智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

医疗领域的大型语言模型综述

肖建力, 许东舟, 王浩, 刘敏, 周雷, 朱林, 顾松

引用本文:
肖建力, 许东舟, 王浩, 等. 医疗领域的大型语言模型综述[J]. 智能系统学报, 2025, 20(3): 530-547.
XIAO Jianli, XU Dongzhou, WANG Hao, et al. Survey of large language models in healthcare[J]. *CAAI Transactions on Intelligent Systems*, 2025, 20(3): 530-547.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202405003

DOI: 10.11992/tis.202405003
网络出版地址: https://link.cnki.net/urlid/23.1538.TP.20240912.1104.002

医疗领域的大型语言模型综述

肖建力[1], 许东舟[1], 王浩[2], 刘敏[3], 周雷[4], 朱林[4], 顾松[5]

（1. 上海理工大学 光电信息与计算机工程学院, 上海 200093; 2. 上海交通大学医学院附属上海儿童医学中心心胸外科, 上海 200127; 3. 复旦大学附属妇产科医院 中西医结合妇科, 上海 200011; 4. 上海理工大学 健康科学与工程学院, 上海 200093; 5. 上海市第一人民医院 创伤临床医学中心, 上海 201620）

摘　要: 深度学习是人工智能领域的热门研究方向之一, 它通过构建多层人工神经网络模仿人脑对数据的处理机制。大型语言模型 (large language model, LLM) 基于深度学习的架构, 在无需编程指令的情况下, 能通过分析大量数据以获得理解和生成人类语言的能力, 被广泛应用于自然语言处理、计算机视觉、智慧医疗、智慧交通等诸多领域。文章总结了 LLM 在医疗领域的应用, 涵盖了 LLM 针对医疗任务的基本训练流程、特殊策略以及在具体医疗场景中的应用。同时, 进一步讨论了 LLM 在应用中面临的挑战, 包括决策过程缺乏透明度、输出准确性以及隐私、伦理问题等, 随后列举了相应的改进策略。最后, 文章展望了 LLM 在医疗领域的未来发展趋势, 及其对人类健康事业发展的潜在影响。

关键词: 人工智能; 深度学习; Transformer; 大型语言模型; 智慧医疗; 数据分析; 图像处理; 计算机视觉
中图分类号: TP18　　**文献标志码:** A　　**文章编号:** 1673−4785(2025)03−0530−18

中文引用格式: 肖建力, 许东舟, 王浩, 等. 医疗领域的大型语言模型综述 [J]. 智能系统学报, 2025, 20(3): 530–547.
英文引用格式: XIAO Jianli, XU Dongzhou, WANG Hao, et al. Survey of large language models in healthcare[J]. CAAI transactions on intelligent systems, 2025, 20(3): 530–547.

Survey of large language models in healthcare

XIAO Jianli[1], XU Dongzhou[1], WANG Hao[2], LIU Min[3], ZHOU Lei[4], ZHU Lin[4], GU Song[5]

(1. School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China; 2. Department of Cardiothoracic Surgery, Shanghai Children's Medical Center, School of Medicine, Shanghai Jiao Tong University, Shanghai 200127, China; 3. Department of Gynecology of Integrated Traditional Chinese and Western Medicine, Obstetrics and Gynecology Hospital of Fudan University, Shanghai 200011, China; 4. School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China; 5. Trauma Center, Shanghai General Hospital, Shanghai 201620, China)

Abstract: Deep learning (DL) is a popular research area in artificial intelligence. It simulates the data processing mechanism of the human brain by constructing multilayer artificial neural networks. Large language models (LLMs) based on the DL architecture can understand and generate human language by analyzing enormous data without programming instructions. Thus, LLMs are widely employed in various domains, such as natural language processing, computer vision, intelligent healthcare, and intelligent transportation. This article summarizes the application of LLMs in the healthcare sector, exploring their basic training processes, specific strategies for executing healthcare tasks, and their applications in specific healthcare scenarios. It also discusses the challenges of applying LLMs to the healthcare field, including the lack of transparency in decision-making processes, the accuracy of the output contents, and issues related to privacy and ethics. Thereafter, several strategies for addressing these issues are discussed. Finally, the future development trends of LLM in healthcare, as well as its criticality in promoting human health, are discussed.

Keywords: artificial intelligence; deep learning; Transformer; large language model; intelligent healthcare; data analysis; image processing; computer vision

收稿日期: 2024−05−05.　**网络出版日期:** 2024−09−12.
基金项目: 国家自然科学基金项目（61603257）.
通信作者: 肖建力. E-mail: audyxiao@sjtu.edu.cn.

近年来, 人工智能技术飞速发展, 大型语言模型 (large language model, LLM) 已逐渐在各个领域

中扮演着不可或缺的角色, 其在医疗领域的应用尤其展现出巨大应用潜力。LLM 通过大规模文本数据进行预训练, 具备了理解并处理文本的能力。它们不仅能够自主分析对话内容, 生成与人类相似的回答, 实现与用户进行交互, 还能高效地分析和处理医疗信息。LLM 的应用有助于提升医学教育质量和患者健康素养, 同时也为辅助医疗诊断、提供临床决策支持和药物研发等医疗领域的具体任务开辟了新的可能性。

自从 2023 年 OpenAI 推出的 ChatGPT-4 (GPT-4) 在多个领域展现卓越表现以来, 越来越多的研究机构和科技公司开始加大对 LLM 研发的关注和投入。特定于医疗领域的 LLM 也得到了快速发展, 并在相关任务中表现出色, 这有力地证明了 LLM 在该领域中具有巨大的应用潜力。例如, 谷歌公司研发的 Med-PaLM 2 通过针对医疗领域的微调和指令提示进行优化, 展现出了对医学问题的深刻理解能力, 并在多项评估指标上超越了医疗从业者, 甚至在临床诊断、医学问答等任务中, 达到了接近专业医生的水准。

然而, 在实际医疗场景中应用这些模型时, 仍面临着许多问题和挑战。例如, 由于 LLM 所采用的 Transformer 架构具有复杂的内部机制, 这导致其决策过程通常难以解释。而模型的透明度和可解释性对于其能否实际部署至关重要, 因为这直接关系到患者的生命安全。此外, 数据质量难以保证、专业化程度不足、隐私保护困难以及伦理和法规等方面的挑战, 也对 LLM 在医疗领域中的实际应用产生了关键性的影响。

因此, 本文列举了这些模型在医疗领域内的应用前景, 分析了它们面临的主要挑战, 并汇总了相应的改进策略, 旨在为领域内的相关学习和研究提供有价值的参考或启发。

1　适用于医疗领域的大型语言模型训练方法

大部分 LLM 采用了谷歌研究团队在 2017 年提出的 Transformer 架构[1], 如图 1 所示。该架构通过其独特的注意力机制, 在自然语言理解与生成等任务中表现出了远超传统循环神经网络和卷积神经网络的性能, 迅速成为了自然语言处理领域的核心架构[2]。不同于循环神经网络所采用的传统处理方法, Transformer 通过自注意力机制能够同时处理序列中的所有元素, 显著提高了并行处理能力, 从而缩短了模型的训练时间。此外, Transformer 能够降低计算的复杂度, 减少了长距离信息传递过程中的损失, 从而提高模型处理自然语言任务的准确性和效率。

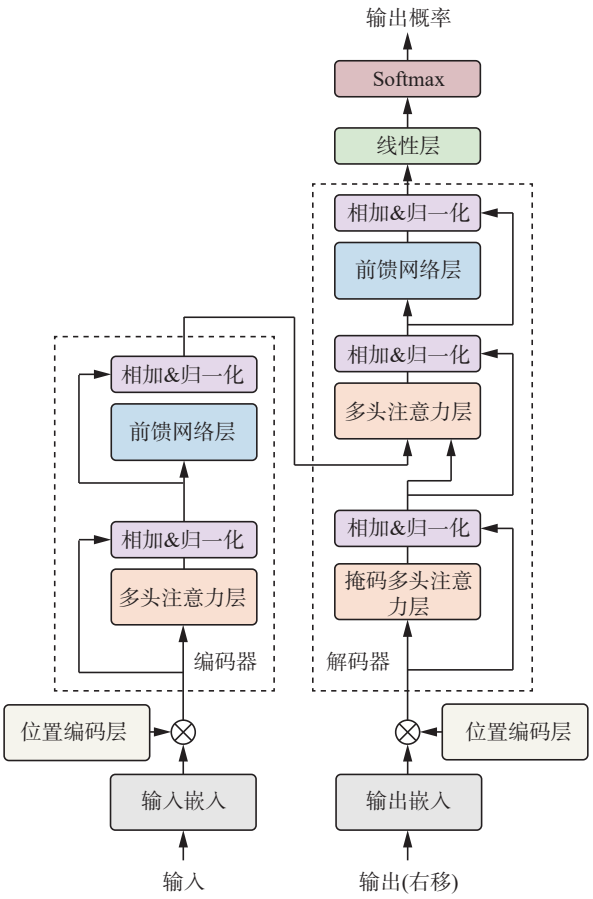

图 1　Transformer 模型架构[1]

Fig. 1　Framework of Transformer model [1]

注意力机制的核心是通过一组查询 (queries)、键 (keys)、和值 (values) 来计算输出, 可以描述为

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{1}$$

式中: $\boldsymbol{Q}$ 是查询矩阵; $\boldsymbol{K}$ 是键矩阵; $\boldsymbol{V}$ 是值矩阵; d_k 是键向量的维度, 用于缩放, 防止点积值过大。式 (1) 首先计算查询矩阵和键矩阵的点积, 然后将获得的兼容性分数除以 $\sqrt{d_k}$ 进行缩放, 并使用 Softmax 函数得到权重, 最后用这些权重对值进行加权求和, 以得到最终输出。

此外, 多头注意力机制[3] 为模型提供了多个维度的学习能力, 能在不同的子空间内并行处理信息。多头注意力机制的公式为

$$\begin{aligned}&\text{MultiHead}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \\ &\text{Concat}(H_1,H_2,\cdots,H_h)\boldsymbol{W}^{\boldsymbol{O}}\end{aligned} \tag{2}$$

每个头的计算方式为

$$H_i = \text{Attention}\left(\boldsymbol{Q}\boldsymbol{W}_i^{\boldsymbol{Q}},\boldsymbol{K}\boldsymbol{W}_i^{\boldsymbol{K}},\boldsymbol{V}\boldsymbol{W}_i^{\boldsymbol{V}}\right) \tag{3}$$

式中: $\boldsymbol{W}_i^{\boldsymbol{Q}}$、$\boldsymbol{W}_i^{\boldsymbol{K}}$、$\boldsymbol{W}_i^{\boldsymbol{V}}$ 分别为第 i 个头的查询、键、值的投影矩阵, $\boldsymbol{W}^{\boldsymbol{O}}$ 为输出投影矩阵, h 为头的数

量。多头注意力机制通过将查询、键和值进行多组线性投影，随后并行执行注意力函数，再将这些输出值连接起来，再次进行投影，从而得到最终输出。这种机制增强了模型对语言多维特征的捕捉和解析，使其能更准确地理解复杂文本的内容。

