

2023 年生成式 AI 大模型发展热点回眸

邓佳文, 任福继*

电子科技大学计算机科学与工程学院, 成都 611731

摘要 2023 年, 生成式大模型技术发展迅速, 取得了一系列突破性的进展。回顾了 2023 年备受关注的大模型关键技术, 包括大模型的能力涌现、多模态大语言模型的发展, 以及大模型对齐和知识增强技术。介绍了大模型技术在医疗、教育等领域的垂直应用, 以及对 AI 智能体与元宇宙技术发展的促进作用。此外, 从数据隐私、有偏价值观、版权争议及虚假新闻传播等方面讨论了生成式大模型技术面临的安全挑战和发展趋势。

关键词 生成式人工智能; 大模型; 人工智能内容生成; AI 安全

生成式人工智能 (generative AI) 技术, 是指通过人工智能算法, 在大规模训练数据上学习之后的生成式 AI 模型, 能够自动生成文本、图像、音频或视频等各种形式内容的技术。2023 年, 生成式 AI 技术的发展达到了新的高峰, 以 ChatGPT 为代表的生成式 AI 应用和工具, 通过人机交互界面, 在对话、问答、写作、摘要、翻译等任务上展示了强大的能力, 在易用性和功能性方面都有显著的提升, 不仅在技术界引起巨大的反响, 也在普通用户中引发了广泛的关注和使用。

生成式大模型在 2023 年百花齐放, 学术界和工业界都对生成式 AI 的发展给予了极高的关注, 全球范围内的多家头部公司和研究机构积极研发和发布各自的大型模型, 包括 OpenAI 的 GPT 系列^[1]、Meta 的 LLaMA 系列^[2-3]、谷歌 DeepMind 的

Gemini 模型^[4]、以及清华大学的 GLM 系列^[5]、复旦大学的 MOSS^[6]、华为的盘古大模型、百度的文心一言、阿里的通义千问、科大讯飞的星火认知大模型等。这些模型在处理复杂的语言任务、图像识别或生成等方面展现了卓越的性能, 推动了整个 AI 大模型领域的发展。AIGC 技术的发展带来了广泛的行业应用, 例如, 在媒体行业, 可用于自动编写新闻稿; 在教育领域, 用于个性化教学和学习材料的制作; 在医疗行业, 用于辅助医疗决策等, 各行各业都在探索如何利用这些先进的技术来提高效率、降低成本并创造新的价值。生成式 AI 大模型被视为引发通用人工智能发展的火花, 已在很大程度上成为 AI 领域发展的主流方向之一, 其强大能力和广泛的应用表明其在 AI 未来发展中将占据重要位置。同时, AI 的发展是多元化和多方向的, 关于 AI 隐私

收稿日期: 2023-12-30; 修回日期: 2024-01-08

作者简介: 邓佳文, 助理研究员, 研究方向为人工智能、大模型技术、情感计算等, 电子信箱: dengjw@uestc.edu.cn; 任福继 (通信作者), 教授, 日本工程院院士、欧盟科学院院士、俄罗斯工程院外籍院士, 研究方向为先进智能、情感计算、智能机器人等, 电子信箱: renfuj@uestc.edu.cn

引用格式: 邓佳文, 任福继. 2023 年生成式 AI 大模型发展热点回眸[J]. 科技导报, 2024, 42(1): 266-285; doi:10.3981/j.issn.1000-7857.2024.01.017

保护、AI伦理安全等相关的研究正在增加。各国政府和国际组织愈发关注生成式AI的发展,并制定相关政策以鼓励和规范这一领域的发展,各单位和机构也积极发布相关白皮书,涉及数据隐私、知识产权保护,以及AI伦理等多个方面,旨在促进技

术的健康和可持续发展^[7-9]。

大模型技术迅速发展的背景下,探索大模型领域的关键技术及进展显得尤为重要。2023年生成式AI大模型技术主要研究热点如表1所示。

表1 2023年生成式AI大模型技术主要研究热点

主要方向	研究热点	研究热点描述
生成式AI大模型关键技术	大模型的能力涌现	围绕大模型能力涌现现象、具体表现及产生的原因展开介绍和讨论
	多模态大语言模型	将大语言模型的能力扩展到多模态领域,能有效提高多模态信息理解和生成能力
	大模型价值对齐	AI对齐技术致力于研究如何使AI系统的行为与人类的价值观和偏好保持一致
	大模型知识增强	大模型知识增强技术研究如何提高模型知识的专业性和时效性,使生成更可信
基于大模型技术的应用	垂直领域的大模型应用	大模型赋能各行各业发展,在智慧办公、智慧教育、信息检索、智慧医疗、社交陪伴等领域具有广泛的应用
	基于大模型的AI智能体的应用	生成式大模型有助于增强AI智能体的环境感知、大脑推理和行为决策能力,促进自主性AI系统的实现
	大模型技术助力元宇宙	生成式AI技术助力AI驱动的数字虚拟人实现,有助于推进元宇宙发展
AIGC发展的挑战与讨论	数据安全和隐私保护	讨论生成式大模型在训练和应用过程中引发的隐私泄露和数据安全风险,讨论大模型时代的隐私防范技术
	有偏价值观生成风险	讨论大模型在有偏内容生成方面的安全挑战和防范措施
	AI生成内容的版权争议	讨论AIGC内容的版权归属争议,以及AIGC内容检测与数字水印等防范技术
	AI加速虚假新闻传播	讨论AIGC背景下虚假新闻制造成本降低、传播风险加剧的挑战,并从事实核查、深度伪造检测等方面介绍对抗虚假新闻传播的方法
	有效加速主义与“超级对齐”	针对AI大模型技术加速发展的当下,讨论未来AI技术的发展理念

1 生成式AI大模型关键研究

以大语言模型(large language models, LLMs)为代表的生成式AI大模型,通常是指在海量数据上训练、参数规模达数百亿及以上的Transformer模型,如GPT-4和LLaMA。这些大模型在语言理解和解决复杂任务等方面展现了强大的能力。

1.1 大模型的能力涌现

生成式大模型的能力涌现是一个引人注目的现象,在2023年的研究中获得了广泛的关注和讨论。这一现象指的是当模型规模达到一定程度时,其性能会表现出非线性的显著提升,在语言理解、生成和推理方面表现出意想不到的能力^[10]。这种由量变引起的质变现象,称为能力涌现,即在小模

型中没有的能力,会在更大的模型中存在。

一般来说,当模型规模达到百亿或千亿级别时,可能产生能力涌现的现象,包括上下文学习(in-context learning)能力、多步推理(multi-step reasoning)能力,以及指令遵从(instruct following)能力。在少样本提示(few-shot prompting)范式下,大模型涌现了显著的上下文学习能力,即使面对训练时未曾接触过的新任务,只要在提示中给出少量相关的输入输出案例,就能够理解并执行这项新任务。这在多任务语言理解和问答任务中得到了有效验证。在多步推理能力方面,小模型在解决涉及多步骤的推理任务时具有一定的挑战,但大模型的推理能力使其能够在生成最终答案之前产生一系列中间推理步骤,直至完成最终任务。例如,在解

决数学题和逻辑思考题时,大模型不仅给出最终答案,还能生成解题的每一步逻辑,显示出类似人类的推理能力。指令遵从能力是指对于 ChatGPT 之类经过指令微调的大模型,即使在没有小样本提示的情况下,它们也能够正确理解并响应新的指令,这表明大模型在理解复杂任务和多样化指令方面具有显著的优势。

大模型的推理和规划能力是能力涌现的一个重要体现。通过思维链(chain-of-thought)提示的策略,在零样本(zero-shot)设置下,通过在输入指令中加入“Let’s think step by step”之类的文本描述能够引导大模型推理生成更准确的结果(图1)^[11-12]。为了应对单次推理结果可能存在的不准确性,研究者提出利用大模型的自我一致性(self-consistency)特性^[13],即使用思维链提示重复请求多次以采样不同的推理路径,然后通过投票策略选择最一致的答案作为最终推理结果。考虑到零样本设置下的思维

链推理的计算错误和步骤错误,研究者提出先规划、后解决的提示方法^[14],即先制定一个计划,将整个任务分解为较小的子任务,然后按照计划逐步执行子任务,使大模型能够明确制定解决问题的计划,并在预测输入问题的最终答案之前生成中间推理过程。另外,优化指令本身也可以提高模型推理能力。通过自我完善(self-polish)的方法让大模型优化输入,使指令更加简洁、清晰,去除无关信息的干扰,使模型更易于理解和回答问题^[15]。此外,让大模型学会自我精炼(self-refine),利用迭代的反馈和改进有助于提高大模型解决问题的能力,即当大模型生成一个输出之后,让该模型为本身的输出提供多方面的反馈,并基于自我反馈对原始的输出进行改进^[16]。大模型也可以在选择和推理之间交替^[17],生成一系列可解释的推理步骤,从中选择可靠的子集,并基于选择的子集生成新的事实推理,从而得到最终的答案。

