

文章编号: 1007-5321(2024)04-0020-09

DOI: 10.13190/j.jbupt.2024-035

大语言模型时代的人工智能: 技术内涵、行业应用与挑战

陈 光, 郭 军

(北京邮电大学 人工智能学院, 北京 100876)

摘要: 大语言模型(LLM)的出现标志着人工智能 LLM 时代的来临。基于海量数据集的预训练, LLM 展现出卓越的适应性和创造力, 正在成为推动社会发展的关键驱动力, 并将在体系化人工智能中扮演重要角色。鉴于既有综述在分析 LLM 面临的挑战、关键属性、工程实现等方面的不足, 笔者从技术内涵、行业应用和主要挑战三个维度重新构建探讨框架。重点阐述了 LLM 在系统架构、训练策略、模型规模、压缩、多模态融合、提示与规划等技术层面的内涵, 以及在教育、科研、医疗、金融、司法等领域的应用前景。同时, 讨论了 LLM 可信性、可控性与安全性的研究现状, 以及 LLM 在技术和社会层面所面临的双重挑战, 展望了 LLM 在体系化人工智能中的角色定位和研究方向的契合点, 以期为 LLM 的研究与应用提供新的视角和思路。

关键词: 大语言模型; 多模态模型; 可信性; 可控性; 体系化人工智能

中图分类号: TP183

文献标志码: A

Artificial Intelligence in the Era of Large Language Models: Technical Significance, Industry Applications, and Challenges

CHEN Guang, GUO Jun

(School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China)

Abstract: The emergence of ChatGPT marks the advent of the era of artificial intelligence powered by large language models (LLM). Based on large-scale datasets for pre-training, LLMs demonstrate exceptional adaptability and creativity, becoming a critical driving force in advancing society and playing a significant role in systematic artificial intelligence. Given the limitations of existent reviews in analyzing the challenges faced by LLMs, their key attributes, and engineering implementation aspects. The framework is rediscussed and reconstructed from three dimensions: technical connotations, industry applications, and major challenges. The focus is on elucidating the connotation on the level of technical aspects of LLMs, including system architecture, training strategies, model scale, compression, multimodal fusion, prompting, and planning. It also explores the application prospects in various fields such as education, scientific research, healthcare, finance, and justice. Additionally, the discussion covers the current state of research on the reliability, controllability, and security of LLMs, as well as the dual challenges LLMs face on both technical and societal levels. It envisions the role of LLMs in systematic artificial intelligence and identifies alignment points in research directions, aiming to provide new perspectives and ideas for the research and application of LLMs.

Key words: large language models; multimodal models; trustworthiness; controllability; systematic artificial intelligence

收稿日期: 2024-02-28

作者简介: 陈 光(1978—), 男, 副教授, 硕士生导师, 邮箱: chenguang@bupt.edu.cn。

在21世纪的第3个十年,人工智能领域迎来了一场划时代的变革——大模型的兴起。这场变革不仅标志着技术上的一大飞跃,更为人工智能在各行各业的应用开辟了新天地。大模型,即大语言模型(LLM, large language models)^[1],如GPT(generative pre-trained Transformer)系列、BERT(bidirectional encoder representations from Transformer)^[2]等,凭借其卓越的语言理解和生成能力,正在重塑人机交互的方式和人工智能的使能方式,使其成为发展新质生产力的核心要素。

人工智能的概念于1956年被正式提出,其发展经历了多次起伏与变革。早期的人工智能研究侧重于规则和逻辑的编程,期望通过明确的指令使机器执行特定任务。20世纪90年代,机器学习逐渐成为人工智能研究的热点。进入21世纪后,深度学习成为人工智能的主要内容。

LLM的出现,使人工智能研究和应用发生了划时代的变化。这些模型通过在大规模数据集上进行基于自监督学习的预训练,掌握了丰富的世界知识和语言规则,在多种任务上都展现出了惊人的适应性和创造性。

LLM具有庞大的体系架构和独特的训练策略,随着模型规模的急剧增长,模型压缩成为现实问题。同时,多模态LLM(MLLM, multimodal LLM)正成为新的技术潮流,并带来能力的显著扩展和跃升。在应用端,如何探索、发现和释放LLM所蕴含的巨大潜能已成为一个新的课题,由此引发了对与LLM对话时采用的提示和规划策略的大量研究。

LLM的快速发展也带来了新的挑战。如何确保这些模型的可信性、可控性和安全性,如何解决训练这些模型所需的巨大计算资源和能源消耗,以及如何应对潜在的伦理和社会问题,都是亟须面对的重要挑战。