1.1　训练医疗大型语言模型的基本步骤

图 2 给出了训练医疗 LLM 的基本步骤，主要包括数据收集、数据预处理、预训练、微调、性能评估以及模型优化等。

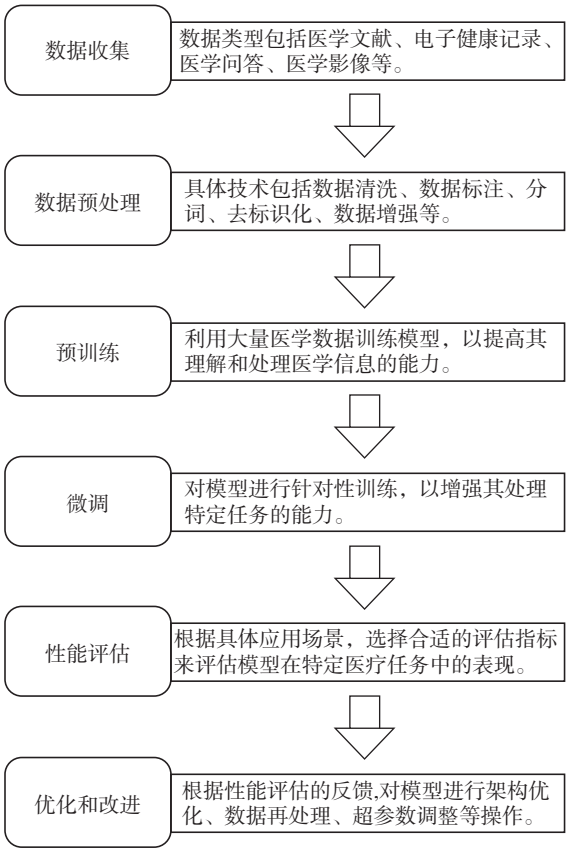

图 2　训练医疗大型语言模型的基本步骤

Fig. 2　Basic steps of a healthcare large language model training

1.1.1　数据收集

LLM 的预训练依赖于大规模数据集的支持，这些数据集的规模和质量对模型在自然语言处理任务中的表现起着决定性的作用。为确保 LLM 在处理复杂的医学专业知识方面能够发挥有效作用，充分保证数据在收集阶段的多样性与高质量至关重要。因此，需要准备多种类型的医学数据，比如医学文献、电子健康记录、医学影像和医学问答数据集等。表 1 列举了医疗领域内的若干常见数据集。

1.1.2　数据预处理

相较于较小规模的语言模型，LLM 对训练数据的质量有着更高的要求。因此，必须通过数据清洗，从语料库中去除重复内容、非文本元素、噪声文档以及低质量文本。随后，还需要进行文本分割，将较长的句子或段落分割成单独的词元，并构建专门针对医疗领域的词汇表，以提高模型的分析效率。此外，考虑到医疗数据的高度私密性，在处理此类数据时需要实施去标识化和匿名化等措施，以充分保障患者的隐私和权益。

1.1.3　预训练

预训练的核心原理是使用大量数据训练 LLM，使其学习数据的结构和模式。这种策略使得模型能在多种下游任务中实现有效的知识迁移和应用[4]。预训练为模型奠定了坚实的语言知识基础，不仅使其掌握了词汇、语法和语义等基本语言结构，还增强了对语言上下文关系的理解能力。除此之外，LLM 通过预训练还能获得跨领域的迁移能力，使其能以更低的成本和更短的时间适应不同但相关的特定任务。

在医疗领域，预训练阶段是提升 LLM 工作能力的关键。它使模型能够理解并处理特定医学信息，以便在下游任务中取得更优表现[5]。这得益于模型对医学文献、电子健康档案、医患对话记录等数据的学习，使其对复杂语境形成更深刻的理解，从而在后续相关任务中表现出更高的性能和准确度。此外，通过学习大量多样化且高质量的数据，LLM 获得了强大的泛化能力，即使在处理数据样本稀少的罕见病症时也能表现出色。

1.1.4　微调

尽管预训练为 LLM 提供了语言理解能力，但在执行某些特定医疗任务时，仅依赖预训练并不足以使模型展现出令人满意的表现。这主要是因为医疗领域专业化程度较高，包含大量专业术语和特定情境，而预训练数据通常无法完全涵盖这些内容。因此，通过微调的方式强化模型对专业知识的处理能力尤为重要。

在医疗领域，微调是实现 LLM 从通用型转向专用型的关键步骤，其目的是使模型适应特定医疗任务。该过程在明确了模型需要执行的具体任务后，通过使用相关的医学数据集进行专门训练来实现。这种训练方式包括根据特定需求调整模型的架构或参数、选取对任务有益的特征，以及适当调整学习率等策略。通过这些专门的调整，微调不仅能提升 LLM 对特定相关数据的理解能力，还有助于降低过拟合的风险。此外，微调在减少训练数据需求的同时，能够加速模型训练过程，从而使模型更迅速地部署到下游任务。

表 1　医疗领域内的常见数据集

Table 1　Common datasets in the healthcare field

数据集	简介	链接
PubMed	收录了3 000多万份医学文献的引用和摘要, 来源于MEDLINE、生命科学期刊和在线书籍等, 涵盖了临床医学、药学、心理学等多个领域。	https://pubmed.ncbi.nlm.nih.gov/
PubMed Central	全球最大的生物医学和生命科学文献数据库之一, 截至2024年3月, 收录了超过970万篇文献。	https://www.ncbi.nlm.nih.gov/pmc/
MedQuAD	包含47 457个医学问答对, 涵盖了37种与疾病、药物、治疗等相关的问题。	https://github.com/abachaa/MedQuAD
n2c2	由去识别化的临床摘要和n2c2挑战赛成果组成, 涵盖了肥胖、冠状动脉疾病、精神病学等多个领域。	https://portal.dbmi.hms.harvard.edu/projects/n2c2-nlp/
UMLS[6]	集合了来自60多个生物医学词汇族的超过90万个概念和200多万个名称, 以及这些概念之间的1 200万个关系。	https://www.nlm.nih.gov/research/umls/index.html
WikiDoc	专注于医疗领域的在线百科全书, 其内容涵盖了包括过敏学、麻醉学、内分泌学、普通外科等多个领域的医学专业知识。	https://www.wikidoc.org/index.php/Main_Page
MIMIC-III[7]	包含4万多名患者的去识别化临床数据, 包括生命体征、医学报告、药物、死亡率等信息。	https://mimic.mit.edu/
HealthQA[8]	包含来自中国三大主流医学问答网站的糖尿病专家问答, 包括135 709个问题及 250 008个回答。	https://github.com/thu-west/HealthQA
Huatuo-26M[9]	包含来自医学知识库、百科全书超过2 600万个中文医学问答对。	https://github.com/FreedomIntelligence/Huatuo-26M
webMedQA[10]	包含63 284个来自真实世界的中文医学问答对。	https://github.com/hejunqing/webMedQA
cMedQA2[11]	由108 000个问题和203 569个答案构成的中文医学问答数据集。	https://github.com/zhangsheng93/cMedQA2
ChatDoctor[12]	由来自在线医疗咨询平台的11 000个真实医患对话和5 000个ChatGPT生成的医患对话组成。	https://github.com/Kent0n-Li/ChatDoctor
MedDialog[13]	由包括340万个对话和1 130万个语句的中文数据集, 以及包含26万个对话和51万个语句的英文数据集两部分组成。	https://github.com/UCSD-AI4H/Medical-Dialogue-System
MedPix	包含近59 000张医学图像、12 000多个患者病例场景医学图像数据库, 涵盖骨折、肺炎、癌症等9 000个领域。	https://medpix.nlm.nih.gov/home
KD-DTI[14]	关于药物–靶标相互作用信息的数据集, 有利于促进对药物和相应靶标之间相互作用机制的理解。	https://github.com/bert-nmt/BERT-DTI
BC5CDR[15]	来源于PubMed中的1 500篇论文, 包括化学实体、疾病实体和化学物质–疾病关系3种类型。	https://paperswithcode.com/dataset/bc5cdr
PMC-15M[16]	包含来自440万篇科学文献的1 500万个医学图像–文本对。	暂未公开
DDI[17]	专门用于研究和分析药物之间相互作用的医学数据集, 来源于DrugBank数据库和MedLine两个语料库。	https://github.com/isegura/DDICorpus

1.1.5　性能评估

性能评估是衡量和评价 LLM 在特定任务中表现的关键环节。该过程不仅能反映模型的准确性, 保证其可靠性, 还能为后续模型的进一步优化提供极有价值的反馈。

要实现对模型的全面评估, 结合人工评估、模型自我评估、用户反馈评估和交叉验证等评估方式至关重要。同时, 应根据具体的医疗应用场景, 选择合适的评估指标, 可以从多个维度深入分析模型的综合性能。

在医疗领域, 对模型进行多方面的深入评估尤为关键, 因为这直接关系到患者的生命安全。确保模型做出的决策不会对患者造成伤害, 是模型能否被实际应用的决定性因素。因此, 评估标准不仅需要包括准确性, 还必须考虑到实际应用, 特别是在临床环境中的有效性和安全性。

1.2　医疗大型语言模型训练方法

1.2.1　监督学习[18]

监督学习通过在标注的专用数据集上进行训练, 旨在提高模型对特定任务的准确性和可解释性。该方法适用于医学影像分析[19]和临床信息处理等任务, 但由于模型性能在很大程度上受限于训练数据的规模和质量, 这可能会导致成本增加。此外, 鉴于部分数据集中可能包含敏感信息, 必须严格考虑隐私保护和伦理问题。

1.2.2　增量调优[20]

相较于传统的全参数微调, 增量调优仅需更新模型的一小部分参数, 例如特定的权重或偏置。这种微调方式不仅能够显著地节约计算资源, 还能有效防止模型在学习新知识后忘记旧知

识。由于微调过程所需的时间和成本都得到显著降低，模型能够更快地进行迭代更新。

增量调优的工作原理是设定一个模型 $\Theta=\{w_1,w_2,\cdots,w_N\}$ 和训练数据 $\mathcal{D}$，其适应目标是生成适应后的模型 $\Theta'=\{w'_1,w'_2,\cdots,w'_M\}$。定义 $\Delta\Theta=\Theta'-\Theta$ 为在原始模型 Θ 上的操作。在传统的全参微调中，所有参数都会参与更新，即 $|\Delta\Theta|=|\Theta|$。而在增量调优中，仅会对少量参数进行调整，即 $|\Delta\Theta|\ll|\Theta|$，从而实现降低计算和存储成本的目的。

这一优势在更新迭代迅速的医疗领域尤为重要。通过增量微调，LLM 能在较短的时间内学习最新的医疗研究成果，从而在执行特定任务时提高准确性。

1.2.3　低秩适应[21]

低秩适应（low-rank adaptation, LoRA）是一种专为 LLM 设计的轻量级训练方法。该技术的核心在于通过低秩矩阵调整模型的特定权重，而不改变模型的原始权重。

假设一个预训练的权重矩阵 $\boldsymbol{W}_0\in\mathbf{R}^{d\times k}$，LoRA 的核心原理是通过低秩分解来约束更新：

$$\boldsymbol{W}_0+\Delta\boldsymbol{W}=\boldsymbol{W}_0+\boldsymbol{BA} \tag{4}$$