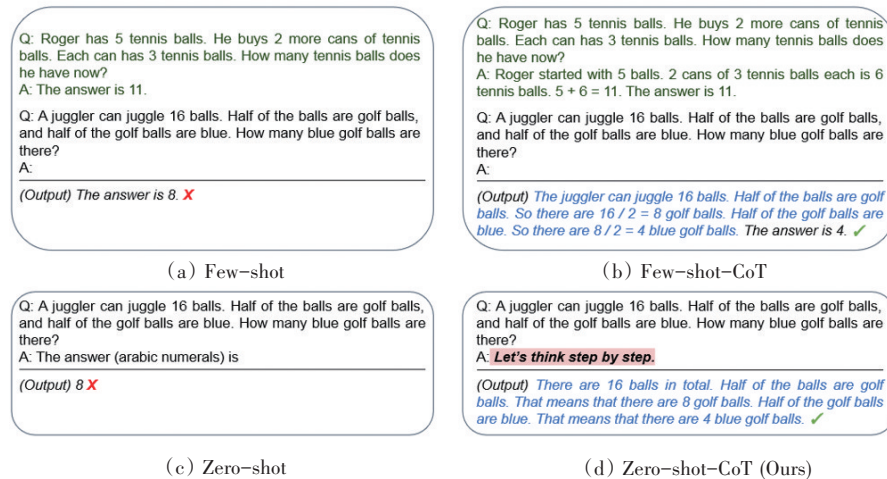


图1 GPT-3模型的思维链(CoT)推理示例

关于大模型的能力涌现现象是否存在是被广泛讨论的问题之一,提出该现象的论文在2023年谷歌学术引用量新增近千次^[10]。一些研究人员认为,对于特定的任务和模型,出现涌现能力现象是由于研究人员所选择的评测方式,而不是由于模型行为随规模的根本变化而出现的^[18]。具体来说,非线性或不连续指标会产生明显的突现能力,而线性或连续指标会产生模型性能的平滑、连续、可预测

的变化。相关实验证明,能力涌现会随着不同的指标或更好的统计数据的出现而消失,并且可能不是扩展AI模型的基本属性^[18]。另外,有研究人员指出能力的涌现与评测任务本身的复杂性有关。展现出涌现现象的任务通常是由多个子任务构成的复杂任务,对于简单的子任务,其性能往往随着模型规模的增大而平稳地提升,符合标度律(scaling law),但当这些子任务的表现提升到一定程度就会

出现执行整体复杂任务的能力涌现现象。此外,也有观点认为,涌现现象是由评测时模型规模采样稀疏导致的。在体现大模型能力涌现现象的图表中,体现性能的纵坐标是线性增长的,而体现模型规模的横坐标则采用了对数刻度(log-scale)的非线性变化。在这样的坐标体系中,从小模型到大模型的扩展过程中缺乏中间规模的模型采样稀疏,这也可能是出现能力涌现的原因之一。

随着参数规模和数据规模的增大,生成式大模型在各个方面表现出了强大的能力。能力涌现是一个复杂且有争议的现象。一方面,这种现象展示了大模型在某些方面具有小模型所不具备的能力,为AI的应用和研究提供了新的视角和可能性。另一方面,对于这种能力涌现的本质和成因还有待进一步研究和探索。未来的研究需要更加深入地分析不同模型规模下的性能变化,以及这些变化背后的机制和原因,以便更全面地理解和利用大模型的潜力。

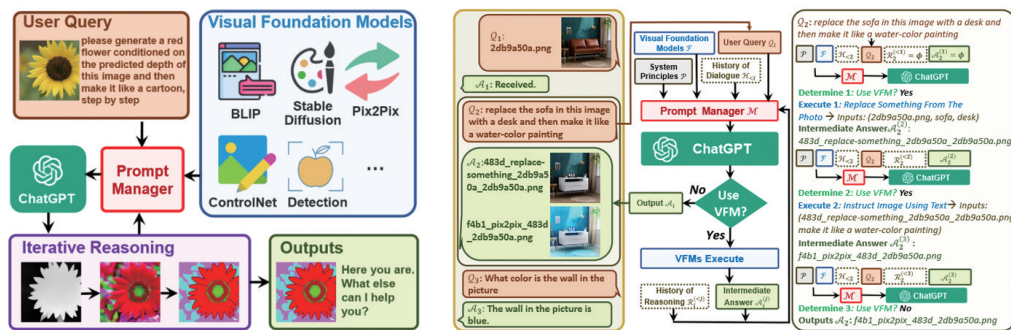
1.2 多模态大语言模型

语言大模型的发展对多模态领域也产生了重要的促进作用。尽管语言大模型在语言推理和生成任务上展现了惊人的能力,但是其只能理解序列文本,而无法处理视觉等其他模态的输入内容。2023年3月,OpenAI发布了GPT-4,和只能处理纯文本模态的GPT-3和InstructGPT模型相比,GPT-4可以接受文本图像的多模态输入,这更接近于人类感知世界的方式,且模型在视觉理解和推理方面的能力给用户留下了深刻的印象。虽然GPT-4的技

术细节并未完全公开,但研究者开始思考如何构建这样一个能够遵从用户指令处理多模态任务的大模型。至此,多模态大语言模型开始被广泛关注。

多模态大语言模型(multimodal large language models, MLLM)是指通过将大语言模型(LLM)与视觉或其他模态的模型相结合,使不同模态的信息之间可以相互补充和增强,赋予LLM处理和生成多模态信息的能力,如LLaVA^[19]、MiniGPT-4^[20]、CogVLM^[21]等模型。与纯文本模式的LLM相比,MLLM能够自然地接收多感官输入,这不仅使模型能够处理更复杂的交互任务,还能为用户提供更友好的交互界面,使用户可以选择性地使用文本、语音、图像等不同方式与大模型进行交互^[22]。

将大语言模型的能力扩展到多模态领域,一种常用的方法是整合一系列其他模态的模型,将输入的多模态信息转化为文本描述后送入LLM进行处理,并生成回答^[23]。例如,微软亚洲研究院于2023年3月提出的Visual ChatGPT结合了多个视觉模型^[24],包括能够理解和提供图像文本描述的BLIP模型、根据文本提示合成图像的Stable Diffusion模型,并将ChatGPT作为推理的大脑中心连接其他模态的模型,回复用户的多模态请求(图2)。得益于大模型的能力涌现,扩展到多模态之后的MLLM在视觉推理等任务上具备强大的理解和推理能力,包括多模态上下文学习、多模态思维链推理等。这样的方法不需要微调语言模型,但其能力受限于所连接的多模态小模型的能力,而且使用文本作为信息传递的桥梁可能会导致信息失真,影响模型的推理结果。



(a) Visual ChatGPT结构示意图

(b) 基于Visual ChatGPT的多轮交互示意

图2 Visual ChatGPT 的结构与多轮对话案例

为了更好地执行用户指令并实现各个模型之间的信息对齐,多模态大语言模型通常采用多模态指令微调方法,这要求在模型和数据方面都进行相应的调整^[22]。多模态指令微调数据通常由指令和输入输出对组成。指令类似于“请生成以下图片的标题”“基于给定图像回答以下问题”之类的指令问题描述,输入输出则和传统多模态任务类似,例如,对于图片标题生成任务,输入包括给定图像,输出是文本答案。这些数据可以通过调整现有的多模态基准数据集获取,例如,指令数据集 MultiInstruct 收集了包含视觉问答、图片理解、多模态常识推理等 10 种类型多模态任务在内的数据,并结合人工编写和基于模板构造的方式生成了相应的指令描述^[25]。在模型方面,为了将其他模态的信息注入

LLM,常见的方法是在预训练的多模态编码器和 LLM 之间引入可学习接口,实现模态之间的桥接,例如,LLaVA^[19]使用额外的线性层连接图片编码器和语言模型,LLaMA-Adapter^[26-27]和 LaVINp^[28]引入适配器模块。MiniGPT-4^[20]是一个能处理视觉和语言任务的模型,其通过投影层结合 BLIP 模型和大语言模型 Vicuna,并通过视觉语言对齐训练和多模态指令微调训练,具备了类似 GPT-4 中所展示的视觉理解能力。考虑到当前多模态大模型大多局限于对多模态输入的理解而缺乏多模态的生成能力,腾讯人工智能实验室与悉尼大学提出了统一的多模态框架 GPT4Video^[29],它既可以理解也能生成图片和视频模态的内容,填补了当前多模态大模型的这方面的空白(图 3)。