笔者重点解读LLM在系统架构、训练策略、模型规模、模型压缩、多模态、提示与规划等方面的技术内涵,以及在教育、科研、医疗、金融和司法等领域的行业应用,讨论LLM在技术层面和社会层面所面临的双重挑战,以及相伴而来的可信性、可控性和安全性要求,展望LLM在体系化人工智能中可能扮演的角色和二者在研究上的共通之处。

1 技术内涵

LLM的出现,标志着人工智能技术的重大突破

和发展。为了更好地理解LLM的技术内涵,需要从多个角度对其进行深入探讨。本节将重点阐述LLM在系统架构、预训练与微调策略、模型规模、模型压缩、多模态以及提示与规划等方面的技术特点和创新之处。

1.1 系统架构

自2017年诞生以来,Transformer^[3]架构因创新的注意力机制及出色的并行处理能力,成为LLM开发的核心。它以编码器-解码器设计取代循环神经网络,利用自注意力机制有效捕捉序列间的长程依赖。编码器通过多头自注意力层深入分析输入序列的内在联系,多层叠加深化了特征提取。解码器增加遮蔽机制维持自回归特性,同时结合编码器输出,确保信息互动与整合。Transformer的高效模型能力和并行计算促进了众多高性能LLM的诞生与成就。

从BERT^[2]到GPT^[4-6]系列,再到T5(text-to-text transfer Transformer)^[7],Transformer架构的应用不仅极大提升了模型的计算效率,也让数百亿参数级的语言模型训练成为现实。它们在语言理解、生成、推理等方面展现出了超越人类的能力,推动了人工智能在认知智能层面的跨越式发展。Transformer强大的建模能力和优越的并行性能,为LLM的发展奠定了坚实的基础。

随着对Transformer架构研究的深入,各种改进和变体模型纷纷涌现。例如,旨在优化计算效率的稀疏注意力机制和增强模型的性能和表达能力的专家混合(MoE, mixture-of-experts)模块^[8]的引入等。特别是稀疏门限MoE等变体模型的出现^[9],为推动LLM的发展起到了重要作用。

模型架构上的另一个重要进展是多模态模型的涌现。最新的架构已经可以进行图像、音频等多种模态信息的集成处理。例如,DALL·E2(dreaming of a generative artificial intelligence with latent embeddings 2)能根据文本描述生成相应的图像。

此外,为了减少对海量训练数据的依赖,提高模型对知识的把握,一些研究将外部知识库直接整合到LLM中。例如,ERNIE(enhanced language representation with informative entities)模型^[10]通过融入知识图谱中的实体和关系,来增强自身的知识表现力。这种知识增强型的模型设计,使得语言模型对现实世界的理解更加深刻和准确。

最近,还出现了将多个LLM进行整合的架构。整合之后,每个模型各司其职,发挥各自的优势,共

同攻关更复杂的问题。这种协同多智能体架构的出现,不仅推动了模型性能的提升,更是向通用人工智能迈进的一个标志性步伐。

1.2 预训练与微调策略

预训练和微调方面的技术突破是 LLM 获得成功的一个关键。

在最初阶段,ELMo (embeddings from language model)^[11]和 GPT 等模型采用单向语言模型的目标函数进行预训练,学习到词的通用向量表示后,通过微调来适应具体的下游任务。

BERT^[2]模型的出现标志着技术的一个飞跃。BERT 采用双向掩码语言模型的预训练目标函数,通过预测上下文中的隐藏词汇获得词向量的双向动态表示,为众多后续的预训练语言模型奠定了基础。

T5^[7]等模型将自然语言处理任务视为从文本到文本的转换问题,利用序列到序列的目标函数将预训练和微调统一起来,为处理各类自然语言处理任务提供了新范式。

GPT-3 之后的模型在模型规模和数据规模上持续扩张。这些模型在进行大规模无监督语言建模预训练的基础上,通过提示学习^[12,13]的方式来适应各种下游任务,使模型能在没有显式微调的情况下适应新任务。

此外,Adapter^[14]等技术通过在预训练的模型参数外附加小型的可微调模块,实现了对下游任务的快速适应,显著降低了模型调整的复杂性和计算成本。

通过自监督预训练方法,语言模型获得了强大的语言表示能力和更好的泛化性。而微调或提示学习则进一步提高了模型的适应性,使预训练语言模型能够更加精准地适应特定任务。

如今,提示学习的方法愈发受到重视,因为它可以使 LLM 在零样本或少样本的情况下迅速适应新任务,免除了大量标注样本的工作,极大地提升了模型的实用性和灵活性。

1.3 模型规模

LLM 借助 Transformer 架构的并行计算优势,不断扩大模型规模。例如 GPT-3 达到了 1 750 亿参数,PaLM (pathways language model)^[15]扩大到了 5 400 亿参数。

