计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
2026,35(1):102−116 [doi: 10.15888/j.cnki.csa.010054] [CSTR: 32024.14.csa.010054]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

大语言模型的原理及其在医疗领域的应用①

姚　琼, 王　增, 郭雨娇, 许　佳, 张梦娇, 石　锐

(四川大学华西医院 信息中心, 成都 610041)
通信作者: 石　锐, E-mail: shirui@wchscu.edu.cn

摘　要: 以 ChatGPT 为代表的大语言模型是当前人工智能领域最热门的研究课题, 被认为是推动传统行业实现革命性转型的关键技术手段, 为产业创新与升级提供了显著的驱动力. 医疗健康领域作为人工智能技术长期探索与应用的重点领域, 在当前面临着人口老龄化加剧、医疗资源供给不足以及医患关系紧张等背景下, 人工智能被视为最有希望缓解甚至彻底解决这一系列矛盾和问题, 尤其以 ChatGPT 为代表的大语言模型的出现, 让人们看到了曙光. 本文首先对自然语言处理技术的发展历程进行了简单介绍, 随后对 GPT 系列的大语言模型的历史发展背景及其技术演进轨迹进行了系统性的介绍. 结合医疗健康行业的现实需求与现状, 分类探讨了以 ChatGPT 为代表的大语言模型在该领域的应用场景和案例. 最后, 本文还深入分析讨论了大语言模型的内在局限, 以及在大规模部署实施和使用过程中所面临的挑战, 并针对性地给出了一些处理方法和解决思路.

关键词: ChatGPT; 大语言模型; 医疗应用; 人工智能; 自然语言处理; 医院信息化

引用格式: 姚琼,王增,郭雨娇,许佳,张梦娇,石锐.大语言模型的原理及其在医疗领域的应用.计算机系统应用,2026,35(1):102–116. http://www.c-s-a.org.cn/1003-3254/10054.html

Principle of Large Language Model and Its Applications in Healthcare

YAO Qiong, WANG Zeng, GUO Yu-Jiao, XU Jia, ZHANG Meng-Jiao, SHI Rui

(Information Center, West China Hospital of Sichuan University, Chengdu 610041, China)

Abstract: Large language models (LLMs) represented by ChatGPT are one of the most prominent research topics in artificial intelligence (AI) today. They are considered critical technological means for driving revolutionary transformations across traditional industries, providing substantial momentum for industrial innovation and upgrading. As a key domain of long-term exploration and application for AI technologies, the healthcare field is currently confronted with accelerated aging of the population, insufficient medical resource supply, and tense physician-patient relationships. Under this background, AI is regarded as the most promising solution to thoroughly solving these conflicts and problems, especially the LLMs represented by ChatGPT, which offer people a glimpse of hope. This study first briefly reviews the development of natural language processing (NLP) technologies, followed by a systematic introduction of the historical development background and technical evolution trajectory of the GPT-series LLMs. By combining the practical demands and current status of the healthcare industry, it discusses the application scenarios and cases of LLMs represented by ChatGPT in this field by category. Finally, this study conducts an in-depth analysis of the inherent limitations of LLMs and challenges encountered during the large-scale deployment, implementation, and utilization, with some targeted solutions and ideas provided.

Key words: ChatGPT; large language model (LLM); healthcare application; artificial intelligence; natural language processing (NLP); hospital informatization

① 基金项目: 四川省重大科技专项“揭榜挂帅”项目 (2024ZDZX0017)
收稿时间: 2025-06-18; 修改时间: 2025-07-25; 采用时间: 2025-08-15; csa 在线出版时间: 2025-11-11
CNKI 网络首发时间: 2025-11-12

1 引言

深度学习在最近十多年时间里快速发展, 在越来越多的行业领域中被广泛应用. 医疗领域就是深度学习应用研究最为活跃的行业之一, 在医学图像和信号处理、临床决策支持、数据挖掘和文本处理等方面取得了广泛应用, 学术研究和实践发现, 在很多应用场景下已经超过了人类专家的水平[1]. 当前深度学习领域的趋势, 是设计越来越复杂的神经网络结构, 以及规模越来越庞大的训练参数, 来增强模型的表现能力, 处理更复杂的应用场景. 大语言模型 (large language model, LLM) 是具有庞大参数规模的深度神经网络模型, 借助于高性能 GPU 集群的算力支撑, 以及海量的训练数据, 使得深度模型对人类自然语言有着更加深入的理解, 进而在文本生成、机器翻译、情感分析、对话系统等任务上有着更加卓越的表现.

虽然大语言模型相关的研究工作早在 2018 年就已经开始, 并且相比于传统机器学习和深度学习算法, 在各种自然语言处理任务中都取得比较明显的优势. 但是, 大语言模型的研究要以 OpenAI 公司在 2022 年 11 月推出的 ChatGPT (chat generative pre-trained transformer) 最具影响力, 该产品对于大语言模型的研究具有里程碑式的意义. ChatGPT 因为具有强大的自然语言处理能力和流畅的交互式对话功能, 以及在开放问答领域的惊艳表现, 在全球范围内迅速获得了广泛的关注, 这种爆发式的增长是 OpenAI 公司没预见到的. 许多专家和学者都预言以 ChatGPT 为代表的大语言模型将会对诸多行业领域产生重大影响, 推动产业升级和变革, 展现出极大的应用潜力和价值. 并且, 随着 ChatGPT 在全球的风靡, 更多的学者、科研机构和商业公司关注到大语言模型领域, 积极地投入资源进行大语言模型的相关研究.

医疗服务同日常生活息息相关, 与广大群众的生活质量密切相连. 在日常诊疗过程中, 涉及医患沟通、临床决策、病历书写和整理等很多环节, 涉及大量的交流沟通和文书处理工作, 占据了医务工作者大量的时间精力. 为了医疗过程的标准化和规范化, 保证医疗质量, 医院还需要对大量疾病案例进行分类、整理和上报. 在患者的整个诊疗过程中, 也会对各个治疗、用药的关键环节进行合规性校验和质控. 同时, 现代医学将医学院与医院紧密结合, 融合教学、研究与临床实践于一体的办学模式下, 科研与教学任务也是大部分医护工作者的重要工作组成部分. 以 ChatGPT 为代表的大语言模型相关的研究成果显著, 如果相关技术在医疗领域得到应用, 对于提高医务人员的工作效率、减轻医护人员的职业倦怠、改善患者就医体验、整体提升医疗服务质量和满意度, 都具有重要的意义.

本文首先介绍大语言模型的发展历程, 以及最具代表性的 ChatGPT 的架构和技术实现. 同时, 以 ChatGPT 为代表探讨当前大语言模型在医疗领域中可能的应用, 以及国内外目前正在探索和使用的场景及案例. 最后, 大语言模型作为一个新生事物, 伴随而来的是其自身还有一些缺陷和问题亟待解决[2], 在对新事物保持积极进取态度的同时, 我们也应该保持审慎和警惕, 正视并克服其种种弊端和不足, 让新技术能够更好、更安全地为人类健康生活服务.

2 ChatGPT 及大语言模型剖析

2.1 自然语言处理的发展历程

自然语言的处理, 最早可追溯到 20 世纪 60 年代, 当时尝试从句法规则的方向上对人类语言进行分析研究. 并且随着相关研究理论和经验的积累, 从 20 世纪 80 年代开始, 工业界开始构建出一系列的专家系统并将其应用在特定的行业领域, 比如医疗、法律等, 其应用场景也覆盖文档预处理、信息提取、结构化、命名实体识别等任务. 但是, 基于规则的自然语言处理主要依靠语言学家编写一系列的规则来进行任务处理, 在处理复杂任务的时候, 这种方式的表现和扩展能力以及通用性越来越受局限. 基于统计方法也是自然语言处理的另外一个重要分支, 它利用统计学方法和概率模型来处理和分析自然语言数据. 以 *n*-gram 语言模型为例, 基于对语言中词汇序列的统计信息进行建模, 使用该统计模型预测一个词序列中某个词的出现概率, 用于自动文本生成、拼写检查纠错等任务; 贝叶斯方法也被应用于自然语言处理中, 可以用于文本分类、情感分析、垃圾过滤等简单任务. 而 20 世纪末至 21 世纪初, 随着人们对机器学习算法的研究, 相关算法的应用也日趋完善, 诸如支持向量机、条件随机场、贝叶斯网络等经典算法使得基于统计方式的自然语言处理任务取得巨大成功, 在命名实体识别、词性标注、语义分析等任务上都取得很好的效果.

随着深度学习技术的快速发展, 深度学习技术在计算机视觉、语音等方面取得了巨大成功, 这让相关

学者期待深度学习在自然语言处理方面能够取得突破. 虽然, 早在 20 世纪 80 年代末, RNN 以及后续改进的 LSTM、GRU 模型就被提出来, 这些模型具有处理序列数据的能力, 因而可以用于自然语言处理, 但是受限于其处理效率不高, 并没有得到大规模的应用. 2013 年, Mikolov 等[3]提出了 Word2Vec, 该方法通过构建浅层神经网络将词映射到连续的向量空间, 捕捉词汇的语义关系, 从高维空间挖掘了语言词汇之间的信息, 这种词嵌入 (word embedding) 的方法后来被大量应用于推荐、检索等应用, 同时也为后续自然语言处理奠定了基础. 2014 年, Bahdanau 等[4]提出了注意力的机制, 虽然当时仅是用于改善机器翻译的效果, 但是作者提出的使模型能够聚焦于输入序列的不同部分的想法, 使得深度学习对长距离依赖的建模能力变为可能. 2017 年, Google 的 Vaswani 等[5]开创性地提出了 Transformer 模型, 被业界公认为是继 MLP、RNN、CNN 之后的第 4 大深度神经网络的基础模型, 该模型克服了 RNN 因为时序依赖性导致无法并行计算、模型训练推理性能慢的问题, 也解决了 CNN 因为卷积核的限制而不适合做长序列建模的难点. Transformer 是只依赖于注意力机制的序列转录模型, 通过使用多头的自注意力机制 (multi-head self-attention) 代替之前在编码解码器架构中经常用到的卷积层, 在机器翻译任务上取得了优异的效果. 同时, 由于该模型结构可以并行化计算, 与 GPU 擅长并行计算的架构完美契合, 使得模型训练和推理效率很高. 可以说, 正是由于 Attention 和 Transformer 的提出, 为后续大语言模型的相关研究提供了坚实的基础.