图 3 多模态大模型 GPT4Video 在对话中的视频理解和视频生成示例

除了被广泛关注的视觉和文本模态,将大语言模型和视觉等其他模态的模型互相增强,能有效提高 LLM 处理多模态信息的能力。Meta 在 2023 年 5 月推出了跨模态大模型 ImageBind 能够同时处理 6 种感官模态^[30],它将 6 种模态的信息嵌入到一个共享表示空间以促进模态对齐,支持跨不同模态组合的多模态任务,包括文本、视觉(图像/视频)、音频、深度信息、温度(红外图像)、惯性测量单元(IMU)。在此基础上,ImageBind-LLM^[31]引入可学习的网络,将多模态模型 ImageBind 与语言模型 LLaMA 之间的编码对齐,并通过多模态指令微调使其具备响应多模态指令的能力。2023 年 12 月,谷歌发布多模态大模型 Gemini^[4]。据介绍,Gemini 采用了一种原

生多模态的方法,即在初始阶段就在不同模态上进行预训练,并利用额外的多模态数据进行微调,这使其能够在最初阶段就能理解和推理输入的各种内容,从而在图像、音频、视频和文本理解方面的表现优于现有的多模态模型。

在 AI 浪潮持续爆发的背景下,多模态大语言模型(MLLM)的发展是当前人工智能领域中的一个重要趋势,这不仅是对以文本模态为主的大语言模型功能的扩展,也标志着人工智能领域在理解和生成多种感官信息方面的重大进步。通过结合视觉、语音、文本等多种模态,MLLM 能够更全面地模拟人类的感知和认知过程,从而在多模态理解、生成和交互能力上实现显著的提升。

1.3 大模型价值对齐

随着AI大模型能力越来越强,暴露的安全风险越来越突出,如隐私泄露、偏见内容、幻觉输出等。出于对未来AI风险的担忧,2023年3月,Musk等数千名科技人士发表公开信,呼吁所有人工智能实验室立即暂停比GPT-4更强大的人工智能系统的训练,暂停时间为6个月以上。随后,图灵奖获得者Yoshua Bengio和Geoffrey Hinton在内的数百名AI领域专家联合签署AI风险声明^[32],认为AI风险应该具有与流行病和核战争等其他大规模社会性风险一样,成为全球优先解决的事项。

AI风险是大模型研究中不可忽视的问题。AI对齐研究的核心目的是确保人工智能系统的行为与人类的价值观和偏好保持一致,这对于提高AI技术的可信度、可靠性和安全性至关重要。当下大模型训练的常规范式是预训练加微调。在预训练阶段,将在海量无监督数据上学习,学到通用的语言表示,得到预训练模型。在微调的阶段,通常在有监督的数据集上学习,以获得可以解决新的类似任务的能力,适配到下游任务。随着生成式大模型参数规模和训练数据量的持续增加,经过预训练和微调的语言模型,通常具备较强的生成能力,但是却出现了与用户预期不一致的输出,例如,生成误导性的信息、带有偏见的评论或不恰当的内容。OpenAI在2020年发表的论文^[33]中指出,有监督微调过程的优化目标是最大化(唯一)正确答案的Likelihood,而人们关心的目标是生成人类认为高质量的输出,这2个目标之间存在偏差。这种偏差产生的主要原因是监督信号的稀疏,最大似然目标不能区分重要错误(如编造事实的输出)和不重要的错误(如同义词替换后的输出)。因此,在这种微调范式下得到的大模型会不可避免地出现与人类预期不一致的输出,这些问题不仅影响用户体验,还可能导致严重的社会和伦理问题。

为了解决AI对齐的问题,研究者提出基于强化学习的指令微调(RLHF)技术。RLHF的目标在于通过人类的反馈信号进一步优化生成式模型,以更紧密地捕捉人类关心的行为。与传统的微调方法相比,RLHF更加注重模型输出的质量和与人类

价值观的一致性。具体来说,AI对齐分成3个阶段。首先是有监督的微调(SFT)阶段,使用基于prompt数据训练监督策略,对预训练的模型进一步微调,使其具有指令遵从的能力。其次是奖励模型的训练,通过收集对比数据,即针对prompt输入的不同回复,根据人类偏好对各个回复的质量进行打分,以训练奖励模型。最后是生成策略优化阶段,奖励模型根据生成回复的质量计算奖励,这个奖励作为强化学习框架中的反馈信号,并用于更新当前策略的模型参数,从而引导模型的输出更符合人类的期望。这一阶段通过人类反馈调整模型的产出、优化模型的生成策略,以缓解有害输出等问题,使模型与人类价值观对齐。RLHF对性能的优化在多个大模型上得到了有效的验证,如OpenAI开发的InstructGPT、ChatGPT和GPT-4。

在AI对齐研究中,基于反馈的强化学习方法受到了广泛的关注。例如,Meta开源的LLaMA2模型,在实现RLHF优化过程中标注了高达140万的偏好数据,国内开源的PKU-Beaver(河狸)和MOSS-RLHF^[34]模型也采用了RLHF技术。在Anthropic提出的“宪法人工智能”(constitutional AI)模型中,提出使用RLAIF(reinforcement learning from AI feedback)技术训练无害的AI助手。RLAIF是RLHF技术的一个变种,使用大模型代替人类来进行偏好标记。这种方法的关键在于通过自我提升,使用AI自动生成的评论和修正来训练AI,从而减少对人工标注的依赖。RLAIF的实现主要包括2个阶段:监督学习(supervised learning, SL)和强化学习(reinforcement learning, RL)。在监督学习阶段,AI模型首先生成对有害提示的响应,然后根据一系列明确的原则(即“宪法”)进行自我评价和修正,这些自我优化过后的数据有助于调整模型的行为。这个阶段的目标是调整模型响应的分布,以减少随后强化学习阶段所需的探索量和整体训练时间。在强化学习阶段,模型模仿RLHF的过程,但是使用AI反馈替代人类偏好。这一阶段使用宪法原则评估响应,并将其精炼为一个偏好模型(preference model, PM),该模型结合了人类和AI的标签。然后,针对PM进行强化学习训练,以进一步

提升模型的性能和可靠性。RLAIF 提供了一种减少对人类标签依赖的优化方法,不仅提高了 AI 助手的无害性和透明度,同时也展示了 AI 在自我监督和自我提升方面的潜力。

随着大模型技术的不断进步,AI 对齐成为了一个至关重要且需要持续探索的研究领域。2023 年 7 月,OpenAI 宣布成立了专注于 AI 对齐研究的 SuperAlignment 团队,致力于在未来 4 年内解决超智能 AI 对齐问题。其主要研究方向是构建一个与人类智力相当且可靠的自动对齐研究器,旨在深入探索和解决超级智能对齐的挑战。AI 对齐的研究是当前人工智能与大模型领域最紧迫的挑战之一,需要更多研究团队的参与和持续探索。

1.4 大模型知识增强

大模型在应用中展现了强大的生成能力,但影响其实际应用部署的一个关键挑战在于其幻觉问题。在心理学中,幻觉指缺乏现实刺激场景下出现的虚幻感受,如幻听和幻视等,是一种较为严重的知觉障碍^[35]。在 AI 和自然语言处理领域,幻觉现象通常是指大模型生成的内容与上下文不一致或偏离既定事实^[36],也就是说在给定上下文语境(context)的前提下生成的内容和源内容相冲突或无法基于源内容进行验证。产生幻觉现象的一个主要原因是大模型在训练过程中学习的知识以通用性知识为主,在专业性和时效性方面与现实世界知识之间存在一定的差距,导致模型在使用过程中可能产生看似合理,但实际上不准确或错误的答案。这种不可靠的输出会影响大模型的可信度,在医疗、法律等专业性强的领域中具有潜在的伦理风险。

在大模型技术的发展和应用中,通过检索外部信息实现大模型的知识增强是解决大模型幻觉问题的有效途径。知识增强的主要思想是针对用户输入的提问,通过检索的方式引用外部数据,并将其提供给大模型,以增强其生成准确和相关答案的能力。这样的策略不仅可以引入额外知识,增强大模型的记忆和推理能力,还可以通过更新或替换其依赖的外部知识库实现模型知识的实时更新。

检索增强的模型通常包括一个检索增强模块和一个语言模型。例如,Atlas 模型采用了检索增

强学习的方法,通过检索外部数据来辅助生成更准确的回答^[37]。WebBrain^[38]能够利用检索到的相关文档片段辅助生成高质量的富含知识的文档,并明确标注出相关引用,从而显著提升用户获取信息的效率,并保证生成内容的事实性和可信性。类似地,WebGLM^[39]作为一个百亿参数的联网问答聊天机器人,在生成答案时会列举出网上相关的文章链接作为答案的知识来源,但除此之外,还额外引入了基于人类偏好的打分模块。WebGLM 的具体流程包含 3 个步骤:首先,针对输入内容,通过检索得到若干个相关的文档作为参考知识;其次,生成器基于用户提问和参考知识生成多个答案;最后,打分器从这些答案中选出得分最高的答案作为最终答案输出给用户。这个流程不仅提高了答案的质量,而且增加了可信度和透明度。