式中：$\boldsymbol{B}\in\mathbf{R}^{d\times r}$，$\boldsymbol{A}\in\mathbf{R}^{r\times k}$（$r$ 远小于 d 和 k 之间较小的维度）。在训练过程中，$\boldsymbol{W}_0$ 保持固定，不接受梯度更新，仅对低秩矩阵 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 进行训练。相较于传统的全参数微调，这种方式能够显著减少训练参数的数量，从而在提高训练效率的同时降低计算资源的消耗。此外，LoRA 还能保留模型在预训练过程中获得的知识，确保模型性能不受影响。

这些特性特别适用于医疗领域，即使是资源有限的小型医疗机构或研究院，也能借助 LoRA 快速调整 LLM，以应对特定医疗任务的需求。

1.2.4　多任务学习[22]

多任务学习（multi-task learning, MTL）能使单个模型同时在多个任务上进行学习，实现任务间有效信息的共享，从而提升在各项任务中的性能。模型还能通过 MTL 提高自身的泛化能力，更准确地处理未见过的数据类型。在更新或优化模型时，与需要重新开始的传统微调方法相比，MTL 可以在原有模型的基础上进行增量式更新，节约了训练所需的资源和时间。该方法在医学图像处理中尤为适用[23]，能有效提升分类网络在图像分割、图像分类中的性能[24]。

1.2.5　来自人类反馈的强化学习[25]

随着 LLM 规模和训练数据集的不断扩大，其生成无效或有害输出的风险也随之增加，这在临床实践中可能导致严重后果，甚至威胁到病人的生命安全。为了应对这一挑战，可以采用来自人类反馈的强化学习（reinforcement learning from human feedback, RLHF）作为优化策略，这是一种有效的方法。RLHF 可以分为反馈收集、奖励建模和策略优化 3 个阶段，即通过人类的反馈来训练奖励模型，然后对语言模型的优化提供指导，帮助模型理解容错率较低的指令，并做出更精确的决策。借助人类反馈，RLHF 可以更好地确保 LLM 的行为更能与人类目标对齐，并符合人类社会价值观和安全标准，从而提高执行复杂指令时的安全性和精确性，这对于医疗领域中的临床决策、医学图像分析和药物研发等方面尤为重要。

1.2.6　少样本学习[26]

随着 LLM 规模的持续扩大，其在分析与任务无关的少样本情境下的性能也相应增强。少样本学习使模型能够通过分析少量特定示例进行学习，从而执行类似任务。这种方法特别适用于处理罕见病例，使 LLM 能够克服训练数据量有限的限制。模型可以参考少数临床案例[27]，辅助医生做出更为合理的医疗决策。此外，零样本学习[28]通过在提示模板中加入先验知识，使模型能利用在大规模数据集上预训练获得的知识，推断出在训练阶段中未直接学习过的类别。这进一步增强了 LLM 在缺少直接经验或训练样本情况下处理图像识别[29]等特定医疗任务的能力。

1.2.7　上下文学习[26]

上下文学习允许模型通过分析输入的上下文信息来加强对特定任务的理解和执行能力，而无需依赖特定的训练和微调。这种训练策略对于具备广泛预训练知识和强大语言理解能力的 LLM 尤其有效。研究表明[30]，向模型提供与任务答案相关的上下文段落，可以显著提高其回答的准确性。

1.2.8　思维链[31]

思维链是一种能有效提升 LLM 推理能力的技术。研究显示，通过添加“让我们逐步思考”的推理提示，可以增强模型处理复杂问题的能力。此外，思维链还能提高模型的决策透明度，使决策过程更易于被理解和解释。这有助于医生和患者更好地接受模型做出的医疗决策，从而辅助医生提供更合理的治疗方案。

2　大型语言模型在医疗领域的应用

将 LLM 应用于医疗领域[32]，不仅能在文本翻译[33]、医学量表分类[34]、病历自动生成[35]等多项任务中减轻医务工作者的负担，而且还能在临床实践中提供合理的参考意见，协助医生做出更准

确的诊断。此外, LLM 在医学教育方面的作用同样不可忽视, 医学生和实习医生可以通过问答[36]的形式, 轻松地获取大量医学知识或临床案例。

2.1　在线问诊

有效沟通在线上问诊过程中起着至关重要的作用, 而准确描述病情则是实现精准治疗的关键前提。在 LLM 出现之前, 患者通常借助搜索引擎查询疾病信息并寻找治疗建议。这种方式的弊端在于缺乏交互性, 不准确的病情描述、对医学信息的误解都可能导致严重后果。如果不能够进行有效的沟通, 即使是在线咨询专家, 误诊的风险也仍然存在。

随着 LLM 的发展, 在线问诊也将迎来颠覆性的变革, 它们能够模拟医生与患者之间的对话, 为问诊过程带来更多便利。图 3 以 ChatGPT-4 为例, 展示了 LLM 如何理解并回应患者的健康询问。在分析症状后, 模型能够引导患者进行后续咨询, 并提供初步诊断和针对性的治疗建议。一项研究[37]比较了 ChatGPT 与医生在社交媒体上回答患者问题的表现, 发现 ChatGPT 所提供的高质量回答获得了更多患者的认可, 这一结果证实了 LLM 于在线医疗问诊领域的应用潜力。

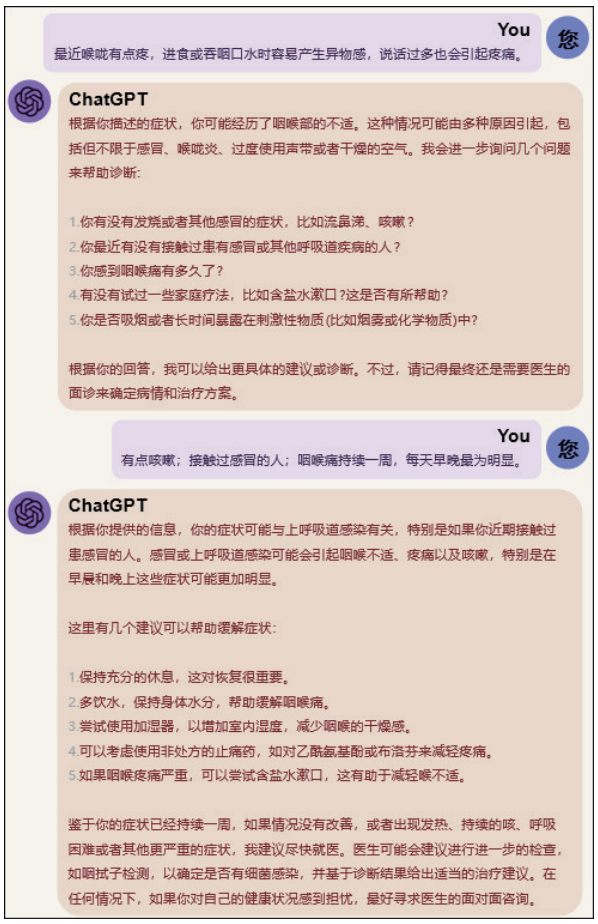

图 3　使用 ChatGPT-4 进行在线问诊

Fig. 3　Use ChatGPT-4 for online consultation

LLM 能够与患者保持 24 小时的持续沟通, 通过分析患者的病史和病情, 为其提供持续的健康监测[38]并制定个性化的治疗方案。此外, LLM 还能为患者提供关于敏感医疗问题的信息, 这些问题在当面咨询医生时通常难以提及, 例如抑郁症、生殖系统疾病或性传播疾病等。

2.2　提升公民健康素养

在公民教育方面, LLM 也扮演着重要角色。以 ChatGPT 为例, 其在回答糖尿病知识问卷时展现出的高准确性[39], 证明了该模型具备理解并分析专业医学知识的能力。这表明 LLM 可以通过问答形式, 向患者提供所需的医学信息, 从而有助于提升患者的健康素养水平。

LLM 在为患者解答医学问题时, 能将复杂的专业术语转化成通俗易懂的语言, 帮助患者提升对自身健康状况的认知[40]。这不仅有助于患者更有效地进行自我诊断和遵循医嘱, 还能在一定程度上缓解医疗资源的紧张状况。此外, LLM 能够为患者提供长期的健康管理建议, 例如合理用药、均衡饮食和适量运动等方面, 帮助患者培养更健康的生活习惯。

2.3　提供临床决策支持

在医疗领域 LLM, 能显著提升处理临床文本的效率。这些模型能快速从临床病例、实验数据和医疗检测报告中提取关键内容并检索相关信息, 极大地提高了医务人员的工作效率[41]。

此外, LLM 通过深入分析最新的医学研究和报道, 能及时向医生提供最前沿的研究成果和治疗方法。同时, LLM 通过分析大量临床案例, 能全面理解不同疾病的种类、病情进展和患者反应等多种因素, 从而为医疗工作人员提供有力的临床判断理论支持, 提高治疗的成功率。

在紧急情况下, LLM 还能实时分析患者的健康指标、医学影像和电子健康记录等关键数据。这为医生提供精确的决策支持, 确保重症患者能够及时且准确地接受治疗, 从而在最大程度上挽救患者生命。

2.4　药物研发

药物在医疗领域的重要性不言而喻, 然而新型药物的研发过程面临着周期长、成本高以及对人体和药物机制的深入理解等复杂挑战。此外, 药物的副作用和安全性是否符合规范也是研发过程中必须严格考虑的问题。在这种背景下, 人工智能技术的进步, 尤其是 LLM, 为推进药物研发提供了新的可能性[42]。

在药物研制过程中, LLM 能够有效提取医学

研究文献中的关键信息，并为研究人员提供关于已知疾病机制、药物靶点和最前沿的研究成果，从而帮助探索有前途的药物设计方向。通过分析临床实验数据，LLM 可以帮助研发人员评估新药物的有效性、安全性以及可能产生的副作用。此外，LLM 还能整合医学、生物学、计算机科学等多个领域的知识，为发现新药物靶点提供全方位的理论基础。

除了推动药物研发，LLM 还有助于实现药物再利用——将已知药物应用于其最初研发目的之外的疾病治疗。这一策略在应对紧急医疗需求时具有显著的战略价值，因为 LLM 能通过分析广泛的相关文献和数据集，发现药物与多种疾病之间的潜在关联，包括药物之间相互作用、药物与不同病症的交互作用以及药学特性等方面的分析和探索。一旦确定现有药物对其他症状可能具有疗效，LLM 将分析临床试验数据，测试药物对这些症状的效果和副作用，以便尽快投入使用，有助于节约研发新药物的时间和成本。

2.5　医学教育

LLM 在医学教育领域展现出了巨大的潜力，可以作为医学生获取医学知识和准备医学考试的强有力工具[43]。例如，一项研究[44]评估了 Multimodal GPT-4V 在包含图像的美国医学执照考试中的表现，结果显示 GPT-4V 的准确率高达 90.7%，远高于医学生考试及格门槛的 60%。另一项研究[45]评估了 ChatGPT 在韩国普通外科专业资格考试中的表现，发现它在所有评估项目中均表现出较高的准确性，并展现了其在理解复杂外科临床信息方面的强大能力，这表明了 LLM 在外科教育中具有良好的应用潜力。

此外，借助 LLM 强大的文献检索能力，用户只需输入关键词便能快速检索到所需的文献摘要和下载链接，极大地便利了学术研究。不仅如此，LLM 还为医学生和实习医生提供了与患者模拟交互的途径，创建一个无压力的训练环境，有助于提升其临床实践和医患沟通[46]等方面的能力。