参数量的显著增加,可使模型具备更强大的表示能力和学习能力,以更好地捕捉输入文本中的语义和上下文信息。更大的模型规模还可以提供更多

的自由度,使模型能够处理更复杂的语言任务,并展现更高的性能。

另外,数据量对于 LLM 的性能也至关重要。随着数据量的增加,模型可以更好地捕捉到语言的复杂性和多样性,从而提高模型的理解和生成能力。LLM 所采用的大型数据集通过包含各种语言形式、话题和领域的文本,使模型能够更全面地学习语言知识,并具备更好的泛化能力。

值得注意的是,模型能力与数据量和模型规模之间存在着相互影响和平衡的关系。增加数据量可以提高模型的泛化能力和鲁棒性,使其更好地适应不同的任务和环境。然而,当数据量达到一定程度后,进一步增加数据量会使收益逐渐减小。另外,增加模型规模可以带来更大的性能提升,但同时也会增加对数据量的需求。

1.4 模型压缩

LLM 庞大的规模带来了对计算资源高需求的挑战。为了让 LLM 能在资源受限的环境下运行,模型压缩的需求十分迫切。模型压缩^[16]的目标是在性能没有感官下降的同时显著减少模型的存储规模和计算负担。模型压缩的方法主要包括模型裁剪、知识蒸馏、参数量化等。

模型裁剪^[17]通过移除模型中非关键的参数,如那些接近零或对模型输出贡献不大的权重,来降低模型的尺度和复杂度。

知识蒸馏^[18]从一个大型的教师模型向一个小型的学生模型迁移知识。通过将教师模型的输出作为目标,学生模型得以学习教师的决策模式,从而获得类似教师模型的性能。

参数量化是指将模型中的浮点数参数转换为低精度的数值表示,如 8-bit 或 4-bit 整数。参数量化可以大幅削减模型的存储需求和计算量,虽然可能会引入一些精度损失,但通过合理的量化策略和校准,精度损失是可控的。

在实践中,通常会结合多种压缩技术来达到最优效果。常见的做法是,先通过模型裁剪去除多余参数,然后利用知识蒸馏来维持性能,最后采用参数量化进一步压缩模型。

这些压缩技术,使得 LLM 在维持其强大功能的同时,变得更加轻便和高效,大大拓宽了应用范围,使智能手机、物联网设备和云计算服务等多种普通环境成为应用环境。

1.5 多模态

MLLM 通过集成处理多种数据类型(模态)的模型来统一理解和生成文本、图像、音频等多种信息。MLLM 的输入和输出均可以是多模态的,覆盖各种多媒体相关任务,如图像描述生成、视频理解、多模态问答等。

MLLM 的构建通常涉及三个步骤,第一步是采用编码器为每种模态生成数据表示(嵌入);第二步是通过一个或多个解码器生成输出;第三步是在必要时通过多模态推理将一种模态转换为另一种模态,或者在复杂推理任务中无缝集成多模态数据。

在计算机视觉方面,MLLM 通过整合视觉信息和文本信息,极大地提升了机器对图像内容的理解能力和泛化性,能在零样本或小样本情况下识别和理解未见过的对象或概念。在视觉推理任务中,MLLM 可使用预训练的视觉模型提取动作等视觉信息来提高描述能力。

在语音识别方面,通过整合语音和文本信息,MLLM 能更准确地识别和转录特定领域的术语和生僻字词,提高了转录的准确率,使语音识别技术适用于更多场景。

1.6 提示与规划

上述进展从多方面提高了 LLM 的能力,但要充分地将能力发挥出来,还需要对模型的对话输入进行精心设计,以及开发新的技术来增强模型的理解和推理能力。为此,“提示”和“规划”方法得到了大量的研究和应用。

所谓提示是指通过手动编写或自动优化一段对话输入来指导模型完成特定的任务。在设计这些提示时,通常会包括任务描述,定义输入数据的格式,以及提供一些上下文信息以帮助模型更好地理解任务。实践表明,提示对于保证模型的输出质量至关重要^[19]。

LLM 所涌现的一个能力是上下文学习^[20]。这一能力是指模型能够从当前给定的上下文中提取信息并答复相应的问题。利用这一能力,可以通过提供例文让 LLM 学会如何回答特定类型的问题。这也是提示的一种有效方法。

“思维链提示”^[21]是一种通过在“提示”中加入一系列逻辑推理步骤来增强模型理解能力的方法。这种方法促进了模型的深入思考,使其在处理需要复杂推理的问题时表现更加出色。通过这