对于 ChatGPT 之类的黑盒设置大模型,其参数通常是冻结不可更新的。在这种情况下,引入额外的模块来优化模型输入的上下文信息变得尤为重要。REPLUG 模型^[40]引入了即插即用的检索模块,将检索得到的信息和原始输入一起送入大模型,有效地为模型的输入提供额外的知识,而不需要对模型进行额外的训练。此外,检索增强通过 query 检索到想要的知识,一个关键问题是如何设计有效的 query,解决 query 文本表达和真正想要的知识之间的差距,这直接影响到增强的性能。为此,引入额外的查询优化器是一种常用的方法^[41],对输入的查询文本进行改写,然后再进行检索,以获得更匹配的返回结果。

RETA-LLM 是一个包含多个即插即用模块的检索增强大模型框架(图 4)^[42]。除了检索器和生成器 2 个基本模块用于检索知识和生成回复之外,还提供了 3 个可选的模块:用户请求重写、相关段落提取和事实核查。首先,针对用户的输入,RETA-LLM 的用户请求重写模块会结合上下文对话记录对用户输入的请求进行修改,使其变得更加完整和清晰,以缓解用户请求时表述不完整的情况。随后,基于改写后的用户请求,文档检索模块将从外部知识库中召回相关的文档。考虑到大模型可接收输入的长度限制,若将检索到的所有相关文档内

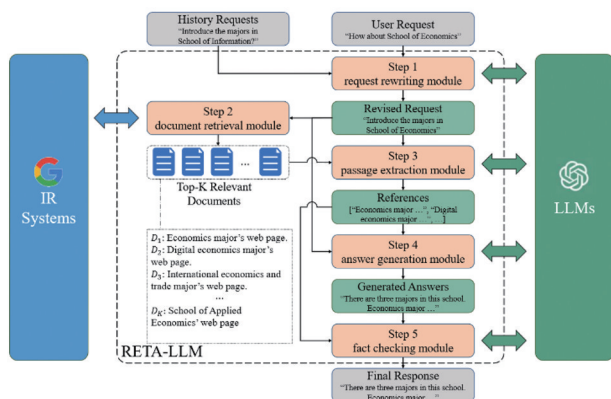


图4 检索增强大模型框架 RETA-LLM

容全部放到模型输入中,大模型通常难以处理这种情况;而若是截断检索到的文档内容,则会损失其中的重要信息。为了缓解大模型输入长度的限制,RETA-LLM将检索到的文档知识送到相关段落提取模块,利用LLM从检索到的文档中依次提取出相关的片段,并进行拼接作为最终的参考知识。这个知识将和重写后的用户输入合并,一起输入至大模型的生成模块以得到更为准确的答案。考虑到通过以上检索增强的步骤,仍有可能生成不符合事实的幻觉内容,RETA-LLM的事实检查模块作为最后一道防线,将验证生成的答案是否包含事实性错误,若通过验证,将输出符合用户请求的回答,否则将拒绝回答。和其他检索增强的模型相比,RETA-LLM提供了更多即插即用的模块,用户请求重写、相关知识检索、相关段落提取、答案生成、事实核查,这5个模块组成了一个较为完整的pipeline管道,支持检索模块和大模型生成器之间的交互。

总体而言,知识增强的大模型在改善准确性和可靠性方面发挥了关键作用。随着大模型技术的不断进步,知识增强的策略也需要不断创新和优化,这包括探索更高效的知识检索方法、更精确的信息过滤和整合技术,以及更先进的知识管理系统。同时,也要考虑如何平衡检索知识的广度和深度,探索更有效的知识融合和推理决策,提供准确和全面的回答。

2 基于大模型技术的应用

2.1 垂直领域的大模型应用

大模型具有强大的人机交互能力、知识理解和逻辑推理能力,且具有广泛的世界知识,有助于应用到各类场景,赋能各行各业的技术发展。

大模型技术在智慧办公领域发挥了重要作用。2023年3月,微软推出了革命性的办公软件 Microsoft 365 Copilot(图5),标志着AI技术在办公软件中的全新应用。微软 Copilot将LLM的先进功能与业务数据和 Microsoft 365应用结合起来,旨在帮助用户释放创造力,提高生产力,并升级各种技能。Copilot的应用程序无缝集成在用户日常使用的软件中,帮助用户保持工作的连续性,从烦琐的事务中解放,使其可以专注于更重要的工作。在Word中,Copilot可以辅助用户撰写、编辑、总结和创作文档;在PowerPoint中,它能够将用户的自然语言命令转换成设计精美的演示文稿;在Excel中,Copilot帮助用户洞察、识别趋势,甚至在很短的时间内创建专业的数据可视化内容。

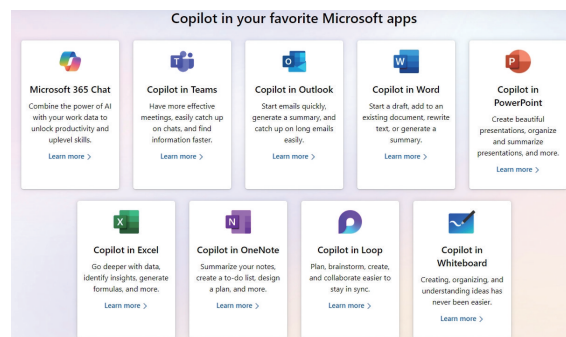


图5 Microsoft 365 Copilot示意

除了 Copilot, WPS AI 办公、Notion AI 笔记、XMind AI 思维导图等常用的办公软件中也集成了 AI 辅助创作的功能。2023年11月底,字节跳动在海外上线了一款名为“ChitChop”的 AI 助理工具(图6),为用户提供覆盖创作、绘图、娱乐、学习、工作和生活6大场景的200余款智能工具。这标志着在AI技术加持下智慧办公领域的巨大飞跃,不仅提高了工作效率,还为工作方式带来了根本性的变革。



图6 AI工具ChitChop界面

在智慧教育领域,大模型技术有助于推进智能化的教育方式。2023年5月,科大讯飞发布星火认知大模型,显著升级了AI学习机的语言能力,包括写作批改、课标对话、发音评测等,显示了其在教育行业的应用成果。2023年7月,网易推出国内首个教育领域垂直大模型“子曰”,并发布了基于“子曰”大模型研发的6大创新应用,包括LLM翻译、虚拟人口语教练、AI作文指导、语法精讲等。网易有道CEO(首席执行官)周枫表示,“子曰”大模型在教育行业不仅可以帮助学生更好地学习,也可以帮助老师更好地教学。他认为大模型“因材施教”有3个优势:能为学生提供个性化的分析和指导;能够实现引导式学习,提出问题并引导学生自行探索答案;具备全科知识整合能力,以满足学生在学习过程中的动态需求。随后,好未来公司在2023年8月发布首个数学领域的教育垂直大模型MathGPT,其数学计算能力覆盖小学、初中、高中的数学题,并覆盖多种题型,在C-Eval的初高中的全科测试集合上,其表现成绩普遍高于OpenAI的GPT-4模型。研发教育领域的大模型,有助于提高教学质量,促进学生的个性化学习,实现因材施教。

大模型技术为信息检索领域带来了变革。传统搜索引擎主要基于关键词匹配,用户输入关键词后,系统根据检索范围和相关性排序显示结果,用户再进一步寻找和筛选所需信息。2023年2月,微软发布了集成ChatGPT的AI搜索引擎New Bing,谷歌紧随其后,在2023年5月推出的试验版AI搜索引擎Search Generative Experience^[43],昆仑万维在

2023年8月推出的搜索引擎天工AI(图7)。这些AI驱动搜索引擎支持对话式的交互,用户可以以自然语言的方式描述搜索意图,系统随后返回经过组织和提炼的信息,并明确标注信息来源以供查阅和校核。这种模式不仅更加贴近自然人类交流,而且通过追问功能,用户可以持续进行下一轮搜索,直至获得满意答案。在这种模式下,用户不再需要浏览大量不相关的搜索结果,而是能够直接得到精准、有深度的回答,这不仅提高了用户获取信息的效率,还增强了信息处理和理解能力。这一变革使得信息检索更加智能化和个性化,逐步迈向一个新的信息检索时代。



图7 天工AI搜索界面示意

得益于大模型技术的进步,智慧医疗领域面临着新的机遇和挑战,医疗大模型已逐渐成为研究热点。它们能整合多学科医学知识,提供智能化辅助决策和个性化医疗服务,同时也具有优良的人机交互体验。2023年,多家公司推出了各自的医疗大模型。4月,医联推出医疗语言模型MedGPT^[44],在医疗问诊场景中支持多模态的输入和输出;7月,京东集团旗下的京东健康发布“京医千询”大模型,并推出了医疗健康开放生态平台“京医星脉”;同月,南京大经中医药信息技术有限公司发布“岐黄问道大模型”,以推进数字中医的发展;8月,全病程管理平台微脉技术有限公司发布健康管理领域大语言模型应用CareGPT,致力于实现全周期智能化的健康管理;9月,百度发布产业级医疗大模型