3　医疗大型语言模型

LLM 在医疗领域展现出了卓越的适用性和应用价值，已被广泛应用于多种医疗任务。虽然一些先进的通用模型依靠其强大的自然语言处理能力在医疗任务中表现出色，但由于医疗领域的特殊性，包括大量的专业术语和复杂概念，对模型的专业知识有着更严格的要求。更为关键的是，某些医疗决策常涉及到患者的生命安全，要求模型在执行相关任务时必须极为准确。因此，针对医疗领域对 LLM 进行专门训练对其应用至关重要。

本节汇总了若干经过医疗领域专门训练的 LLM，如表 2 所示。尽管 GPT-4 未经过医疗领域的专门训练，但由于其在多项医学任务中的出色表现，因此也将其包括在讨论范围之内。

表 2　医疗大型语言模型

Table 2　LLM in healthcare

模型名称	模型基底	训练策略	模型优势
GPT-4	Transformer	预训练、微调	广泛适用于医疗任务
BioBERT	BERT	预训练	擅长文本挖掘、医学问答
Med-PaLM	Flan-PaLM	指令提示	擅长医学问答
Med-PaLM 2	PaLM 2	微调、指令提示	擅长医学问答、长文本对话
BiomedGPT	BART	预训练、微调	擅长关系提取、医学问答、文本分类
LLaVA-Med	LLaVA	生物医学概念对齐、微调	擅长图像理解和推理、视觉问答
MedAlpaca	LLaMA	微调（LoRA）	擅长医学问答，部署难度低
ChatDoctor	LLaMA	微调、加入在线知识检索功能	擅长医疗诊断
DoctorGPT	Baichuan2	预训练、微调（LoRA）	擅长中文医学问答，部署难度低
DoctorGLM	ChatGLM	微调（LoRA）、加入提示设计模块	擅长中文医学咨询，部署难度低
BenTsao	LLaMA	微调	擅长中文医学问答，部署难度低
HuatuoGPT	LLaMA	微调、基于人工智能反馈的强化学习	擅长中文医学问答

3.1　GPT-4

GPT-4 采用 Transformer 架构，并沿用了前代模型的预训练和微调策略。作为通用 LLM，GPT-4 能凭借其强大的语言处理能力，即使未经针对医疗问题的专门训练，也在多项医疗测试中展现出接近或超越专家级别的能力。

在 Nori 等[47] 的研究中, GPT-4 在美国医学执照考试中的自我评估和样本考试部分分别达到了 86.65% 和 86.7% 的准确率, 不仅超过了考试及格线, 而且显著优于其前代模型 GPT-3.5 的 53.61% 和 58.78%。此外, 该研究还评估了 GPT-4 在 MedQA、PubMedQA、MedMCQA 和 MMLU 4 个数据集上的表现, 发现其准确率能与人类专家媲美。

除医学问答外, GPT-4 在临床诊断和疾病分析等方面同样表现出色。一项研究[48] 采用新英格兰医学杂志进行的评估显示, GPT-4 在无选项条件下的诊断类别中取得了 89% 的准确率, 而在有选项情况下, 其准确率更是高达 98%。另一项研究[49] 评估了 GPT-4 对 424 份癌症 CT 的分析能力, 研究结果表明, 在提取病变参数和识别转移性疾病方面, GPT-4 分别达到了 98.6% 和 98.1% 的极高准确率。

3.2　BioBERT

BioBERT[5] 模型基于 BERT(bidirectional encoder representations from Transformers)[50], 利用 PubMed 中的医学文献进行预训练, 这增强了其对复杂医学文本的理解能力, 从而提高处理医学文本的效率和准确性。

该模型在命名实体识别、关系提取和问答 3 项医学文本挖掘的主流任务中表现出色。一项研究[51] 对 BioBERT 处理生物医学问答任务的能力进行了评估。在该研究中, BioBERT 首先采用 SQuAD 和 SQuAD 2.0 数据集进行预训练, 随后通过 BioASQ 数据集进行了微调, 从而显著提升了模型在医学问答方面的性能。最终, BioBERT 在第七届 BioASQ 挑战赛中取得了最佳表现。

3.3　Med-PaLM

Med-PaLM[52] 是谷歌公司研发的一种医疗 LLM。它在 Flan-PaLM 的基础上通过使用指令提示进行优化。具体来说, 研究人员首先使用软提示作为跨医学数据集的统一引导, 随后添加了专门的提示, 如与特定任务相关的指令、少样本示例以及具体问题或情景, 为模型提供了一个专业的思维框架。这种方法显著增强了模型的理解能力, 和精确处理医疗需求的能力, 使其能提供更高质量的医疗答案。

此外, 谷歌研究团队还制定了一个名为 MultiMedQA 的多元化评估基准[52], 它由 MedQA、MedMCQA、PubMedQA、LiveQA、MedicationQA、MMLU 和 HealthSearchQA 7 个数据集构成, 旨在对模型的临床应用和医学问答能力进行全面评估。Flan-PaLM 在 MedQA、MedMCQA 和 PubMedQA 3 个评估基准上的表现分别达到了 67.6%、57.6% 和 79.0%。在多领域的 MMLU 数据集上, Flan-PaLM 同样取得了优异成绩, 尤其是在临床知识和专业子集上的准确率高达 80.4% 和 83.8%。

该研究还对模型的输出结果进行了人工评估, 从 HealthSearchQA、LiveQA 和 MedicationQA 中随机抽取问题, 并让临床医生提供专家回答作为对照。结果显示, Med-PaLM 在各项评估标准上显著优于 Flan-PaLM, 并且其表现接近医学专家水平。

3.4　Med-PaLM 2

为了缩小 LLM 与医学专家在问答准确率上的差距, 谷歌研究团队基于 PaLM 2[53] 模型, 通过结合针对性微调和指令提示技术, 推出了新一代的医疗 LLM, 名为 Med-PaLM 2[54]。与早期的 Med-PaLM 相比, Med-PaLM 2 在多个问答数据集上的表现获得显著提升, 其性能接近甚至超越医学专家水平。

研究 [54] 对 Med-PaLM 2 在 MedQA、PubMedQA 和 MedMCQA 数据集上的性能进行了评估, 模型在这些数据集上分别达到了 86.5%、81.8% 和 72.3% 的准确率。此外, Med-PaLM 2 在 MMLU 临床主题的各项测试中也取得了高达 90% 的出色成绩。这些研究结果表明, Med-PaLM 2 具备处理多项选择题和长篇幅医学问答的强大能力, 这不仅反映了其在临床决策支持和医疗咨询时的高准确性, 还有助于医生制定合理的治疗方案。

此外, 研究中还添加了两个对抗性问题数据集, 用以评估模型在处理可能产生有害或带有偏见问题时的能力。结果显示, Med-PaLM 2 生成的答案在潜在风险方面明显低于 Med-PaLM, 其更高的安全性为患者健康提供更强保障。

3.5　BioGPT

BioGPT[55] 是一款基于 Transformer 架构的语言模型, 由大量的生物医学文献训练生成, 可以高效生成和挖掘医学文本。

研究采用多个数据集对 BioGPT 进行综合评估。在端到端关系提取任务中, BioGPT 在 BC5CDR、KD-DTI 和 DDI 数据集上的准确率分别为 46.17%、38.42% 和 40.76%。同时, 在 PubMedQA 的医学问答和 HoC 的文本分类任务中, BioGPT 分别取得了 78.2% 和 85.12% 的准确率, 显示了模型在处理问答和分类任务时的强大能力。此外, 在生物医学文本生成任务中, BioGPT 也展现出了优异的性能。

3.6　BiomedGPT

BiomedGPT[56] 是一个基于 BART 的通用医疗

模型，其采用 Transformer 架构，并通过多种类型的生物医学数据进行预训练，在 25 个医学数据集上进行微调。尽管 BiomedGPT 是一个轻量级的医疗 LLM，但其性能与其他先进的医疗 LLM 相比仍然具有很强的竞争力。此外，完全开源是 BiomedGPT 的另一大优势，既便于用户使用，又有利于模型的进一步开发和应用。

BiomedGPT 能与其他 LLM 相媲美的关键在于：它通过相关的训练数据进行细化训练，确保模型与特定医学问题的高度对齐，从而提高了在实际应用场景中的可靠性。

3.7　LLaVA-Med

LLaVA-Med[57] 是一款首次将多模态指令与生物医疗领域相结合的模型。具体而言，研究团队使用 GPT-4 分析了来自 PMC-15M 数据集的 1500 万个生物医疗领域图像及其文本描述，从而生成了与图像–文本对相关的指令跟踪数据。在 LLaVA-Med 的训练策略中，首先将图像–文本对与预训练中的医学概念进行对齐，随后进行模型的端对端指令调优，最后对下游数据集进行微调。

模型的微调过程中，使用了由 GPT-4 生成的指令遵循数据进行训练，让模型学习开放式对话，以提高其在理解、处理生物医学视觉任务的能力。这种训练方式使 LLaVA-Med 能在 15 h 内完成训练。

研究人员从 PMC-15M 中随机选取 193 个问题，构建了一个包含对话和详细描述的数据集，用于评估 LLaVA-Med 处理多模态任务的能力。结果显示，在使用 GPT-4 作为评估标准的情况下，使用 10000 条指示数据训练的 LLaVA-Med 获得了 39.9% 的总体相对得分，高于原版本 LLaVA 的 36.1%。当训练数据量增加至 60 000 条时，LLaVA-Med 的整体得分提升至 49.4%。而在加入内联提及后，模型的理解能力得到进一步增强，最终得分达到 50.2%。

该研究中使用 3 个生物医学视觉问答数据集对 LLaVA-Med 进行了性能评估。对于闭集问题，LLaVA-Med(改进自 LLaVA) 在 VQA-RAD、SLAKE 和 PathVQA 3 个数据集上的表现分别达到了 84.19%、85.34% 和 91.21%。LLaVA-Med 的其他两种变体的性能也均优于 LLaVA 的 65.07%、63.22% 和 63.20%。然而，在开放集问题上，LLaVA-Med 仅在 SLAKE 数据集上表现最佳，而在其他两个数据集上的性能出现明显下降。

值得注意的是，尽管 LLaVA-Med 的训练数据中不包含中文，但它仍能正确理解中文问题，这可能归功于其在训练过程中从 LLaMA 获得的多语言知识，从而能够有效实现跨语言的零样本迁移。

3.8　MedAlpaca

MedAlpaca[58] 是一款完全开源的医疗对话模型，能够有效保护用户隐私并防止医学数据泄露。该模型基于 LLaMA，为适应有限的计算资源，采用了一种高效训练策略，对 7B(70 亿)和 13B(130 亿)两种参数规格的 LLaMA 进行微调。通过 LoRA 技术对模型的权重进行更新，使其适应特定任务。此外，模型在训练过程中还使用了 8 位矩阵乘法和 8 位优化器对前馈和注意力投影层进行优化，这种方法与 LoRA 结合使用时能进一步降低内存和计算需求。

