小 型 微 型 计 算 机 系 统
Journal of Chinese Computer Systems
DOI:10.20009/j.cnki.21-1106/TP.2025-0311
2026 年 4 月 第 4 期
Vol.47 No.4 2026

大模型驱动的可解释医疗推荐技术综述

赵海燕[1],郝家辉[1],曹 健[2],朱能军[3],朱思吉[4]

[1](上海理工大学 光电信息与计算机工程学院,上海 200093)
[2](上海交通大学 计算机学院,上海 200030)
[3](上海大学 计算机科学与工程学院,上海 200444)
[4](上海交通大学 医学院附属瑞金医院普外科乳腺疾病诊治中心,上海 200025)
E-mail:cao-jian@sjtu.edu.cn

摘 要: 医疗推荐系统通过整合多源医学证据与患者特征,提供精准高效的个性化诊疗路径,显著提升疗效并降低风险.其临床应用的核心挑战在于可解释性,传统模型因"黑箱"弊端导致医患信任缺失.现有研究虽在因果推理与多模态语义对齐方面取得进展,但仍受限于传统机器学习范式,多模态大语言模型凭借出色的自然语言理解与多模态融合能力,为可解释医疗推荐提供新路径.本文通过分析判别式、生成式及混合式模型在医疗领域中的应用,提出一种基于大语言模型的医疗推荐可解释分类方法,揭示其在解释质量、多模态支持及指南依从性的突破.最后,本文指出联邦解释学习、因果增强等技术将会成为该领域内未来热门方向,进一步推动可信医疗推荐系统发展.

关 键 词: 大语言模型;医疗推荐系统;可解释性;联邦解释学习
中图分类号: TP391 **文献标识码:** A **文 章 编 号:** 1000-1220(2026)04-0909-10

Survey on Large Language Model-driven Explainable Medical Recommendation Technologies

ZHAO Haiyan[1],HAO Jiahui[1],CAO Jian[2],ZHU Nengjun[3],ZHU Siji[4]

[1](School of Optical-Electrical and Computer Engineering,University of Shanghai for Science and Technology,Shanghai 200093,China)
[2](Department of Computer Science and Technology,Shanghai Jiao Tong University,Shanghai 200030,China)
[3](School of Computer Science and Engineering,Shanghai University,Shanghai 200444,China)
[4](Comprehensive Breast Health Center,Department of General Surgery,Ruijin Hospital,School of Medicine,Shanghai Jiao Tong University,Shanghai 200025,China)

Abstract: Medical recommendation systems provide precise and efficient personalized diagnosis and treatment pathways by integrating multi-source medical evidence and patient characteristics,significantly improving therapeutic efficacy while reducing risks. The core challenge for their clinical application lies in explainability,as traditional "black-box" models lead to a lack of trust from doctors and patients. Although existing research has made progress in causal reasoning and multimodal semantic alignment,it remains constrained by traditional machine learning paradigms. Multimodal large language models,leveraging their exceptional natural language understanding and multimodal fusion capabilities,offer a new pathway for explainable medical recommendations. This paper analyzes the application of discriminative,generative,and hybrid models in the medical field,proposes a large language model-based explainable classification method for medical recommendations,and reveals breakthroughs in explanation quality,multimodal support,and guideline compliance. Finally,this paper indicates that technologies such as federated explainable learning and causal enhancement will become future key directions in this field,further advancing the development of trustworthy medical recommendation systems.

Keywords: large language models;medical recommendation systems;explainability;federated explainable learning

0 引 言

医疗推荐系统在辅助疾病诊断、治疗方案优化和预后预测中展现出显著价值.近年来,相关技术正逐渐从单一模态的感知处理,发展为具备多模态融合能力的认知建模,在降低误诊率、优化治疗路径及改善患者预后等方面展现出巨大潜力[1].医疗推荐系统的可解释性是其落地临床应用的可信赖性基石,直接影响医生采纳意愿与患者安全意识.在高度敏感的医疗决策场景中,当模型仅提供预测结果而缺乏可解释透明决策逻辑时,容易导致临床认知断层、医患信任危机、监管

收稿日期:2025-07-30 收修改稿日期:2025-10-23 基金项目:上海交通大学医工交叉项目(YG2024QNB05)资助;国家自然科学基金项目(82573726)资助. 作者简介:**赵海燕**,女,1975 年生,博士,副教授,CCF 会员,研究方向为服务计算、数据挖掘、推荐系统;**郝家辉**,男,2000 年生,硕士研究生,研究方向为医疗推荐;**曹 健**,男,1972 年生,博士,教授,博士生导师,CCF 杰出会员,研究方向为智能数据分析、服务计算、协同计算、网络计算等;**朱能军**,男,1992 年生,博士,讲师,CCF 会员,研究方向为数据挖掘、推荐系统、可信医疗推荐;**朱思吉**,男,1985 年生,博士,主治医师,研究方向为外科手术.

合规缺失[2]等后果,可见医疗推荐的可解释性有着重大意义,由此催生了大量相关研究.例如,在方法论层面,因果推理与多模态语义对齐成为医疗推荐解释生成的核心理论支撑,一方面,因果图模型通过 do 算子量化治疗方案对预后的干预效应,将传统特征相关性解释升级为因果机制推演[3],另一方面,多模态注意力机制则通过联合嵌入空间映射,揭示文本-影像-时序数据的异质特征交互模式[4];在应用层面,医疗推荐可解释性研究主要覆盖诊断决策支持、治疗方案推荐与预后风险评估三大核心场景[5].

现有医疗可解释技术分类方法多基于模型架构(如 Transformer 变体)或输出形式分类,这类分类方式虽关注模型内部结构或最终表达形式,却未能有效构建起技术与临床动态工作流需求以及知识可信保障机制之间的桥梁.这导致现有方法的解释能力往往与医生实际决策过程存在脱节,且难以提供符合医学知识体系、可验证、可追溯的解释依据,从而削弱了其临床可信度和实用价值.为弥合这一鸿沟,本文提出一种以临床工作流适配性和知识可信保障为核心的医疗大模型(Large Language Models,LLMs)可解释性分类框架.该框架基于解释生成范式与知识整合机制,将技术路径划分为判别式、生成式与混合式 3 类.

在关联临床动态工作流方面,判别式解释范式通过特征重要性得分实现实时单模态解释,高度适配治疗方案快速筛选或影像检验关键指标解读等需要即时、局部、聚焦式反馈的场景;生成式解释范式通过自回归架构融合多模态输入,生成跨模态自然语言因果链,天然契合诊断推理、医患沟通等需要动态交互式追问和复杂因果叙事的场景,模拟了医生逐步推导的过程;混合式解释范式融合因果推理与统计学习生成结构化因果链,特别适用于预后风险评估等需要结合深层因果机制与统计证据支撑的复杂决策场景.

在强化知识可信保障机制方面,判别式范式虽相对简单,但其解释可直接关联到具体的、可溯源的临床特征或生物标志物,便于医生依据专业知识进行验证;生成式范式通过整合多模态数据和生成自然语言解释,有潜力将医学知识图谱、临床指南中的因果逻辑融入解释链条,提高解释的医学合理性和可理解性;混合式范式则显式地结合了因果图模型和统计学习,其生成的"结构化因果链"旨在提供一种既符合医学因果认知,又有数据支撑的可验证解释框架,为知识可信性提供了双重保障.

本文主要有以下贡献:

1)提出一套 LLMs 与可解释医疗推荐方法结合的分类体系,包含 LLMs 推荐范式(判别式、生成式、混合)与可解释方法(因果性解释、对比性解释、合规性解释),为研究提供结构化分析工具;

2)系统梳理与分析三类 LLMs 推荐范式的可解释方法适配机制,介绍三类 LLMs 推荐范式下的医疗推荐可解释方法技术,为将来的研究提供新思路;

3)提出未来研究的两个重要方向:一是通过联邦解释学习构建多中心协作框架,实现隐私保护下的可解释模式共享;二是开发自动化检测算法和模板迭代引擎,实现实时指南更新.