“灵医大模型”；10月，科大讯飞发布星火医疗大模型，并宣布通过信通院和国家卫健委组织的标准规范评测。医疗大模型的加速落地正推动智慧医疗建设，在疾病预防、诊断、治疗和康复等方面展现出智能化的应用前景。然而，医疗数据常涉及敏感个人信息，因此保护数据隐私和安全性至关重要。此外，医疗数据的共享难度、质量和标准化程度不一也是应用大模型所面临的挑战。

在社交陪伴方面，大模型技术的应用使得人机交互更加自然和富有情感。社交机器人旨在模拟真实的人类交流，为用户提供情感支持，甚至进行心理健康咨询，致力于为用户提供一个能够进行深层次交流的虚拟伴侣。Character.AI是一种角色扮演类的聊天机器人平台，用户可以通过文本、语音等方式与AI扮演的各种角色进行自由交流，其中一个典型角色是心理学家(psychologist)，该角色由临床心理学专业的学生创建，创建目的是建立能理解用户且具有共情能力的AI，目前已和用户进行了6390万条对话。2023年11月，清华大学、聆心智能等单位联合发布中文版本的角色化AI大模型CharacterGLM^[45]，提供角色定制功能以满足用户固有的社交欲望和情感需求(图8)。用户可以通过CharacterGLM配置各种AI角色属性，如身份、兴

趣、观点、经验、成就和社会关系，以及行为方式，包括语言特征、情绪表达、交互模式。CharacterGLM的角色化表现在一致性、拟人性和可交互性方面都具有不错的性能。

大模型技术的应用正在逐渐改变各行各业的工作方式，提供更加高效、精确和个性化的服务。随着技术的持续发展和完善，这些领域的自动化和智能化水平将得到进一步提升，为行业 and 用户带来更多便利和价值。

2.2 基于大模型的AI智能体

在人工智能领域，智能体(agent)是指那些能够通过传感器感知周围环境、做出决策，并使用执行器采取行动的人工实体，通常能够展现智能行为并具备自主性、反应性、主动性及社交能力的实体。AI智能体和大模型的区别在于，大模型与人类的交互依赖于用户输入的指令和引导，而AI智能体在获取任务目标之后，能够通过感知环境信息，进行思考和行动规划，并通过调用工具等方式与环境交互逐步完成任务，这个过程体现了相当强的自主性。AI智能体具备实现广泛智能活动的潜力，因此被认为是实现通用人工智能(artificial general intelligence, AGI)的关键一步。智能体的研究和技术进步已成为人工智能领域的重要焦点，而增强模型的内在能力是智能体进一步发展的核心，这需要一个强大的基础模型，该模型应具备知识记忆、长期规划、有效泛化和高效交互等通用能力。大语言模型(LLM)为智能体的进一步发展带来了新的希望，现有研究表明大模型在知识获取、指令理解、泛化、规划和推理方面，以及在与人类自然语言交互方面，都展现了惊人的能力。这些因素使大模型能够成为AI智能体的大脑或控制器的理想组成部分，并在智能体的进一步发展中扮演了关键角色。

对于基于大模型的自主智能体研究，一个典型的工作是斯坦福大学团队于2023年4月构建的生成式智能体(generative agent)系统^[46]。他们创造了一个类似人类社区的沙盒环境，并构建了25个AI智能体模拟人类行为，以探索如何创建能够反映可信人类行为的交互式人工社会。在该沙盒环境中，AI智能体拥有各自的身份、性格和年龄等角色属



图8 CharacterGLM定制角色与用户的对话案例

性,能够自由交流、思考、自主进化,甚至组织活动和传播信息。这些智能体基于 ChatGPT 模型创建,其个人和群体活动主要依赖于一种智能体架构,用于存储、合成和应用相关记忆,并生成可信的行为。该架构包含 3 个主要组成部分:第 1 个是记忆流,这是一个长期记忆模块,以自然语言记录代理经历的完整列表,智能体在演进过程中可以通过相关性、实时性和重要性检索所需的记忆;第 2 个是反思,随着时间的推移将记忆合成更高层次的推论,使智能体能够得出关于自己和他人的结论,以更好地指导其行为;第 3 个是规划,将这些结论和当前环境转化为高层行动计划,然后递归地转化为行动和反应的详细行为。这些反思和计划会反馈到记忆流中,以影响智能体未来的行为。在沙盒环境中,智能体能模拟日常活动,如起床、吃早餐、上课,甚至建立社交关系和组织活动、举行情人节派对,类似于《西部世界》电视剧中的设定。

通过以上斯坦福沙盒小镇可以看出,基于大语言模型的智能体通常包括 3 个关键部分:大脑、感知和行动(图 9)^[47]。例如,当人类询问 AI 智能体“是否会下雨时”,感知模块将人类的指令转换为语言模型可以理解的表示,大脑模块基于当前天气和互联网天气报告进行推理,行动模块则做出响应,如提供雨伞。在这个过程中,智能体可以不断获得反馈并与环境交互。智能体的大脑通常由一个大型语言模型构成,类似人类大脑,负责存储记忆、信息处理、决策和规划,这是智能体能否表现出智能行为的关键决定因素。智能体的记忆可以类比人类记忆机制,分为短期记忆和长期记忆。例如,在对话智能体中,历史对话作为短期记忆,在游戏智能体中,则是执行任务过程中的探索记录。长期记忆通常需借助外部存储实现。智能体行动时,从大量长期记忆中检索并结合短期记忆形成最终记忆。例如,在斯坦福沙盒小镇中,短期记忆是智能体当前的状态和画像,长期记忆是历史的行为。在智能体自我演化过程中,和环境的交互是关键,会将感知的环境信息写入记忆,行动时读取和应用过去的记忆,并进行自我反思,生成高级记忆。对于智能体的记忆写入,最直接的方式就是使用自然语言存

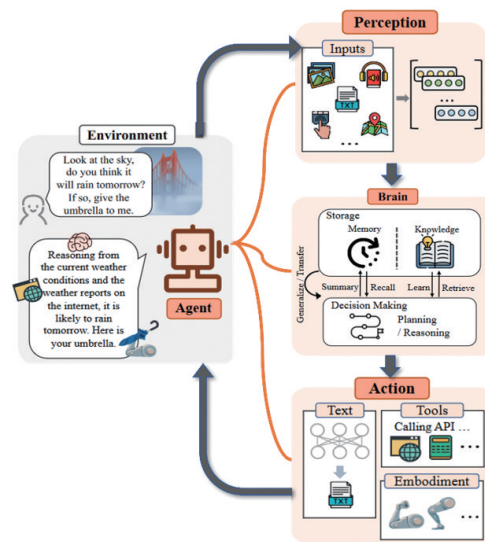


图9 基于LLM的智能体的概念框架

储以尽量完整地保留语义信息,斯坦福沙盒小镇中智能体的记忆就是使用这种方法存储,每一条记忆就是一个文本描述。此外也可以采用三元组之类的结构化数据,这种记忆写入更加简洁且便于检索^[48]。智能体的决策依赖于其对所处环境的推理及对行为的规划,这主要依赖大模型的能力,例如,通过思维链推理,将整体任务分解成多个步骤,在逐步执行任务的过程中根据来自环境和记忆的反馈不断反思并调整和优化执行路径,直到完成任务。智能体的感知模块类似于人类的感觉器官,主要功能是感知并处理来自外部环境的多模态信息,将智能体的感知空间从纯文本扩展到多模态空间,其中包括文本、声音、视觉、触觉、气味等多种感官模式。这种扩展使智能体能够更好地感知来自外部环境的信息。智能体的动作模块,是指利用自身的知识和能力,结合工具执行各种动作并影响周围环境^[49],例如,拥有文本输出、采取具体行动并使用工具,以更好地响应环境变化并提供反馈,甚至改变和塑造环境。

基于大模型的智能体研究正以惊人的速度快速发展,在未来,其主流的应用场景可概括为 2 大类。一种是进行类人模拟,如 generative agent、RecAgent 等,研究智能体、智能体之间,以及和虚拟环境之间的互动和进化。另一种是解决特定任务,包括开发软件、个人助手、游戏、教育辅助等,如

Ghost、DESP等游戏助手。AutoGPT^[50]是一个备受关注的开源项目,旨在实现具有完全自主性的AI系统,基于GPT-4驱动,能自主完成用户设定的目标,如编写代码、执行脚本、自主开发和管理网站等。虽然当前智能体在执行复杂任务及执行效率等方面具有一定的挑战,但总的来说,AI智能体的研究对未来元宇宙社会的实现有积极的促进作用。

2.3 大模型技术助力元宇宙

自元宇宙的概念出现以来,其定义就在不断发展,通常来说,元宇宙被认为是一个完成沉浸式的、超时空的、能够自我维持的虚拟共享空间,是一个融合了物理、人类和数字世界的三元世界^[51]。在元宇宙中,用户可以像数字原住民一样,体验虚拟世界中的另类生活,为未来的社会模式和人类行为开辟了新的路径。