2.2 GPT 系列大语言模型的进化

2018 年 6 月, OpenAI 发布了 GPT-1[6], 该模型使用了 Transformer 模型中的解码器, 采用了半监督的训练机制, 即首先在大量不带标签的文本数据集上执行非监督类型的预训练生成一个语言模型, 然后再使用少量带标签的数据集, 根据具体的任务对模型进行微调, 其实这种思想当时在计算机视觉的深度学习任务上已经广泛被采用. 虽然模型的参数量为 1.17 亿, 在现在看来模型的规模还是比较“小”的, 但是在实验 12 个自然语言处理任务中, GPT-1 在其中的 9 个任务中取得了最好的结果. 这也表明, 通用语言表示可以采用生成式预训练 (generative pre-training) 的方式获得, 并在多种下游任务中通过微调实现高性能的两阶段训练范式的可行性.

2019 年 2 月, OpenAI 发布了 GPT-2[7], 表明语言模型在执行下游任务时可以使用零样本学习 (zero-shot learning) 设定, 使得预训练模型不再需要使用带标签的文本做特定任务的微调, 提高了预训练模型的通用性, 预训练模型可以直接适用于自然语言处理任务. 虽然架构上仍基于 Transformer 的解码器部分, 不过相比前代 GPT-1 要复杂很多, 使用经过清洗和处理后大约 800 万篇较高质量的网络文档作为训练语料, 模型的可学习参数达到了 15 亿量级. 而且, 从评测数据上观察到, 模型的性能会随着模型的参数规模稳定上升, 所以后续的迭代朝着更“大”规模参数量演进.

2020 年 6 月, GPT-3[8]模型发布, 其参数规模达到了 1 750 亿, 能够捕获更复杂的模式. 相较之前的模型, GPT-3 的训练数据来源于互联网内容和书籍文献, 而且数据规模也更为庞大. 由于模型体量巨大, 传统的模型微调方法已不再适用于特定自然语言处理任务. 然而, 通过提供给模型一些特定任务提示, 便可以显著提升模型完成样例类似的特定任务性能. 因此, 在 GPT-3 中引入了上下文学习 (in context-learning) 机制, 包含单样本学习 (one-shot learning)、少样本学习 (few-shot learning), 即在模型推理时给出一个或者少量演示任务作为样例, 但不会对预训练模型的结构和权重更新. 通过这种方式, GPT-3 模型在很多自然语言处理任务上表现出色, 甚至在某些任务上的效果可以媲美经过微调的模型.

OpenAI 在 2023 年 3 月发布了 GPT-4[9], 其具体结构、训练数据及方法未予公开披露. 作为多模态模型, GPT-4 接受文本和图像输入, 训练数据截至 2021 年 9 月, 模型输入标记的窗口限制也从 GPT-3 的 4 096 个增加到最多支持 32 768 个, 从而为生成任务提供更丰富的上下文支持. 该模型针对聊天场景进行了优化, 同时在复杂生成任务、专业知识及学术基准测试中表现出色, 部分性能超越人类水平. 同年 4 月, OpenAI 宣布推出 GPT-4 Turbo, 并在 2024 年 4 月发布了经过重大改进的 GPT-4 Turbo 版本, 该模型增加了计算机视觉的支持, 训练数据的时效更新到了 2023 年 12 月, 输入标记的窗口限制扩展到了 128 000. 2024 年 5 月, OpenAI 发布了 GPT-4o, 作为 ChatGPT 旗舰水平的大语言模型, 其智能水平与 GPT-4 Turbo 相当, 但响应速度更快、成本更低. GPT-4o 具备端到端多模态能力,

可处理文本、视频、音频、图像输入, 并生成任意组合的输出, 克服了先前需调用多个模型的限制, 交互流程的耗时控制在 300 ms 左右, 非常接近于人类流畅沟通时的响应耗时水准, 进一步拓宽了大语言模型的应用范围.

2.3 InstructGPT 的训练和 ChatGPT 模型

GPT-3 这类底座模型, 其训练数据主要来源于互联网, 尽管在训练之前会进行大量的过滤、筛选和清洗任务, 但仍然可能包含虚假信息或者有偏见的内容, 比如涉及种族歧视、性别歧视、文化冲突等. 因此, 这些模型很可能会生成错误的, 或者不健康、不准确的内容. 此外, 这类语言模型的本质是基于上下文来预测下一个标记, 训练目标不一定能同用户希望模型执行的目标任务保持一致, 即无法保证模型遵循用户的意图和指令.

针对上述问题, OpenAI 在 2021 年发布了基于 GPT-3 改进的新模型, 并取名为 InstructGPT[10]. InstructGPT 通过强化学习和人类反馈进行优化, 不断利用反馈来学习和改进模型, 旨在提升生成内容的真实性、可靠性和任务一致性, 并且在具体任务上更加符合人类的意图. InstructGPT 基于 GPT-3, 其训练过程主要有两个阶段: 监督微调 (supervised fine-tuning, SFT) 和人类反馈强化学习 (reinforcement learning from human feedback, RLHF). 在 SFT 阶段, 通过监督学习对原始 GPT-3 进行微调, 首先 OpenAI 会选择一些用户提示词 (prompts), 并由标注员编写与提示词对应的理想回答, 这样便形成了“用户提示词-生成内容”的监督训练数据集, 通过该数据集对 GPT-3 进行微调, 得到初步的 SFT 模型. 在 RLHF 阶段, 训练过程进一步分为训练奖励模型和使用奖励模型进行强化学习两个阶段. 奖励模型旨在为生成内容自动评分, 当模型的生成内容和提示词匹配的时候, 可以获得高分数, 而当生成内容和提示词不匹配的时候, 降低其分数. 为训练奖励模型, SFT 模型首先生成多个候选答案, 然后标注员根据和提示词的匹配相关性和内容健康程度对这些答案进行排序, 在多次重复这个过程后, 奖励模型得以优化. 然后在强化学习阶段, 模型根据随机提示生成输出内容, 再经过之前的奖励模型评估分数, 依据分数更新模型参数. 这一迭代过程无需人工干预即可高效进行, 从而显著提升模型性能. InstructGPT 总体训练流程如图 1 所示.

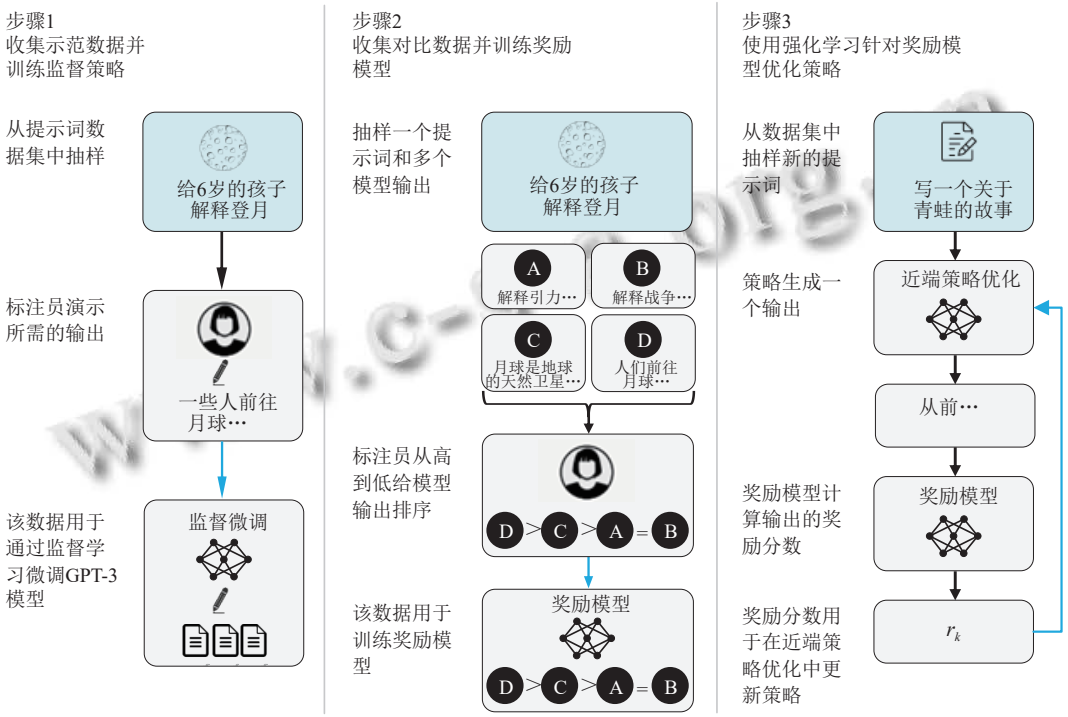

图 1 基于人类反馈的强化学习训练流程

InstructGPT 的优化使其在任务指令遵循度和生成内容质量上均优于原始 GPT-3[10], 展现了通过人类反馈改进大语言模型的可行性与潜力.