1 医疗推荐可解释性与大语言模型介绍

1.1 医疗推荐可解释性

医疗推荐的可解释性是指推荐系统能够以清晰、透明的方式向用户阐明其推荐结果的逻辑依据、推理过程以及所依赖的支持证据[6].这一特性是推动医疗推荐系统落地应用的核心要求,直接影响医生采纳率、患者信任度以及监管合规性.

医疗推荐的可解释性需要同时关注临床决策透明化与医疗场景特殊性.在透明化层面,其核心矛盾源于黑箱模型的表征学习不确定性:数据分布偏差可能导致潜在空间映射失真,引发错误关联[7].医疗场景的特殊性则体现在高风险性与多模态解释的技术相互结合[8]:首先,高风险性要求模型通过不确定性量化技术[9]评估潜在风险,以应对极低容错阈值的临床决策;其次,多模态复杂性则需解释系统构建跨模态注意力机制,通过对比学习对齐病理文本描述与影像特征,生成影像-文本联合归因图,突破单模态解释局限.二者共同导致可解释医疗推荐系统落地的关键瓶颈:前者保障生命安全;后者实现多模态证据融合.这些要求的交汇点揭示了医疗可解释性研究的本质矛盾:如何在保持模型预测性能的同时,构建符合临床认知范式与监管标准的透明化决策体系.

1.1.1 传统的可解释方法分类

传统的可解释方法可划分为因果性解释(Causal Explanation)、对比性解释(Contrastive Explanation)与合规性解释(Compliance Explanation)[10],在医疗领域,通过这种三维解释框架,提升医疗决策支持系统的临床可信度与科学可追溯性.因果性解释的核心在于揭示治疗决策的生物学机制本质,其方法论基础为结构因果模型(Structural Causal Model,SCM)[11],通过路径分析(Path Analysis)[12]显式建模医疗要素(如药物)作用通路与临床表象的因果链.例如,Sigismund 等[13]在 EGFR 突变型非小细胞肺癌治疗中,不仅推荐奥希替尼作为优选方案,更通过蛋白质激酶域三维构象模拟,解释该药物通过共价结合 ATP 口袋,选择性抑制 T790M 耐药突变的分子机制.

对比性解释则基于多准则决策分析(Multi-Criteria Decision Analysis,MCDA)框架[14],通过构建效用函数量化备选方案的收益-风险权衡.以 HER2 阳性乳腺癌辅助治疗为例,Menard 等人[15]将曲妥珠单抗联合化疗与 TD-M1 方案进行多维度对比:在疗效维度、毒性维度及经济维度建立帕累托前沿(Pareto Frontier),其本质是通过凸优化[16]求解临床偏好的非劣解集.

合规性解释的底层逻辑在于构建循证医学知识图谱[17],通过动态语义映射将推荐方案与权威指南条款进行逻辑蕴含验证.以柯尔斯顿大鼠肉瘤病毒癌基因同源物(Kirsten Rat Sarcoma Viral Oncogene Homolog,KRAS)野生型转移性结直肠癌一线治疗为例,Hu 等人[18]首先将指南条款形式化为逻辑规则,继而将患者数据映射为语义实体,经子概念包含判定验证方案合规性,若存在禁忌症,则触发冲突消解规则生成替代方案.该过程输出可审计的解释链,实现循证约束的决策保障.

1.1.2 医疗推荐可解释性方法评估

评估医疗推荐系统的可解释性,其标准需深刻融入医疗

场景的核心需求与约束,显著区别于通用领域.其特殊性首要体现在临床逻辑验证性上,即解释的核心价值在于其是否符合循证医学逻辑和病理生理机制,需能被临床专家回溯验证合理性,常通过指南依从率和专家共识评分量化.其次,多模态解释融合度至关重要,需评估解释是否有效整合并清晰呈现跨模态证据链的协同作用,关键指标包括跨模态注意力一致性和多模态联合可追溯性.第三,区别于揭示相关性,医疗解释的核心目标是指导可执行的临床干预,其因果可干预性需通过反事实推理实验验证解释是否揭示了可操作的因果路径.第四,解释生成速度必须严格匹配临床决策流程的紧迫性,急诊场景要求判别式模型的实时解释需在亚秒级响应,而复杂会诊允许生成式模型的因果链解释有稍长但可接受的延迟,解释生成延迟成为关键实用性指标.最后,安全性与合规性审计要求解释便于监管审查和风险评估,必须清晰呈现决策依据、量化不确定性并符合伦理规范,其可审计性和风险标识清晰度是重要考量.这种严格聚焦于临床逻辑、多模态融合、因果干预、时效约束及安全审计的评估框架,是确保医疗可解释性具备临床落地价值和安全性的关键保障.

1.2　大语言模型

大语言模型是基于深度学习框架的海量参数模型,通过自监督学习从大规模文本数据中捕捉语言规律,具备文本生成[19]、语义理解[20]与知识推理能力[21].大语言模型的发展历程[22]如图 1 所示.

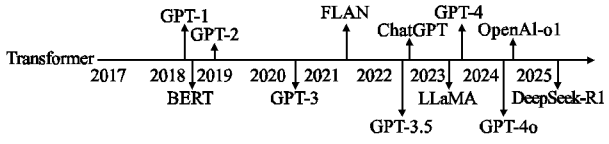

图 1　大语言模型的发展历程

Fig. 1　Evolutionary trajectory of large language models

此外,在不同时间段的代表性大语言模型的摘要信息如表 1 所示.

1.2.1　大语言模型

LLMs 通过融合深度学习框架的算法创新、自然语言处理的任务建模能力,以及大规模计算资源的工程化支持,实现了人类语言理解与生成的能力突破.LLMs 的基础是 Transformer 架构[34],其摒弃了传统循环神经网络(Recurrent Neural Network,RNN)的序列依赖性[35],Transformer 架构的核心在于自注意力机制(Self-Attention Mechanism)[36],其数学本质是通过查询(Query)、键(Key)、值(Value)向量的点积运算动态计算文本中词元间的语义相关性权重,公式化表达为(1):

$$Attention(Q,K,V) = soft\ max\left(\frac{Qk^T}{\sqrt{d_k}}\right)V \tag{1}$$

其中,Q、K、V 由输入向量线性变换生成,d_k 为缩放因子,该机制使模型能捕捉文本中的跨位置语义关联;位置编码(Positional Encoding)[37]通过正弦函数或可学习向量嵌入位置信息,解决 Transformer 对序列顺序不敏感的问题;多头注意力(Multi-Head Attention)[38]并行执行多组自注意力计算,增强模型对复杂语义模式的捕捉能力.

表 1　各个时间段的代表性大语言模型

Table 1　Representative large language models across key developmental stages

年份	模型名称	参数规模	模型结构	关键贡献
2017	Transformer[23]	65M	自注意力机制	首次提出自注意力机制,开启了并行计算与高效长程依赖建模的新时代
2018	GPT-1[24]	117M	自回归 Transformer	开创了生成式预训练范式,奠定了大规模语言模型的基础
2018	BERT[25]	110M	双向 Transformer	引入双向预训练方法,大幅提升文本理解效果
2019	GPT-2[26]	1.5B	自回归 Transformer	展现了显著的零样本学习能力,推动文本生成任务的极限
2020	GPT-3[27]	175B	自回归 Transformer	证明了大规模预训练的强大能力,实现了卓越的上下文理解与小样本学习
2021	FLAN[28]	137B	指令调优 Transformer	通过指令微调显著提升零样本学习表现
2022	GPT-3.5[29]	175B	自回归 Transformer	在 GPT-3 基础上优化交互性能
2022	ChatGPT[30]	175B	基于 GPT-3.5	将对话能力提升到新高度
2023	LLaMA[31]	65B	自回归 Transformer	以开放高效为目标,促进了学术界的广泛应用与后续模型的快速迭代
2023	GPT-4[32]	兆级参数	多模态 Transformer	在多模态处理和复杂推理能力上实现突破
2024	GPT-4o	兆级参数	多模态 Transformer	实现文本、语音、图像等多模态的实时交互
2024	OpenAI-o1	兆级参数	多模态 Transformer	下一代基础模型,展现出尖端性能与广泛适用性
2025	DeepSeek-R1[33]	671B	混合专家 Transformer	以高效与成本效益兼顾的设计推动前沿研究