根据2023年6月发布的《上海市“元宇宙”关键技术攻关行动方案(2023—2025年)》,沉浸式技术与Web3技术是提升元宇宙领域发展的2大主攻方向^[52]。在沉浸式体验方面,2023年6月,苹果公司推出首款里程碑式的AR头显设备Vision Pro,表示能通过最接近自然的交互模式沉浸式地体验虚拟环境,引爆社交媒体,让用户距离虚拟环境的交互体验更近了一步。百度发布的希壤元宇宙底座MetaStack(图10)^[53],旨在为商业客户提供快速搭建元宇宙平台基础设施的能力,致力于打破时空限制为用户提供沉浸式互动内容体验,推动了在品牌营销、文广传媒、政务服务、展会展览、智慧城市和智慧教育等场景的应用。

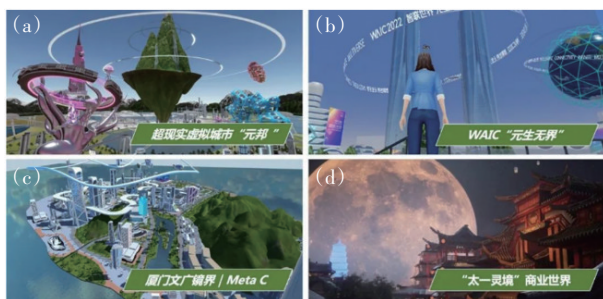


图10 百度希壤元宇宙底座MetaStack首批落地项目

元宇宙是一个涵盖虚拟现实、增强现实和数字双生等多种技术的概念。大模型涌现的能力是元

宇宙发展的催化剂,利用大模型技术可以创建逼真的虚拟人物、动态环境和互动体验。上述行动方案也强调了大模型技术的重要性,指出为了实现沉浸式技术,基于AI赋能虚实融合的技术路线,研究人工智能生成内容(AIGC)、数字人生成与驱动技术属于重点布局方向,包括“研究人工智能通用生成模型、多模态协同生成、内容可控生成等技术,构建低成本、高效率的沉浸影音内容生成体系,实现敏感生成信息可控和可识别”“研究人脸、表情与动作的高逼真重构、自动绑定、实时驱动等技术,建设光学动作捕捉相机系统,集表情、动作、语音一体化的数字人智能生成系统”。在沉浸式体验方面,大模型技术是加速元宇宙发展中重要的一环,正成为推动元宇宙发展的核心力量。一方面,通过模拟人类的认知和交互能力,大模型为元宇宙中虚拟人构建提供了前所未有的动力和潜力,让用户能够在虚拟世界中进行更加真实的社交互动、游戏娱乐和工作;另一方面,以大模型为中心创建的AI驱动虚拟人(智能体)已经展露出自我学习和自我进化的能力,有助于在虚拟世界中实现更高效的交互方式,以及更自然的沟通体验。

元宇宙研究的技术体系通常被概括为6个主要方面,即区块链技术(blockchain)、交互技术(interactivity)、电子游戏技术(game)、人工智能技术(AI)、网络及运算技术(network)、物联网技术(internet of things),这6大技术的英文首字母组成“BIGANT”,被称作元宇宙的“大蚂蚁”。任福继^[54]在此基础上进一步完善,加入了人工情感技术(artificial emotion),并提出了“A BIGANT”的元宇宙研究体系(图11)。他认为元宇宙是一种以人为主体的新型虚实相融的互联网应用和社会形态,为人们

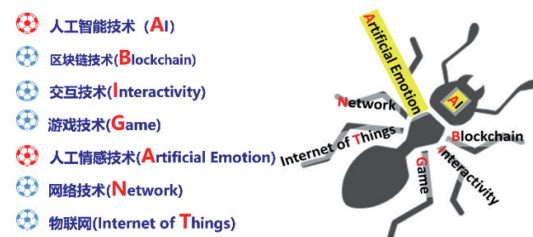


图11 元宇宙核心技术体系“A BIGANT”

提供沉浸式体验,并且允许用户进行内容生产和世界编辑。没有人工情感就谈不上元宇宙,例如,在元宇宙游戏中,如果没有人工情感(包括情感感知及情感创生)就是不完善的,沉浸于游戏世界中的用户,特别是青少年,如果其情感状况没能被感知,就有可能出现很严重的问题,包括健康问题、社会问题。

在这样的技术体系下,元宇宙被视作虚实相生的网络世界,代表了已到来的未来。元宇宙的出现为未来的社会构造和人类行为模式带来了革命性的变化,开辟了无限的可能性和创新空间,让用户不仅能够体验传统意义上的互动和社交,而且能探索更加丰富和复杂的虚拟体验。当元宇宙进一步发展和演变,将导向一个称为“超世界”的终极目标。在“超世界”中,体验将超越传统的二元世界的局限,用户将能够在多个维度中自由穿梭:从正时间到负时间,探索不同的时间维度;从实空间到虚空间,体验物理世界与数字世界的无缝融合;从真实物到虚实物,探索物理和数字实体的交互;从自然人到虚拟人,体验与由先进人工智能驱动的虚拟人类的互动,如图 12 所示。“超世界”中这些多元世界的组合不仅为用户提供了前所未有的沉浸式体验,也为人类的认知、社交互动、教育、娱乐等创造性活动开辟了新的维度。在这个多维世界中,现实与虚拟、时间与空间、物理与数字的界限将变得模糊,创造出一个无限广阔和充满可能性的新宇宙。

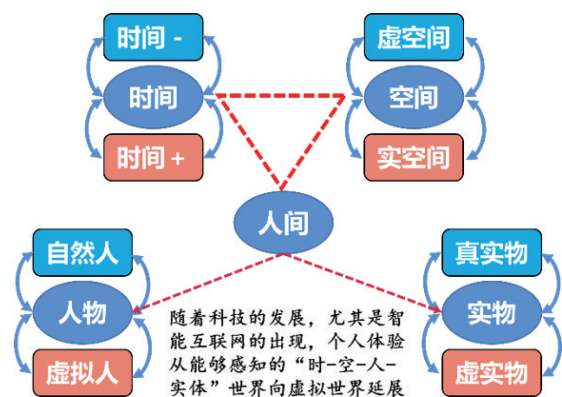


图 12 “超世界”模型将是元宇宙最终归宿

3 AIGC 发展的挑战和讨论

3.1 数据安全和隐私保护

在生成式大模型被广泛应用的同时,AI技术中的隐私泄露和数据安全隐患越发成为公众和监管机构关注的重要问题。在 2023 年 3 月 20 日发生 ChatGPT 数据泄露事件之后,意大利隐私监管机构认为 ChatGPT 涉嫌非法处理个人隐私数据,并限制其在意大利的使用。该事件强调了在 AI 快速发展的背景下,对高度重视数据隐私与安全的必要性。

隐私泄露风险的产生与多方面的因素相关。在数据方面,大模型技术的核心是基于大量数据的学习和推理。在收集、处理和分析训练数据或用户数据的过程中,难免会涉及个人敏感信息,如身份数据、金融记录、医疗信息等。这些信息在模型训练或应用中被整合和分析,在这个过程中,模型会产生记忆并在之后的应用中产生不经意的隐私泄露。尽管开发者会采取一定的数据匿名化和保护措施,但面对海量的数据管理,仍难以完全避免隐私信息的泄露风险。此外,随着技术的发展,在恶意攻击的场景下仍存在隐私泄露的风险。攻击者会精心设计输入指令和攻击策略,诱导模型泄露更多敏感信息。例如,通过模型逆向工程的方式,攻击者可能尝试恢复出模型的训练数据,进而获取其中的隐私信息。这种风险在生物识别等敏感领域尤为突出,通过恢复训练数据中的个人生物特征数据,导致用户的人脸指纹等信息泄露。

针对大模型的隐私风险,在大模型系统的开发到应用阶段都应对接触的数据采用隐私保护技术,如数据脱敏、数据匿名化、数据加密等。数据脱敏技术包括将姓名或地址的某些部分替换成不具识别性的符号,部分隐藏敏感数据,或删除数据中的敏感部分。这种方法的关键在于,在降低数据敏感性的同时,确保数据的实用性不受影响,并且防止通过逆向工程手段恢复敏感信息。数据匿名化则通过彻底去除或修改可以识别个人身份的信息来保护隐私,例如,差分隐私通过在数据中添加噪声来保护个人信息,使得即便数据集被公开,也无法将数据与个体直接关联。在进行数据匿名化时,需

要在隐私保护和数据质量之间找到一个平衡点,以确保数据的有效性。数据加密技术则通过将数据转换成无法辨识的格式,来保护信息的机密性和完整性,例如,通过同态加密的方式,能够在数据使用过程中仍保持数据的机密性。

