文章编号:1671-4598(2026)08-0001-08 DOI:10.16526/j.cnki.11-4762/tp.2026.08.001 中图分类号:TP182 文献标识码:A

大模型技术及其在垂直领域应用综述

刘梦瑶[1]，王武军[1]，彭继阳[1]，徐基法[1]，贾 康[2]，贺 凯[1]，范琳琳[1]
(1. 浪潮电子信息产业股份有限公司，济南 250000；
2. 山东大学 前沿交叉科学青岛研究院，山东 青岛 266000)

摘要：大语言模型是一类由具有大量参数的人工神经网络构成的语言模型，通常通过自监督学习或半监督学习方式，在大量未标注文本上进行训练，是当前生成式人工智能技术的核心组成部分，在机器翻译、问答系统、对话生成等多个任务中表现出色；然而，现有的综述性研究多侧重于 LLMs 的理论框架与训练方法，对垂直领域的介绍相对不足，且局限于特定领域；为更全面地应对大模型在自然语言处理领域的发展及其对通用任务和行业应用带来的影响，在系统梳理 LLMs 的基础架构、训练技术与发展脉络的基础上，通过对其在若干具有广阔前景的垂直领域——如科学研究、医疗健康、智慧教育、现代农业等——的应用现状进行系统分析，深入探讨了大模型在实际应用中面临的挑战，提出相应的应对策略，并展望未来的研究方向。

关键词：大语言模型；垂直领域应用；检索增强生成；强化学习；自然语言处理

Review of LLM Technology and Applications in Vertical Fields

LIU Mengyao[1], WANG Wujun[1], PENG Jiyang[1], XU Jifa[1], JIA Kang[2], HE Kai[1], FAN Linlin[1]
(1. Inspur Electronic Information Industry Co., Ltd., Jinan 250000, China;
2. Institute of Frontier and Interdisciplinary Science in Qingdao, Shandong University, Qingdao 266000, China)

Abstract: Large language models (LLMs) are a type of language model composed of artificial neural networks with a large number of parameters, which are typically trained on a vast amount of unlabeled text through self-supervised or semi-supervised learning methods. As a core component of current generative artificial intelligence technology, LLMs exhibit well in multiple tasks such as machine translation, question-answering systems, and dialogue generation. However, existing reviews mostly focus on the theoretical framework and training methods of LLMs, with relatively insufficient coverage of vertical fields and limitation to specific domains. To comprehensively address the development of LLMs in the field of natural language processing and their impact on general tasks and industry applications, this paper systematically reviews the basic architecture, training techniques, and development trajectory of LLMs, comprehensively makes an analysis of their application status in several vertical fields with broad prospects such as scientific research, medical health, smart education, and modern agriculture, deeply discusses the challenges faced by LLMs in practical applications, and proposes corresponding strategies and future research directions.

Keywords: LLM; vertical domain applications; retrieval augmented generation; reinforcement learning; natural language processing

0 引言

在人工智能技术快速进步的背景下，大语言模型(LLMs，large language models)逐渐成为关键技术研究的重要方向[1]。由 OpenAI 推出的 ChatGPT 凭借对大规模文本及对话语料的深度学习，展现出卓越的自然语言处理能力，不仅能实现多语言流畅对话、精准应答复杂问题，还可完成任务计划书撰写、程序代码编写等特

收稿日期:2025-09-23； **修回日期**:2025-11-02。
基金项目:山东省自然科学基金(ZR2019LZH006)。
作者简介:刘梦瑶(1997-),女,博士,工程师。
通讯作者:贾 康(1997-),男,博士后。
引用格式:刘梦瑶,王武军,彭继阳,等.大模型技术及其在垂直领域应用综述[J].计算机测量与控制,2026,34(8):1-8,33.

定专业任务。该模型的问世在全球范围内掀起大模型研发热潮，国内科技企业迅速跟进，形成“百模大战”的竞争格局：智谱华章的 ChatGLM 系列[2]、百川智能的 Baichuan 系列[3]、阿里巴巴的 Qwen[4] 系列等产品相继问世，推动实际技术应用进入集中探索阶段。随着技术迭代，国内科技企业将研发重心从通用大模型转向垂直领域深入研究，从单纯追求参数规模与性能优化，转为关注实际应用场景需求。与通用大模型在广泛应用场景中展现强大语言理解能力不同，垂直领域大模型更侧重于满足特定行业的个性化需求，在深入适配用户具体应用场景方面具有显著优势。

目前，国内外针对大模型的研究综述主要可以划分为以下几个方向：1）技术维度的系统性梳理，如文献［5］全面总结了参数高效微调技术在大规模预训练模型应用中的具体方法及核心优势；2）关注自然语言处理的细分应用研究，文献［6］深入探讨了检索增强生成技术的关键进展，并针对检索器、生成器及各环节的优化方向提出可行性方案；3）侧重垂直领域的实践分析，文献［7］的研究验证，在医疗领域 LLMs 稳健青光眼筛查具有可行性且表现接近人类专家水平；4）围绕评估体系的科学性构建，文献［8］从可靠性与鲁棒性双维度，对当前 LLMs 性能评估方法进行了系统性分析，为评估体系的优化提供了理论依据。从上述梳理可见，现有综述研究大多针对 LLMs 的理论框架与训练方法，并且局限于单一视角或特定领域：在垂直领域应用综述中，这类局限尤为突出。