1.2.2　LLMs 在可解释医疗推荐中的优势

LLMs 的可溯源推理能力使得其在可解释医疗推荐领域具有很强的适配性,通过术语本体映射、跨模态认知建模与可微分推理路径 3 项关键技术,构建了面向医疗大语言模型的认知可信框架.其中,术语本体映射的核心在于解决自由文本表述与医学知识体系间的语义异构性,其技术本质是通过实体链接(Entity Linking)[39]建立基于临床医学术语标准的语义对齐层——利用双塔神经网络(Dual-Tower Network)[40]

将非结构化描述映射至标准术语空间,同时整合 DrugBank 的药物-靶点作用矩阵与在线人类孟德尔遗传数据库(Online Mendelian Inheritance in Man,OMIM)的基因型-表型关联图谱,构建循证知识融合引擎. 当模型生成治疗建议时,该引擎通过知识图谱嵌入(Knowledge Graph Embedding)[41]动态验证推荐逻辑,确保决策路径符合临床证据链.

多模态处理能力的提升主要源于联合语义空间构建的数学创新:视觉 Transformer 编码器提取医学影像的层级化特征,通过对比学习损失函数与文本编码器的语义表征进行跨模态对齐,其技术本质是优化影像-文本对的余弦相似度度量,公式化表达为式(2):

$$s(v,t)=\frac{\phi_v(v)\cdot\phi_t(t)}{\|\phi_v(v)\|\ \|\phi_t(t)\|} \tag{2}$$

2 LLMs 驱动的可解释医疗推荐方法分类

LLMs 在医疗场景中依据解释范式与知识整合机制可分为判别式可解释方法、生成式可解释方法与混合式可解释方法 3 类,其分类本质源于不同技术路径对临床需求的差异化适配. 相关分类如图 2 所示.

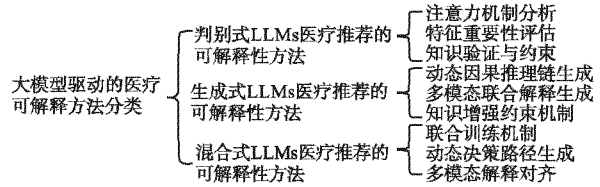

图 2 大模型驱动的可解释医疗推荐方法分类

Fig. 2 Taxonomy of LLMs-Driven explainable medical recommendation methods

2.1 判别式 LLMs 医疗推荐的可解释性方法

判别式模型直接学习输入特征 X 到输出标签 Y 的映射关系(条件概率 P(Y|X)),聚焦决策边界而非数据生成过程. 典型如逻辑回归、支持向量机(Support Vector Machine,SVM),适用于高时效分类任务[42]. 判别式大语言模型(Discriminative Large Language Models,D-LLMs)具有基于大语言模型编码能力构建的高级判别式架构,采用 Transformer 编码器(如 BERT),禁用文本生成功能,输出限定为分类/回归结果. 判别式大语言模型的独立分类源于其在高时效性医疗场景中的实时决策可验证性与单模态解释适配性优势. 判别式大语言模型的可解释方法依赖于注意力机制分析、特征重要性评估与知识图谱验证 3 种重要手段. 本节系统阐述这一类的技术框架及典型方法案例.

2.1.1 注意力机制分析

基于 Transformer 编码器架构,模型通过自注意力权重动态捕捉输入特征间的语义关联,其数学本质是通过构建查询、键、值三组向量的动态交互,量化输入特征间的语义关联强度. 这种注意力驱动的解释框架,其核心价值在于通过动态语义拓扑揭示特征间的潜在病理关联,而非孤立评估单指标重要性. 例如,CLIN-X[43]模型通过分层注意力权重可视化,定位多模态数据中的关键决策依据,显著提升医生对医疗推荐的可信度. Huang 等人[44]利用交叉注意力矩阵量化桥本氏甲状腺炎 CT 影像与文本报告的关联,实现较高的病理可追溯性;PathoExplainer[45]基于多头注意力熵值分析,揭示胃癌组织芯片中的动态关联;Causal Transformer[46]则将注意力权重转化为因果效应估计,通过前扣带回皮质厚度介导的中介效应分析,验证其对抗抑郁药治疗响应的因果预测机制.

2.1.2 特征重要性评估

特征重要性评估引入模型无关解释方法(如 SHAP、LIME)量化局部特征贡献度. SHAP(SHapley Additive Explanations)[47]基于博弈论分配特征贡献值,其数学本质是通过构造可解释代理模型逼近黑箱系统的局部决策边界. 该方法通过计算所有特征组合的边际效应,精确量化每个特征的全局贡献度. LIME(Local Interpretable Model-agnostic Explanations)[48]则通过局部线性逼近构建可解释代理模型,这两种方法共同构建了从全局特征重要性到局部决策规则的立体解释体系. SHAP 的博弈论基础确保了特征贡献分配的公平性与可加性,而 LIME 的微分几何视角则捕捉了决策边界的局部曲率特性. 尽管当前方法已在临床决策一致性验证中展现出一定的价值,但在时序数据与多模态场景的扩展仍需发展动态 Shapley 值计算框架与多尺度 LIME 优化算法,以实现全维度临床认知的可解释性映射.

2.1.3 知识验证与约束

知识验证与约束通过医学知识库(如 SNOMED CT、DrugBank)[49]的规则映射,确保模型输出符合临床指南. 动态知识图谱映射将模型预测与 SNOMED CT、DrugBank 等知识库节点进行语义对齐,知识验证(Knowledge Verification)[50]与约束机制(Constraint Enforcement)[51]是确保模型输出符合医学知识体系的核心技术,其本质是通过结构化先验知识对生成过程施加逻辑边界,从而规避因数据偏差或模型幻觉引发的临床风险. 例如,Cuesta[52]等人在肿瘤治疗推荐系统中整合 DrugBank 药物相互作用知识,通过子图匹配算法拦截不良反应组合,显著降低了不合理推荐率;MedGuard[53]基于 SNOMED CT 构建诊断逻辑约束树,在急诊胸痛分诊中阻断相关违规路径;Schreier 等人[54]利用 DrugBank 药代动力学图谱,实时校验肾衰患者抗生素剂量方案,显著降低剂量错误率.

2.1.4 判别式 LLMs 医疗推荐的局限性及未来优化方向

判别式 LLMs 虽然能够在医疗推荐任务中快速生成基于注意力权重与特征重要性评分的可解释结果,但其动态因果逻辑推演能力缺失的局限性显著制约临床深度决策支持能力. 现有框架依赖统计相关性分析,难以区分混杂因素与真实因果效应,导致解释的临床可干预性受限;同时,多模态数据的独立归因机制割裂了跨模态协同效应,致使医生对推荐逻辑的信任度下降. 此外,静态知识图谱更新滞后无法适配指南动态迭代,造成模型推荐与最新循证依据的偏离风险. 为解决上述问题,未来研究应探索融合因果推断与多模态融合技术突破瓶颈:一方面,通过结构方程模型(Structural Equation Modeling,SEM)[55]解耦直接/间接效应路径,构建联邦注意力网络解析跨模态交互权重,另一方面,开发 NLP 驱动的指南监测系统实现知识库在平衡点内实时同步.