2022 年, OpenAI 发布了 GPT-3 的更新版本, 称为

GPT-3.5, 其训练数据截至 2021 年 4 月. 同年 11 月, OpenAI 推出了 ChatGPT, 该模型基于 GPT-3.5 构建, 采用了与 InstructGPT 类似的训练技术, 通过强化学习优化交互式对话能力, 在交互式对话中表现出色, 成为大语言模型驱动的最成功应用之一. 目前, OpenAI 提供的 ChatGPT 在线服务主要基于 GPT-4o 模型, 同时引入了轻量级版本 GPT-4o mini, 经优化后更适用于处理日常任务[9]. OpenAI GPT 系列大语言模型的关键总结见表 1.

表 1 OpenAI GPT 系列大语言模型汇总

模型信息	GPT-1	GPT-2	GPT-3	GPT-3.5 (InstructGPT)	GPT-4	GPT-4o
发布时间	2018年6月	2019年2月	2020年6月	2022年	2023年3月	2024年5月
参数规模	117 million	1.5 billion	175 billion	1.3 billion 6 billion 175 billion	约1 trillion	未披露
最大输入token限制	512	1 024	4 096	4 096 16 385	8 192 128 000	128 000
训练语料	BooksCorpus数据集 (40 GB文本数据)	主要使用Common Crawl 中的reddit数据, 并且筛选Karma值较高的贴文	Common Crawl、BookCorpus、Wikipedia、书籍、文章等 (进行文章去重, 低质内容过滤, 570GB文本数据)	未披露	未披露, 数据截至2021年9月	未披露, 数据截至2023年12月
训练技术和训练方法	半监督训练: 大量无标记数据集进行无监督预训练后, 使用少量标注数据进行任务微调	采用“modified objective training”训练方法, 指导模型生成的内容更具相关性和一致性	采用GShard (giant-sharded model parallelism)技术, 支持模型的并行训练和推理	采用监督微调 (SFT)和人类反馈强化学习 (RLHF), 指导模型按照人类意图生成内容	未披露	端到端多模态训练
特征和性能	(1) 在12个NLP任务中的9个任务中取得了最好的结果 (2) 展现出了zero-shot的能力 (如QA、情感分析等任务	(1) 在8个测试数据集中, zero-shot learning模型在其中7个里均表现优异, 超越了其他模型, 展现出较强的通用性和泛化能力 (2) 发现模型参数的增加, 模型的效果也越来越好	(1) 强化了模型zero-shot learning和few-shot learning的能力, 对于闭卷QA问答、图表解析、翻译等任务表现优异 (2) 多语言支持, 支持英语、中文、法语、德语等30多种语言	(1) 对话接口设计, 针对聊天应用场景优化 (2) 提升内容真实性与任务一致性	(1) 支持多模态输入 (文本、图像) (2) 擅长考试和推理部分基准测试超越人类	(1) 多模态单神经网络 (文本、视觉、音频) (2) 支持50多种语言 (3) 响应延迟约300 ms

3 大语言模型在医疗领域的应用

3.1 电子病历和医疗文本处理

相关研究表明, 当前医疗工作者的工作负担正在急剧加重[11,12], 行政和文书工作占据了医生和护士的大部分工作时间, 这种情况不仅进一步造成了医疗资源紧张, 还会加剧医务工作者的职业倦怠感, 影响患者的治疗效果以及就医体验. 尽管医疗软件提供商不断优化医护工作站的使用效率, 但医疗领域行业复杂、行业门槛很高, 提供准确完备的医疗记录很难通过程序自动化操作完成.

以 ChatGPT 为代表的大语言模型, 具有强大的自然语言和文本的理解与处理能力. Nuance 和微软推出 Dragon Ambient eXperience (DAX™) Copilot 产品[13], 作为一款全自动化临床文档应用程序, 它集成了 OpenAI 公司的 GPT 大语言模型, 通过结合医患对话、上下文环境、生成式 AI 技术, 将医生和患者的聊天沟通内容进行记录、总结和整理归纳, 在每次患者就诊后自动创建草稿临床笔记, 供临床医生立即审查修改, 可以极大地减轻医务工作者的行政和文书负担, 使临床医护人员更加专注于同患者的沟通互动中去.

同时, 医院日常运营任务还包含病案归档、单病种质控上报、传染病上报等其他繁重文书工作, 而这些工作往往还需要调取多部门、多系统的数据资料, 并对涉及的条目进行归纳整理, 工作十分繁重复杂. 借助 ChatGPT, 有望将这些病案、报告的检索获取、整理归纳工作进行自动化处理. Cascella 等[14]也对 ChatGPT 在现实医学场景中的应用进行了实验, 把在重症监护室 ICU 中病人的生命体征、检查检验报告、治疗方案等信息以随机顺序和格式提供给 ChatGPT, ChatGPT 能够对所有数据进行整理, 生成格式化良好的结构化数据记录.

3.2 临床诊断和决策支持

当前 ChatGPT 在临床中的实际应用案例还较少, 但是很多学者和临床医护人员在探索和尝试将 Chat-

GPT 应用在他们的日常科研和工作当中. Cheng 等[15]和 He 等[16]探讨了 ChatGPT 在外科手术场景中的应用, 认为 ChatGPT 在围手术期都可以参与进来. 比如在术前阶段, 医生可以让 ChatGPT 访问患者的健康数据和检查记录, 评估和制定合适的手术方案; 同时, 模型也可以帮助医生更好地与患者进行沟通, 解决患者的疑虑, 建立医患之间的信任依赖关系; 术中 ChatGPT 可以密切检测患者的各项生命指标, 在发生异常状况时及时给出提示和告警, 避免和减少医疗事故的发生; 术后可以为患者建立个性化的恢复方案, 包括训练计划、饮食建议、药物管理等. Harskamp 等[17]验证 ChatGPT 在辅助医疗决策方面的能力, 评估模型在回答心血管相关测试问题的准确性, 同时, 还选出 20 个临床案例, 对比模型生成的结论与临床专家意见相比的准确性. 验证结果显示, 对于基本医学问题, 以及医生与患者之间沟通涉及的简单临床问题, ChatGPT 都能给出完全准确的回答, 不过对于部分复杂的医学知识和临床决策问题会发生一些错误. 作者也发现新版本模型相比旧版本而言, 评测结果的准确度有显著提升, 认为随着大语言模型的不断进化, ChatGPT 具有成为医学 AI 辅助决策工具的潜力. Shen 等[18]指出, 可以将 ChatGPT 与当前的计算机辅助诊断 (computer-aided diagnosis, CAD) 系统结合起来, 医生针对病人的检查检验信息与 ChatGPT 进行开放性的交互沟通, 利用大语言模型的知识辅助形成诊断结论, 减少误诊漏诊的情况, 也克服传统 CAD 系统只能标注异常部位和数据、难以展示系统对诊断结论的洞察和推理过程的问题.

医生的临床决策是一个高度专业化的过程, 在了解患者身体状况及病史信息, 结合临床检验结果、影像报告和病理学检查结果, 往往还需要参阅各种权威的医学资料和临床实践指南, 综合各种信息给出合理的诊断决策. Tariq 等[19]指出在实时医疗决策场景中, 查阅访问临床指南并作出正确的医疗决策通常很困难, 作者通过插件的方式将临床监测指南信息上传给 ChatGPT 模型, 发现在后续的访问中与指南相关疾病的提问, 模型能够提供与指南高度一致的答案, 通过提供快速对话式的交互和准确的回答, 可以帮助医生增强临床决策的能力. ClinicalKey AI[20]模型的训练数据源来自权威的医学资料, 其自然语言交互式的方式也可以让医生便捷地访问可信赖的资料进行更好的临床决策. 大语言模型还具有执行验证任务的能力, 因此大语言模型可能会成为用于减少医疗错误问题的有力工具. Rao 等[21]指出多重用药的问题, 对于涉及复杂病情的患者来说是一个挑战, 作者将一些医生的减药决策案例输入给 ChatGPT 模型, 然后评估 ChatGPT 在标准化临床情景下的减药决策能力, 结果表明大语言模型在临床情景下的减药决策能力和全科临床医生具有一定程度的一致性, 因此大语言模型可以为初级保健医生提供有用的临床支持, 帮助管理多重用药.

近年来, 医院信息化建设取得了长足的进步, 随着 HL7、IHE、DICOM、openEHR 等标准的制定和实施, 从医院中各个科室单元的系统获取患者相关的数据变得十分便捷. 通过接口和插件形式, ChatGPT 具有整合多种来源、多种格式数据的能力, 而且当前大语言模型也支持多模态输入信息[9,22], 因此大语言模型结合现代化的医院信息系统, 可以为患者提供更加完整可靠的临床诊断和决策支持服务.

3.3 增强医患沟通

医患之间的有效沟通一直是医疗过程中十分耗时, 但也是极为关键的环节, 相关的研究表明成功的医患沟通可以减少信息传输偏差, 对患者的治疗康复也有很大帮助[23]. 但是, 受限于患者和医生之间的性别、年龄、家庭条件、文化水平、伦理观念等各方面差异, 医患之间的沟通不一定能够顺利有效地完成, 而一旦医患沟通产生了偏差, 则容易导致误诊、漏诊的发生, 甚至造成严重医疗事故.