据此，本文通过构建更具整体性的分析框架，整合多行业的实践案例，对 LLMs 在垂直领域中的研究现状进行综述：首先系统梳理 LLMs 与自然语言处理的核心技术背景，构建相应的理论认知框架；而后系统分析大模型的垂直化技术特点；然后重点关注科学研究、医疗、教育等领域，通过实际案例分析大模型在垂直场景中的应用模式和技术发展前景；同时从实践的角度全面探讨大模型所面临的多重挑战，并提出相应的应对策略；最后，总结全文的研究成果，并对未来的发展趋势进行展望，为大语言模型下一步的相关研究提供参考和启示。

1 大模型背景

1.1 自然语言处理

自然语言处理（NLP，natural language processing）[9]是人工智能领域的重要研究方向，旨在利用计算机技术对自然语言进行深入分析与理解，进而实现对语言信息的有效处理。其核心任务包括自然语言理解与生成两部分：前者旨在使计算机具备解析人类语言语义、逻辑及语境的能力，使其能够准确解释并响应自然语言输入；后者则着重将非语言形式的结构化数据，例如图表、数据库信息等转化为符合人类表达习惯的文本内容，以此实现人工智能系统与用户的自然语言交互。

1.2 大模型

预训练语言模型（PLM，pre-training language model）[10]指在大规模数据集上预先训练的神经网络模型，其在通用任务中学习到的特征可迁移至特定下游任务。该方法的核心思想是利用大规模数据蕴含的丰富信息初始化模型参数，再通过微调或迁移学习策略，使模型有效适配具体目标任务，从而提升模型性能与训练效率。

LLMs[11]是在 PLM 基础上发展而来的进阶形态，主要指基于 Transformer 框架、参数规模达数百亿至数千亿级的神经网络语言模型，如 PaLM[12]、LLaMA[13]、GPT-4[14]等。与 PLM 相比，LLMs 的差异不仅体现在模型参数规模和训练数据量的大幅增加，语言理解与生成能力显著增强；更核心的区别在于，当参数与数据突破临界阈值时，LLMs 会涌现出小模型不具备的“涌现能力”[15]。

1.3 大模型架构

随着深度学习的不断发展，传统的卷积神经网络和循环神经网络（RNN，recurrent neural network）等模型在处理序列数据时逐渐显现出其局限性。特别是 RNN 的逐词顺序处理方式，不仅影响模型对上下文的整体把握，还严重制约了并行计算的效率。为解决上述问题，《Attention is All You Need》中引入了全新的 Transformer 架构[16]，该架构依赖自注意力机制，以编码器－解码器为基础框架，深度集成注意力机制，其中自注意力机制使每个 token 能直接关注序列中所有 token，轻松捕捉长距离依赖关系，解决了长期语境建模难题；多头注意力机制则通过多组独立注意力头的并行运算，从多重语义空间强化复杂关联特征的提取，进一步提升语义表征的准确性。当前 Transformer 框架主要包含 3 种基础架构[16]，如图 1 所示。3 种架构的比较如表 1 所示，它们的核心区别在于注意力机制中掩码矩阵的配置策略，编码器架构依赖双向注意力机制，支持对完整上下文的双向语义建模；解码器架构采用自回归掩码，限制模型仅能访问当前位置之前的词序列；编码器－解码器架构则通过跨注意力机制，在编码器生成的语义表征与解码器的生成过程之间建立交互桥梁。

2 模型垂直化技术

LLMs 构建包含预训练、监督微调、奖励建模及强化学习 4 个阶段，但预训练成本较高，目前多借助基座模型微调以降低应用门槛。通用大模型的泛化能力强，可跨领域处理自然语言任务，而垂直领域大模型一般基

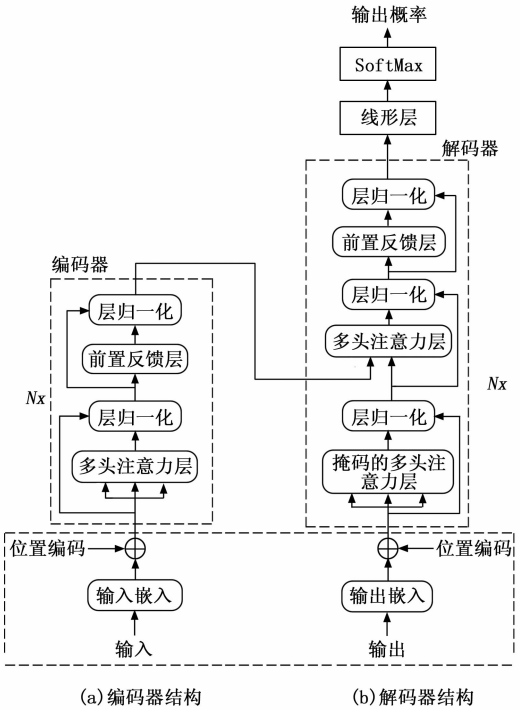

图 1　Transformer 架构图

表 1　LLMs 三种架构之间的比较。

架构类型	代表模型	架构特点	主要应用
编码器架构	BERT[17]	由多个编码器组成，擅长理解文本语义和上下文关系。	自然语言理解任务，例如文本分类、问答系统、情感分析等。
解码器架构	GPT[14]	由多个解码器组成，擅长文本生成。	文本生成任务，例如文本生成、对话系统和故事创建等。
编码器—解码器架构	T5[18]	由编码器和解码器组成，编码器处理输入文本，解码器生成输出文本。	翻译任务、摘要总结和文本到文本的转换任务等。