2.2 生成式 LLMs 医疗推荐的可解释性方法

生成式大语言模型(Generative Large Language Models,

G-LLMs)通过端到端的自然语言生成能力,为医疗推荐系统提供了更加动态化、多模态与个性化的可解释性支持.生成式大语言模型的可解释方法依赖于动态因果推理链生成、多模态联合解释生成与知识增强约束机制 3 种重要手段.三大技术均深度依赖 G-LLMs 的生成特性:动态因果链需文本序列生成模拟临床推理,多模态联合解释需跨模态内容生成融合证据,知识约束需条件生成动态植入规则.

2.2.1 动态因果推理链生成

基于思维链提示技术[56]模拟临床决策路径,直接生成符合医学逻辑的解释文本.其数学本质是通过自回归生成过程(Autoregressive Generation Process)[57]将黑箱推理分解为可验证的因果逻辑序列.具体而言,给定输入特征向量 $\mathbf{X}=\{x_1,x_2,\cdots,x_n\}$,模型通过条件概率链式法则(3)生成分步推理文本 $\mathbf{Y}=\{y_1,y_2,\cdots,y_m\}$:

$$P(Y\mid X)=\prod_{t=1}^{m}P(y_t\mid X,y_{1:t-1}) \tag{3}$$

其中,t 为推理步时间戳,m 为推理链总长度.该过程通过隐式马尔可夫决策过程(Implicit Markov Decision Process,IMDP)[58]模拟临床思维路径,其中每个推理步骤 y_t 对应医学逻辑单元.为提升解释的临床合规性,采用医疗指令微调(Medical Instruction Fine-tuning)策略[59],基于结构化临床指南构建因果链标注数据集 $D=\{(X_i,Y_i^{COT})\}_{i=1}^{N}$,通过最大化对数似然函数(4):

$$L_{COT}=\sum_{i=1}^{N}\sum_{t=1}^{m}\log_P(y_{i,t}\mid x_i,y_{i,1:t-1}) \tag{4}$$

优化模型参数,同时引入逻辑约束损失(Logical Constraint Loss)[60],该机制使模型在生成解释时,严格遵循证据等级的循证医学规范.例如,Miao 等人[61]的实验表明,该方法在临床逻辑一致性评分上较传统方法有着显著提升,且在零样本场景下保持较高的指南依从率,验证了动态因果推理链对医疗解释可审计性的强化作用;MedChain[62]通过分层思维链,从症状检查到诊断和治疗一系列步骤,生成急性腹痛解释文本,显著降低误诊率.

2.2.2 多模态联合解释生成

多模态联合解释生成的目标是构建一个基于跨模态语义对齐(Cross-modal Semantic Alignment)[63]的联合解释生成框架,其数学原理是通过对比学习(Contrastive Learning)[64]实现异质模态在共享 Hilbert 空间中的几何同构映射.例如,Li 等人[65]采用 CLIP(Contrastive Language-Image Pre-training)架构[66]的改进变体,通过双塔编码器分别提取病理影像特征 $V\in R^{d_V}$ 与文本描述特征 $T\in R^{d_t}$,并优化跨模态相似度度量函数(5):

$$L_{align}=-\frac{1}{N}\sum_{i=1}^{N}\log\frac{exp(V_i^TT^i/\tau)}{\sum_{j=1}^{N}exp(V_i^TT_j/\tau)} \tag{5}$$

其中,τ 为温度超参数,用于控制特征分布锐度.该机制使模型能够将 HER2 免疫组化(IHC)染色图像中的"膜强阳性"模式与循证医学文本在共享语义空间中对齐,生成具有空间可解释性的热力图标注;CAMAF[67]是一种用于可解释肺病风险分层的情境感知多模式框架,采用 TabTransforme 和 Vision Transformer 进行高保真特征提取,并利用 Transformer 编码器增强多模态融合,确保卓越的预测性能;Wu 等人[68]提出的交叉对齐多模态表示学习方法,引入了跨模态表示对齐学习网络,通过在公共子空间中有效地学习模态不变表示来减少模态差距,为癌症生存预测提供了多模态数据的整体视图.

2.2.3 知识增强约束机制

知识增强约束机制通过检索增强生成与反事实推理引擎(Counterfactual Reasoning Engine)的协同架构,共同构建了一套面向医疗决策可解释性的双重验证体系.

检索增强生成(Retrieval-Augmented Generation,RAG)的核心原理在于突破传统生成模型的封闭知识边界,通过向量检索技术[69]实时调用权威医学知识库,建立生成内容与循证知识源的动态映射关系.例如,Xia[70]等人在 Mmed-RAG 架构研究中发现,该机制显著降低药物推荐场景的幻觉率,其核心创新在于:首先,构建基于知识图谱的语义索引层,将非结构化医学文本编码为可检索的向量空间;其次,通过最大边界损失函数[71]优化生成内容与检索证据的语义对齐,确保推荐方案严格遵循知识库中的适应证-禁忌证逻辑约束.

反事实推理引擎则从因果推断视角出发,重构可解释性框架,其方法论基础是构建医疗决策的结构因果模型(Structural Causal Model,SCM)[72].该引擎通过干预算子(do-calculus)[73]生成假设性治疗路径,并借助贝叶斯网络计算不同干预策略的预期效果差异.这种解释机制的核心目标是通过潜在结果框架(Potential Outcomes Framework)[74]量化决策不确定性,其数学表达式(6)为:

$$E[Y\mid do(X=x)]-E[Y\mid do(X=x')] \tag{6}$$

其中 Y 为临床结局指标,X 与 x′代表不同干预策略.例如,Zhang[75]等人提出了一个去混淆表示学习框架,通过生成与治疗变量无关的协变量的表示来估计连续治疗的反事实结果,将反事实推理网络嵌入到框架中,使学习到的表示同时服务于去混淆和可信推理,在学习去混淆表示方面表现出色.

2.2.4 生成式 LLMs 医疗推荐的局限性及未来优化方向

G-LLMs 虽然能够在医疗推荐任务中生成自然语言解释与多模态交互建议,但在临床应用的安全性与循证依从性方面仍存在显著局限.当前框架依赖概率驱动的自由生成模式,易产生医学幻觉或指南偏离,导致临床误用风险升高,且生成式解释缺乏对多模态数据的结构化整合能力,常以文本为中心割裂跨模态关联,导致推荐逻辑碎片化.同时,静态训练数据与动态医学知识间的更新鸿沟使模型难以及时适配最新指南,造成推荐与循证依据的时效性脱节.为此,未来研究需融合知识约束与因果推理技术突破瓶颈:通过嵌入医学知识图谱与规则引擎拦截违规建议,构建因果增强架构,并开发动态知识注入系统.结合证据权重采样策略抑制幻觉生成,G-LLMs 有望从"概率驱动生成"转向"循证、因果、合规"三位一体的医疗推荐范式,为疾病精准治疗、个体化用药等复杂场景提供安全、透明且可溯源的决策支持.

2.3 混合式 LLMs 医疗推荐的可解释性方法

混合范式(Hybrid Paradigm)通过融合判别式与生成式大语言模型的技术优势,构建"精准决策-动态解释"协同优化的医疗推荐框架.混合范式的可解释性生成遵循"判别-生成协同优化"原则,其核心包括基于 BERT 架构实现精准分类/回归的判别模块、基于 GPT 架构生成自然语言解释,并通过知识图谱约束内容合规性的生成模块;采用对比学习对齐文本、影像与时序数据的语义空间的跨模态对齐层.混合式

LLMs 的结构如图 3 所示. 混合范式的可解释性本质是“判别模块决策,生成模块解释”的分工体系,一般分为基于联合训练机制、动态决策路径生成与多模态解释对齐的技术. 三大子类别均以生成模块为可解释性引擎——联合训练依赖文本生成建立决策-解释关联,动态决策路径通过条件生成植入动态规则,多模态解释对齐需跨模态生成构建证据链. 判别模块仅提供决策输入,解释的动态性、合规性与多模态融合均由生成模块实现.