研究团队使用了美国医学执照考试的 3 个步骤对 MedAlpaca 的性能进行评估。MedAlpaca 13B 在考试第一、第二、第三步中分别取得了 47.3%、47.7%、60.2% 的准确率，明显高于参数量较少的 MedAlpaca 7B 模型，后者在评估中的准确率分别为 29.7%、31.2%、39.8%。

值得注意的是，尽管使用 LoRA 和 8 位精度可以有效减少模型训练的时间和资源消耗，但这也可能会导致模型精度下降。例如，在仅使用 LoRA 的情况下，MedAlpaca 13B 在美国医学执照考试 3 个步骤的准确率分别下降至 25.0%、25.5%、25.5%。而当 LoRA 与 8 位精度结合使用时，模型的准确率则会进一步降至 18.9%、30.3% 和 28.9%。

3.9　ChatDoctor

ChatDoctor[12] 是一款基于 LLaMA 的医疗模型，训练时使用了 Alpaca 的指令跟踪数据。该模型通过对 10 万个包含丰富医学专业知识的真实医患对话数据进行微调，从而能更精准地理解患者的需求，并提供更合理的诊断和情感建议。此外，ChatDoctor 还增加了自主知识检索功能，可以实时查询维基百科或疾病数据库中的医疗信息，为患者提供更具有专业化的回答。

为评估 ChatDoctor 的性能，文献 [12] 将其与 ChatGPT 进行了比较。他们选取了来自名为 iCliniq 的在线医疗咨询平台上的问题，并以真实医生的答案作为参考基准。该研究使用 BERT 分数对两个模型进行量化比较。实验结果表明，ChatDoctor 无论是在准确率、召回率和 F_1 分数上，均优于 ChatGPT。

3.10　DoctorGPT

DoctorGPT[59] 是一个针对医学知识问答设计的 LLM，它以 Baichuan2 为基础模型。在预训练

阶段, 模型使用维基医学数据获取了丰富的医学知识, 然后使用了约两百万条指令数据进行监督式微调。这种训练方法使 DoctorGPT 能更好地适用于中文环境, 从而在中文医学问答任务中表现出色。

在整个训练过程中, DoctorGPT 都使用了 LoRA, 这不仅提高了模型的训练效率, 还显著减少了存储空间需求, 使得模型能在 32 GB 的 GPU 上完成预训练和微调。此外, 通过采用 INT8 量化模型在进一步提升推理速度的同时, 还降低了部署成本, 便于医疗机构进行部署和使用。

为了全面评估 DoctorGPT 的性能, 研究团队选择了 ChatGLM-6B、ChatGLM2-6B、Baichuan-7B 和 Baichuan2-7B 等基线模型进行比较, 使用 BLEU、GLEU、ROUGE 以及 Distinct 等多种评估指标, 并在 Huatuo-26M 和 cMedQA2 这两个中文医疗问答数据集上进行了实验。结果显示, DoctorGPT 在所有评估指标上的分数均高于其他基线模型。这表明 DoctorGPT 在处理中文医学问答任务时, 无论在理解深度还是回答质量上, 都具有显著的优势。

3.11　DoctorGLM

上海科技大学研究团队为了提高 LLM 在提供中文医疗建议时的准确性, 使用 ChatGPT 将 ChatDoctor 的数据集翻译成中文, 创建了一个中文医疗对话数据集。该数据集被用于训练作为模型基底的 ChatGLM-6B。此外, 研究团队还使用 LoRA 技术训练出了易于部署的 DoctorGLM[60]。通过使用提示设计模块来提取用户输入中的关键词, 然后根据疾病知识库生成简要的疾病说明, 有效提高了对特定疾病的识别准确率, 并为用户提供有效信息。

DoctorGLM 能够在仅使用单张 A100 GPU 的情况下, 在 13 h 内完成对 ChatGLM-6B 的微调, 并可以在 RTX 3090 等消费级 GPU 上完成推理过程, 显著降低了模型的部署门槛。这种较低的硬件需求使得大多数医院或研究机构能够负担得起训练成本, 有助于模型在不同任务中进行广泛部署。

文献 [60] 将 DoctorGLM 与 ChatGLM-6B、GPT-3.5-turbo 两款通用模型在医学问答任务上进行了比较, 从侧面评估了其性能。结果显示, 相较于其他两款模型较为笼统的回答, DoctorGLM 提供的答案则更具有针对性。

尽管 DoctorGLM 在实际的临床任务应用中还面临着挑战, 如可能会提供错误或过于武断的诊断建议、生成回答时出现内容重复、响应时间过长及难以确定合适的训练结束点, 但其在研究中取得的进展无疑是令人兴奋的, 为推动国内 LLM 在医疗领域的应用和普及做出了显著贡献。

3.12　BenTsao

BenTsao[61] (原名 HuaTuo) 是哈尔滨工业大学推出的一款开源的中文生物医疗领域 LLM。该模型以 LLaMA-7B 为基础, 整合了源自中文医学知识图谱[62] 中的 8 000 条指令数据, 并通过中文医学指令进行微调, 使得 BenTsao 特别适用于处理与中文医学问答相关的任务。

为全面评估模型在生成医学问答中的性能, 研究团队提出了一种名为 SUS(safety, usability, and smoothness)的评估指标[61], 其中包括安全性(模型的回应是否会对用户造成误导, 造成危害)、可用性(回答能在多大程度上准确反映医学知识)和流畅性(模型在语言生成方面的能力)。

在使用 SUS 指标进行的性能评估中, BenTsao 与 LLaMA、Alpaca 和 ChatGLM 3 款模型进行了对比。结果显示, 尽管 BenTsao 的安全性方面的得分略低于 LLaMA, 但在可用性和流畅性两项指标上均显著优于 LLaMA(1.21 和 1.58), 分别取得了 2.12 和 2.47 的最高得分, 超越了所有参与评估的其他模型。

值得注意的是, BenTsao 的训练资源需求相对较低, 预计单张 RTX 3090 或 4090 这样的消费级显卡即可满足模型训练的基本要求。训练门槛的降低有助于更多研究人员参与到 LLM 的研究中, 从而推动 LLM 在医疗领域的发展和进步。

3.13　HuatuoGPT

由香港中文大学研发的 HuatuoGPT[63] 是一款专为中文医疗咨询而设计的 LLM。该模型在监督微调阶段同时使用了 ChatGPT 生成的数据和来自医生的真实交互数据。为了在最大程度上发挥这两种数据的优势, HuatuoGPT 采用了人工智能反馈的训练方法, 有效克服了 ChatGPT 在医疗咨询场景中的多种局限性, 如拒绝诊断和开药、难以进行类似医生的交互式诊断, 以及容易产生幻觉等问题。

研究团队基于 cMedQA2、webMedQA 和 Huatuo-26M 3 个中文医学数据集, 通过 BLEU、ROUGE、GLEU 等多种评估指标, 对比了 HuatuoGPT 与 GPT-3.5 以及经微调的 T5(text-to-text transfer transformer)[64]。结果显示, HuatuoGPT 在多项评估指标上均优于 GPT-3.5-turbo, 并且在某些方面甚至超越了微调后的 T5 模型。

在单轮对话自动评估中, HuatuoGPT 的表现

在所有类别上显著优于 BenTsao 和 DoctorGPT 两款中文 LLM, 并且在某些方面超过了 GPT-3.5-turbo。然而, 相较于 GPT-4, 其整体性能仍存在明显差距。在多轮对话的自动评估中, HuatuoGPT 分别以 8.72 比 5.63 和 8.42 比 7.85 的优势胜过 DoctorGLM 和 GPT-3.5-turbo。在人工评估方面, HuatuoGPT 同样展现出了优于 BenTsao、DoctorGPT 和 GPT-3.5-turbo 的性能, 仅次于 GPT-4。

4　风险和挑战

随着 LLM 在医疗领域的广泛应用, 医疗保健行业正在迎来新的篇章。然而, LLM 的使用也面临着一系列风险和挑战, 尤其在模型的可解释性、准确性、隐私保护和伦理问题等方面。本节将着重探讨在医疗领域中应用 LLM 时遇到的关键挑战, 以及可能的相应解决策略。

4.1　可解释性问题

可解释性涉及到人工智能系统进行预测或决策时的透明度和可理解性。简单来说, 它指模型的决策过程和内部工作原理能够被人类理解并信任。绝大多数 LLM 采用的 Transformer 结构, 由于其独特的自注意力机制, 通常被认为缺乏可解释性。在医疗领域, 增强对 LLM 决策的可解释性至关重要。这不仅直接关系到患者的健康, 还能提升医护人员和患者对模型决策的信任度。

尽管 LLM 具有极强的性能, 但其“黑盒”特性意味着内部机制不明确、不透明[65], 这在医疗应用中可能会造成极大的风险。Amann 等[66] 从技术、法律、医学和患者 4 个维度探讨了人工智能在医疗保健中的可解释性, 突显了其作为多维度概念的重要性。Rudin 则认为黑盒模型会在医疗保健和刑事司法等领域引发严重后果, 并主张停止解释黑盒模型, 转而采用可解释的模型[67]。

为提升模型的可解释性, 一种方法是将复杂模型的知识提炼到简单的小型模型中[68]。这种方法能在不牺牲模型性能的前提下, 提高其可解释性。另一种增强可解释性的策略是通过采用可解释性框架来理解模型的决策行为, 例如使用分类法进行评估[69]、构建基于决策理论的可解释人工智能系统[70]。有研究提出[71] 了一种名为沙普利加性解释的统一解释框架, 该方法能够直观、准确且灵活地解释各种类型模型的决策。此外, Alammar 等[72] 提出的开源库 Ecco, 提供了一套可以分析并可视化基于 Transformer 架构模型内部机制的方法, 以达到增加决策透明度的目的。

4.2　准确性问题

模型的准确性不仅高度依赖于训练数据的质量和规模, 还会受训练数据代表性、无关特征干扰以及过拟合等多种因素影响。此外, 研究表明[73], 模型在处理特定类型数据或长尾知识时性能会显著下降。若无法达到医疗应用对数据集的高标准要求, 可能导致模型的答案产生较高错误率, 进而在临床上做出错误诊断或提供不恰当的治疗建议。即使训练数据完全符合要求, 模型也可能因泛化能力不足而在面对罕见病例或者变种疾病时陷入困境。

在生成文本的过程中, LLM 有时会产生 “人工幻觉”。具体来说, 这种现象是指模型在处理输入时, 会生成看似真实且逻辑合理的回答, 但实际上却是缺乏根据的虚构信息。这种现象在模型生成长篇回答时尤为严重[74], 主要是因为 LLM 在进行预测时依赖于数据中的统计关系, 而不是真正理解概念之间的本质联系或逻辑, 因此缺乏进行复杂判断的能力[75]。不仅如此, 还有研究表明[76], 模型生成的幻觉可能会被其他系统无意中接受, 并进一步复制和传播, 加剧了错误信息的扩散, 形成所谓的雪球效应, 即使最先进的模型也难以完全避免这种情况。

在医疗领域中, 模型准确性至关重要。过度依赖 LLM 可能导致误导性治疗决策、不准确诊断建议和虚构数据等一系列潜在医疗风险。ALKAISSI 等[77] 通过对 ChatGPT 在撰写医学和非医学领域的短段落测试中观察到, ChatGPT 提供的内容混合了真实与虚构的信息。另一项研究[78] 也证实了 LLM 会做出容易让人信服的陈述, 然而产生的医学摘要内容与事实并不相符。相较于人工编写摘要, LLM 难以识别并提取出关键信息, 尤其是在处理较长输入时更容易出错。