于基座模型，使用高质量领域数据微调和对齐以适配特定领域，但模型存在知识盲区、准确性不足等问题，需结合外部知识库提升实用性。因此，模型垂直化需模型训练、模型微调、知识库集成、策略优化等关键技术支撑。

2.1　模型训练

模型训练需完成技术阶段的构建和优化。首先是数据预训练与任务设计。在此阶段，对大规模语料进行清洗和结构化处理，以提升数据质量和多样性；通过语言建模捕捉语言统计特征，并结合去噪自编码增强语义理解，将语言规则与领域知识编码至模型参数，为后续训练奠定基础。然后实施高效参数优化策略，采用大规模批量训练以提高计算资源利用效率，并结合动态批量调整技术，在初期使用较小批量，随着模型稳定性提升逐步增加批量大小，从而保障训练精度并避免震荡问题。目前主流的优化器包括 Adam[19] 及其改进版本 AdamW[20] 和 Adafactor[21] 等，这些优化器通过自适应调整学习率，加快模型收敛速度，同时保持参数更新稳定。随着 LLMs 参数量和训练数据规模迅速增长，模型训练面临计算资源紧张和内存效率受限的双重挑战。为应对这些硬件制约并进一步提升训练效率，目前采取多种技术手段进行优化，包括内存优化技术 ZeRO、3D 并行技术以及混合精度训练等，以实现高效且稳定的模型训练。

近年来，混合专家（MoE，mixture of experts）模型训练方法在高效预训练领域展现出显著优势。MoE 核心架构如图 2 所示，主要由两个关键组件构成：1）稀疏 MoE 层，该层取代传统 Transformer 中的前馈网络，由多个相互独立的“专家网络”组成。每个专家网络负责处理特定类型的输入信息，通常为轻量级前馈网络，也可扩展为层级化结构或其他复杂形式。其核心在于动态选择机制，即在处理不同输入 token 时，仅激活少数相关专家，从而实现计算资源的高效分配。2）门控网络，作为路由决策的核心模块，门控网络为每个输入 token 生成专家分配的概率权重。根据任务需求，该机制可将 token 分配至单个专家或多个专家，并通过软权重或硬选择策略实现输入专家的动态映射，使模型能够根据不同语义特征调用专属的处理单元，在保持模型容量的同时避免冗余计算。例如医疗多模态

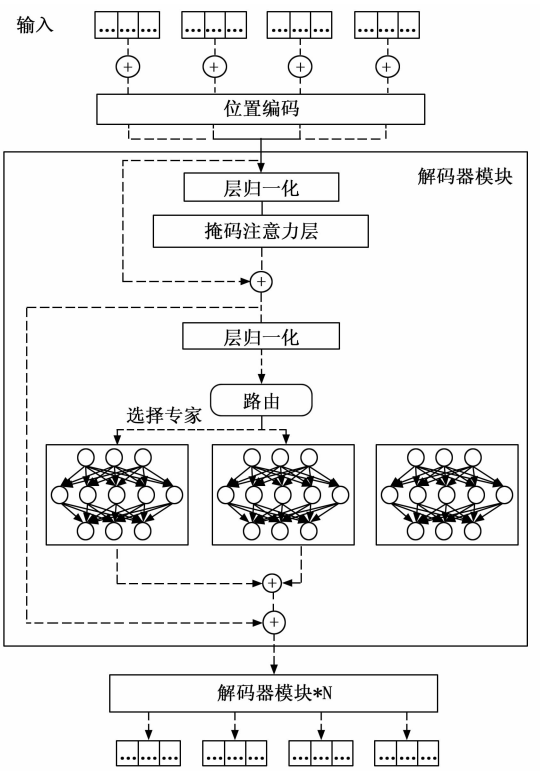

图 2　混合专家模型架构示意图

LLMs 框架 Med-MoE[22] 通过稀疏激活机制，在参数规模仅为传统 Transformer 30%～50%的情况下，医疗视觉问答准确率为 82.72%～91.98%，开放式问答召回率为 58.55%～85.06%，超越 LLaVA-Med 的各项对应指标，但训练推理成本显著降低。

2.2 模型微调

大模型的微调是指在已预训练好的 LLMs 基础上，使用特定领域或任务的数据集进行进一步训练，以使其适应新任务并提升在该任务上的表现的过程[23]。尽管经过预训练的 LLMs 在通用领域已展现出较强的泛化能力，但在各垂直应用场景中，仍需通过微调过程引入领域知识，以实现有效的场景适配和对齐使用者偏好的效果。监督微调作为核心技术之一，包含指令微调[24]、基于人类反馈的强化学习（RLHF，reinforcement learning from human feedback）[25]以及参数高效微调技术（PEFT，parameter-efficient fine-tuning）[26]等。其中，指令微调通过构建由指令及其对应输出构成的领域专用数据集，有助于提升模型对特定任务指令的理解与执行能力；RLHF 通过引入人工标注的质量评分或排序偏好作为奖励信号，利用强化学习算法优化模型输出；PEFT 包括提示微调[27]和低秩自适应微调（LoRA，low-rank adaptation）[28]，提示微调的核心思想是将输入实例重构为一种被称为“提示（prompt）”的填空式表达形式，从而缩小模型在语义理解方面与人类认知之间的差异。借助提示学习，模型能够在不更新全部参数的前提下，充分利用有限的标注数据生成高质量的推理输出，例如 Prefix Tuning[29]和 P-Tuning[30]；LoRA 方法通过引入低秩矩阵对 LLMs 的权重参数进行调整，在保留其原有通用知识的基础上，能够高效学习特定任务相关的信息，从而显著减少微调过程中所需的计算资源和参数规模[28]。