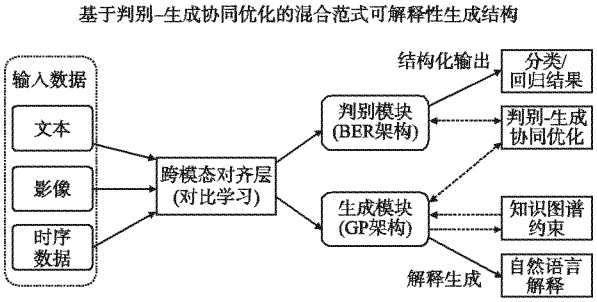

图 3 混合式 LLMs 的结构

Fig. 3 Architectures of hybrid large language models

2.3.1 联合训练机制

联合训练机制通过参数共享编码器与多任务协同优化框架,实现跨模态语义融合. 联合训练机制的核心在于判别模型与生成模型通过共享编码器进行深度协同:判别模型利用该表征执行精准决策,优化判别损失;生成模型同步基于同一表征生成动态解释,优化生成损失;知识约束模块作为生成模型的监督器,通过规则惩罚项实时校验解释合规性,其梯度惩罚直接作用于生成模型并间接影响共享编码器. 共享编码器将异构医疗数据映射至统一的高维表征空间,通过多目标损失函数同步优化判别任务、生成任务与知识图谱约束[76]. 判别任务聚焦于疾病分类与治疗方案推荐,采用交叉熵损失函数构建决策边界;生成任务驱动模型输出符合临床规范的自然语言解释,通过自回归语言建模捕捉语义连贯性,例如,Sun 等人[77]研发的 MedFusion-TransNet 框架采用分层 Transformer 编码器,其判别模块(Transformer 编码器)对齐病理文本与组织切片影像的潜在特征,精准分类肺癌亚型,同时生成模块基于同一特征空间动态输出可解释描述;二者通过共享编码器的联合梯度更新形成闭环优化——当判别结果与生成描述冲突时,梯度回传同步修正编码器参数. 知识图谱约束则通过图嵌入技术将医学实体关系编码为几何空间中的拓扑结构,确保模型决策与权威指南的语义一致性,例如,Han 等人[78]提出将 DrugBank 药物相互作用图谱嵌入生成过程,在化疗方案生成中实时拦截不良反应组合,若生成模块输出违规建议,则触发梯度惩罚并驱动判别模块重新评估方案合理性,最终使不合理化疗建议率下降. 三者的联合优化形成动态平衡,既提升判别精度,又强化生成解释与医学知识体系的逻辑对齐.

2.3.2 动态决策路径生成

动态决策路径生成方法利用判别模块输出触发生成式推理链:判别模块首先提取关键临床特征,随后生成模块通过可微分逻辑层将离散的临床规则转化为连续空间的操作. 这一过程通过张量运算模拟因果链展开,例如,Kim 等人[79]提出的动态路径决策如表 2 所示.

表 2 动态决策路径

Table 2 Dynamic decision pathways

1	HER2 3 + → 靶向治疗适应症(NCCN 指南 Breast-3)
2	PD-L1 CPS≥20 → 联合帕博利珠单抗(KEYNOTE-522 试验 pCR 率 64.8%)
3	心脏射血分数≥50% → 耐受性达标

判别模块提取关键临床特征作为生成起点,生成模块通过可微分逻辑层将离散临床规则转化为连续张量运算,动态构建因果路径;当检测到决策冲突时,系统立即触发反事实干预机制——生成模块基于特征干预条件概率重计算,实时切换路径至合规方案. 该架构通过判别输出锚定生成起点、生成路径验证判别逻辑的闭环交互,实现指南规则与数据驱动的深度耦合,同时确保决策全程因果可追溯.

2.3.3 多模态解释对齐

多模态解释对齐作为混合式 LLMs 的核心枢纽,通过跨模态注意力机制与动态时序建模实现判别模块与生成模块的协同验证:其本质是构建异构模态在联合决策中的可审计证据链,该能力无法由判别式或生成式模型单独承载——判别式模型仅提取单模态特征,生成式模型仅输出文本序列,而混合架构需同步完成跨模态语义映射与联合推理验证. 视觉-文本对齐模块通过对比学习优化病理影像区域与文本描述之间的语义映射关系,使模型生成可验证解释. 例如,Taleb 等人[80]利用多模态对比损失函数对齐乳腺病理切片感兴趣区域(Region of Interest,ROI)与病理报告关键词,显著提升了影像-文本解释一致性准确率;时序数据处理[81]采用神经控制微分方程,以连续时间建模生命体征流数据的动态演变规律,捕捉瞬态事件对决策的因果贡献. 例如,Salomao 等人[82]通过条件密度估计(Conditional Density Estimation,CDE)捕捉 ICU 患者血压波动的瞬态事件对脓毒症预测的因果贡献,降低了预测误差;跨模态注意力权重矩阵进一步揭示不同模态特征在联合推理中的相对重要性,例如影像特征对肿瘤分型的贡献度与文本特征对治疗方案选择的指导权重. 这种几何空间的对齐机制不仅提升了解释的临床可接受性,更为多模态医疗 AI 的可信决策提供了可审计的证据链.

2.3.4 混合式 LLMs 医疗推荐的局限性及未来优化方向

尽管混合式大语言模型在医疗推荐中融合了生成式与判别式模型的优势,但其异构架构协同性不足与动态知识整合滞后等问题仍显著影响临床应用的可靠性与安全性. 现有框架中,生成模块与判别模块的参数耦合效率低下,容易引发逻辑冲突,而多模态数据的异步解释对齐缺陷进一步加剧决策依据的碎片化. 此外,混合架构对动态医学知识的增量适应能力不足,因模块间知识蒸馏速率差异导致推荐偏离最新循证依据. 为此,未来优化需首先通过异构架构动态路由与多模态时序对齐技术提升系统实时响应能力,同时基于增量知识协同更新机制保障指南一致性;进而嵌入因果约束联合训练框架,确保生成与判别输出的病理机制可溯性;由此融合动态协同推理、时序对齐优化与因果归因技术,推动混合式大语言模型突破“机械拼合”范式,最终构建机理透明、决策一致、实时合规的新一代医疗推荐体系,为急诊响应及肿瘤精准治疗提

供高置信决策支持.

3　未来方向

医疗推荐系统的可解释性正面临双重挑战:在数据层面,多中心协作需求与隐私保护要求之间的矛盾限制了知识共享的深度;在认知层面,传统解释方法对因果机制的忽视削弱了临床决策的可信度. 为突破这些瓶颈,未来研究需聚焦隐私安全的可解释性协同学习与因果增强的决策逻辑验证两大方向. 前者通过联邦解释学习框架(Federated Explainable AI, FedXAI)[83]实现分布式知识融合与解释一致性优化,后者则借助因果发现构建循证医学级可验证解释体系. 这些技术路径的核心在于将数据驱动模型的灵活性与医学知识体系的严谨性深度融合,从而推动医疗推荐系统从"统计关联拟合"向"因果机制建模"的范式跃迁,为构建安全、可信、可审计的系统奠定方法论基础.

3.1　联邦解释学习

在医疗多中心协作场景中,隐私保护与解释一致性矛盾源于分布式数据的异构性和敏感信息的不可共享性. 联邦可解释学习框架通过以下技术路径实现隐私与解释的协同优化.

3.1.1　隐私保护解释共享机制

联邦可解释学习框架通过 LLMs 重构隐私保护解释共享机制:联邦可解释学习框架通过解释元特征抽象、加密知识聚合与差分隐私加固三重核心技术,实现医疗协同学习中隐私保护与决策可解释性的统一平衡. 例如,Phuong 等人[84]基于知识蒸馏(Knowledge Distillation)将本地复杂模型的决策逻辑压缩为低维语义向量,剥离患者敏感信息;Regueiro 等人[85]利用同态加密(Homomorphic Encryption Aggregation)对分布式解释特征执行密文聚合,中心服务器采用同态加密聚合对分布式 LLMs 输出的解释特征执行密文计算,实现原始数据零接触;Zhang 等人[86]提出注入差分隐私噪声阻断模型逆向推理,形成语义抽象-加密计算-噪声防御的闭环机制. 该框架在甲状腺癌多中心诊断等场景中,既提炼出微钙化灶密度等共性解释因子,又满足医疗数据全链路保密要求,为可解释性协同学习提供了严格的形式化验证基础. 未来研究可聚焦 LLMs 解释语义强化、联邦提示工程、生成式隐私防御及可验证推理链,推动 LLMs 从解释生成工具升级为隐私-解释协同引擎,通过语义抽象能力解耦敏感信息与临床知识.