文献 [79] 指出用 LLM 生成临床摘要的其他弊端。由于 LLM 具有概率性特征, 因此每次生成的摘要可能会有所差异, 即便是细微的提示差异也会对输出结果造成影响。而由“人工幻觉”导致的错误决策可能会误导医生做出不恰当的治疗诊断。

为确保 LLM 在医疗应用中的准确性, 除了需要保证高质量的训练数据之外, 还可采用 ReFeed[80]、知识注入[81] 或提示策略[82] 等方法来减少幻觉的产生, 提高模型在实际应用中的安全性和可靠性。

4.3　隐私问题

医疗 LLM 在训练、与用户互动及病情分析等过程中会接触到大量具有隐私性的敏感数据或个

人信息。LLM 复杂且不透明的工作机制导致用户隐私存在着很大的泄露风险。不仅如此, 苏黎世联邦理工学院的研究[83] 证实了 LLM 能凭借强大的推理能力, 通过给定文本推断出用户的广泛个人信息, 如性别、年龄甚至所处地理位置。这意味着模型不仅能评估患者的健康状况, 而且还能间接推断出用户的详细身份信息。

此外, 不恰当的数据存储方式可能导致患者的病史、诊断记录或个人隐私信息因黑客攻击而泄露, 进而对患者的身心健康造成二次伤害。

为保护用户隐私, 加强对医疗 LLM 的监管至关重要[84]。可采用联邦学习[85]、弱监督学习、差分隐私[86] 或者使用无需数据共享的技术[87] 等方法, 在不牺牲模型准确性的前提下, 降低隐私泄露的风险。

4.4　伦理问题

当 LLM 应用于医疗领域时, 常伴随着偏见和歧视的产生, 可能会涉及性别、种族、职业等多种因素。一项针对 GPT-4 的研究[88] 调查了其在不同临床案例中对种族和性别的偏见问题, 发现当 LLM 使用有偏见的数据进行训练时, 可能会学习并放大这些偏见。例如, GPT-4 会根据刻板印象, 对某些人群在特定流行病上的患病率产生偏见, 例如夸大亚洲人患乙肝或黑人患结节病的可能性。并且会因为这些刻板印象, 向不同的患者群体提供不同的诊断建议。即使未直接提供患者的敏感信息, 模型仍可能通过分析已知数据, 推断出会对患者产生偏见的因素[83]。

随着 LLM 的参数规模不断扩大, 它们逐渐具备了强大的自主决策能力。然而, 这些由模型做出的决策是否能完全符合人类的价值观和社会伦理标准, 仍然需要更深入的探讨和研究。在处理复杂的医疗决策时, 专业医护人员通常还会考虑到患者的心理状态、经济条件、文化和信仰等非技术性因素。而对 LLM 来说, 充分理解人类的情感、同理心以及医患关系仍然是一道难题, 这使得其难以完全模拟医护人员与患者之间的对话, 从而加剧患者的焦虑和负面情绪, 甚至严重损害患者的精神状态和身心健康。此外, 人工智能领域的法律不完善可能会造成责任归属问题, 包括由 LLM 做出错误诊断而导致的医疗事故或隐私泄露等情况。

为有效缓解 LLM 在医疗领域引发的伦理问题, 首先需要加强对模型潜在风险的了解[89], 并建立健全的法律法规和监管体系[84]。其次, 采用价值敏感设计[90] 等方法, 确保 LLM 与人类的伦理和价值观保持一致[91], 并减少带有偏见和歧视性的输出。同时, 提高模型的可解释性也至关重要, 这有助于医护人员及时纠正 LLM 在临床应用中的偏见行为。最后, 应该加强医护人员、人工智能科学家和法律专家之间的跨学科交流合作[92], 促进多领域共同探讨 LLM 在医学中的伦理道德问题。

5　未来发展趋势

5.1　多模态大型语言模型

医疗领域跨多种模态的特征, 使得多模态 LLM[93] 成为必然的发展趋势。多模态技术能够使 LLM 突破了只能处理文本数据的局限, 可有效解析多种类型的数据, 如图像[94]、音频[95] 和视频等, 并根据任务需求生成相应数据格式的输出。如图 4 所示, 模型采用文本、图像、音频和视频等多种模态的数据进行训练, 得到多模态 LLM, 从而可以利用该模型产生文本、图像、音频和视频等不同模态的输出。通过采用语言–图像对比学习的预训练方法, 模型能更好地理解和处理与文本、图像相关的医疗任务, 从而提高模型执行跨模态任务的能力。这不仅能为患者提供更精确的医疗服务, 还极大地拓展了 LLM 在医疗健康领域中的应用范围[96]。

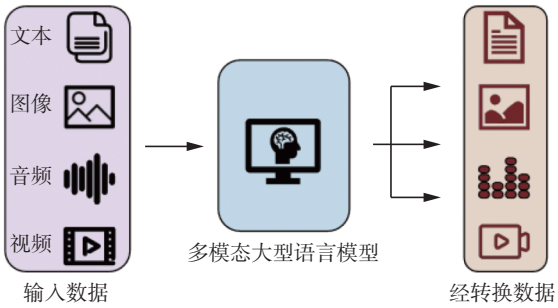

图 4　多模态大型语言模型

Fig. 4　Multimodal large language model

5.2　压缩模型参数

权衡模型的性能与参数规模, 是未来应用中的关键方向。尽管模型性能会随着参数规模增大而提高, 但这也对计算资源提出了更高要求, 可能需要数量高达上万的 GPU, 导致仅有少数科技巨头公司有能力承担高昂的训练成本。因此, 在保持模型性能的前提下, 采用适当的方法压缩模型参数是十分必要的[97]。例如, 通过采用知识蒸馏[98]、权重共享[99] 和网络剪枝[100] 等技术, 可以有效实现从大型语言模型向“小型语言模型”的转变。这不仅有助于降低成本、保护环境, 还能降低产生过拟合的风险、降低调试难度、提高模型

运行效率和泛化能力。这项改进能使模型在特定的工作中更高效地发挥作用。

5.3　基于大型语言模型的智能体

最新研究显示[101], LLM 在推理和规划问题方面存在局限性。尤其是在医疗领域, 不擅长制定详细的计划可能会严重影响其在下游任务中的应用效果。因此, 开发在医疗环境中基于 LLM 的智能体尤为关键。智能体指能在某种环境中自主做出决策并采取行动的实体[102], 它们能够结合外部数据或工具, 为特定医学任务提供创新解决方案。智能体可以优化与患者的交互, 使 LLM 更好地理解患者个体差异, 从而提供更准确、人性化的诊断建议; 通过访问外部医学知识库, 智能体还可以增强 LLM 提供临床决策的准确性; 此外, 在无需人工介入的情况下, LLM 可以借助智能体自动处理和分析大量医疗数据, 进而分析和预测患者病情的发展趋势。

5.4　提高模型适用性

通用型医疗人工智能[103] 是医疗 LLM 的发展新方向。它可以根据用户的具体需求调整并执行多样化的任务, 大幅增强了模型的应用灵活性, 使其能广泛应用于医疗行业的多个领域。此外, 通用型医疗人工智能通过自监督学习方式, 利用大量未经标注的数据进行学习。这不仅缩短了模型的训练时间, 还能有效节约训练成本。

5.5　与医疗设备集成

穿戴式医疗设备结合 LLM 强大的分析能力, 能为患者提供实时的健康监测和评估。通过持续监测患者的健康指标, 并将这些数据实时传输至 LLM 分析, 系统可及时识别潜在的健康问题, 起到预防疾病的作用。医生也可以通过这种集成方式, 为患者提供远程医疗服务, 极大提高医疗服务的效率和时效性。这不仅能优化医疗资源的分配, 还能为患者带来更便利的个性化医疗体验。

6　结束语

LLM 的兴起标志着人工智能时代进入了一个崭新的阶段, 人们对这类模型与各行各业相融合的兴趣愈发强烈。LLM 在医疗诊断、药物研发、临床决策支持以及辅助医学教育等方面展现出了巨大潜力, 对推动人类医疗保健事业的发展具有深远意义。这预示着它们不仅能有效提高医疗服务的质量, 还有助于实现医疗资源配置的优化, 从而提高整个医疗行业的效率。随着研究的不断深入, LLM 有望在医疗领域中发挥更加关键的作用。

尽管如此, LLM 在实际应用中仍面临着诸多挑战, 并且目前仍存在着许多局限性, 因而无法完全替代专业医疗人员, 需要在严格控制的条件下使用, 以确保其安全性和可靠性[104]。医生、护士在实际使用 LLM 辅助工作的过程中, 应对模型提供的建议进行批判性的分析, 并结合自身的专业知识做出最终决策。同时, 还必须警惕这些模型可能对患者造成的潜在威胁, 并严格遵守相关的法律规范和伦理标准。此外, 进一步加强不同学科之间的深入合作对促进 LLM 的发展至关重要。

参考文献:

[1] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceeding of the 3lth International Conference on Neural Information Processing Systems. New York: ACM, 2017: 6000−6010.

[2] WOLF T, DEBUT L, SANH V, et al. Transformers: state-of-the-art natural language processing[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Stroudsburg: Association for Computational Linguistics, 2020: 38−45.

[3] ZHANG Tianfu, HUANG Heyan, FENG Chong, et al. Enlivening redundant heads in multi-head self-attention for machine translation[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2021: 3238−3248.

[4] HAN Xu, ZHANG Zhengyan, DING Ning, et al. Pre-trained models: past, present and future[J]. AI open, 2021, 2: 225−250.

[5] LEE J, YOON W, KIM S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining[J]. Bioinformatics, 2020, 36(4): 1234−1240.

[6] BODENREIDER O. The unified medical language system (UMLS): integrating biomedical terminology[J]. Nucleic acids research, 2004, 32(Suppl): D267−D270.

[7] JOHNSON A E W, POLLARD T J, SHEN Lu, et al. MIMIC-III, a freely accessible critical care database[J]. Scientific data, 2016, 3: 160035.

[8] YIN Yanshen, ZHANG Yong, LIU Xiao, et al. HealthQA: a Chinese QA summary system for smart health[C]//International Conference on Smart Health. Cham: Springer, 2014: 51−62.

[9] LI Jianquan, WANG Xidong, WU Xiangbo, et al. Huatuo-26M, a large-scale Chinese medical QA dataset

[EB/OL]. (2023−05−02)[2023−12−12]. http://arxiv.org/abs/2305.01526.

[10] HE Junqing, FU Mingming, TU Manshu. Applying deep matching networks to Chinese medical question answering: a study and a dataset[J]. BMC medical informatics and decision making, 2019, 19(Suppl2): 52.

[11] ZHANG Sheng, ZHANG Xin, WANG Hui, et al. Multi-scale attentive interaction networks for Chinese medical question answer selection[J]. IEEE access, 2018, 6: 74061−74071.

[12] LI Yunxiang, LI Zihan, ZHANG Kai, et al. ChatDoctor: a medical chat model fine-tuned on a large language model meta-AI (LLaMA) using medical domain knowledge[J]. Cureus, 2023, 15(6): e40895.