2.3 检索增强生成

检索增强生成（RAG，retrieval-augmented generation）技术的核心思想是将语言生成模型与外部知识库有效融合。尽管传统的 LLMs 在多种任务中展现出强大的生成能力，但其内部知识受限于训练数据的时效性，难以应对动态更新的信息需求。RAG 通过引入外部检索机制，使模型在推理过程中动态获取与当前任务相关的最新信息，从而提升生成结果的准确性与相关性。该技术的实现流程如图 3 所示。首先，进行索引构建，采用如 Text2Vec 等嵌入方法，将文档切分后的片段转化为固定维度的向量表示，以支持高效的语义匹配；其次，将用户输入的问题编码为向量，并与向量数据库中的文档片段进行相似度计算，筛选出最相关的上下文内容；最后，LLMs 结合检索所得的外部信息作为补充上下文，生成更加精准、信息丰富的响应结果。

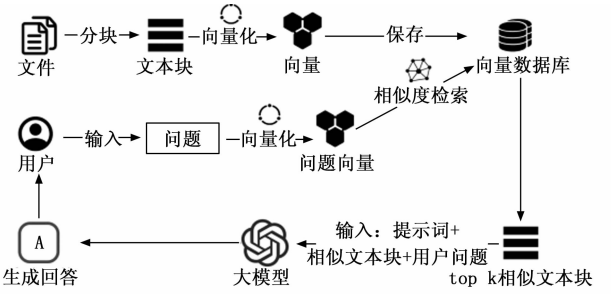

图 3 检索增强生成技术工作流程

2.4 智能涌现

LLMs 的涌现能力是指在模型规模扩展过程中，仅在较大规模模型中出现而在较小规模模型中缺失的特定能力。该能力的核心特征体现在存在性差异与非线性突变两个方面[31]：在模型规模尚未达到某一临界阈值之前，其在特定任务上的表现通常接近随机水平；而一旦跨越该阈值，性能则会迅速提升，显著优于随机表现，并呈现出质的飞跃，而非渐进式的改进。此外，涌现能力具有不可外推性，即无法通过小规模模型的性能趋势及其缩放规律进行准确预测，也无法依据小模型的行为推断其在大规模模型中是否会出现或何时出现。因此，智能涌现能力已成为当前 LLMs 研究的重要热点。目前，主流研究方向包括量化对涌现能力的影响，上下文学习中的涌现能力，以及如何借助提示策略和微调策略来激发 LLMs 的智能涌现能力，例如思维链提示（CoT，chain-of-thought prompting）[32]和指令微调等，具体内容在表 2 中详细介绍。

表 2 对比各项技术对涌现能力的影响

技术类型	定义及操作	对涌现能力的影响	结论
量化技术	通过降低神经网络参数数值精度实现模型压缩，可搭配量化后微调等恢复策略。	随着量化精度下降逐步退化，低比特量化会显著抑制涌现能力，恢复技术可缓解该影响。	存在精度临界阈值，低于阈值则涌现能力基本丧失；4 位量化可维持多数涌现能力，2 位量化时涌现能力近乎丧失[33]。
上下文学习	不调整模型参数，通过少样本/零样本提示引导模型完成新任务。	涌现能力的重要支撑，缺少样本提示时，模型涌现能力仅略优于随机水平。	提示词的示例选择、格式规范、意图清晰度、示例顺序等要素，直接影响涌现能力相关性能[34]。
思维链提示	引入中间推理过程，通过结构化示例展示逐步递进推理，连接输入与输出。	引导模型多步骤推理，间接促进与推理相关的涌现能力发挥。	在需要逐步推导的任务中效果显著，可通过定制化提示策略提升模型性能及对应涌现能力[15]。
指令微调	基于指令数据微调模型，优化模型整体表现。	更多作用于模型整体表现优化，不直接驱动涌现能力生成。	仅在千亿级以上大规模模型中，可显著提升模型性能[35]。

2.5 知识蒸馏

LLMs 中的知识蒸馏是一种通过轻量化模型压缩技术，将 PLM 中隐含的复杂知识迁移至小模型的过程，旨在使小模型在保持较高性能的同时降低计算成本与部署门槛[36]。其核心机制是利用“教师模型”的输入，如软标签概率分布、中间层特征表示、注意力权重等作为监督信号，引导学生模型学习更高效的知识表征。在具体实现过程中，通常通过构建组合形式的损失函数，迫使学生模型模仿教师模型的推理模式[36]。知识蒸馏在 LLMs 的应用中展现出多方面的价值：一方面，该技术可将拥有千亿参数的大型模型压缩为百亿或十亿级参数的轻量化版本，从而满足边缘设备上的推理需求；另一方面，通过多教师蒸馏方法——即整合多个大模型的知识，或采用自蒸馏策略——实现同一模型不同版本之间的知识迁移，有助于进一步增强小模型的泛化性能，甚至在某些特定任务中接近原始大模型的表现。