种方法,模型不仅能给出答案,还能提供到达答案的逻辑路径,对于提高模型透明度和可解释性具有重要意义。

相对“提示”,“规划”^[22]是针对更为复杂任务的方法。涉及将大任务分解成一系列更小、更易管理的子任务,然后为模型制定一个行动计划来逐一完成这些子任务。这种方法特别适用于解决需要多步操作或决策的复杂问题。

上述这些技术进展共同推动了 LLM 在各种应用场景中的有效利用,使其在理解语言、解决问题以及生成人类风格的语言方面变得更加高效和精准。

2 行业应用

LLM 在技术层面取得了突破性进展,展现出卓越的性能和广阔的应用前景。接下来探讨 LLM 在教育、科研、医疗、金融、司法等重点领域的应用现状和潜力。

2.1 教育

LLM 在教育领域具有广阔的应用前景,有望对传统的教学模式和学习方式产生深远影响。LLM 在教育领域的典型应用包括以下几个方面。

智能辅导与个性化学习。LLM 可以根据学生的学习能力、知识水平和兴趣爱好,提供个性化的学习内容推荐和智能辅导服务^[23]。

教学内容生成与课程设计。LLM 可以辅助教师进行教学内容的生成和课程设计。基于海量的教育资源和知识库,LLM 能够自动生成教案、课件、习题等教学材料,减轻教师的备课负担。

作业批改与评估自动化。LLM 可以自动批改学生的文字作业,如论文、报告等,并给出评分和修改建议。

智能写作与阅读助手。LLM 可以作为学生的写作和阅读助手,提供写作指导和阅读辅导。

跨学科知识整合与探究式学习。LLM 具备强大的知识整合和逻辑推理能力,可以将不同学科的知识进行关联和融合。这为跨学科学习和探究式学习提供了支持。

尽管 LLM 在教育领域有诸多应用价值,但其在实践中也面临一些挑战和问题,如学生抄袭、人工智能生成内容的潜在偏见、过度依赖 LLM、非英语国家和地区的使用障碍等^[1]。未来,教育工作者需要在充分利用 LLM 优势的同时,也要注重加强学生的

独立思考能力和学术诚信意识,确保 LLM 成为促进教育公平和提高教学质量的有益工具。

2.2 科研

LLM 在科学研究领域展现出了巨大的应用潜力,有望成为推动科研进展的重要助力。

在文献综述阶段,LLM 可以帮助研究者全面了解特定研究领域的进展情况^[24]。在研究思路生成阶段,LLM 展现出了提出有趣科学假设的能力^[25]。基于对已有研究成果的深入理解,LLM 可以通过知识的重组和推理,提出新颖的研究思路和假设,激发研究者的灵感,开拓创新性的研究方向。

在数据分析阶段,LLM 可用于对数据特征进行自动分析,包括数据探索、可视化和分析结论的推导^[26]。LLM 能够快速识别数据中的模式和趋势,生成直观的可视化图表,并根据分析结果提供洞见和建议,提高数据分析的效率和质量。

在论文写作阶段,研究者同样可以受益于 LLM 在科学写作方面的辅助。LLM 可以通过多种方式科学写作提供宝贵的支持,如总结现有内容、润色文字表达等^[27]。

此外,LLM 还可以辅助自动化的论文评审过程,包括错误检测、清单验证和候选论文排序等任务。

尽管 LLM 在科研领域取得了这些进展,但要成为真正有用、值得信赖的科研助手,还有很大的改进空间。一方面,需要进一步提高 LLM 生成科学内容的质量,确保其准确性、逻辑性和创新性;另一方面,要减少 LLM 产生有害幻觉的风险,避免误导研究方向或得出错误结论。

2.3 医疗

在医疗行业,LLM 的应用正在开启一场覆盖临床诊断、治疗护理、医学研究等方面的技术革命^[28]。

在临床诊断方面,将 LLM 用于对患者病历、检查报告和生理参数等进行深入的自然语言处理和整合医疗大数据分析,辅助医生快速准确地诊断疾病并制定治疗方案。

在治疗护理方面,Google 的 Med-PaLM 2 和 EPFL(École polytechnique Fédérale de Lausanne)的 Meditron 等工具展示了在提高临床效率和患者护理水平方面的应用潜力。