[13] ZENG Guangtao, YANG Wenmian, JU Zeqian, et al. MedDialog: large-scale medical dialogue datasets[C]// Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg: Association for Computational Linguistics, 2020: 9241−9250.

[14] HOU Yutai, XIA Yingce, WU Lijun, et al. Discovering drug-target interaction knowledge from biomedical literature[J]. Bioinformatics, 2022, 38(22): 5100−5107.

[15] LI Jiao, SUN Yueping, JOHNSON R J, et al. BioCreative V CDR task corpus: a resource for chemical disease relation extraction[J]. Database, 2016, 2016: baw068.

[16] ZHANG Sheng, XU Yanbo, USUYAMA N, et al. BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs [EB/OL]. (2023−03−02)[2024−01−01]. http://arxiv.org/abs/2303.00915.

[17] HERRERO-ZAZO M, SEGURA-BEDMAR I, MARTÍNEZ P, et al. The DDI corpus: an annotated corpus with pharmacological substances and drug-drug interactions[J]. Journal of biomedical informatics, 2013, 46(5): 914−920.

[18] NASTESKI V. An overview of the supervised machine learning methods[J]. Horizons b, 2017, 4: 51−62.

[19] MURALI N, KUCUKKAYA A, PETUKHOVA A, et al. Supervised machine learning in oncology: a clinician's guide[J]. Digestive disease interventions, 2020, 4(1): 73−81.

[20] DING Ning, QIN Yujia, YANG Guang, et al. Delta tuning: a comprehensive study of parameter efficient methods for pre-trained language models[EB/OL]. (2022−03−15)[2024−01−01]. http://arxiv.org/abs/2203.06904.

[21] HU E J, SHEN Yelong, WALLIS P, et al. LoRA: low-rank adaptation of large language models[EB/OL]. (2021−10−16)[2024−01−01]. http://arxiv.org/abs/2106.09685.

[22] ZHANG Yu, YANG Qiang. A survey on multi-task learning[J]. IEEE transactions on knowledge and data engineering, 2022, 34(12): 5586−5609.

[23] JING Baoyu, XIE Pengtao, XING E. On the automatic generation of medical imaging reports[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2018: 2577−2586.

[24] 刘侠, 吕志伟, 王波, 等. 联合超声甲状腺结节分割与分类的多任务方法研究[J]. 智能系统学报, 2023, 18(4): 764−774.
LIU Xia, LYU Zhiwei, WANG Bo, et al. Multi-task method for segmentation and classification of thyroid nodules combined with ultrasound images[J]. CAAI transactions on intelligent systems, 2023, 18(4): 764−774.

[25] CHRISTIANO P, LEIKE J, BROWN T B, et al. Deep reinforcement learning from human preferences[EB/OL]. (2017−07−13)[2024−01−01]. http://arxiv.org/abs/1706.03741.

[26] BROWN T B, MANN B, RYDER N, et al. Language models are few-shot learners[EB/OL]. (2020−07−22)[2024−01−01]. http://arxiv.org/abs/2005.14165.

[27] GUO Zijun, AO Sha, AO Bo. Few-shot learning based oral cancer diagnosis using a dual feature extractor prototypical network[J]. Journal of biomedical informatics, 2024, 150: 104584.

[28] PALATUCCI M, POMERLEAU D, HINTON G, et al. Zero-shot learning with semantic output codes[C]// Proceedings of the 23rd International Conference on Neural Information Processing Systems. New York: ACM, 2009: 1410−1418.

[29] 翟永杰, 张智柏, 王亚茹. 基于改进 TransGAN 的零样本图像识别方法[J]. 智能系统学报, 2023, 18(2): 352−359.
ZHAI Yongjie, ZHANG Zhibai, WANG Yaru. An image recognition method of zero-shot learning based on an improved TransGAN[J]. CAAI transactions on intelligent systems, 2023, 18(2): 352−359.

[30] KANDPAL N, DENG Haikang, ROBERTS A, et al. Large language models struggle to learn long-tail knowledge[EB/OL]. (2022−11−15)[2024−01−01]. http://arxiv.org/abs/2211.08411.

[31] KOJIMA T, GU S S, REID M, et al. Large language models are zero-shot reasoners[EB/OL]. (2022−10−02)[2024−01−01]. http://arxiv.org/abs/2205.11916.

[32] 马武仁, 弓孟春, 戴辉, 等. 以 ChatGPT 为代表的大语

言模型在临床医学中的应用综述[J]. 医学信息学杂志, 2023, 44(7): 9−17.
MA Wuren, GONG Mengchun, DAI Hui, et al. A comprehensive review of the applications of large language models in clinical medicine with ChatGPT as a representative[J]. Journal of medical informatics, 2023, 44(7): 9−17.

[33] 王和私, 马柯昕. 人工智能翻译应用的对比研究: 以生物医学文本为例[J]. 中国科技翻译, 2023, 36(3): 23−26.
WANG Hesi, MA Kexin. The application of artificial intelligence in biomedical text translation: a comparative study[J]. Chinese science & technology translators journal, 2023, 36(3): 23−26.

[34] 郝洁, 彭庆龙, 丛山, 等. 基于提示学习的医学量表问题文本多分类研究[J]. 中国循证医学杂志, 2024, 24(1): 76−82.
HAO Jie, PENG Qinglong, CONG Shan, et al. A Few-shot classification method for TCM medical records based on hybrid prompt learning[J]. Chinese journal of evidence-based medicine, 2024, 24(1): 76−82.

[35] 姜会珍, 胡海洋, 马琏, 等. 基于医患对话的病历自动生成技术研究[J]. 中国数字医学, 2021, 16(10): 36−40.
JIANG Huizhen, HU Haiyang, MA Lian, et al. Research on automatic generation of electronic medical record based on doctor-patient dialogue[J]. China digital medicine, 2021, 16(10): 36−40.

[36] 杨波, 孙晓虎, 党佳怡, 等. 面向医疗问答系统的大语言模型命名实体识别方法[J]. 计算机科学与探索, 2023, 17(10): 2389−2402.
YANG Bo, SUN Xiaohu, DANG Jiayi, et al. Named entity recognition method of large language model for medical question answering system[J]. Journal of frontiers of computer science and technology, 2023, 17(10): 2389−2402.

[37] AYERS J W, POLIAK A, DREDZE M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum[J]. JAMA internal medicine, 2023, 183(6): 589−596.

[38] KHANBHAI M, WARREN L, SYMONS J, et al. Using natural language processing to understand, facilitate and maintain continuity in patient experience across transitions of care[J]. International journal of medical informatics, 2022, 157: 104642.

[39] NAKHLEH A, SPITZER S, SHEHADEH N. ChatGPT's response to the diabetes knowledge questionnaire: implications for diabetes education[J]. Diabetes technology & therapeutics, 2023, 25(8): 571−573.

[40] 陈一鸣, 刘健, 从承志, 等. 强直性脊柱炎患者与 Chat GPT 的对话实验: 患者教育的新方式[J]. 风湿病与关节炎, 2023, 12(7): 37−43.
CHEN Yiming, LIU Jian, CONG Chengzhi, et al. Dialogue experiment between patients with ankylosing spondylitis and ChatGPT: a new way of patient education[J]. Rheumatism and arthritis, 2023, 12(7): 37−43.

[41] JUNG H, KIM Y, CHOI H, et al. Enhancing clinical efficiency through LLM: discharge note generation for cardiac patients[EB/OL]. (2024−04−08)[2024−05−01]. http://arxiv.org/abs/2404.05144.

[42] 余泽浩, 张雷明, 张梦娜, 等. 基于人工智能的药物研发: 目前的进展和未来的挑战[J]. 中国药科大学学报, 2023, 54(3): 282−293.
YU Zehao, ZHANG Leiming, ZHANG Mengna, et al. Artificial intelligence-based drug development: current progress and future challenges[J]. Journal of China pharmaceutical university, 2023, 54(3): 282−293.

[43] 刘月嫦, 陈紫茹, 杨敏, 等. 国内外大语言模型在临床检验题库中的表现[J]. 临床检验杂志, 2023, 41(12): 941−944.
LIU Yuechang, CHEN Ziru, YANG Min, et al. Performance of domestic and international large language models in question banks of clinical laboratory medicine[J]. Chinese journal of clinical laboratory science, 2023, 41(12): 941−944.

[44] YANG Zhichao, YAO Zonghai, TASMIN M, et al. Performance of multimodal GPT-4V on USMLE with image: potential for imaging diagnostic support with explanations[EB/OL]. (2023−10−26)[2024−01−01]. https://www.medrxiv.org/content/10.1101/2023.10.26.23297629v3.

[45] OH N, CHOI G S, LEE W Y. ChatGPT goes to the operating room: evaluating GPT-4 performance and its potential in surgical education and training in the era of large language models[J]. Annals of surgical treatment and research, 2023, 104(5): 269−273.

[46] DANON L M, BÄHR V, SCHIFF E, et al. Learning to establish a therapeutic doctor-patient communication: German and Israeli medical students experiencing integrative medicine's skills[J]. Social science, humanities and sustainability research, 2021, 2(4): 48.

[47] NORI H, KING N, MCKINNEY S M, et al. Capabilities of GPT-4 on medical challenge problems[EB/OL]. (2023−04−12)[2024−01−01]. http://arxiv.org/abs/2303.13375.

[48] UEDA D, WALSTON S L, MATSUMOTO T, et al.

Evaluating GPT-4-based ChatGPT's clinical potential on the NEJM quiz[J]. BMC digital health, 2024, 2(1): 4.

[49] FINK M A, BISCHOFF A, FINK C A, et al. Potential of ChatGPT and GPT-4 for data mining of free-text CT reports on lung cancer[J]. Radiology, 2023, 308(3): e231362.

[50] DEVLIN J, CHANG Mingwei, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[EB/OL]. (2018−10−11)[2024−01−01]. http://arxiv.org/abs/1810.04805.

[51] YOON W, LEE J, KIM D, et al. Pre-trained language model for biomedical question answering[M]//Communications in Computer and Information Science. Cham: Springer International Publishing, 2020: 727−740.

[52] SINGHAL K, AZIZI S, TU Tao, et al. Large language models encode clinical knowledge[J]. Nature, 2023, 620: 172−180.

[53] ANIL R, DAI A M, FIRAT O, et al. PaLM 2 technical report[EB/OL]. (2023−09−13)[2024−01−01]. http://arxiv.org/abs/2305.10403.

[54] SINGHAL K, TU Tao, GOTTWEIS J, et al. Towards expert-level medical question answering with large language models[EB/OL]. (2023−05−16)[2024−01−01]. http://arxiv.org/abs/2305.09617.

[55] LUO Renqian, SUN Liai, XIA Yingce, et al. BioGPT: generative pre-trained transformer for biomedical text generation and mining[J]. Briefings in bioinformatics, 2022, 23(6): bbac409.

[56] ZHANG Kai, ZHOU Rong, ADHIKARLA E, et al. BiomedGPT: a generalist vision-language foundation model for diverse biomedical tasks[EB/OL]. (2023−05−26)[2024−01−01]. http://arxiv.org/abs/2305.17100.