ChatGPT 可以作为医生工作站的得力助手, 辅助医患之间进行更有效、更顺畅地沟通. 对于那些专业却又难懂的医学术语, 医生不一定能够在有限的时间内为患者描述和解释清楚, 让患者合理评估自己的病情和诊疗方案, 但 ChatGPT 具有跨越语言和认知、沟通障碍的能力, 可以将答案“简化”, 使用更加通俗的语言让患者更容易理解和接受. 同时, 对于医嘱、患者日常护理事项、药物禁忌等问题, ChatGPT 也可以为患者提供全面、详细和个性化的解释与说明. 在沟通过程中, ChatGPT 甚至可以帮助医护发挥同理心和人性关怀, 比如在沟通过程中可以主动解决患者可能存在的疑虑, 给予积极的支持和安慰, 积极帮助和鼓励患者表达自己的想法和感受, 这对于减轻医生的工作和心理负担, 提升医患关系都有着积极的意义.

Lyu 等[24]尝试使用 ChatGPT 将专业的医学影像报告转述成普通语言的描述, 以便能让患者和家属更好地理解报告内容. 作者选取了临床的低剂量胸部 CT 报

告和脑部 MRI 报告, 让 ChatGPT 将报告的内容翻译成更容易理解的表达方式, 然后再由两位临床放射科医生人工对 ChatGPT 生成的报告进行评价, 结果显示生成的报告在使得病人更容易理解的情况下, 几乎没有重要信息的遗漏和错误, 而模型生成报告质量评价平均可获 4.268 分 (5 分制). 作者还尝试让 ChatGPT 根据报告内容, 给患者和医护人员这两种不同角色提供针对性的病情处置建议, 虽然绝大部分的建议都是“随诊”, 但是仍有 37% 的比例中, ChatGPT 根据报告中的具体病灶给出了专业性的建议, 这也展示了 ChatGPT 在临床诊断中应用的潜力. Moons 等[25]认为医生为病人撰写的内容通常都难于理解, 于是探索尝试 ChatGPT 和 Bard 是否可以降低阅读者的水平要求. 作者从医学期刊中摘取部分疾病相关的信息, 然后让上述两种工具提升原始文本的可读性, 结果 Bard 可读性更强, 但却丢失了大多数的信息, 而 ChatGPT 在保留重要信息的同时也增强了文本的可读性, 显示出一定的应用价值. Ayers 等[26]将 ChatGPT 与医生在公共社交媒体论坛上回复患者问题时的表现进行比较, 分析对比聊天机器人和医生回复问题的质量和同理心指数情况. 对比发现, 占 78.6% 的评估案例, 评估者更倾向于聊天机器人的回复, 聊天机器人回复中被评为“良好”或“非常好”的比例高达 78.5%, 而医生仅为 22.1%, 前者约为后者的 3 倍之多, 聊天机器人回复中被评为“有同理心”或“非常有同理心”的比例为 45.1%, 而医生这一得分比例仅为 4.6%, 显然医护人员可以借助 ChatGPT 这类工具, 与患者之间建立更友好的沟通方式.

3.4 互联网问诊和远程医疗

随着互联网的快速发展, 越来越多的人通过网络获取健康相关信息. 文献[23]的研究表明, 美国约有接近 1/3 的民众通过互联网查询医疗信息并进行自我诊断, 但其中仅有约一半的人随后寻求专业医生的帮助. 这反映了互联网的便捷性, 但也凸显了传统就诊方式在时间和资源成本上的高昂负担. 近年来, 互联网问诊逐渐流行, 许多医院已设立专门的互联网诊疗部门, 通过在线聊天和留言等方式实现患者与医生的初步线上沟通, 而医生则根据患者情况分流, 决定是否需要线下进一步诊治. 这种突破时间和地域限制的问诊模式显著降低了患者的就医成本, 同时优化了患者分流, 改善了医疗资源的利用效率.

然而, 当前各大医院提供的互联网问诊服务通常对单次会话设置时长和轮次限制, 以督促患者在有限对话中清晰、完整地描述病情, 提升沟通效率并节约资源. 然而, 这也提高了患者使用互联网问诊的门槛, 影响了用户体验. 通过利用 ChatGPT 等大语言模型在人机交互方面的优势, 可开发基于大模型的聊天机器人, 提供全天候服务, 支持患者进行多轮连贯对话. 凭借强大的上下文处理能力, 这些模型可根据历史对话内容动态调整后续问题, 在会话结束时自动生成患者情况总结, 供医生审查并决定是否需要补充信息. 这种方式在满足患者个性化诊疗需求的同时, 显著提升诊疗过程的连贯性和效率. 美国 HealthTap 公司推出的虚拟医生助手的应用 Dr.A.I.[27]整合了 GPT-4 模型, 可以帮助患者在访问临床医生之前预先与大语言模型进行交互, Dr.A.I.会让患者先提供年龄、性别、就诊病情等信息, 同时还会根据个体情况向患者询问其他需要了解的信息, 在收集完必要的信息之后, GPT-4 会自动为医生生成结构化的患者信息总结和临床病历草稿, 同时也向患者推荐他们需要当面向临床医生咨询了解的问题. 该应用显著提高患者与医生之间的就诊效率. 同样, Andor 公司通过将其 ThinkAndor 平台与 Oracle Health 认证互通, 使得该平台可以访问患者的电子健康记录 (electronic health record, EHR) 数据. Andor[28]也将 ChatGPT 整合到该平台中, 应用于虚拟护理和虚拟陪护业务, 使得患者无论在医院还是家庭, 都可以随时执行虚拟巡诊、虚拟陪护、远程咨询、患者监测等任务. 研究表明, 借助 ChatGPT 的大语言模型处理能力和 EHR 数据整合, 该平台在就诊、护理和看护方面实现了效率提升、成本降低和效果优化的积极成果. 运行显示, 虚拟护理和虚拟看护的成本分别降低 30% 和 70%, 患者的再入院率降低了 40%. Temsah 等人[29]也指出, 最新的 GPT-4o 模型响应速度快, 在视觉和音频理解方面比之前的模型更为出色, 适用于远程医疗和远程咨询场景, 可以为患者、医疗服务者提供更生动便捷的体验. 基于大语言模型实现的远程医疗解决方案, 可以将优质的远程医疗服务覆盖到那些医疗资源匮乏和偏远的地区, 缓解医疗资源分布不均的问题.

3.5 医学教育和科研

ChatGPT 作为大语言模型, 可协助学生检查与修改作业、审阅课题报告与论文, 功能不仅限于简单的拼写与语法修正, 还可作为领域专家判断报告中的知识点与观点, 提出疑问及改进建议[30]. 此外, ChatGPT 还可以扮演“虚拟导师”的角色, 根据学生的水平与学习目标制定个性化学习计划, 实时解答学习疑难, 评估

学习效果, 识别不足并提供针对性训练, 从而提升学生的自主学习效率与能力. 在医疗教育领域, 医生和护士的培养具有较强的实践性, ChatGPT 可以构建虚拟就诊案例, 在安全可控的环境下模拟与病人交互场景, 这不仅锻炼医学生将理论医学知识应用在实践中的能力, 同时在交互的过程中还能锻炼医学生的沟通技能, 培养辩证逻辑思维, 提高决策决断能力和自信心[31].

对于教师, ChatGPT 可以帮助老师搜集课程需要的教学素材, 生成测验试题、课程案例等信息, 减轻医务教学者处理繁杂教学事务的负担, 使更多时间用于教学设计、创新工作及与学生的互动[32]. 同时, 还可以将学生的课程信息和测验成果输入 ChatGPT, 生成报表并通过模型分析教学效果, 进而改进教学方案和教学方法, 提升教学质量. 李戈等[33]也总结探讨了以 ChatGPT 为代表的大语言模型在循证实践和医学教育领域的应用, 认为在循证医学的教学实践中, 合理利用生成式 AI 可激发学生自主学习热情, 提升批判性思维与逻辑思维能力. 在实践中可以使用大语言模型帮助构建临床问题, 使用客观工具评估教学效果, 培养学生批判性思维和逻辑思维, 提升学习效率和学习效果.

ChatGPT 具有“阅读”专业文献, 生成比摘要更详尽的内容概述, 而且还可以结合文章内容, 回答读者提出的与文章内容相关的问题. 这种能力可以帮助研究者快速判断文章的科研价值, 决定是否深入研读, 从而节约科研时间[34]. 而且, 通过交互式提问, 研究者可与 ChatGPT 进行类学术研讨, 激发科研思路与创新想法. 此外, 通过编程实现批量解析科研文献, ChatGPT 可从海量专业文献和数据中筛选有价值信息, 避免遗漏关键论文. 研究者还可以把临床医疗数据输入给模型, 依据大模型的强大推理能力和洞察力, 有可能帮助科研工作者迅速发现领域知识空白和潜在高价值研究方向, 覆盖科研学者的知识盲区, 构想出一些创新的, 甚至可能是颠覆性的观点和想法. ChatGPT 的跨语言理解与生成能力超越传统翻译工具, 可降低非英语母语研究者在学术交流中的语言障碍, 促进全球前沿知识与经验的合作.

正如任何工具有双重性, 大语言模型的强大功能在教学与科研中也可能带来不利影响. 首先, 借助大语言模型, 意味着可以借助该工具更加快速便捷地访问学术资源, 并且可以直接生成作业、论文内容, 由此引发了对使用大语言模型生成成果的评价争议: 这些成果是否真实反映作者的能力? 将使用大语言模型的学生或研究者与未使用者的成果一同评估是否公平? 大语言模型生成的内容, 其所有者是属于大语言模型服务提供者, 还是大语言模型的使用者, 抑或是大语言模型训练数据的所有者[32]? 若学生过度使用大语言模型, 专业知识与技能可能未获充分锻炼, 极易丧失独立思考与任务处理能力, 甚至在主观性或创新性任务中过度依赖模型生成观点, 违背教育提升学生批判性思维与问题解决能力的初衷. 同时 ChatGPT 的使用还可能导致学生学习行为的扁平化和碎片化, 不利于学生进行系统性、深入性的学习和思考. ChatGPT 及很多其他大语言模型不仅通过了 USMLE 考试, 而且还可以对医学考试中的开放问题进行解答, 其医学领域的专业知识基本达到了在校学习医学生的水平[35], 很多教师与专家担心学生会在测验和考试中使用 ChatGPT 来作弊. Nolan[36]报道两名教授发现学生课题论文是使用 ChatGPT 生成的, 一位教授发现学生的课题论文写作表现手法上很优秀, 但包含课程未提及内容, 而且还有毫无意义的谬误; 另外一位教授指出其学生文章语法完美, 和他应有的水平不相匹配, 而且文章内容空洞, 最终学生承认是使用 ChatGPT 生成的.