在医学研究方面,将 LLM 用于扫描和综合分析海量的科学论文,识别潜在的药物靶点和功效因素,加速了新药的研发^[29]。

在蛋白质设计和新药发现方面,LLM 已经展示出蛋白质序列处理和结构预测的强大能力,极大地提升了蛋白质设计的效率和准确性。

2.4 金融

LLM 在金融领域展现出了广阔的应用前景。

LLM 可用于从海量金融文本数据中自动提取关键信息,如金融实体、事件、关系等,并进行情感分析和观点挖掘。这有助于金融机构快速洞察市场动态,把握投资机会。

LLM 可通过对金融文本数据进行深入分析,识别潜在的金融风险,并做出预测。这包括对上市公司财务报告的分析,以发现财务造假等违规行为;对社交媒体等非结构化数据的分析,以预测市场情绪和股价走势等。

LLM 可用于辅助投资决策和组合管理。通过对金融市场的各种信息进行综合分析,LLM 可以提供投资建议和风险提示,帮助投资者优化投资组合,控制投资风险。

LLM 可应用于金融欺诈检测和反洗钱领域。通过对大量金融交易数据进行分析,LLM 可以发现异常交易模式和可疑行为,及时预警潜在的欺诈风险。在反洗钱方面,LLM 可协助金融机构进行客户尽职调查,识别高风险客户和交易,提高反洗钱的有效性。

总的来说,LLM 正在金融领域展现出巨大的应用潜力,有望成为驱动金融行业变革的重要技术力量。但与此同时,LLM 在金融领域的应用也面临着准确性、安全性、合规性等方面的挑战。

2.5 司法

LLM 在法律领域的应用正在蓬勃发展,为法律专业人士提供了强大的辅助工具。

近期,众多研究将 LLM 应用于法律文档分析^[30]、法律判决预测以及法律文书写作^[31]等任务。一项最新研究^[32]发现,LLM 表现出了强大的法律解释和推理能力。更令人瞩目的是,最新的 GPT-4 模型在模拟律师资格考试中取得了超过人类考生前 10% 的成绩。

总的来说,LLM 已经能够作为法律专业人士的得力助手,协助他们处理各种法律事务。然而,LLM 在法律领域的应用也引发了一些法律挑战,如版权问题、个人信息泄露以及偏见和歧视^[33]等,这些问题亟需在技术发展的同时予以重视和应对。

3 可信性、可控性与安全性研究现状

3.1 可信性

LLM 在可信性方面面临着诸多挑战。首先,LLM 可能生成包含歧视、仇恨言论、误导信息等有害内容,对个人和社会造成潜在危害。其次,由于训练数据通常来源于互联网,LLM 易继承其中存在的社会、文化、性别、种族等多维度偏见,导致模型产生刻板印象、不公平待遇或错误结论,加剧社会不公。

为提升 LLM 的可信性,研究者采取了多种策略。一方面,通过构建精细的风险分类体系,并采用红队测试等方法对模型进行全面的安全性评估^[34],以识别模型弱点并指导改进。另一方面,研究者探索角色扮演与情境模拟、数据增强与对抗性训练等技术,以检测和减少模型的潜在偏见,提高其在不同社会群体视角下的表现。此外,个性化反馈与定制化训练也有助于提升模型在特定任务上的可信度。

3.2 可控性

LLM 的可控性挑战主要体现在对复杂场景的应对不足以及难以完全遵循人类意图。当处理涉及伦理道德、法律规范、敏感话题等复杂场景时,LLM 往往难以准确判断和回应,尤其在缺乏明确指令或上下文约束的情况下,容易出现误判或有害输出。此外,攻击者可能利用精心设计的输入诱导 LLM 生成非预期内容,绕过内置的安全机制。

为增强 LLM 的可控性,研究者提出了多种解决方案。通过引入辅助 LLM 对主模型生成的响应进行分类和评分,可以为每个风险类别提供详细的安全性见解。同时,采用数据过滤与预处理策略,如关键词匹配和零样本分类,可以筛选和去除训练数据中可能引发危害的提示。此外,强调 LLM 需遵循明确的指导原则和评估框架,确保模型设计与部署符合负责任人工智能的原则,也是提高可控性的重要举措。

3.3 安全性

LLM 在安全性方面面临隐私泄露、数据合规以及对抗攻击^[35]等风险。模型可能无意间存储并泄露训练数据中的敏感信息,甚至通过生成模式暴露未公开数据集的内容^[34]。此外,LLM 在使用过程中需要遵守不同国家和地区的法律法规,确保数据处理和结果输出符合隐私保护和数据安全要求。

为提升 LLM 的安全性,研究者采取了多方面措施。首先,开发针对 LLM 的基准测试集,以符合政

府标准和法律法规的要求,同时鼓励研究者合作建立安全港,确保安全研究不受利益冲突影响。其次,对 LLM 进行定期安全评估和更新,监测新出现的风险并及时调整模型策略以适应变化的法规环境和社会期望。最后,公开发布评估数据集和代码,鼓励研究界持续探索 LLM 安全性的改进措施,推动安全技术的发展和共享。