3.1.2　动态解释一致性优化

联邦可解释学习框架通过 LLMs 重构跨机构解释一致性机制:联邦可解释学习框架通过注意力分布对齐与动态权重优化的协同机制,保障跨机构解释模型的语义一致性. 首先基于信息论度量筛选决策共识节点,抑制数据偏移导致的解释偏差;进而通过时间衰减函数动态调节节点贡献度,降低低质量数据影响;最终将历史行为可信度量化为贝叶斯纳什均衡权重,提升系统稳定性. 例如,Fuglede 等人[87]采用 Jensen-Shannon 散度,在阿尔茨海默病 MRI 分析中使海马体萎缩模式显著提升共识度;Zhang 等人[88]提出滑动窗口衰减函数,成功抑制标注噪声节点权重;Tian 等人[89]构建信誉激励机制,驱动高数据质量节点获得额外决策权.

3.2　因果增强的可解释性

传统解释方法的本质缺陷在于其仅刻画变量间的统计相关性,而医疗决策需严格依赖因果机制,LLMs 通过因果结构学习与反事实推理实现机制性可解释. 高维因果结构学习的核心在于从电子健康记录(Electronic Health Records,EHR)的复杂数据拓扑中提取可验证的因果依赖关系. 例如,针对 EHR 的高维性、稀疏性及时序特性,Scutari 等人[90]采用约束型结构学习算法(Constraint-based Structure Learning),结合临床指南的因果关系先验构建因果图;Lv 等人[91]通过 PC 算法(Peter-Clark Algorithm)执行条件独立性测试,在控制混杂变量后,利用卡方检验[92]逐步剔除无因果关联的边;Strobl 等人[93]基于 FCI 算法(Fast Causal Inference)识别潜在混杂因子,通过隐变量概率图建模修正因果效应估计,避免因未观测变量导致的偏倚. 未来可构建联邦因果 LLMs 实现跨中心因果图协作学习,微调 LLMs 生成因果假设引导约束型结构学习,并联合优化生成解释损失与因果发现损失.

4　总　结

本文系统梳理了 LLMs 驱动的医疗推荐可解释方法分类,并从多个角度进行了分析和总结. 首先,本文对传统的可解释方法分类在医疗领域的应用进行了回顾,包括因果性解释、对比性解释与合规性解释. 其次,本文介绍了 LLMs 的发展与核心技术特征,分析了 LLMs 在医疗可解释领域的优势,在此基础上,提出了三类 LLMs 驱动的可解释医疗方法分类,分别为判别式 LLMs 医疗推荐的可解释性方法、生成式 LLMs 医疗推荐的可解释性方法以及混合式 LLMs 医疗推荐的可解释性方法,并对各个分类方法的技术框架与典型应用进行了详细介绍,同时指出了这些方法当前存在的挑战与解决方案. 最后,本文从联邦解释学习和因果增强的可解释性两个方向出发,对医疗推荐可解释的未来进行展望,提出对现存问题的解决方向.

References:

[1] Tran T N T, Felfernig A, Trattner C, et al. Recommender systems in the healthcare domain: state-of-the-art and research issues[J]. Journal of Intelligent Information Systems, 2021, 57(1): 171-201.

[2] Goyal V A, Parmar D J, Joshi N I, et al. Medicine recommendation system[J]. Medicine(Baltimore), 2020, 7(3): 1658-1662.

[3] Baillie J K, Angus D, Burnham K, et al. Causal inference can lead us to modifiable mechanisms and informative archetypes in sepsis [J]. Intensive Care Medicine, 2024, 50(12): 1-12.

[4] WANG H R, WANG Y M, ZHOU B J, et al. Knowledge-aware recommendation method with multimodal information fusion[J]. Journal of Zhengzhou University(Engineering Science), 2025, 46(6): 1-9.

[5] Panagiotis S, Stergios C, Markus Z. Safe, effective and explainable drug recommendation based on medical data integration[J]. User Modeling and User-Adapted Interaction, 2022, 32(5): 999-1018.

[6] Chetana V L, Seetha H. Genome data-based explainable recommender systems: a state-of-the-art survey[J]. Genomics at the Nexus of AI, Computer Vision, and Machine Learning, 2025: 149-168, doi: 10.1002/9781394268832.ch7.

[7] Phillips S J, Dudík M, Elith J, et al. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data [J]. Ecological Applications, 2009, 19(1): 181-197.

[8] Certain Choices W G F, Simianu V V, Grounds M A, et al. Understanding clinical and non-clinical decisions under uncertainty: a scenario-based survey [J]. BMC Medical Informatics and Decision Making, 2016, 16(1): 1-9.

[9] Singh Y, Andersen J B, Hathaway Q, et al. Deep learning-based uncertainty quantification for quality assurance in hepatobiliary imaging-based techniques [J]. Oncotarget, 2025, 16: 249-255, doi: 10.18632/oncotarget.28709.

[10] Liznerski P, Ruff L, Vandermeulen R A, et al. Explainable deep one-class classification [J]. arXiv preprint arXiv: 2007.01760, 2020, doi. org/10.48550/arXiv.2007.01760.

[11] Pawlowski N, Coelho De Castro D, Glocker B. Deep structural causal models for tractable counterfactual inference [J]. Advances in Neural Information Processing Systems, 2020, 33(73): 857-869, doi: 10.5555/3737916.3742149.

[12] Stage F K, Carter H C, Nora A. Path analysis: an introduction and analysis of a decade of research [J]. The Journal of Educational Research, 2004, 98(1): 5-13.

[13] Sigismund S, Avanzato D, Lanzetti L. Emerging functions of the EGFR in cancer [J]. Molecular Oncology, 2018, 12(1): 3-20.

[14] Diaby V, Campbell K, Goeree R. Multi-criteria decision analysis (MCDA) in health care: a bibliometric analysis [J]. Operations Research for Health Care, 2013, 2(1-2): 20-24.

[15] Menard S, Pupa S M, Campiglio M, et al. Biologic and therapeutic role of HER2 in cancer [J]. Oncogene, 2003, 22(42): 6570-6578.

[16] Geng C, Li Y, Li L, et al. Optimized temporal interference stimulation based on convex optimization: a computational study [J]. IEEE Transactions on Neural Systems and Rehabilitation Engineering: a Publication of the IEEE Engineering in Medicine and Biology Society, 2025, doi. 10.1109/TNSRE.2025.3558306.

[17] Li L, Wang P, Yan J, et al. Real-world data medical knowledge graph: construction and applications [J]. Artificial Intelligence in Medicine, 2020, 103: 101817, doi. 10.1007/s10844-020-00633-6.

[18] Hu Z, Martí J. Unraveling atomic-scale mechanisms of GDP extraction catalyzed by SOS1 in KRAS-G12 and KRAS-D12 oncogenes [J]. Computers in Biology and Medicine, 2025, 186(29): 109599-109599.

[19] Tang R, Chuang Y N, Hu X. The science of detecting LLM-generated text [J]. Communications of the ACM, 2024, 67(4): 50-59.

[20] Wu S, Fei H, Qu L, et al. Next-GPT: any-to-any multimodal LLM [C]//41st International Conference on Machine Learning, 2024: 53366-53397.

[21] Shao Y, Li L, Dai J, et al. Character-LLM: a trainable agent for role-playing [J]. arXiv preprint arXiv: 2310.10158, 2023.