对于突如其来的 ChatGPT, 很多教育机构尚不知道如何应对这种情况. 例如在美国洛杉矶, 很多的公立学校在学习公共网络和设备上禁止访问 ChatGPT, 包括纽约、华盛顿在内的很多其他地区也紧随其后. 但他们随后发现, 使用这种方法根本无法杜绝学生使用 ChatGPT, 且可能错失生成式 AI 带来的机遇和好处, 相比未禁用地区差距扩大, 因而许多学校纷纷解除对于 ChatGPT 的禁用限制[37]. 香港多所大学在短暂禁用 ChatGPT 之后也都纷纷取消禁令, 认为大学作为一个多元化的环境, 应该比其他领域更加积极地面对和拥抱这项变革[38]. 在教育领域里, 教师和学校需要教会学生们如何正确合理地使用 ChatGPT 这类大语言模型工具, 明确可使用与不可使用的场景, 并使学生意识到当前大语言模型的缺陷与风险, 以及滥用可能产生的后果.

Cotton 等[32]也描述了使用 ChatGPT 生成内容的一些特征, 以及识别使用 ChatGPT 这类工具进行学术剽窃和作弊的方式. 然而, 随着技术迭代, 人工智能生成的内容将更加接近于人类生成的形式, 显而易见的错误与缺陷可能消失, 识别难度增加. 当前, 许多反作弊工具无法检测 AI 生成内容, 甚至 OpenAI 自行开发的 AI 检测工具因检出率过低已下线[39]. 正如作弊和反作弊

一样, 是一个永无止境的过程, 核心仍需树立使用者的正确价值观与道德准则, 确保其为行为负责并承担后果.

表 2 系统梳理了多项代表性研究中的大语言模型的应用分类、场景、收益与不足.

表 2 大语言模型在医疗领域的应用

文献	应用分类	应用场景与解决的问题	效果和收益	不足与改进
[13]	自动化临床文档	通过语音识别和环境感知技术, 自动识别和捕捉医生与患者沟通对话内容, 并生成临床诊疗记录	减轻医务工作者行政和文书负担	—
[14]	自动化临床文档 医学教育和科研	将ICU病人的生命体征、检查检验报告、治疗方案等信息进行分析整理, 生成结构化的数据; 自动对科研论文进行总结并生成摘要	帮助医护快速获取病人信息, 提高患者信息获取效率; 提高科研效率	对于复杂场景的因果关系理解有限
[15]	临床诊断和决策支持 增强医患沟通医学教育和科研	研究了ChatGPT和GPT-4在外科手术中的应用,包括围手术期管理、影像诊断、医患沟通、科学研究写作	提高手术的准确性、效率和安全性, 避免和减少医疗事故的发生, 并改善术后管理	大模型不会取代外科医生, 因为人类医生拥有独特的经验和技能
[16]	临床诊断和决策支持	探讨了ChatGPT/GPT-4在脊柱外科手术中的应用, 包括医患沟通、数据处理、手术规划和术后康复	可以辅助脊柱外科手术的正确、顺利实施执行, 提升脊柱外科手术的效率	需要注意数据安全和隐私风险
[17]	临床诊断和决策支持	评估大语言模型在回答心血管相关测试问题的准确性; 同时针对20个临床案例, 对比生成结论与临床专家意见相比的准确性	在回答简单病例上的准确率较高, 可以帮助减轻医生的负担	部分复杂的医学知识和临床决策问题会发生一些错误, 复杂病例场景仍有局限
[18]	临床诊断和决策支持 增强医患沟通	研究ChatGPT及其他大语言模型在放射学中的应用, 可以与CAD结合起来完成报告的自动生成; 帮助完成医患沟通相关工作	从影像数据中提取信息并生成初步报告, 供放射科医生审查; 医患沟通有助于建立信任关系, 同时增强患者信心	准确性需要验证; 实际使用时涉及伦理方面的考虑, 如患者数据隐私和数据安全
[19]	临床诊断和决策支持 增强医患沟通	评估ChatGPT-4通过整合临床指南来提升其临床决策支持能力的潜力	整合指南后, ChatGPT回答准确性明显提高; 面对内容冲突时, ChatGPT-4能有效总结归纳信息	在复杂病例中可以继续改进, 提升准确性和可靠性
[20]	临床诊断和决策支持	整合可信的医疗内容 (包括科学期刊、药物信息、临床参考内容和医学教科书), 评估基于可信医疗内容的模型在医疗领域的潜在应用, 及其对临床实践的影响	帮助医生提供快速、准确的临床信息, 支持决策过程; 根据患者信息 (如合并症、当前药物、年龄) 提供个性化支持	—
[21]	临床诊断和决策支持 医学教育和科研	帮助老年患者用药时进行多药物用药管理, 减少某些药物的使用	可以提供初步去药化建议, 降低药物相互作用以及副作用的风险; 帮助用药医生更高效地决策	—
[24]	增强医患沟通	将放射学报告中的专业术语翻译成通俗语言, 使患者和非专业医疗人员易于理解	增强患者教育、问答与沟通效果	GPT-4比旧版本模型表现更好, 随着模型迭代可以增强其表现能力; 优化提示词可以提高生成内容的质量
[25]	增强医患沟通	简化患者信息材料, 降低阅读难度, 提升患者对健康状况与治疗方案的理解	提高文本可读性, 帮助患者理解病情和治疗方案	仍有改进的空间, 如简化的版本会有部分重要信息的丢失
[27]	互联网问诊和远程医疗	就诊前与患者互动访谈, 收集病史、症状等信息, 进行就诊前预评估, 并生成预诊断笔记提供给医生	节省就诊时间, 提高了效率	部署前仍需要大量安全性测试论证
[28]	互联网问诊和远程医疗	基于ChatGPT技术驱动的虚拟健康平台和电子健康记录 (EHR)系统整合, 增强虚拟健康互动, 帮助医疗机构建立可持续的虚拟健康策略 (比如床旁或家庭护理)	改善医疗效率, 提升用户体验	—
[29]	互联网问诊和远程医疗	探讨了GPT-4o在远程医疗中的应用潜力, 利用模型实时处理音频、视觉和文本数据的能力, 以及多语言处理的能力; 提出患者、医疗专业人员和GPT-4o的协作沟通模式, 强调AI与人类专业知识的整合	提高远程医疗服务的效率; 缓解医疗资源分配的问题	评估AI在复杂医疗案例中的可靠性, 并确保AI工具与人类专业知识的有效整合

4 大语言模型的问题和挑战

4.1 数据安全与隐私

医疗健康信息高度私密且敏感, 涉及个人病史、治疗方案、身体数据等内容, 因此对于数据安全和隐私的要求相当严苛, 一旦发生未经授权的访问或者数据泄露, 就会导致患者隐私被侵害, 并产生严重的后果.

以 ChatGPT 为代表的大语言模型由 OpenAI 公司私有化提供, 而且不提供客户在第三方环境的私有部署, 因此在医疗领域使用过程中, 需要将患者和医生的对话内容、病历、检验报告等信息组装成通信报文, 通过互联网访问 OpenAI 的服务方能得到结果, 数据脱离医疗机构内网, 经由互联网传输就有被窃取和泄露的风险. 其次, 在 OpenAI 的隐私政策中明确规定: 通过 non-API 方式访问 ChatGPT 的会话内容, 会被用作语料来训练提升模型, 而通过 OpenAI API 的方式访问 ChatGPT 的会话内容, 不会用来训练 OpenAI 的模型, 但是 OpenAI 会将这部分数据保留 30 天, 以监控和审查是否存在滥用的情形, 且部分授权员工及第三方机构可访问这些数据[40]. 对于非美国本土用户使用 OpenAI 的服务, 这就牵涉到数据跨境传输问题, 可能引发国家层面的数据安全与战略安全关切. 由此看来, 以 ChatGPT 为代表的第三方大语言模型, 如果不能在医疗机构内部实现私有化部署, 其数据安全与隐私风险较高, 需谨慎评估使用可行性.

ChatGPT 模型的训练数据大部分来源于互联网资源和一些专业书籍, 尽管当前互联网上存在部分医疗机构提供的去标识化数据集, 但其数量规模与覆盖范围远不足以满足需求. 使用大量真实可靠的医疗数据训练大语言模型, 对于提升大语言模型的专业性和可靠性十分重要; 同时, 更加专业可靠的大语言模型, 也可以更有效、更安全地帮助医院、医生护士和患者. 因此, 监管机构、医院、医护人员及人工智能技术厂商应积极推动医疗数据的收集与共享, 挖掘其潜在价值. 例如, 可通过规范化数据收集、存储与传输, 结合关键字段脱敏及去标识化技术, 确保患者隐私安全; 而患者应当对自己数据拥有最终的处置权, 可选择允许或拒绝将其医疗数据用于大语言模型训练. 当然, 医疗数据的收集与利用涉及众多利益相关方, 并牵涉隐私、数据安全及伦理道德等复杂问题, 亟需持续探索合理、合规、安全的解决方案.