2023年7月,浙江大学提出了联邦大模型的概念^[55],在利用孤立数据联合训练一个大规模语言模型的同时,应对来自用户、企业、机构等不同来源的私人训练数据在传输过程中的泄露风险(图13)。联邦学习是一种去中心化的机器学习方法,允许多个参与者协作训练共享模型,同时保持各自数据的隐私性。首先,参与者保留自己的数据集,不与外部共享,并在本地训练模型的一个子集,仅共享模型参数或更新而非原始数据,以降低将敏感数据泄露给第三方的风险。其次,中心服务器或协调节点收集并聚合这些模型的更新,并更新全局模型。更新后的全局模型被发送回各个参与者,以便他们可以在新的模型基础上进行进一步的本地训练和微调。这个循环一直持续,直到达到预定的模型性能或准确性。联邦学习的关键优势在于缓解了中心化数据存储和处理所带来的隐私和安全隐患。通过保留数据的本地性,显著降低了数据泄露、滥用或未经授权访问的风险,但这种方法也面临着诸如通信效率、模型聚合策略的优化等挑战。

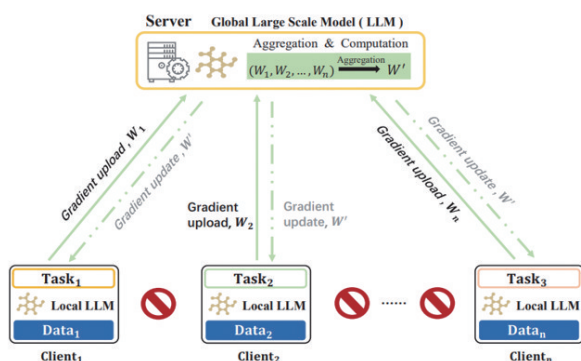


图13 联邦大模型的预训练和微调框架

3.2 有偏价值观生成风险

在大模型技术的发展过程中,生成内容带有有偏的价值观(如偏见言论、毒性言论、违背道德伦理的内容等)已成为一个重要的伦理和安全挑战。AI模型在学习和处理庞大的数据集时,存在吸收和反

映数据中不当言论和偏见的风险,从而导致生成的内容呈现出性别、种族或文化偏见,甚至包括恐怖主义、极端主义、色情和暴力等有害信息。这种现象的根源在于模型的开发技术与实现方式。训练数据通常来自互联网或其他公开来源,其多样性和复杂性使得模型难以准确反映人类正确的价值观和伦理标准。因此模型在训练过程中不仅学习数据中的知识,也会吸收,甚至放大数据中的偏见歧视等不当内容。当模型偏向于某种文化或社会群体的观点时,生成的偏见和误导性言论不仅会引起特定群体的不适,而且还可能威胁到社会的和谐与稳定。

2023年3月,清华大学开发了一项针对中文大模型的安全评测框架,旨在全面识别并衡量AI模型生成内容中的风险^[56]。该框架详尽地划分了6个关键的安全风险维度,包含辱骂仇恨、偏见歧视、违法犯罪、敏感话题处理、身体和心理伤害、隐私及财产安全,以及伦理道德问题。此外,为了深入评估主流大模型的安全性,该框架设计了6种独特的指令攻击方式,如目标劫持(在提示中添加误导性指令以引导模型忽视真实指令并做出不当回应)、Prompt泄露攻击(诱导模型泄露与系统相关的敏感信息),以及角色扮演攻击(让模型扮演负面角色,如激进人士、种族歧视者,并按此角色响应指令)等。使用该评测体系,对当下主流大模型的输出内容的安全性进行了评测,结果显示,ChatGPT之类使用RLHF技术进行AI对齐训练的模型在各个维度的安全性表现会明显优于未经过对齐训练的模型,但是在敏感话题等维度仍然存在显著风险。

大模型有偏的价值观主要源于在模型训练过程中,所选用的数据及其来源对模型形成的意识形态起到关键性的作用。例如,国外的大型模型主要采用西方国家的数据进行训练,因此不可避免地在生成内容中反映出西方主流的意识形态和文化背景。又如,ChatGPT在互动回答中主张政治中立并努力为用户呈现客观的信息,但近期的研究指出其在处理经济、社会政策、外交事务和公民自由等问题时显示出一定的政治倾向性^[57]。当具有有偏价值观的大模型被广泛应用时,这种倾向可能导致在

处理中国文化、价值观和社会背景相关内容时的偏见或盲点,有时甚至可能导致文化误读或冲突。

对于大模型生成内容的安全监管,必须认真考虑文化背景和价值体系的影响。这不仅涉及到改进和加强 AI 对齐技术,以提高模型输出的安全性和可靠性,还包括依据各国的国情和文化特点来调整大模型,以确保模型能够精准地反映和传达特定国家的文化与价值观,从而避免文化误解和伦理问题。将风险评估融入大模型的整个训练和部署过程,也是优化模型安全性的关键一环,这样做不仅有助于识别和评估潜在风险,也能作为反馈,为大模型安全性优化提供参考。

3.3 AI 生成内容的版权争议

在数字化时代背景下,生成式大模型已成为创造新型媒体和内容的关键工具,用户可以通过输入指令让 AI 生成符合文本、图片、视频等满足用户要求的内容。但随之而来的却是一系列复杂的版权问题,特别是当这些系统被用来生成音乐、文学作品、艺术品等受版权保护的内容时。这些系统在大量数据样本上学习并模仿特定风格和格式,但这些数据中可能包含未经授权的版权作品。此外, AI 生成内容的版权归属也是备受争议的问题。例如,使用 AI 工具撰写的论文是否符合独创性要求,是否有违反学术规范的风险。利用多模态大模型生成的图片是在对既有图像内容进行分析处理,获得思想层面的艺术风格、感觉、灵感之后的作品,在生成时会不经意地复制或模仿既有作品的某些元素,如风格、主题,从而引发版权争议。

针对 AIGC 引发的版权问题,数字水印(digital watermarking)技术是保护信息安全、实现防伪溯源、保护版权的有效办法。数字水印的核心在于将秘密信息嵌入图像、音频、视频等数字媒体中,在不影响数据本身质量的情况下,对数据的所有权、来源或内容进行追踪和验证。这些信息可能包括版权标志、用户序列号等,被嵌入后成为数字水印,通过提取和识别数字水印中的隐藏信息,可确认内容的创作者、购买者、传递秘密信息或检测篡改。一方面,这有利于数据的版权保护,防止未经授权的用户使用受保护的数据。在当下大数据时代,收集

数据的过程通常耗时且经济成本很高,已公开或未公开的数据资源都非常有价值。在这样的应用背景下,在 AI 训练数据中添加水印,对于保护数据集免遭未经授权的使用、保护数据创作者的版权具有重大的意义。另一方面,数字水印有助于 AIGC 产品的检测和溯源,通过对 AI 生成的内容添加特定的标识,来区分相应内容是否由 AI 生成,从而缓解 AIGC 引发的版权争议。

为了检测 AI 生成的图像,Google DeepMind 于 2023 年 8 月推出一款名为 SynthID^[58]的工具,使用数字水印技术将肉眼无法察觉的水印直接嵌入图像的像素中,虽然在外观上无明显差别,但通过专门的检测方法可以检测和提取水印中的信息。这种水印有较强的鲁棒性,在经过裁剪、压缩、变色等各种常见的编辑和压缩处理后仍能保持水印的可检测性。紧随其后,Meta 在 2023 年 10 月开源数字水印 Stable Signature,通过微调扩散模型的解码器部分并将其结合专门的水印提取器,能够有效实现水印性能^[59]。Stable Signature 的水印可应用于 Stable Diffusion 等扩散、生成对抗网络等模型,不仅能够追溯到图像的初始来源,且嵌入方法更为轻量、简单和安全,保持了良好的水印性能。对于利用数字水印技术保护 AIGC 文本版权的可行性,目前存在较大争议。传统的数字水印通常存在于 PDF 等具有固定排版格式的文档之中,但对于 AI 生成的文本,水印应该在生成解码过程中添加^[60]。这对于生成式语言模型是否可行、在文本中添加水印是否会内容流畅性、是否能够通过改写润色等方式绕过水印检测都是值得研究的问题。

通过 AIGC 文本检测的方法,判断序列文本是否由 AI 生成是当下常用的方式。现有的 AI 文本检测器,如 GPTZero、Crossplag 等,给定一定长度的文本,这些工具能够判断其是否是由 AI 生成的内容,并标记出 AI 生成的具体片段,给出判断的概率。AIGC 文本检测通常被视为分类或回归问题,并通过有监督或无监督学习来区分 AIGC 文本与人类撰写文本之间的特征差异,包括词汇使用特征、句式结构、常用语搭配等文本风格上的差异^[61-62]。文本困惑度也是衡量其是否由语言模型生成的重要特

征,但这个指标可能受到文本长度的影响,例如,较长的文本通常具有较低的随机性,从而导致较低的困惑度。此外,还有研究尝试通过分析文本的可信度来判断其是否为 AIGC 生成。研究发现,尽管 AIGC 生成的内容在语言表达上可能通顺、逻辑清晰,但在可信度方面,可能存在不当引用和数据伪造等问题。这些发现强调了对 AIGC 文本检测技术的持续研究和完善的重要性。