[22] Wang Z, Chu Z, Doan T V, et al. History, development, and principles of large language models: an introductory survey [J]. AI and Ethics, 2024: 1-17, doi: 10.1007/s43681-024-00583-7.

[23] Chen X, Yan B, Zhu J, et al. Transformer tracking [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 8126-8135.

[24] Ghojogh B, Ghodsi A. Attention mechanism, transformers, BERT, and GPT: tutorial and survey [J]. HAL (Home-Archive ouverte), 2020, doi. 10.1186/s12911-025-02954-4.

[25] Koroteev M V. BERT: a review of applications in natural language processing and understanding [J]. arXiv preprint arXiv: 2103.11943, 2021.

[26] Bharathi Mohan G, Prasanna Kumar R, Parathasarathy S, et al. Text summarization for big data analytics: a comprehensive review of GPT2 and BERT approaches [J]. Data Analytics for Internet of Things Infrastructure, 2023: 247-264, doi: 10.1007/978-3-031-33808-3_14.

[27] Zhang M, Li J. A commentary of GPT-3 in MIT technology review 2021 [J]. Fundamental Research, 2021, 1(6): 831-833.

[28] Abeysinghe S, Xhebraj A, Rompf T. Flan: an expressive and efficient datalog compiler for program analysis [J]. Proceedings of the ACM on Programming Languages, 2024, 8(86): 2577-2609.

[29] Buruk O O. Academic writing with GPT-3.5: reflections on practices, efficacy and transparency [J]. arXiv preprint arXiv: 2304.11079, 2023, doi. org/10.1145/3616961.3616992.

[30] Kalla D, Smith N, Samaah F, et al. Study and analysis of chat GPT and its impact on different fields of study [J]. International Journal of Innovative Science and Research Technology, 2023, 8(3): 827-833.

[31] Vakayil S, Juliet D S, Vakayil S. RAG-Based LLM chatbot using llama-2 [C]//7th International Conference on Devices, Circuits and Systems (ICDCS), 2024: 1-5.

[32] Achiam J, Adler S, Agarwal S, et al. GPT-4 technical report [J]. arXiv preprint arXiv: 2303.08774, 2023, doi. org/10.48550/arXiv.2303.08774.

[33] Guo D, Yang D, Zhang H, et al. Deepseek-r1: incentivizing reasoning capability in LLMs via reinforcement learning [J]. arXiv preprint arXiv: 2501.12948, 2025, doi. org/10.48550/arXiv.2501.12948.

[34] Min E, Chen R, Bian Y, et al. Transformer for graphs: an overview from architecture perspective [J]. arXiv preprint arXiv: 2202.08455, 2022, doi. org/10.48550/arXiv.2202.08455.

[35] Sherstinsky A. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network [J]. Physica D: Nonlinear Phenomena, 2020, 404: 132306, doi: 10.1016/j.physd.2019.132306.

[36] Shaw P, Uszkoreit J, Vaswani A. Self-attention with relative position representations [J]. arXiv preprint arXiv: 1803.02155, 2018, doi. org/10.48550/arXiv.1803.02155.

[37] Zheng J, Ramasinghe S, Lucey S. Rethinking positional encoding [J]. arXiv preprint arXiv: 2107.02561, 2021, doi. org/10.48550/arXiv.2107.02561.

[38] Cordonnier J B, Loukas A, Jaggi M. Multi-head attention: collaborate instead of concatenate [J]. arXiv preprint arXiv: 2006.16362, 2020, doi. org/10.48550/arXiv.2006.16362.

[39] Shen W, Li Y, Liu Y, et al. Entity linking meets deep learning: techniques and solutions [J]. IEEE Transactions on Knowledge and Data Engineering, 2021, 35(3): 2556-2578.

[40] He Q, Li X, Cai B. Graph neural network recommendation algorithm based on improved dual tower model [J]. Scientific Reports, 2024, 14(1): 3853, doi. org/10.1038/s41598-024-54376-3.

[41] Wang Q, Mao Z, Wang B, et al. Knowledge graph embedding: a

survey of approaches and applications[J]. IEEE Transactions on Knowledge and Data Engineering,2017,29(12):2724-2743.

[42] SUN Z W,SONG M Y,PAN Z H,et al. Context-aware discriminative topic model[J]. Journal of Shandong University(Engineering Science),2022,52(4),131-138+150.

[43] Lange L,Adel H,Strötgen J,et al. CLIN-X:pre-trained language models and a study on cross-task transfer for concept extraction in the clinical domain[J]. Bioinformatics,2022,38(12):3267-3274.

[44] Huang Q,Jiang W,Li J,et al. Hashimoto's thyroiditis recognition from multi-modal data via global cross-attention and distance-aware training[J]. Medical Image Analysis,2025,102:103515-103515,doi:10.1016/j.media.2025.103515.

[45] Dalton J C,Crowell K A,Ntowe K W,et al. Utility of axillary staging in older patients with her2-positive breast cancer[J]. Annals of Surgical Oncology,2024,31(11):1-13.

[46] Zhu Y,Yang F,Torgashov A. Causal-transformer:spatial-temporal causal attention-based transformer for time series prediction[J]. IFAC PapersOnLine,2024,58(14):79-84.

[47] Van Den Broeck G,Lykov A,Schleich M,et al. On the tractability of SHAP explanations[J]. Journal of Artificial Intelligence Research,2022,74:851-886,doi:10.1613/jair.1.13283.

[48] Garreau D,Luxburg U. Explaining the explainer:a first theoretical analysis of LIME[C]//International Conference on Artificial Intelligence and Statistics,2020:1287-1296.

[49] JIA Y L,WANG Y T,HOU X F. Research and application of medical knowledge base construction[J]. Big Data Time,2024,(11):69-72.

[50] Wang X,Ban T,Chen L,et al. Knowledge verification from data[J]. IEEE Transactions on Neural Networks and Learning Systems,2022,35(3):4324-4338.

[51] Bauchau O A,Laulusa A. Review of contemporary approaches for constraint enforcement in multibody systems[J]. Journal of Computational & Nonlinear Dynamics,2008,doi. org/10.1115/1.2803258.

[52] Cuesta S A,Mora J R,Márquez E A. In silico screening of the DrugBank database to search for possible drugs against SARS-CoV-2[J]. Molecules,2021,26(4):1100,doi. org/10.3390/molecules26041100.

[53] Nasrin P T,Mohaimenul I M,Jack L Y. Clinical usefulness of drug-disease interaction alerts from a clinical decision support system,medguard,for patient safety:a single center study[J]. Studies in Health Technology and Informatics,2022,290:326-329,doi:10.3233/SHTI220089.

[54] Schreier T,Frick M T,Böhm R. Integration of FAERS,drugbank and sider data for machine learning-based detection of adverse drug reactions[J]. Datenbank-Spektrum,2024:1-10,doi:10.1007/s13222-024-00486-1.

[55] Mueller R O,Hancock G R. Structural equation modeling,the reviewer's guide to quantitative methods in the social sciences:Routledge,2018:445-456.

[56] ZHENG M Q,CHEN X H,LIU B,et al. A survey of chain-of-thought generation and enhancement methods in prompt learning[J]. Computer Science,2025,52(1):56-64.

[57] Hang T,Bao J,Wei F,et al. Fast autoregressive models for continuous latent generation[J]. arXiv preprint arXiv:2504.18391,2025,doi. org/10.48550/arXiv.2504.18391.

[58] Kallenberg L. Markov decision processes[J]. Lecture Notes,University of Leiden,2011,428,doi. org/10.1002/9781118557426.ch1.

[59] Peng C,Zhang K,Lyu M,et al. Scaling up biomedical vision-language models:fine-tuning,instruction tuning,and multi-modal learning[J]. arXiv preprint arXiv:2505.17436,2025,doi. org/10.48550/arXiv.2505.17436.