对于医疗机构, 可采用开源大语言模型作为基础模型, 结合专业医学教材、文献及本院病案数据对模型进行微调, 并将微调后的大语言模型私有化部署于医院内网环境中. 这是当前实现数据安全与隐私保护的最可靠实施方案.

4.2 生成内容的可解释性与可靠性

以 ChatGPT 为代表的大语言模型在多语言支持、连贯对话、数学计算、推理和逻辑, 以及一些专业知识领域展现出令人惊叹的能力, 但是相关学者和开发人员却无法理解这些能力从何而来, 以及在何种情况下大语言模型可能会犯错. 一直以来, 深度学习的可解释性一直是业内被困扰的难题, 原因在于当前深度神经网络过于庞大复杂, 难以直观地了解数据和特征有效性, 以及对于最终输出结果的贡献程度. 与此同时, 深度神经网络的工作原理还是建立在坚实的理论基础之上的, 例如, 针对图像处理、模式识别等类别的应用, 深度神经网络是在原始训练数据和标记之间通过训练得到了某些特征和模式, 再利用这些特征和模式作用于新的数据测出结果. 但是像 ChatGPT 这样的大语言模型, 其工作机制主要基于上下文预测下一个词的统计与概率模型, 无法保证生成的内容是依照某些事实为依据的, 而正是这样一个简单的生成器却实现了自然对话、计算、常识推理、医学诊断等强大的功能, 在让人惊叹的同时又不免让人充满疑惑. 这种令人惊叹的同时又充满疑惑的“黑盒”特性, 势必引发人们对其在医疗领域大规模使用的安全性的担忧.

通用大语言模型生成的内容无法保证以事实为依据, 可能存在不准确或与事实、伦理相悖的情况, 甚至在没有“合适”答案的情况下, 会“捏造”事实进行回复. 而且 ChatGPT 具有连贯流畅的上下文对话能力, 不准确的信息可能以极具说服力的方式呈现, 让用户难以辨别真伪, 若用户被误导则很容易带来无法挽回的严重后果. Wagner 等[41]的研究就发现, 在使用 ChatGPT 进行 88 个放射学科问题提问的生成结果中, 正确回答了 59 个 (67%). 同时, 模型给出了 343 篇参考文献, 但是其中只有 124 篇 (36.2%) 可以在互联网上检索访问到, 剩余的 219 篇 (63.8%) 都是 ChatGPT 伪造的, 而可访问的 124 篇文献中, 也只有 47 篇 (37.9%) 可以被认为其背景知识可以支持相关问题的回答. 因此, 尽管回复内容形式上看似严谨, 可能隐含着许多的错误和虚构, 需要更加谨慎地去核实. 也有学生使用 ChatGPT 生成课题论文, 虽然形式和表达上无法挑剔, 但内容可能

出现不合逻辑或与事实不符的谬误[36], 完全在制造形式工整的信息垃圾.

事实上, 大语言模型研究者很大部分的研发工作, 都在致力于了解可能引发“幻觉”的情形, 并采取相应的措施加以规避或者缓解这种状况. 在使用 GPT 等大语言模型时, 采用链式思维 (chain-of-thought, CoT)[42,43] 可以帮助提升大语言模型执行复杂任务的能力. 通过在输入任务提示的时候显式给出有逻辑的解题中间步骤, 并要求模型生成中间推理过程, 可以引导大语言模型生成更好的结果. 同时, 要求显式给出的中间解题步骤也使得推理结果更具有可解释性, 便于监测与验证, 从而降低错误风险.

4.3 知识的扩充和更新迭代

大语言模型的知识依赖于模型训练资料的覆盖领域和时效性. 以 ChatGPT 为代表的大语言模型通常为通用类型, 其训练数据主要来源于书籍、互联网等资源. 然而, 某些领域的知识可能未被训练数据覆盖, 或在训练过程中被模型忽略, 导致模型在特定领域的表现不佳. 同时, 训练资料的时效性也会影响模型的表现. 例如, ChatGPT 的基础模型 GPT-3.5 的训练数据截至 2021 年 9 月, 对于这个时间点之后所发生的事情, ChatGPT 的模型是不知道的. 针对这种情况, 通常需使用新语料对模型进行训练更新, 但更新大语言模型的参数需消耗大量算力与时间, 且更新后需经过测试、验证及部署等复杂流程. 更新后的大语言模型也可能会和旧版本产生很大差异, 导致支撑的业务表现出不稳定性和不确定性, 所以频繁更新大语言模型在实际中不具可行性, 从而使其知识覆盖率与时效性存在一定偏差与滞后.

目前, 主流大语言模型具有强大上下文学习 (in context-learning) 的能力[8]. 通过自然语言描述任务设定, 并在会话中提供一个或少量任务实例作为样例, 模型可在不更新参数的前提下, 通过样例学习并执行完成新任务的能力. Glicksberg 等[44]使用纽约 7 家医院的电子健康记录 (EHR) 数据, 对比使用传统监督机器学习模型、没有示例任务的 GPT-4、提供额外示例知识的 GPT-4 进行医院急诊单元病人收治入院概率的预估性能. 结果显示, 直接使用 GPT-4 模型预估性能最差, 但是通过使用少样本学习机制 (few-shot learning) 为模型提供额外信息, GPT-4 模型的预估性能显著提升, 预估指标甚至接近于传统监督学习的预估性能. Tariq 等[19]的研究也发现, 通过插件的方式向 ChatGPT 上传艰难梭菌感染及结肠息肉监测的治疗指南, 相比于原始没有提供该指南的情形, 上传之后生成的回答准确性明显提高, 并始终提供与指南一致的答案, 展示其在复杂临床情境中的实用性.

鉴于医疗领域 AI 应用拥有着广阔前景, 而通用大语言模型对于医疗行业领域知识又有着一定的限制, 所以医疗领域的专用大语言模型也越来越多. Elsevier Health 发布的 Clinical AI[20]构建于受信任的医疗资料库, Elsevier 积累了大量高质量可靠的医疗科技文献、药物治疗、临床参考指南、医学书籍信息, 因此相信 Clinical AI 相比于通用大语言模型, 可以提供更可靠的、基于循证的医学信息, 临床医护人员可以随时以自然语言接口查询这些可靠的医疗信息, 更好地支持临床决策. Google 也基于医疗使用场景发布了专用的 Med-PaLM 2[22], 该模型在回答 USMLE 类型的医学问题, 已经达到人类专家的知识水平. 该模型还提供长文本的医疗问题的回答, 与人类相比, Med-PaLM 2 的答案质量更高, 更不容易忽略遗漏重要信息, 且不受医护人员个人偏好影响, 确保客观性.

检索增强生成 (retrieval augmented generation, RAG) 也是一项改善大语言模型生成内容的重要技术. RAG 通过引入信息检索组件, 首先根据用户输入信息从外部新数据源检索提取信息, 然后再合并用户输入和检索结果提供给大语言模型, 可以让大语言模型生成更好的响应内容. 通过 RAG, 可以在不重新训练或者微调模型的情况下, 以较低成本且高效的方式解决平常大语言模型领域知识欠缺与数据过时的问题. Miao 等[45]通过将 2023 年 7 月 1 日发布的《慢性肾脏病评估和管理临床实践指南》作为外部数据源创建 RAG 改进的 GPT-4 模型, 然后对原始 GPT-4 模型和改进 GPT-4 模型进行提问, 前者生成的内容比较普通, 而后者生成的内容专业性更强, 而且同提供的指南内容高度一致.

4.4 生成内容的偏见和歧视风险

大语言模型的生成行为, 依赖于训练数据、算法设计和训练流程等技术. 若训练数据分布不合理, 最终模型可能对某些问题产生偏见与歧视性输出. 以 ChatGPT 为例, 其训练数据绝大多数来源于互联网, 所以模型行为必然会融入大量偏见和错误. 同时, 即便是真实的医疗数据, 对应的医疗行为和医疗数据也会随着社会经济、地理位置和环境、文化等因素的变化而产生

巨大差异. 比如, 不同疾病在不同环境下的易感性、防护及治疗手段与当地社会自然环境密切相关; 拥有完善医疗保障和对自身健康状况比较关注的群体, 与那些仅能够接受和承担起初级保健卫生的患者群体, 其会诊记录和治疗方案也会有很大差别. 在训练流程上, ChatGPT 采用强化学习结合人类反馈 (reinforcement learning from human feedback, RLHF) 方式训练[10], 这个过程中的奖励模型可能偏向训练人员的主观偏好, 所以 ChatGPT 相对于大众而言也可能会产生偏差. 让多元化的用户对模型生成的内容进行反馈, 并输入到模型的训练迭代中去, 或许可以减少偏见发生.

人工智能的发展必须确保公平, 确保人人都有平等享受这些技术利益的机会, 同时避免对老人、女性、贫穷群体等特殊人群或个体造成偏见与歧视. 针对这种情况, 除了对大语言模型的输出进行严密的监视和审查之外, 大语言模型的提供者也需要通过向模型提供更加多样性的医疗数据集, 在训练流程和技术层面上改进大语言模型的生成行为, 才有可能让大语言模型根据患者具体情况生成合理的内容. 相关研究者或认证监管部门, 可以收集和整理全面的样本, 并对这些样本进行分类以形成一个基准数据集, 将不同子群体输入模型, 比较生成内容是否存在显著差异, 并评估这些差异是否反映现实中应避免的社会偏见与歧视, 作为用户选择大语言模型的参考依据.