3 垂直领域应用

垂直领域大模型通过针对特定领域的专业需求，构建具有专业差异化的技术优势，从而实现对特定领域专业知识的精准调用与复杂问题的智能化处理。目前，这类模型已在自然科学、医学、金融及农业等专业壁垒高、知识体系复杂的领域开展实践探索，其显著的应用成效展现出广阔的发展前景。为系统梳理垂直领域大模型的技术价值与应用潜力，本章将从不同领域维度深入分析其在细分场景中的创新实践，旨在为行业的技术实现与应用拓展提供参考依据。

3.1 科学研究

人工智能科学（AI4S，artificial intelligence for science）是一种将人工智能技术与科学研究深度融合的新兴科研模式，引发了科学研究方法的创新性变革。这一转变不仅在物理学、化学、生物学、天文学等基础自然科学领域，而且在材料科学、环境科学等应用科学领域实现了技术突破，为跨学科与交叉学科研究开辟了全新的研究方向。

2024 年诺贝尔物理学奖与化学奖均授予人工智能在相关学科取得的重要研究成果，标志着人工智能技术已不再局限于工具层面的辅助性创新，而是逐步演进为推动科学革命的核心动力。AI4S 助力科学家从突破劳动力瓶颈和实验维度瓶颈的阶段，过渡到更高层次的信息瓶颈乃至认知瓶颈的重大挑战，例如中国科学院高能物理研究所推出的 Xiwu 大模型[37]，是首个面向高能物理（HEP，high energy physics）领域的 L2 级大语言模型，其核心优势体现在两方面：1）基础模型具备良好的灵活性，可从 LLaMA 平滑升级至 LLaMA2、Vicuna 等开源模型以实现能力持续增强；2）具备高效的可学习性，能够通过基于向量存储的即时学习系统，实现低成本领域知识的快速更新，有效解决科学领域大模型训练成本高、难以跟进开源模型迭代的问题。在性能方面，在 5 100 条 HEP 提示词评估中，Xiwu-13B 相较于 Vicuna-13B 的胜率及平局比例达到 95%，且其整体性能可达 ChatGPT-175B 的 65%，而 Vicuna-13B 不足 10%。在具体应用中，Xiwu 在 BOSS 粒子衰变模拟代码生成任务中能够准确提供衰变卡与配置参数，而 GPT-4 则出现概念性错误；在 HEPS 术语问答场景中，Xiwu 能够准确识别用户意图且回答准确无误，展现出较强的鲁棒性；而 GPT-4 不仅存在事实性偏差，且缺乏自我纠正机制。因此，该模型不仅为高能物理领域提供了定制化大模型解决方案，可有效支撑文献综述、代码生成等关键科研任务；也为其他科学领域大模型应用提供了可借鉴的方法论。

3.2 医疗

随着健康与医疗产业的信息化进程的不断演进，LLMs 深刻推动医疗领域的变革，可应用于医疗数据与知识管理、临床诊断与决策支持、患者交互与服务优化，以及生物医疗与药物研发等多个方面，为提升医患双方的诊疗效率、优化医疗服务管理流程，并助力科研机构开展高水平医学研究提供智能化解决方案。

在传统中医药领域，文献［38］通过解析中医防疫古籍等结构复杂、规模庞大的医学文献，构建了用于疫情防控用药相关智能问答系统 EpidemicCHAT，提升了模型在中医疫病场景下的问答准确性和推理能力。其关键方法包括 3 个方面：1）数据构建，收集 194 部中医古籍中的疫病相关内容，结合中医疫病防治知识图谱，经 ChatGPT 转化为现代语言并生成指令数据，最终形成 690 万条训练数据集；2）知识融合，采用自生成指令将知识图谱中的实体与关系转化为训练指令，补充模型专业知识；3）高效调优，运用 LoRA 技术，在冻结原始 ChatGLM 模型参数的同时添加低秩矩阵，配合量化技术，降低计算成本并避免过拟合。经后续测试验证，该系统在中医疫病方剂生成任务中的 BLEU-4、ROUGE-L 和 METEOR 指标分别达到了 44.02、61.10 和 59.40，显著超越了 ChatGLM、ChatGPT 等基线模型。在案例测试中，该系统能够准确回答关于古籍的专业问题，有效规避“幻觉”现象。此外，该系统高效整合了中医古籍知识与知识图谱，为构建适用于中医领域的大语言模型提供了一条可行路径，并推动了中医疫病知识的数字化传播。

在健康领域筛查与诊断方面，文献［39］提出了一种与 LLMs 集成的医疗物联网系统 AutoHealth，旨在通过智能穿戴设备实现帕金森病的早期检测、持续监测及康复指导，从而解决传统诊疗中依赖人工评估、主观

性强和依从性低的问题。该系统融合了多模态感知与人工智能算法，通过运动与语音信息实现帕金森生物标志物的检测。同时，它集成 LangChain 框架与 ChatGPT API 以构建聊天机器人，实现语音和文本之间的交互，并整合谷歌翻译及文字转语音功能以突破语言障碍。此外，该系统可存储医疗数据并提供个性化饮食和运动建议。研究表明，该系统的运动震颤与冻结步态的检测准确率达到 95.4%，语音震颤检测准确率达 96.9%，均优于多款传统模型。其跨平台和低成本特性有助于降低医疗支出，提高诊疗可及性。通过多模态传感器融合以及对 LLMs 的深度应用，该系统不仅实现了对帕金森病患者实时监测和个性化指导，还具备扩展至其他慢性病管理的潜力，使患者能够自主掌控健康管理。