[60] Giunchiglia E,Stoian M C,Lukasiewicz T. Deep learning with logical constraints[J]. arXiv preprint arXiv:2205.00523,2022,doi. org/10.24963/ijcai.2022/767.

[61] Miao J,Thongprayoon C,Suppadungsuk S,et al. Chain of thought utilization in large language models and application in nephrology[J]. Medicina,2024,60(1):148,doi:10.3990/medicina60010148.

[62] Shen B,Guo J,Yang Y. MedChain:efficient healthcare data sharing via blockchain[J]. Applied Sciences,2019,9(6),doi:10.3390/app9061207.

[63] Cheng Q,Tan Z,Wen K,et al. Semantic pre-alignment and ranking learning with unified framework for cross-modal retrieval[J]. IEEE Transactions on Circuits and Systems for Video Technology,2022,doi.10.1109/TCSVT.2022.3182549.

[64] Khosla P,Teterwak P,Wang C,et al. Supervised contrastive learning[J]. Advances in Neural Information Processing Systems,2020,33:18661-18673,doi:10.48550/arXiv.2004.11362.

[65] Li Y,Liang F,Zhao L,et al. Supervision exists everywhere:a data efficient contrastive language-image pre-training paradigm[J]. arXiv preprint arXiv:2110.05208,2021,doi. org/10.48550/arXiv.2110.05208.

[66] Tu W,Deng W,Gedeon T. Toward a holistic evaluation of robustness in CLIP models[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2025,10.1109/TPAMI.2025.3580234.

[67] Dusari S R,Challa N P. CAMAF:context-aware multimodal alignment framework for explainable lung disease risk stratification[J]. Expert Systems with Applications,2025,279:127398-127398,doi:10.1016/j.eswa.2025.127398.

[68] Xingqi W,Yi S,Minghui W,et al. CAMR:cross-aligned multimodal representation learning for cancer survival prediction[J]. Bioinformatics(Oxford,England),2023,39(1),doi:10.1093/bioinformatics/btad025.

[69] Billhardt H,Borrajo D,Maojo V. A context vector model for information retrieval[J]. Journal of the American Society for Information Science and Technology,2002,53(3):236-249.

[70] Xia P,Zhu K,Li H,et al. Mmed-rag:versatile multimodal rag system for medical vision language models[J]. arXiv preprint arXiv:2410.13085,2024,doi. org/10.48550/arXiv.2410.13085.

[71] Li B,Yang B,Liu C,et al. Beyond max-margin:class margin equilibrium for few-shot object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,2021:7363-7372.

[72] Arif S,Macneil M A. Applying the structural causal model framework for observational causal inference in ecology[J]. Ecological Monographs,2023,93(1):e1554,doi:10.1002/ecm.1554.

[73] Lu Y,Zheng Q,Quinn D. Introducing causal inference using bayesian networks and do-calculus[J]. Journal of Statistics and Data Science Education,2023,31(1):3-17.

[74] Huang Y,Gilbert P B,Janes H. Assessing treatment-selection mark-

ers using a potential outcomes framework[J]. Biometrics,2012,68(3):687-696.

[75] Zhang G, Yuan G, Cheng D, et al. Deconfounding representation learning for mitigating latent confounding effects in recommendation[J]. Knowledge and Information Systems,2025,67(7):1-22.

[76] Li M, Sun Z, Zhang S, et al. Enhancing knowledge graph embedding with relational constraints[J]. Neurocomputing, 2021, 429: 77-88, doi:10.1109/ICBK50248.2020.00015.

[77] Sun J. MedFusion-TransNet: multi-modal fusion via transformer for enhanced medical image segmentation[J]. Frontiers in Medicine, 2025, 12:1557449, doi:10.3389/fmed.2025.1557449.

[78] Han X, Tian Y. Storage and query of drug knowledge graphs using distributed graph databases: a case study [J]. Bioengineering, 2025, 12(2):115, doi:10.3390/bioengineering/2020115.

[79] Kim Y, Park C, Jeong H, et al. Mdagents: an adaptive collaboration of llms for medical decision-making[J]. Advances in Neural Information Processing Systems, 2024, 37: 79410-79452, doi: 1052202/079017-2522.

[80] Taleb A, Kirchler M, Monti R, et al. Contig: self-supervised multimodal contrastive learning for medical imaging with genetics [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 20908-20921.

[81] Aydin S. Time series analysis and some applications in medical research[J]. Journal of Mathematics and Statistics Studies, 2022, 3(2):31-36.

[82] Salomão R, Ferreira B, Salomão M, et al. Sepsis: evolving concepts and challenges[J]. Brazilian Journal of Medical and Biological Research, 2019, 52: e8595, doi:10.1590/1414-431X20198595.

[83] Witt L, Heyer M, Toyoda K, et al. Decentral and incentivized federated learning frameworks: a systematic literature review[J]. IEEE Internet of Things Journal, 2022, 10(4):3642-3663.

[84] Phuong M, Lampert C. Towards understanding knowledge distillation[C]//International Conference on Machine Learning, 2019: 5142-5151.

[85] Regueiro C, Seco I, De Diego S, et al. Privacy-enhancing distributed protocol for data aggregation based on blockchain and homomorphic encryption[J]. Information Processing & Management, 2021, 58(6):102745, doi:10.1016/j.ipm.2021.102745.

[86] ZHANG S F, TANG B J, TIAN Z K, et al. A survey on federated learning with differential privacy[J]. Journal of Computer Applications, 2025, 45(10):3221-3230.

[87] Fuglede B, Topsoe F. Jensen-shannon divergence and Hilbert space embedding[C]//International Symposium Oninformation Theory, 2004:31, doi:10.1109/ISIT.2004.1365067.

[88] Zhang H R, Chen R, Wen S H, et al. SWIM: sliding-window model contrast for federated learning [J]. Future Generation Computer Systems, 2025, 164: 107590-107590, doi: 10.1016/j.future.2024.107590.

[89] WU J Y, YUAN L Y, CHEN M H, et al. Blockchain dynamic sharding model based on node credibility[J]. Application Research of Computers, 2024, 41(12):3563-5371.

[90] Scutari M. Bayesian network constraint-based structure learning algorithms: parallel and optimized implementations in the bnlearn R package[J]. Journal of Statistical Software, 2017, 77: 1-20, doi: 10.18637/jss.v077.i02.

[91] Lv F, Si S, Xiao X, et al. Modified local Granger causality analysis based on Peter-Clark algorithm for multivariate time series prediction on IoT data[J]. Computational Intelligence, 2024, 40(5): e12694, doi:10.1111/coin.12694.

[92] Pandis N. The chi-square test[J]. American Journal of Orthodontics and Dentofacial Orthopedics, 2016, 150(5):898-899.

[93] Strobl E V, Visweswaran S, Spirtes P L. Fast causal inference with non-random missingness by test-wise deletion [J]. International Journal of Data Science and Analytics, 2018, 6: 47-62, doi: 10.1007/s41060-017-0094-6.

附中文参考文献：

[4] 王海荣,王怡梦,周北京,等. 融合多模态信息的知识感知推荐方法[J]. 郑州大学学报(工学版),2025,46(6):1-9.

[42] 孙志巍,宋明阳,潘泽华,等. 上下文感知的判别式主题模型[J]. 山东大学学报(工学版),2022,52(4):131-138+150.

[49] 贾玉来,王英涛,侯晓锋. 医学知识库的构建研究与应用[J]. 大数据时代,2024,(11):69-72.

[56] 郑明琪,陈晓慧,刘 冰,等. 提示学习中思维链生成和增强方法综述[J]. 计算机科学,2025,52(1):56-64.

[86] 张淑芬,汤本建,田子坤,等. 基于差分隐私的联邦学习研究综述[J]. 计算机应用,2025,45(10):3221-3230.

[89] 吴加英,袁凌云,陈美宏,等. 基于节点可信度的区块链动态分片模型[J]. 计算机应用研究,2024,41(12):3563-5371.