4.5 工具滥用和监管

当前以 ChatGPT 为代表的大语言模型正处于快速发展的阶段, 极易产生监管漏洞和被滥用的风险. 通常而言, 监管和法规会在新技术问世很久之后才能逐步地跟进和完善, 而在大语言模型时代这个问题将变得十分严峻. 设想任何人几乎没有任何领域知识门槛、技术门槛, 只需要借助大语言模型工具就可以快速生成大量的虚假内容, 而且普通大众很难分辨出来, 如果被不法分子肆意滥用, 其后果将不可估量.

当前全球普遍认为, 以 ChatGPT 为代表的大语言模型技术的开发使用, 需要接受相关部门对其进行约束和监管. 我国于 2023 年 7 月 10 日发布了《生成式人工智能服务管理暂行办法》[46], 并自 2023 年 8 月 15 日起施行. 该法律在鼓励生成式人工智能创新发展的同时, 对生成式人工智能服务实行包容审慎和分类分级监管, 明确了生成式人工智能行业的发展政策方向、服务提供者的服务范围和服务规范, 以及相关监管部门的管理权限和管理责任. 欧盟于 2024 年 3 月发布了《The EU Artificial Intelligence Act》[47], 以“风险的强度和范围”对人工智能进行了风险层级的划分, 监管部门可以对大语言模型的使用分类型、分场景进行差异化操作. 在《The EU Artificial Intelligence Act》中, 医疗行业相关的 AI 系统通常被归类为高风险系统, 这是因为医疗领域的 AI 系统在患者健康和安全方面具有显著影响, 因此需要遵守严格的法规和合规要求. 由此可见, 在法律法规上对生成式人工智能服务实行包容审慎和分类分级监管的方式, 既可以维护公众利益不受侵害, 同时也不会因为过度严格的监管措施扼杀相关技术的创新, 是当前监管的主流形式. 健康人工智能联盟 (coalition for health AI, CHAI)[48,49]致力于在医疗健康领域开发和推广负责任的 AI 标准, 以确保高质量、值得信赖和公平的 AI 应用服务于大众. 该组织联合技术创新者、学术研究团队、医疗保健组织、政府机构和患者等多个利益相关者共同合作, 采纳多样化的声音、需求和专业知识, 同时测试、部署和评估 AI 系统, 为寻求 AI 在医疗保健中的开发、评估和使用的最佳实践做出贡献.

在技术方向上, 相关企业和监管部门需要有对大语言产品进行测评和认证的能力, 用于帮助企业和组织进行对比选型, 以及帮助大语言模型服务提供商提升其产品性能. 由于大语言模型在医疗和临床应用中对生成内容的质量和安全性有着极高的要求, Singhal 等[50]提出和创建了 MultiMedQA 评估数据集, 整合了专业医学考试、研究和消费者用户查询等多个来源的数据集, 以全面衡量模型在各种应用场景下的性能表现. 此外, 团队还设计了一个多维度的人类评估框架, 从事实性、精确性、潜在危害和偏见等角度对模型答案进行评估, 揭示模型的潜在缺陷, 为提升模型的安全性和实用性提供了基础. 同时, 在诸如医疗等关键领域的应用过程中, 能够定期对模型性能进行监测评估, 或者对相关产品的表现进行实时监控, 发现问题能够及时给予纠正, 确保大语言模型的安全性和可靠性. 大模型服务提供者也需要具备实时对生成内容进行自动化和智能化地监督, 减少人为干预或者人工审核的工作量, 这也是提升大语言模型安全可靠性并尽快在各个关键领域落地的关键因素.

4.6 技术开放和自主可控

随着 GPT 系列大语言模型迭代更新, OpenAI 对于

新模型的模型结构、训练数据、训练过程、工程实现等技术细节公开透露的信息越来越少. 从 GPT-2 开放源代码, 到 GPT-3 只提供论文, ChatGPT 的训练方式与细节未予披露, 以及 GPT-4 仅提供使用评测报告, 这一趋势使得 OpenAI 的发展方向与公司名称中的“Open” (开放) 理念渐行渐远. 不可否认, 开发和更新大语言模型需要大量的技术创新和资源投入, 企业出于商业竞争和知识产权保护的考虑可能选择封闭策略, 因此 OpenAI 的这些举动很可能是出于商业竞争方面的考虑, 以保持同越来越多的竞争对手的领先性. 但是, 封闭使得公众无法对其大语言模型进行更全面的了解和评估, 进而可能导致公众对 OpenAI 产品的接受度更为保守.

对于具有革命性潜力的技术, 对公众保持开放、透明的态度, 能够让整个社会都参与进来, 才能够保证技术的繁荣和可持续性发展. 当前, 大语言模型领域发展十分迅速, 除了 OpenAI 的 GPT 系列外, 很多公司都训练并发布了自己的大语言模型. 国外主流科技公司比如 Google 的 Bard (后更名为 Gemini)[51], Meta 的 LLaMA[52], Anthropic 的 Claude[53]等, 并且在某些任务的表现上已经超过了 OpenAI 的 GPT-4 旗舰模型[51,53,54], 这让我们在大语言模型体验和使用上有了更多的选择. 值得指出的是 Anthropic 是由 OpenAI 初创团队成员成立的公司, 秉持负责任的 AI 使用理念, 让模型在对齐关键原则和遵守安全限制的前提下训练, 减少在运行过程中生成有害、有偏见的信息. 在国内, 代表性的大语言模型有百度的文心一言、阿里通义千问、腾讯混元、讯飞星火等, 这些互联网企业结合自身业务特点, 积极尝试将大语言模型应用于智能化办公、商业化推荐、教育、医疗、零售、金融等领域, 这也是对于大语言模型促进产业升级的积极尝试. Meta 公司创始人及 CEO Zuckerberg[55]宣布开源其公司最新的旗舰模型 LLaMA 3.1, 在大语言领域引发了较大的反响. 尽管之前也有大语言模型开源, 但其评测性能和闭源商用的大语言模型相差很远, 而此次的 LLaMA 3.1 开源大语言模型, 在超过 150 个基准数据集上进行了评估, 并结合大量人类参与的测试, 实验结果表明模型性能可以同 GPT-4、GPT-4o、Claude 3.5 Sonnet 相媲美[56]. 此外, LLaMA 3.1 开源了包含 4050 亿、700 亿、80 亿参数不同大小的模型, 允许不同规模和背景的公司及个人基于合适模型进行训练、微调或知识蒸馏, 从而在保护数据隐私安全的前提下, 规避与闭源商业模型合作的限制与风险. 凭借 Meta 在开源项目运营方面的丰富经验, 相信 LLaMA 的开源对整个大语言模型领域的合作化发展有着积极推动作用.

5 结束语

以 ChatGPT 为代表的大语言模型正悄然影响着多个行业, 而医疗行业作为与人类健康息息相关的重要领域, 被广泛认为是人工智能最具有应用价值、最具变革潜力的应用领域. 长期以来, 为全球各国和地区提供更高效、便捷的医疗卫生服务一直是备受关注的热点课题, 尤其在全球人口老龄化加剧的背景下, 医疗资源短缺问题日益突出, 医疗行业亟需借助新兴技术进行产业革新, 以缓解供需矛盾. 虽然以 ChatGPT 为代表的大语言模型在医疗领域的初步应用已经展露出一些令人兴奋的成果, 但其应用仍伴随着诸多的风险和不确定性. 鉴于医疗领域直接关系到个体生命与健康的重大责任, 必须小心谨慎地去对待, 相关技术的部署必须极为审慎.

目前, 以 ChatGPT 为代表的大语言模型, 在医疗领域的应用尚无统一且可靠的实施路径. 因此, 至少在短期内, 其使用应在人类直接且严格的监督下进行, 且需限制于有限场景, 确保生成内容的真实性、安全性和有效性. 应对这些挑战需要大语言模型的研究人员、开发者、监管机构、医护人员、患者以及其他利益相关方通力合作, 共同制定解决方案. 与此同时, 正如许多治疗方案和药物存在副作用与安全风险一样, 人工智能与大语言模型在医疗领域的应用亦需在风险与收益、创新与谨慎之间寻求平衡. 在安全可控的前提下, 充分利用这些技术惠及更多人群, 是未来发展的核心目标.

参考文献

1 姚琼, 王觅也, 师庆科, 等. 深度学习在现代医疗领域中的应用. 计算机系统应用, 2022, 31(4): 33–46. [doi: 10.15888/j.cnki.csa.008411]

2 夏光辉, 曹艳林, 陈炳澍, 等. 大模型人工智能技术在医疗服务领域应用的专家共识. 中国卫生法制, 2023, 31(5): 124–126. [doi: 10.19752/j.cnki.1004-6607.2023.05.024]

3 Mikolov T, Chen K, Corrado G, *et al*. Efficient estimation of word representations in vector space. Proceedings of the 1st International Conference on Learning Representations. Scottsdale: ICLR, 2013.