3.3 教育

LLMs 正深刻影响教育教学方式，文献［40］针对现有教育类 AI 主要为单一任务型且缺乏真实课堂互动感的问题，提出了基于 LLMs 的多智能体课堂模拟框架 SimClass，以探索多智能体协作模拟动态课堂的可行性。SimClass 框架定义了两类核心智能体角色：教学类负责知识传授与课堂管理；学生类用于模拟同伴间的互动，并支持用户自定义角色。同时，该框架设计了课堂会话控制器，通过状态感知、功能执行和管理 3 大模块，实现对课堂流程与发言逻辑的动态调控。该框架以 GLM-4 作为基础模型，选取两门真实课程开展实验，并招募 400 余名学生参与，同时结合弗兰德斯互动分析系统与探究社区理论进行评估。此外，还设计了消融实验以验证不同角色及其互动的重要性。在性能方面，SimClass 展现出优异的课堂模拟效果：师生之间以及学生之间的互动频繁，其中学生主动发起互动比例达 89.6%～91.7%，显著超过传统课堂；对于知识类课程学生最终考核平均分为 0.65，与章节测验平均分持平，并且用户交互频次、消息长度与成绩呈显著正相关。此外，消融实验证明，学生类智能体使得用户发言长度提升 26.5%～45.2%，显著增强认知存在感和社交存在感。SimClass 框架通过对多智能体角色设计及动态课堂控制机制，有效地模拟了生动活泼的教学场景，也验证了基于 LLMs 的多智能体系统在教育领域应用的可能性。

3.4 金融

人工智能作为数字经济时代的核心技术，推动了金融业的技术革新。在金融智能领域，彭博机器学习产品与研究团队推出了首个面向金融垂直领域的 LLMs——BloombergGPT[41]，该模型采用混合数据训练策略，其中金融领域的数据占比为 51.27%，而通用公开数据占 48.73%，旨在实现金融领域的专业性与通用能力之间的平衡。BloombergGPT 的模型架构基于 BLOOM 的纯解码器因果语言模型，包含 506 亿个参数，并使用 Unigram 分词器及 ALiBi 位置编码技术。在训练过程中应用了 ZeRO 优化、混合精度训练、动态调整策略等多项技术；同时，通过数据去重和时间序列验证集划分等方法，以确保数据质量。因此，BloombergGPT 在金融任务和通用 NLP 任务中均展现出卓越的性能。在金融领域，其在 ConvFinQA、FiQA SA 等外部金融任务中的平均得分达到 62.51，显著高于同类模型如 GPT-NeoX 和 OPT-66B。在内部情感分析任务中，其平均 F1 值达到了 62.47，而在命名实体识别与消歧义任务中更是以明显优势领先。此外，在通用任务方面，BloombergGPT 在 BIG-bench Hard、知识评估及阅读理解等基准测试中的表现优于相近规模的其他模型。部分指标甚至接近或超越了参数更多的 BLOOM-176B。并且在语言学相关任务中，该模型以 85%的胜率领先其他同类模型，实现了金融领域专精能力与通用能力之间良好的平衡。

哥伦比亚大学开发的开源、轻量化 LLMs 的 FinGPT[42]，采用数据驱动的方法，构建了包含数据源层、数据工程层、LLMs 层和应用层的端到端框架。数据源层整合了金融新闻、公司申报文件及社交媒体讨论等多种类型的数据，涵盖包含 SEC 官网在内的多个渠道，并进行严格的数据预处理；数据工程层则通过实时 NLP 技术过滤噪声并提取关键信息，以应对金融数据高时间敏感性和低信噪比的挑战；LLMs 层运用 LoRA 以及强化学习等轻量级微调技术，大幅减少可训练参数，控制训练成本远低于 BloombergGPT；应用层则支持智能投顾、量化交易及低代码开发等多种金融场景。FinGPT 通过轻量化微调与开源框架设计，在降低金融 LLMs 应用门槛的同时，也确保了模型的时效性和适应性，为金融领域 AI 创新提供了一种高效且低成本的解决方案。

3.5 农业

农业是人工智能技术应用最为重要且前景广阔的领域之一。农业专用大模型 AgroLLM[43]能够从丰富的开源农业资源中提取准确且符合上下文的信息，旨在解决通用 LLMs 在农业领域适配不足的问题，弥补专业词汇理解、区域农业特性适应及数据质量不均等短板，为农业领域提供教育资源、互动式教程以及多语言支持，以确保来自不同地区的农民能够获取关键的农业信息。该模型集成了 RAG 与向量数据库技术，通过生成式 AI 提取文本嵌入向量，并将其存储于 FAISS 向量数据库，从而实现高效的相似性检索。在模型选型阶段，对 Gemini 1.5 Flash、ChatGPT-4o Mini 和 Mistral-7B-Instruct-v0.2 共 3 个模型进行了比较，从嵌入质量、检索效率和响应连贯性等多个维度开展评估。在训练过程中，首先将农业文本拆分为带有元数据的重叠片段，然