4 挑战及应对

LLM 技术展现出了巨大的潜力和广阔的前景。然而,伴随着 LLM 的快速发展和日益广泛的应用,研究者们也不得不正视其在技术和社会层面所面临的诸多挑战。这些挑战不仅限于技术层面,更涉及伦理、法律等社会层面。只有正视并积极应对这些挑战,LLM 才能真正造福人类社会。

在技术层面,计算资源瓶颈、训练效率低下、部署成本高昂等问题亟需突破。为此,研究者需要在算法创新、架构优化、分布式训练、高效推理等方面持续发力。绿色人工智能理念的兴起为这一领域指明了方向——通过算法和硬件的协同设计,在保证性能的同时最小化能耗。这不仅有利于降低成本,更是 LLM 走向可持续发展的必由之路。

数据是 LLM 的生命之源,但高质量数据的匮乏已成为制约其发展的关键因素。尤其在医疗、法律等专业领域,隐私保护和知识产权的壁垒使数据获取难上加难。数据合成技术的进步为缓解数据饥渴提供了新思路,但仍面临着可能引入噪声和偏见的风险。未来,高效、可靠的数据增强算法将是研究的重点。同时,少样本学习和零样本学习等技术有望减少对大规模数据的依赖,但如何在小数据上达到甚至超越大数据训练的效果仍是一个开放的难题。

泛化能力是衡量 LLM 是否具备真正智能的关键指标。当前 LLM 虽然在特定任务上已显示出超人的能力,但面对全新的领域或复杂的语境时往往难以适从。究其原因,一方面在于模型缺乏对语言和世界的深层理解,另一方面也暴露了当前 LLM 在常识推理、类比思维等高层认知能力上的不足。为了突破这一局限,研究者需要在认知科学、神经科学等学科的启发下,开发出更加接近人脑机制的模型架构和学习范式。与此同时,构建覆盖更广泛语境的高质量数据集将是提升模型鲁棒性的重要举措。

在社会层面,伦理和偏见问题成为 LLM 应用的最大绊脚石。训练数据中的偏见会被模型放大,并

在应用中产生难以预估的负面影响。“黑盒”的决策过程更是引发了公众对于算法公平性和可解释性的强烈质疑。解决之道在于,加强伦理意识,将伦理准则融入模型设计的每个环节;加大力度开展去偏见研究,通过算法创新和数据增强最大限度地消除偏见;加快可解释人工智能技术突破,赋予模型“自我解释”的能力。与此同时,行业自律和外部监管并重,共同构建 LLM 健康发展的长效机制。

法律法规是确保 LLM 造福社会的重要保障。当前,LLM 在数据隐私、知识产权等方面还存在诸多模糊地带。数据主权、模型版权归属等问题亟待厘清。这就要求法律工作者与技术专家通力合作,以技术发展为牵引,以伦理为导向,以民意为依托,加快相关法律法规的制定和完善。唯有法律红线和伦理底线并重,LLM 才能在创新与规范的道路上行稳致远。

5 结束语

从 2021 年 ChatGPT 问世以来,LLM 的发展速度势如破竹,令人目不暇接。伴随着各种大胆的尝试和探索,LLM 的能力边界还会大幅延伸,从而进一步向经济、社会、科技、军事、文化等各个领域强力渗透,并产生更多的颠覆性作用。

同时,所引发的挑战和问题也与日俱增,对价值伦理、公平公正、社会信任、社会分工等带来和即将带来的冲击难以估量。因此,对 LLM 的快速成长产生关切和忧虑是理所当然的。然而,这些关切和忧虑不应也不能阻挡 LLM 的发展。从 LLM 的本质属性和技术特征来看,人类社会一定能够将其始终置于正确的发展道路上,为自身带来更多福祉。

展望 LLM 的下一步发展趋势,笔者认为,伴随着各种大胆的尝试和探索,LLM 的能力边界还会大幅延伸,从而进一步向经济、社会、科技、军事、文化等各个领域强力渗透,并产生更多的颠覆性作用。

眼下来看,多专家混合 LLM 架构正在成为新的潮流,其中的路由机制是一个研究热点。LLM 智能体机器人领域已成为统领性技术,不仅极大地提高了机器人感知、判断、规划和决策等能力,还通过支撑主动视觉、主动听觉、主动触觉、自主抓取、自主行走等关键要素成为促进具身智能发展的主要动力。人工智能科学正成为科学研究的一个新范式,其在蛋白质结构预测、气象预测、新药发现等方面的惊人进展预示着不可估量的发展前景。