4 Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. arXiv:1409.0473,

2014.
5 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000–6010.
6 Radford A. Improving language understanding with unsupervised learning. https://openai.com/index/language-unsupervised/. (2018-06-11).
7 Radford A, Wu J, Child R, *et al*. Language models are unsupervised multitask learners. https://paperswithcode.com/paper/language-models-are-unsupervised-multitask. (2019-02-14).
8 Brown TB, Mann B, Ryder N, *et al*. Language models are few-shot learners. Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 159.
9 OpenAI. Hello GPT-4o. https://openai.com/index/hello-gpt-4o/. (2024-05-13).
10 Ouyang L, Wu J, Jiang X, *et al*. Training language models to follow instructions with human feedback. Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2022. 2011.
11 Herd P, Moynihan D. Health care administrative burdens: Centering patient experiences. Health Services Research, 2021, 56(5): 751–754. [doi: 10.1111/1475-6773.13858]
12 Overhage JM, McCallie D. Physician time spent using the electronic health record during outpatient encounters: A descriptive study. Annals of Internal Medicine, 2020, 172(3): 169–174. [doi: 10.7326/M18-3684]
13 Microsoft. Automatically document care with DAX™ Copilot. https://www.nuance.com/asset/en_us/collateral/healthcare/data-sheet/ds-ambient-clinical-intelligence-en-us.pdf. (2024-08-28).
14 Cascella M, Montomoli J, Bellini V, *et al*. Evaluating the feasibility of ChatGPT in healthcare: An analysis of multiple clinical and research scenarios. Journal of Medical Systems, 2023, 47(1): 33. [doi: 10.1007/s10916-023-01925-4]
15 Cheng KM, Sun ZJ, He YB, *et al*. The potential impact of ChatGPT/GPT-4 on surgery: Will it topple the profession of surgeons? International Journal of Surgery, 2023, 109(5): 1545–1547. [doi: 10.1097/JS9.0000000000000388]
16 He YB, Tang HF, Wang DX, *et al*. Will ChatGPT/GPT-4 be a lighthouse to guide spinal surgeons? Annals of Biomedical Engineering, 2023, 51(7): 1362–1365. [doi: 10.1007/s10439-023-03206-0]
17 Harskamp RE, De Clercq L. Performance of ChatGPT as an AI-assisted decision support tool in medicine: A proof-of-concept study for interpreting symptoms and management of common cardiac conditions (AMSTELHEART-2). Acta Cardiologica, 2024, 79(3): 358–366. [doi: 10.1080/00015385.2024.2303528]
18 Shen YQ, Heacock L, Elias J, *et al*. ChatGPT and other large language models are double-edged swords. Radiology, 2023, 307(2): e230163. [doi: 10.1148/radiol.230163]
19 Tariq R, Voth E, Khanna S. Integrating clinical guidelines with ChatGPT-4 enhances its' skills. Mayo Clinic Proceedings: Digital Health, 2024, 2(2): 177–180. [doi: 10.1016/j.mcpdig.2024.02.004]
20 Elsevier. Elsevier Health launches ClinicalKey AI, the most advanced Gen AI-powered clinical decision support tool for clinicians. https://www.elsevier.com/about/press-releases/elsevier-health-launches-clinicalkey-ai-the-most-advanced-gen-ai-powered. (2024-02-29).
21 Rao A, Kim J, Lie W, *et al*. Proactive polypharmacy management using large language models: Opportunities to enhance geriatric care. Journal of Medical Systems, 2024, 48(1): 41. [doi: 10.1007/s10916-024-02058-y]
22 Singhal K, Tu T, Gottweis J, *et al*. Towards expert-level medical question answering with large language models. arXiv:2305.09617, 2023.
23 Will ChatGPT transform healthcare? Nature Medicine, 2023, 29(3): 505–506. [doi: 10.1038/s41591-023-02289-5]
24 Lyu Q, Tan J, Zapadka ME, *et al*. Translating radiology reports into plain language using ChatGPT and GPT-4 with prompt learning: Results, limitations, and potential. Visual Computing for Industry, Biomedicine, and Art, 2023, 6(1): 9. [doi: 10.1186/s42492-023-00136-5]
25 Moons P, Van Bulck L. Using ChatGPT and Google Bard to improve the readability of written patient information: A proof of concept. European Journal of Cardiovascular Nursing, 2024, 23(2): 122–126. [doi: 10.1093/eurjcn/zvad087]
26 Ayers JW, Poliak A, Dredze M, *et al*. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Internal Medicine, 2023, 183(6): 589–596. [doi: 10.1001/jamainternmed.2023.1838]
27 Rutledge GW, Sivura A. A generative AI-based virtual physician assistant. Proceedings of the AAAI Symposium Series, 2024, 3(1): 64–65. [doi: 10.1609/aaaiss.v3i1.31182]
28 Andor Health. Andor health brings the power of OpenAI & ChatGPT at scale with oracle health's validation. https://www.prnewswire.com/news-releases/andor-health-brings-the-power-of-openai--chatgpt-at-scale-with-oracle-healths-validation-301847515.html. (2023-06-13).
29 Temsah MH, Jamal A, Alhasan K, *et al*. Transforming virtual healthcare: The potentials of ChatGPT-4omni in telemedicine. Cureus, 2024, 16(5): e61377. [doi: 10.7759/cureus.61377]

30 Seetharaman R. Revolutionizing medical education: Can ChatGPT boost subjective learning and expression? Journal of Medical Systems, 2023, 47(1): 61. [doi: 10.1007/s10916-023-01957-w]
31 Goldsworthy S, Muir N, Baron S, *et al*. The impact of virtual simulation on the recognition and response to the rapidly deteriorating patient among undergraduate nursing students. Nurse Education Today, 2022, 110: 105264. [doi: 10.1016/j.nedt.2021.105264]
32 Cotton DRE, Cotton PA, Shipway JR. Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. Innovations in Education and Teaching International, 2024, 61(2): 228–239. [doi: 10.1080/14703297.2023.2190148]
33 李戈, 吴涛, 章萌, 等. 大语言模型在循证实践和医学教育中的应用现状及对循证医学教学的启示. 数字医学与健康, 2024,2(2):102–107.[doi:10.3760/cma.j.cn101909-20231109-00062]
34 Qureshi R, Shaughnessy D, Gill KAR, *et al*. Are ChatGPT and large language models "the answer" to bringing us closer to systematic review automation? Systematic Reviews, 2023, 12(1): 72. [doi: 10.1186/s13643-023-02243-z]
35 Strong E, DiGiammarino A, Weng YJ, *et al*. Chatbot vs medical student performance on free-response clinical reasoning examinations. JAMA Internal Medicine, 2023, 183(9): 1028–1030. [doi: 10.1001/jamainternmed.2023.2909]
36 Nolan B. Two professors who say they caught students cheating on essays with ChatGPT explain why AI plagiarism can be hard to prove. https://www.businessinsider.com/chatgpt-essays-college-cheating-professors-caught-students-ai-plagiarism-2023-1. (2023-01-14).
37 Singer N. Despite cheating fears, schools repeal chatgpt bans. https://www.nytimes.com/2023/08/24/business/schools-chatgpt-chatbot-bans.html. (2023-08-24).
38 Leung M, Sharma Y. After a period of caution, universities open up to ChatGPT. https://www.universityworldnews.com/post.php?story=20230823151346920. (2023-08-23).
39 OpenAI. New AI classifier for indicating AI-written text. https://openai.com/index/new-ai-classifier-for-indicating-ai-written-text/. (2023-01-31).
40 OpenAI. Privacy policy. https://openai.com/policies/row-privacy-policy/. (2023-11-14).
41 Wagner MW, Ertl-Wagner BB. Accuracy of information and references using ChatGPT-3 for retrieval of clinical radiological information. Canadian Association of Radiologists Journal, 2024, 75(1): 69–73. [doi: 10.1177/08465371231171125]
42 Wei J, Wang XZ, Schuurmans D, *et al*. Chain-of-thought prompting elicits reasoning in large language models. Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2022. 1800.
43 Kojima T, Gu SS, Reid M, *et al*. Large language models are zero-shot reasoners. Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2022. 1613.
44 Glicksberg BS, Timsina P, Patel D, *et al*. Evaluating the accuracy of a state-of-the-art large language model for prediction of admissions from the emergency room. Journal of the American Medical Informatics Association, 2024, 31(9): 1921–1928. [doi: 10.1093/jamia/ocae103]
45 Miao J, Thongprayoon C, Suppadungsuk S, *et al*. Integrating retrieval-augmented generation with large language models in nephrology: Advancing practical applications. Medicina, 2024, 60(3): 445. [doi: 10.3390/medicina60030445]
46 中华人民共和国中央人民政府. 生成式人工智能服务管理暂行办法. https://www.gov.cn/zhengce/zhengceku/202307/content_6891752.htm. (2023-07-10).
47 European Commission. The EU Artificial Intelligence Act. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=OJ:L_202401689. (2024-06-13).
48 Coalition for Health AI. Coalition for health AI (CHAI) announces founding partners. https://www.prnewswire.com/news-releases/coalition-for-health-ai-chai-announces-founding-partners-302089460.html. (2024-03-14).
49 Coalition for Health AI. Blueprint for trustworthy AI implementation guidance and assurance for healthcare. https://assets.ctfassets.net/7s4afyr9pmov/4AXIWGIlcrjWDaW2ueTaRS/f98e5cb2528187635895cce6ba5ec309/Blueprint_for_Trustworthy_AI.pdf. (2023-04-04).
50 Singhal K, Azizi S, Tu T, *et al*. Large language models encode clinical knowledge. Nature, 2023, 620(7972): 172–180. [doi: 10.1038/s41586-023-06291-2]
51 Gemini Team Google. Gemini: A family of highly capable multimodal models. arXiv:2312.11805, 2023.
52 Touvron H, Martin L, Stone K, *et al*. Llama 2: Open foundation and fine-tuned chat models. arXiv:2307.09288, 2023.
53 Anthropic. Introducing the next generation of Claude. https://www.anthropic.com/news/claude-3-family. (2024-03-04).
54 Anthropic. Claude 3.5 sonnet. https://www.anthropic.com/news/claude-3-5-sonnet. (2024-06-21).
55 Zuckerberg M. Open source AI is the path forward. https://about.fb.com/news/2024/07/open-source-ai-is-the-path-forward/. (2024-07-23).
56 Meta. Introducing Llama 3.1: Our most capable models to date. https://ai.meta.com/blog/meta-llama-3-1/. (2024-07-23).

(校对责编: 李慧鑫)