后通过多模型生成嵌入向量，再利用 FAISS 实现快速检索。最后结合 RAG 技术，将相关文本片段整合为提示词和上下文信息输入至 LLMs，同时设置 fallback 机制以保障在无相关文档时的响应能力。研究结论显示，ChatGPT-4o Mini 结合 RAG 技术表现最佳，其准确率达到 93%，并在 MRR、Recall@10 及 BLEU 等指标上全面领先。AgroLLM 通过领域专属数据库与 LLMs 的融合，实现了对农业领域的专业化适配，提高了知识传递的精准性与实用性。在提高农作物产量方面，文献[44]对基于 LLMs 的环境控制系统进行了更为深入的探讨，其研究涵盖了如何将多类传感器、自动化技术与人工智能有效整合至生长室中，以实现对光照、温度、湿度、营养供应和二氧化碳浓度等关键参数进行精确评估与调控，从而为农作物增产与产量预测提供了切实可行的技术路径。

4　挑战与展望

当前 LLMs 已在诸多领域取得显著成果，但其在实际应用中的有效性与可持续发展仍受到一系列问题的制约。首先，训练数据的质量和数量仍是限制模型性能进一步提升的关键因素。目前，模型主要依赖公开数据集进行训练，而这些数据往往存在偏差，难以充分反映现实世界中的多样性与复杂性。尤其在某些特定领域或行业应用场景中，可用数据极为有限，导致模型的泛化能力受限。同时，训练数据在数量和质量上的欠缺，直接影响了模型的学习效果与实际应用表现，致使许多任务的准确率难以达到实际应用的要求。其次，是模型幻觉问题，LLMs 在内容生成过程中会出现事实错误、逻辑无关或输出不一致的现象。尽管 LLMs 具备生成高质量自然语言的能力，但在实际运行中仍难以准确辨别真实信息与虚构内容，容易做出不符合事实的推断，这一缺陷严重削弱了模型在生成任务中的可信度。然后，是价值观对齐问题，现有 LLMs 主要基于大规模文本数据进行训练，而这些数据往往缺乏系统性的道德、伦理及价值观层面的筛选与规范。因此，模型的输出可能带有偏见、不适宜甚至违背社会主流价值观念的信息，导致其行为与人类期望的价值导向产生偏差。最后，是高成本的部署与训练问题，LLMs 在训练和实际应用过程中依赖大量的计算资源与存储空间，导致整体开销巨大，许多机构和研究人员难以承担。随着模型参数规模的持续扩展，所需的算力和基础设施投入呈指数级上升，严重制约了模型的广泛推广与长期可持续发展。

针对上述挑战，可采取以下策略进行有效缓解：1）提升数据质量与多样性，可着重发展数据增强技术，利用多种手段合成具有代表性和多样性的数据样本。例如，针对医疗领域标注数据的稀缺性，尤其是罕见病病例较少，可以基于生成对抗网络生成模拟病例；在农业领域，由于数据存在显著的地域差异，例如南方水稻与北方小麦种植的数据，可结合迁移学习和地域适配进行微调；2）消除模型幻觉，缓解幻觉问题的关键在于将模型生成能力与外部知识有效结合，使用 RAG 技术可引入实时更新的外部知识库，为模型提供可靠的事实支撑，使得在 LLMs 中显著降低虚构内容的产生，增强输出结果的准确性与一致性；3）推动模型价值观与人类伦理对齐，可通过加强其在伦理与道德层面的设计。比如，训练数据的筛选需坚守合法性、去标识化与偏见过滤标准，明确剔除敏感信息与歧视性样本；建立严格的内容审查与反馈机制，制定具象化审核规则，建立分级响应流程，将反馈数据定期回灌训练，保障模型在生成过程中遵循普遍接受的社会价值观和法律法规，防止输出有害或误导性内容；4）发展小型化模型，采用知识蒸馏、量化等技术，将 LLMs 中的知识迁移到轻量级模型中，可在保持较高任务性能的同时降低资源消耗。同时，还可以设计更加优化的模型结构，并结合软硬件协同改进，推动模型向实用化与可持续化方向发展。

5　结束语

本文系统梳理了大语言模型的发展脉络及其在垂直领域中的应用现状。首先，介绍了自然语言处理技术和大模型的原理。随后，对当前大模型的垂直化技术进行了分析与阐述，包括模型微调、检索增强生成、知识蒸馏等。在此基础上，深入探讨了大模型在科学研究、医疗、教育、金融等专业领域的实际应用情况，并归纳了各领域中具有代表性的模型实例。最后，文章进一步剖析了大模型在实际部署过程中所面临的主要挑战，包括训练数据质量、部署成本、模型幻觉以及伦理与安全风险等问题，并针对上述问题给出了相应的应对措施。总体而言，大模型的垂直化应用是一个不断发展演进的过程，涉及多个技术体系。每个技术体系的研究与发展又与自然语言处理技术、大模型架构、深度学习等领域密切相关。未来，随着大模型技术的不断进步，其垂直化应用将愈加广泛，推动垂直领域技术趋向更高水平的智能化和自动化。

参考文献：

[1] RAIAN M A K. A review on large language models: architectures, applications, taxonomies, open issues and challenges [J]. IEEE Access, 2024, 12: 26839 - 26874.

[2] GLM T, ZENG A, XU B, et al. ChatGLM: a family of large language models from GLM-130B to GLM-4 all tools [J]. ArXiv Preprint ArXiv: 2406.12793, 2024.