毫无疑问,LLM 也必将在体系化人工智能中发挥决定性作用。从角色上看,它既可以作为体系化人工智能的一类组件分别发挥专门作用,形成一个面向开放场景、复杂环境的超级智能体,也可以作为基础 LLM,通过与外围组件的协同工作实现更加通用的人工智能。从研究上看,体系化人工智能所要求的可信性、可控性和安全性将主要通过 LLM 的相关技术和方法实现。体系化人工智能的网络化部署问题将首先是 LLM 的网络化部署问题。体系化人工智能所推动的闭环服务也可基于开发和利用 LLM 的关键机制——人工反馈强化学习^[36]加以研究和实现。

参考文献:

- [1] KASNECI E, SESSLER K, KÜCHEMANN S, et al. ChatGPT for good? On opportunities and challenges of large language models for education [J]. *Learning and Individual Differences*, 2023, 103: 102274-102283.
- [2] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding [C] // NAACL HLT 2019-2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies Proceedings of the Conference. Stroudsburg: ACL, 2019: 4171-4186.
- [3] VASWANI A, SHAZEER N M, PARMAR N, et al. Attention is all you need [C] // *Advances in Neural Information Processing Systems 30 (NIPS 2017)*. New York: Curran Associates, 2017: 5998-6008.
- [4] FLORIDI L, CHIRIATTI M. GPT-3: its nature, scope, limits, and consequences [J]. *Minds and Machines*, 2020, 30: 681-694.
- [5] OUYANG L, WU J, JIANG X, et al. Training language models to follow instructions with human feedback [C] // *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*. New York: Curran Associates, 2022: 27730-27744.
- [6] QIN C W, ZHANG A, ZHANG Z S, et al. Is ChatGPT a general-purpose natural language processing task solver? [C] // *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Stroudsburg: ACL, 2023: 1339-1384.
- [7] RAFFEL C, SHAZEER N M, ROBERTS A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer [J]. *Journal of Machine Learning Research*, 2019, 21(140): 1-67.

- [8] DU N, HUANG Y P, DAI A M, et al. GLaM: efficient scaling of language models with mixture-of-experts[C]// Proceedings of the 39th International Conference on Machine Learning (ICML 2022). New York: PMLR, 2022: 5547-5569.
- [9] SHAZEER N, MIRHOSEINI A, MAZIARZ K, et al. Outrageously large neural networks: the sparsely-gated mixture-of-experts layer[C]// 5th International Conference on Learning Representations (ICLR 2017) - Conference Track Proceedings. New York: ICLR, 2017: 2035-2054.
- [10] ZHANG Z Y, HAN X, LIU Z Y, et al. ERNIE: enhanced language representation with informative entities[C]// ACL 2019-57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2019: 1441-1451.
- [11] PETERS M E, NEUMANN M, IYYER M, et al. Deep contextualized word representations[C]// Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: ACL, 2018: 2227-2237.
- [12] LESTER B, AL-RFOU R, CONSTANT N. The power of scale for parameter-efficient prompt tuning[C]// Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2021: 3045-3059.
- [13] TANG T Y, LI J Y, ZHAO W X, et al. Context-tuning: learning contextualized prompts for natural language generation[C]// Proceedings of the 29th International Conference on Computational Linguistics. New York: International Committee on Computational Linguistics, 2022: 6340-6354.
- [14] HOULSBY N, GIURGIU A, JASTRZCZBSKI S, et al. Parameter-efficient transfer learning for NLP[C]// Proceedings of the 36th International Conference on Machine Learning. New York: PMLR, 2019: 2790-2799.
- [15] CHOWDHURY A, NARANG S, DEVLIN J, et al. PaLM: scaling language modeling with pathways[J]. Journal of Machine Learning Research, 2023, 24(240): 1-113.
- [16] DENG B L, LI G Q, HAN S, et al. Model compression and hardware acceleration for neural networks: a comprehensive survey[J]. Proceedings of the IEEE, 2020, 108(4): 485-532.
- [17] FANG G F, MA X Y, SONG M L, et al. DepGraph: towards any structural pruning[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Los Alamitos: IEEE Computer Society, 2023: 16091-16101.
- [18] GOU J P, YU B S, MAYBANK S J, et al. Knowledge distillation: a survey[J]. International Journal of Computer Vision, 2021, 129(6): 1789-1819.
- [19] CLAVIÉ B, CICEU A, NAYLOR F, et al. Large language models in the workplace: a case study on prompt engineering for job type classification[C]// Natural Language Processing and Information Systems. Cham: Springer Nature Switzerland, 2023: 3-17.
- [20] BROWN T B, MANN B, RYDER N, et al. Language models are few-shot learners[C]// Advances in Neural Information Processing Systems 33 (NeurIPS 2020). New York: Curran Associates, 2020: 1877-1901.
- [21] WEI J, WANG X Z, SCHUURMANS D, et al. Chain of thought prompting elicits reasoning in large language models[C]// Advances in Neural Information Processing Systems 35 (NeurIPS 2022). New York: Curran Associates, 2022: 24824-24837.
- [22] SONG C H, WU J M, WASHINGTON C, et al. LLM-Planner: few-shot grounded planning for embodied agents with large language models[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway, NJ: IEEE Press, 2022: 2986-2997.
- [23] KAMALOV F, CALONG D S, GURRIB I. New era of artificial intelligence in education: towards a sustainable multifaceted revolution[J]. Sustainability, 2023, 15(16): 12451-12486.
- [24] AYDIN Ö, KARAARSLAN E. OpenAI ChatGPT generated literature review: digital twin in healthcare[J]. Emerging Computer Technologies, 2022: 22-31.
- [25] PARK Y J, KAPLAN D, REN Z C, et al. Can ChatGPT be used to generate scientific hypotheses? [J]. Journal of Materiomics, 2024, 10(3): 578-584.
- [26] CHENG L Y, LI X X, BING L D. Is GPT-4 a good data analyst? [C]// Findings of the Association for Computational Linguistics (EMNLP 2023). Stroudsburg: ACL, 2023: 9496-9514.
- [27] BURUK O. Academic writing with GPT-3.5: reflections on practices, efficacy and transparency[C]// Mindtrek 23: Proceedings of the 26th International Academic Mindtrek Conference. New York: ACM Press, 2023: 144-153.
- [28] HE K, MAO R, LIN Q, et al. A survey of large language models for healthcare: from data, technology, and