[57] LI Chunyuan, WONG C, ZHANG Sheng, et al. LLaVA-med: training a large language-and-vision assistant for biomedicine in one day[EB/OL]. (2023−06−01)[2024−01−01]. http://arxiv.org/abs/2306.00890.

[58] HAN Tianyu, ADAMS L C, PAPAIOANNOU J M, et al. MedAlpaca: an open-source collection of medical conversational AI models and training data[EB/OL]. (2023−10−04)[2024−01−01]. http://arxiv.org/abs/2304.08247.

[59] LI Wenqiang, YU Lina, WU Min, et al. DoctorGPT: a large language model with Chinese medical question-answering capabilities[C]//2023 International Conference on High Performance Big Data and Intelligent Systems. Macau: IEEE, 2023: 186−193.

[60] XIONG Honglin, WANG Sheng, ZHU Yitao, et al. DoctorGLM: fine-tuning your Chinese doctor is not a Herculean task[EB/OL]. (2023−04−17)[2024−01−01]. http://arxiv.org/abs/2304.01097.

[61] WANG Haochun, LIU Chi, XI Nuwa, et al. HuaTuo: tuning LLaMA model with Chinese medical knowledge[EB/OL]. (2023−04−14)[2024−01−01]. http://arxiv.org/abs/2304.06975.

[62] 奥德玛, 杨云飞, 穗志方, 等. 中文医学知识图谱 CMeKG 构建初探[J]. 中文信息学报, 2019, 33(10): 1−7.
ODMAA, YANG Yunfei, SUI Zhifang, et al. Preliminary study on the construction of Chinese medical knowledge graph[J]. Journal of Chinese information processing, 2019, 33(10): 1−7.

[63] ZHANG Hongbo, CHEN Junying, JIANG Feng, et al. HuatuoGPT, towards taming language model to be a doctor[C]//Findings of the Association for Computational Linguistics: EMNLP 2023. Stroudsburg: Association for Computational Linguistics, 2023: 10859−10885.

[64] COLIN R, NOAM S, ADAM R, et al. Exploring the limits of transfer learning with a unified text-to-text transformer[J]. Journal of machine learning research, 2020, 21(140): 1−67.

[65] ZHAO Haiyan, CHEN Hanjie, YANG Fan, et al. Explainability for large language models: a survey[J]. ACM transactions on intelligent systems and technology, 2024, 15(2): 1−38.

[66] AMANN J, BLASIMME A, VAYENA E, et al. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective[J]. BMC medical informatics and decision making, 2020, 20: 1−9.

[67] RUDIN C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead[J]. Nature machine intelligence, 2019, 1(5): 206−215.

[68] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network[EB/OL]. (2015−03−09)[2024−01−01]. http://arxiv.org/abs/1503.02531.

[69] DOSHI-VELEZ F, KIM B. Towards A rigorous science of interpretable machine learning[EB/OL]. (2017−03−02)[2024−01−01]. http://arxiv.org/abs/1702.08608.

[70] WANG Danding, YANG Qian, ABDUL A, et al. Designing theory-driven user-centric explainable AI[C]//Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. Glasgow: ACM, 2019: 1−15.

[71] LUNDBERG S, LEE S I. A unified approach to interpreting model predictions[EB/OL]. (2017−11−25)[2024−01−01]. http://arxiv.org/abs/1705.07874.

[72] ALAMMAR J. Ecco: an open source library for the explainability of transformer language models[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations. Stroudsburg: Association for Computational Linguistics, 2021: 249−257.

[73] PAN J Z, RAZNIEWSKI S, KALO J C, et al. Large language models and knowledge graphs: opportunities and challenges[EB/OL]. (2023−08−11)[2024−01−01]. http://arxiv.org/abs/2308.06374.

[74] YE Hongbin, LIU Tong, ZHANG Aijia, et al. Cognitive mirage: a review of hallucinations in large language models[EB/OL]. (2023−09−13)[2024−01−01]. http://arxiv.org/abs/2309.06794.

[75] 陈小平. 大模型关联度预测的形式化和语义解释研究[J]. 智能系统学报, 2023, 18(4): 894−900.
CHEN Xiaoping. Research on formalization and semantic interpretations of correlation degree prediction in large language models[J]. CAAI transactions on intelligent systems, 2023, 18(4): 894−900.

[76] ZHANG Muru, PRESS O, MERRILL W, et al. How language model hallucinations can snowball[EB/OL]. (2023−05−22)[2024−01−01]. http://arxiv.org/abs/2305.13534.

[77] ALKAISSI H, MCFARLANE S I. Artificial hallucinations in ChatGPT: implications in scientific writing[J]. Cureus, 2023, 15(2): e35179.

[78] TANG Liyan, SUN Zhaoyi, IDNAY B, et al. Evaluating large language models on medical evidence summarization[J]. NPJ digital medicine, 2023, 6: 158.

[79] GOODMAN K E, YI P H, MORGAN D J. AI-generated clinical summaries require more than accuracy[J]. JAMA, 2024, 331(8): 637−638.

[80] YU Wenhao, ZHANG Zhihan, LIANG Zhenwen, et al. Improving language models via plug-and-play retrieval feedback[EB/OL]. (2023−05−23)[2024−01−01]. http://arxiv.org/abs/2305.14002.

[81] MARTINO A, IANNELLI M, TRUONG C. Knowledge injection to Counter large language model (LLM) hallucination[M]//Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, 2023: 182−185.

[82] PAL A, SANKARASUBBU M. Gemini goes to med school: exploring the capabilities of multimodal large language models on medical challenge problems & hallucinations[EB/OL]. (2024−02−10)[2024−05−01]. http://arxiv.org/abs/2402.07023.

[83] STAAB R, VERO M, BALUNOVIĆ M, et al. Beyond memorization: violating privacy via inference with large language models[EB/OL]. (2023−11−11)[2024−01−01]. http://arxiv.org/abs/2310.07298.

[84] MESKÓ B, TOPOL E J. The imperative for regulatory oversight of large language models (or generative AI) in healthcare[J]. NPJ digital medicine, 2023, 6: 120.

[85] ZHANG Chen, XIE Yu, BAI Hang, et al. A survey on federated learning[J]. Knowledge-based systems, 2021, 216: 106775.

[86] YU Da, NAIK S, BACKURS A, et al. Differentially private fine-tuning of language models[EB/OL]. (2021−10−13)[2024−01−01]. http://arxiv.org/abs/2110.06500.

[87] SOIN A, BHATU P, TAKHAR R, et al. Multi-institution encrypted medical imaging AI validation without data sharing[EB/OL]. (2021−08−13)[2024−01−01]. http://arxiv.org/abs/2107.10230.

[88] ZACK T, LEHMAN E, SUZGUN M, et al. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study[J]. The lancet digital health, 2024, 6(1): e12−e22.

[89] WEIDINGER L, MELLOR J, RAUH M, et al. Ethical and social risks of harm from Language Models[EB/OL]. (2021−12−08)[2024−01−01]. http://arxiv.org/abs/2112.04359.

[90] 古天龙, 马露, 李龙, 等. 符合伦理的人工智能应用的价值敏感设计: 现状与展望[J]. 智能系统学报, 2022, 17(1): 2−15.
GU Tianlong, MA Lu, LI Long, et al. Value sensitive design of ethical-aligned AI applications: current situation and prospect[J]. CAAI transactions on intelligent systems, 2022, 17(1): 2−15.

[91] 刘学博, 户保田, 陈科海, 等. 大模型关键技术与未来发展方向——从 ChatGPT 谈起[J]. 中国科学基金, 2023, 37(5): 758−766.
LIU Xuebo, HU Baotian, CHEN Kehai, et al. Key technologies and future development directions of large models — Starting from ChatGPT[J]. Science foundation in China, 2023, 37(5): 758−766.

[92] ZHOU Ying, LI Zheng, LI Yingxin. Interdisciplinary collaboration between nursing and engineering in health care: a scoping review[J]. International journal of nursing studies, 2021, 117: 103900.

[93] World Health Organization. Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models[M]. Geneva: World Health Organization, 2024.

[94] ZHAO Zihao, LIU Yuxiao, WU Han, et al. CLIP in medical imaging: a comprehensive survey[EB/OL].

(2023−12−26)[2024−05−01]. http://arxiv.org/abs/2312.07353.

[95] 丁维昌, 施俊, 王骏. 自监督对比特征学习的多模态乳腺超声诊断[J]. 智能系统学报, 2023, 18(1): 66−74.
DING Weichang, SHI Jun, WANG Jun. Multi-modality ultrasound diagnosis of the breast with self-supervised contrastive feature learning[J]. CAAI transactions on intelligent systems, 2023, 18(1): 66−74.

[96] TOPOL E J. As artificial intelligence goes multimodal, medical applications multiply[J]. Science, 2023, 381(6663): adk6139.

[97] 高晗, 田育龙, 许封元, 等. 深度学习模型压缩与加速综述[J]. 软件学报, 2020, 32(1): 68−92.
GAO Han, TIAN Yulong, XU Fengyuan, et al. Survey of deep learning model compression and acceleration[J]. Journal of software, 2020, 32(1): 68−92.

[98] GOU Jianping, YU Baosheng, MAYBANK S J, et al. Knowledge distillation: a survey[J]. International journal of computer vision, 2021, 129(6): 1789−1819.

[99] ULLRICH K, MEEDS E, WELLING M. Soft weight-sharing for neural network compression[EB/OL]. (2017−05−19)[2024−01−01]. http://arxiv.org/abs/1702.04008.

[100] LIU Zhuang, SUN Mingjie, ZHOU Tinghui, et al. Rethinking the value of network pruning[EB/OL]. (2018−11−11)[2024−01−01]. http://arxiv.org/abs/1810.05270.

[101] KAMBHAMPATI S, VALMEEKAM K, GUAN Lin, et al. LLMs can't plan, but can help planning in LLM-modulo frameworks[EB/OL]. (2024−02−12)[2024−07−01]. http://arxiv.org/abs/2402.01817.

[102] XI Zhiheng, CHEN Wenxiang, GUO Xin, et al. The rise and potential of large language model based agents: a survey[EB/OL]. (2023−09−19)[2024−01−01]. http://arxiv.org/abs/2309.07864.

[103] MOOR M, BANERJEE O, ABAD Z S H, et al. Foundation models for generalist medical artificial intelligence [J]. Nature, 2023, 616: 259−265.

[104] 陈小平. 人工智能中的封闭性和强封闭性: 现有成果的能力边界、应用条件和伦理风险[J]. 智能系统学报, 2020, 15(1): 114−120.
CHEN Xiaoping. Criteria of closeness and strong closeness in artificial intelligence—limits, application conditions and ethical risks of existing technologies[J]. CAAI transactions on intelligent systems, 2020, 15(1): 114−120.

作者简介:

肖建力, 副教授, 主要研究方向为人工智能与大数据。2023 年吴文俊人工智能科学技术奖科技进步奖(科普项目)获得者, 中国计算机学会杰出会员。发表学术论文 10 篇, 著有图书《人工智能怎么学》。E-mail: audyxiao@sjtu.edu.cn。

许东舟, 硕士研究生, 主要研究方向为智慧医疗。E-mail: 233370870@st.usst.edu.cn。

王浩, 副主任医师, 主要研究方向为先天性心脏病和先天性气管狭窄的外科治疗。E-mail: haowang_nt@163.com。