[3] YANG A, XIAO B, WANG B, et al. Baichuan 2: open largescale language models [J]. ArXiv Preprint ArXiv:

2309.10305, 2023.
[4] AI J Z, BAI S, CHU Y F, et al. Qwen technical report [J]. ArXiv Preprint ArXiv, 2023.
[5] LEI S F, HUA Y, ZHIHAO S F. Revisiting fine-tuning: a survey of parameter-efficient techniques for large AI models [J]. ArXiv Preprint. ArXiv, 2025.
[6] GAO Y, XIONG Y, GAO X, et al. Retrieval-augmented generation for large language models: a survey [J]. ArXiv Preprint. ArXiv, 2023.
[7] DE VENTE C. AIROGS: Artificial intelligence for robust Glaucoma screening challenge [J]. IEEE Transactions on Medical Imaging, 2024, 43 (1): 542 - 557.
[8] KUMAR P, MISHRA S. Robustness in large language models: a survey of mitigation strategies and evaluation metrics [J]. ArXiv Preprint. ArXiv, 2025.
[9] JEGAN R, HENRICH A. A structured literature review on traditional approaches in current natural language processing [J]. ArXiv Preprint. ArXiv, 2025.
[10] ZANGARI L, GRECO C M, PICCA D, et al. A survey on moral foundation theory and pre-trained language models: current advances and challenges [J]. AI &Society, 2025, 40 (6): 1 - 26.
[11] MINAEE S, MIKOLOV T, NIKZAD N, et al. Large language models: a survey [J]. ArXiv Preprint. ArXiv, 2025.
[12] CHOWDHERY A, NARANG S, DEVLIN J, et al. PaLM: scaling language modeling with pathways [J]. ArXiv Preprint. ArXiv, 2022.
[13] TOUVRON H, LAVRIL T, IZACARD G, et al. LLaMA: Open and efficient foundation language models [J]. ArXiv Preprint. ArXiv, 2023.
[14] OPENAI, ACHIAM J, ADLER S, et al. GPT-4 technical report [J]. ArXiv Preprint. ArXiv, 2023.
[15] BERTI L, GIORGI F, KASNECI G. Emergent abilities in large language models: a survey [J]. Arxiv Preprint. Arxiv, 2025.
[16] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [J]. ArXiv Preprint. ArXiv, 2017.
[17] AHMADOU F, GHAFFARZADEGAN S, NOUR B, et al. Automated attack testflow extraction from cyber threat report using BERT for contextual analysis [J]. ArXiv Preprint. ArXiv, 2025.
[18] RAFFEL C, SHAZEER N, ROBERTS A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer [J]. ArXiv Preprint. ArXiv, 2019.
[19] KINGMA D P, BA J. Adam: a method for stochastic optimization [J]. ArXiv Preprint. ArXiv, 2014.
[20] LOSHCHILOV I, HUTTER F. Decoupled weight decay regularization [J]. ArXiv Preprint. ArXiv, 2017.
[21] SHAZEER N, STERN M. Adafactor: adaptive learning rates with sublinear memory cost [J]. ArXiv Preprint. ArXiv, 2018.
[22] JIANG S, ZHENG T, ZHANG Y, et al. Med-MoE: mixture of domain-specific experts for lightweight medical vision-language models [J]. ArXiv Preprint. ArXiv, 2024.
[23] ZHANG S, DONG L, LI X, et al. Instruction tuning for large language models: a survey [J]. ArXiv Preprint. ArXiv, 2023.
[24] CHUNG H W, HOU L, LONGPRE S, et al. Scaling instruction-finetuned language models [J]. ArXiv Preprint. ArXiv, 2022.
[25] OUYANG L, WU J, JIANG X, et al. Training language models to follow instructions with human feedback [J]. ArXiv Preprint. ArXiv, 2022.
[26] HAN Z, GAO C, LIU J, et al. Parameter-efficient fine-tuning for large models: a comprehensive survey [J]. ArXiv Preprint. ArXiv, 2024.
[27] LESTER B, AL-RFOU R, CONSTANT N. The power of scale for parameter-efficient prompt tuning [J]. ArXiv Preprint. ArXiv, 2021.
[28] HU E J, SHEN Y, WALLIS P, et al. LoRA: low-rank adaptation of large language models [J]. ArXiv Preprint. ArXiv, 2021.
[29] LI X L, LIANG P. Prefix-tuning: optimizing continuous prompts for generation [J]. ArXiv Preprint. ArXiv, 2021.
[30] LIU X, JI K, FU Y, et al. P-tuning v2: prompt tuning can be comparable to fine-tuning universally across scales and tasks [J]. ArXiv Preprint. ArXiv, 2021.
[31] SCHAEFFER R, MIRANDA B, KOYEJO S. Are emergent abilities of large language models a mirage? [J]. Arxiv Preprint. Arxiv, 2023.
[32] WEI J, WANG X, SCHUURMANS D, et al. Chain-of-thought prompting elicits reasoning in large language models [J]. ArXiv Preprint. ArXiv, 2023.
[33] LIU P, LIU Z, GAO Z F, et al. Do emergent abilities exist in quantized large language models: an empirical study [J]. Arxiv Preprint. Arxiv, 2023.
[34] WEI J, WEI J, TAY Y, et al. Larger language models do in-context learning differently [J]. Arxiv Preprint. Arxiv, 2023.
[35] LU S, BIGOULA E, SACHDEVA R, et al. Are emergent abilities in large language models just in-context learning? [J]. Arxiv Preprint. Arxiv, 2023.
[36] MANSOURIAN A M, AHMADI R, GHAFOURI M, et al. A comprehensive survey on knowledge distillation [J]. Arxiv Preprint. Arxiv, 2025.

(下转第 33 页)