- applications to accountability and Ethics [EB/OL]. (2023-10-09) [2024-04-15]. <https://doi.org/10.48550/arXiv.2310.05694>.
- [29] NARGANES-CARLÓN D, CROWTHER D J, PEARSON E R. A publication-wide association study (PWAS), historical language models to prioritise novel therapeutic drug targets[J]. *Scientific Reports*, 2023, 13(1): 8366-8385.
- [30] BLAIR-STANEK A, HOLZENBERGER N, VAN DURME B. Can GPT-3 perform statutory reasoning? [C]// 19th International Conference on Artificial Intelligence and Law (ICAIL 2023). New York: ACM Press, 2023: 22-31.
- [31] CHOI J H, HICKMAN K E, MONAHAN A, et al. ChatGPT goes to law school[J]. *Journal of Legal Education*, 2022, 71: 387-403.
- [32] NAY J. Law informs code: a legal informatics approach to aligning artificial intelligence with humans[J]. *Northwestern Journal of Technology and Intellectual Property*, 2022, 20: 309-374.
- [33] ABID A, FAROOQI M, ZOU J. Persistent anti-muslim bias in large language models [C]// AIES 21: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York: ACM Press, 2021: 298-306.
- [34] PEREZ E, HUANG S, SONG F, et al. Red teaming language models with language models [C]// Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2022: 3419-3448.
- [35] WEI A, HAGHTALAB N, STEINHARDT J. Jailbroken: how does LLM safety training fail? [C]// Advances in Neural Information Processing Systems 36 (NeurIPS 2023). New York: Curran Associates, 2023: 80079-80110.
- [36] CHRISTIANO P F, LEIKE J, BROWN T B, et al. Deep reinforcement learning from human preferences [C]// Advances in Neural Information Processing Systems 30. New York: Curran Associates, 2017: 4302-4310.

(上接第19页)

- [28] 刘建强, 卢为党, 黄国兴, 等. 基于深度学习的互联网虚假信息识别研究[J]. *情报工程*, 2022, 8(5): 86-100.
- LIU J Q, LU W D, HUANG G X, et al. Identifying Internet fake information based on deep learning[J]. *Technology Intelligence Engineering*, 2022, 8(5): 86-100.
- [29] 冯文杰, 熊翱. 基于新闻时效性的协同过滤推荐算法[J]. *计算机系统应用*, 2018, 27(5): 193-197.
- FENG W J, XIONG A. Collaborative filtering recommendation algorithm based on news timeliness[J]. *Computer Systems & Applications*, 2018, 27(5): 193-197.
- [30] 张勇, 李丹丹, 韩璐, 等. 隐私保护的群体感知数据交易算法[J]. *通信学报*, 2022, 43(5): 1-13.
- ZHANG Y, LI D D, HAN L, et al. Privacy-protected crowd-sensed data trading algorithm[J]. *Journal of Communications*, 2022, 43(5): 1-13.