep learning for cardiovascular medicine: a practical primer [J]. Eur Heart J, 2019, 40: 2058-2073.

[66] Yu J, Xu Z, He X, et al. DIA-TTS: deep-inherited attention-based text-to-speech synthesizer [J]. Entropy, 2022, 25: 41.

[67] Vo TH, Nguyen NTK, Kha QH, et al. On the road to explainable AI in drug-drug interactions prediction: a systematic review [J]. Comput Struct Biotechnol J, 2022, 20: 2112-2123.

[68] Anisuzzaman DM, Malins JG, Friedman PA, et al. Fine-tuning large language models for specialized use cases [J]. Mayo Clin Proc Digit Health, 2025, 3: 100184.

[69] Benary M, Wang XD, Schmidt M, et al. Leveraging large language models for decision support in personalized oncology [J]. JAMA Netw Open, 2023, 6: 2343689.

[70] Maeda K, Hisaka A, Ito K, et al. Classification of drugs for evaluating drug interaction in drug development and clinical management [J]. Drug Metab Pharmacokinet, 2021, 41: 100414.

[71] Collerton D, Barnes J, Diederich NJ, et al. Understanding visual hallucinations: a new synthesis [J]. Neurosci Biobehav Rev, 2023, 150: 105208.

[72] Yilmaz BE, Gokkurt Yilmaz BN, Ozbey F. Artificial intelligence performance in answering multiple-choice oral pathology questions: a comparative analysis [J]. BMC Oral Health, 2025, 25: 573.

[73] Shi B, Chen L, Pang S, et al. Large language models and artificial neural networks for assessing 1-year mortality in patients with myocardial infarction: analysis from the medical information mart for intensive care IV (MIMIC-IV) database [J]. J Med Internet Res, 2025, 27: 67253.

[74] Freyer O, Wiest IC, Kather JN, et al. A future role for health applications of large language models depends on regulators enforcing safety standards [J]. Lancet Digit Health, 2024, 6: 662-672.

[75] Zayed AM, Janssens A, Mamouris P, et al. lab2clean: a novel algorithm for automated cleaning of retrospective clinical laboratory results data for secondary uses [J]. BMC Med Inform Decis Mak, 2024, 24: 245.

[76] Kilic Y, Tulu CN. TurkSentGraphExp: an inherent graph aware explainability framework from pre-trained LLM for Turkish sentiment analysis [J]. PeerJ Comput Sci, 2025, 11: 2729.

[77] Abramson J, Adler J, Dunger J, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3 [J]. Nature, 2024, 630: 493-500.

[78] Sun D, Hadjiiski L, Gormley J, et al. Outcome prediction using multi-modal information: integrating large language model-extracted clinical information and image analysis [J]. Cancers, 2024, 16: 2402.

[79] Mahajan A, Tran A, Tseng ES, et al. Performance of trauma-trained large language models on surgical assessment questions: a new approach in resource identification [J]. Surgery, 2025, 179: 108793.

[80] Downing GJ, Tramontozzi LM, Garcia J, et al. Harnessing internet search data as a potential tool for medical diagnosis: literature review [J]. JMIR Ment Health, 2025, 12: 63149.

[81] Olawade DB, Teke J, Adeleye KK, et al. Artificial intelligence in *in-vitro* fertilization (IVF): a new era of precision and personalization in fertility treatments [J]. J Gynecol Obstet Hum Reprod, 2025, 54: 102903.

[82] Olawade DB, Weerasinghe K, Mathugamage MDDE, et al. Enhancing ophthalmic diagnosis and treatment with artificial intelligence [J]. Medicina, 2025, 61: 433.

[83] Sousa-pinto B, Vieira RJ, Mmrques-cruz M, et al. Artificial intelligence-supported development of health guideline questions [J]. Ann Intern Med, 2024, 177: 1518-1529.

[84] Diaz O, Rodriguez-ruiz A, Sechopoulos I. Artificial intelligence for breast cancer detection: technology, challenges, and prospects [J]. Eur J Radiol, 2024, 175: 111457.

[85] Risin DJ, Monda KL, Gagne JJ, et al. Implementing accuracy, completeness, and traceability for data reliability [J]. JAMA Netw Open, 2025, 8: 250128.

[86] Williams CYK, Miao BY, Kornblith AE, et al. Evaluating the use of large language models to provide clinical recommendations in the emergency department [J]. Nat Commun, 2024, 15: 8236.

[87] Nazar W, Nazar G, Kamińska A, et al. How to design, create, and evaluate an instruction-tuning dataset for large language model training in health care: tutorial from a clinical perspective [J]. J Med Int Res, 2025, 27: 70481.

[88] Chen Y, Zhang H, Cameron M, et al. Predictive sequence learning in the hippocampal formation [J]. Neuron, 2024, 112: 2645-2658.

[89] Bui N, Nguyen G, Nguyen N, et al. Fine-tuning large language models for improved health communication in low-resource languages [J]. Comput Methods Programs Biomed, 2025, 263: 108655.

[90] Zhu H, Wang R, Qian J, et al. Leveraging large language models for predicting postoperative acute kidney injury in elderly patients [J]. BME Front, 2025, 6: 0111.