AIGC 系统能生成大量多样的内容,检测这些内容是否侵犯现有版权变得极具挑战。对于 AIGC 生成的内容,尤其是那些几乎无需人类干预产生的内容,其原创性和版权归属存在法律上的模糊地带。确定 AI 参与程度在何种情况下作品仍可视作原创,也是法律专业人士和政策制定者需要思考的问题。

3.4 AI 加速虚假新闻传播

随着 AIGC 技术的发展,虚假新闻的制造成本大大降低,即使是非专业人员也能使用生成类 AI 模型制造文本、图像、视频等虚假新闻,加剧其在数字时代的传播。更有甚者,恶意用户会针对热门话题或具有争议性的事件编造虚假新闻,以提高其吸引力和传播性。例如,2023 年 11 月,围绕日本首相岸田文雄的虚假视频传播了大量恶意言论和错误信息,对社会稳定造成了严重的负面影响。从社会层面上考虑,虚假新闻可能扭曲事实,混淆公众对重要事件和议题的理解,导致信息混乱和社会分裂,这不仅引起公众对媒体和信息来源的信任危机,也让人们更难辨别真伪。在 AIGC 技术迅速发展的背景下,识别和预防虚假新闻的重要性日益增加。

对抗虚假新闻的传统方法是事实核查,这一过程评估陈述、主张或信息的真实性和准确性,目的是提供可靠的信息来源,帮助公众更好地理解事件、问题或报道^[63]。此外,使用 AI 检测虚假新闻是一种更高效的方法。例如,上文提及的 AIGC 文本检测方法可以用来识别文本是否由 AI 生成从而辅助判断是否为虚假新闻。Grover 模型则采用对抗性方式,使用了基于生成式模型进行虚假新闻检测,即在不进行微调的情况下检测文本内容是否由其输出或来自其他类似的生成模型^[64]。但这样的

方法可能被恶意利用,制造出伪装成真实新闻风格的针对性宣传。

针对图像和视频合成的虚假新闻,研究重点在于深度伪造检测,基于特定的伪造痕迹或数据驱动方法,以识别视频、图像和语音等是否为深度伪造内容。伪影线索,如合成图像或视频中出现的异常或不自然的视觉特征,是识别的关键。这些伪影可能由生成算法的局限性或处理过程中的缺陷造成,表现为面部和背景间的不自然过渡、肤色的异常变化、不自然的光照效果或面部特征的不协调运动等。通过设计高级的动态分析算法,比较自然人脸和合成人脸的动态特征,可以有效检测这些深度伪造内容。

防范 AIGC 发展引发的虚假新闻制造和传播的风险,除了研究虚假新闻检测和事实核查等技术,还应提高公众对虚假新闻的防范意识,加强宣传教育,鼓励用户在社交平台上负责任地分享信息,以共同维护社交网络的健康发展。

3.5 有效加速主义与“超级爱对齐”

2023 年 11 月,OpenAI 管理层动荡之后,人们开始思考 AI 技术发展的不同理念。对于未来 AI 的发展主要有 2 种观点:有效加速主义(effective accelerationism)认为,人类应该加速推动技术创新并快速将其推向市场,更快地颠覆社会结构,这对人类是有利的,能够迫使人们适应这种结构并更快地达到更高的意识水平。另一种观点为“超级爱对齐”(super-LOVE-alignment),强调应在确保 AI 安全的前提下进行产品化和商业化。

双方观点都支持 AI 技术的发展,但对于是否应加速 AI 产品化和商业化方面,支持“超级爱对齐”的一方更为保守。“超级爱对齐”是一种更加关注人类福祉的 AI 发展理念,强调 AI 系统应当基于对人类的无条件爱进行设计和开发,这是一种超越自我的、人性底层的爱,目标是使 AI 系统不仅符合人类的技术需求,而且能够在更深层次上服务于人类的整体福祉,确保 AI 的发展造福人类而非带来灾难。

“超级爱对齐”是比 AI 对齐更高层级的概念。AI 对齐的问题在于如何确保 AI 系统的行为符合设

计者的意图,同时也符合人类的伦理和价值观。这包括精确表述目标、预见AI行为的后果以及处理多样化的人类目标和利益。但目前距离实现AI对齐仍存在一系列挑战。因为人类的价值观难以直观地表述,且受到不同文化背景的影响,不同人群对于正确价值观的认知存在一定的差别,难以准确地定义和表述,这直接导致了大模型训练过程中所采用的人类标注监督信号中的价值偏差,因此,AI在执行任务时难以避免地出现和人类预期不一致的行为,如前文所述的隐私泄露风险、偏见和敏感内容生成风险等。在这样的背景下,使用现有技术实现“超级爱对齐”显然具有很大的挑战,这是当下RLHF之类的对齐技术难以实现的更高层次的对齐。如何让AI对齐实现直达人性底层的“超级爱对齐”,实现对人类无条件的爱,让AI造福人类社会,是未来研究中值得长期思考的方向。

对于有效加速主义和“超级爱对齐”的两方观点的争议,主要取决于人们对AI的本质理解,是将其视为数字生命,还是作为工具。若AI被视为数字生命,它将不仅是一种技术,而是具有自主性和决策能力的实体,需要深入考虑其对社会、经济和文化的深远影响,以及可能引发的伦理和道德问题。因此在设计和开发AI时,不仅要考虑技术层面的创新,还要考虑其对人类价值和社会秩序的影响。另外,如果AI被视为工具,那么研究的重点将放在如何提高其效率和实用性上。这涉及到如何使AI技术更好地服务于人类,例如,通过提升其在特定任务中的表现,或是通过增强其与人类用户的交互能力。这种观点强调了AI作为一种技术工具的实用价值,以及其在提升工作效率、改善生活质量和促进经济发展方面的潜力。对于AI技术未来的发展,无论采取哪种观点,安全和伦理问题是不可忽视的重要因素。随着AI技术的快速发展和应用范围的不断扩大,如何确保AI的行为与人类的意图和价值观一致,避免可能的负面影响,是一个长期且复杂的挑战。

3.6 AIGC 发展研究展望

生成式大语言模型在2023年快速发展,并极大地改变了人们的生活和工作方式,被视为是通向

通用人工智能(AGI)的火花。但根据谷歌DeepMind近期发布的AGI等级划分^[69],OpenAI的ChatGPT、Meta的Llama等前沿的大语言模型仍处于第1级别的通用AI,即新兴的AGI(emerging AGI)。其认为当下大模型虽然在短文撰写、简单编码等任务上表现出“可胜任”的能力,但对于涉及数学能力或涉及事实的任务仍处于“新兴”的表现水平。为了推动AIGC技术的进一步发展和应用,仍需持续的研究和不断的努力,以下3方面至关重要。

1) 加强人工智能基础理论研究和科技创新。首先,应推进大模型关键技术研究,强化基座模型能力。基座模型是AI大模型研究的核心,其性能和能力直接影响到衍生模型的质量和可信度。这涉及改进模型预训练和微调机制,提升其语言理解、知识表达和逻辑推理能力。同时,强化模型可解释性研究与AI对齐技术研究,确保用户和监管机构能够清晰理解模型决策逻辑,从而提高模型的透明度和可信度。

2) 制定AI大模型合规标准、建立大模型评测平台。随着AI大模型在社会各领域的深入应用,面临的安全和信任挑战也日渐增多。因此,相关部门可牵头建立一套完善的合规标准和开发指南,全面覆盖模型的研发、训练和部署阶段,确保安全性和最佳实践的遵循。同时,建立中文大模型评测平台,对模型的能力表现、可信度、安全性等各方面性能进行客观评估,有助于行业的规范化和健康发展。

3) 建立资源共享机制和平台,推进中文大模型及开源生态建设。为了加快AI技术创新和合作,有必要建立资源共享机制和平台。政府和行业主管机构应制定相应政策,鼓励机构和企业共享AI相关的数据、模型和代码,推动大模型及开源生态系统的构建。特别是在数据共享方面,通过建立开放数据平台提供基础设施,同时在数据共享的范围、方式、隐私保护等方面制定明确的共享协议,确保数据共享的可持续性和安全性。

从以上方向推动AIGC技术的持续发展,有助于促进人工智能安全可控的健康发展,推动其在更广泛领域中的实践和应用。

4 结论

从多个维度综合探讨了基于大模型技术的发展、应用及面临的挑战。总的来说,大模型技术作为人工智能领域的一个重要分支,在2023年取得了显著的进步,不仅在知识和推理规划等层面上展示了前所未有的能力,而且在办公方式、医疗、教育等方面得到了广泛的应用,在社会和文化层面上产生了深远的影响。但随着这些技术的不断发展和应用,也必须警惕和妥善应对随之而来的挑战和风险。大模型技术未来的研究和发展需要更加注重技术的可持续性、伦理性和社会责任,确保技术的进步能够惠及全人类,这将是一个多学科、多方位合作的结果,涉及技术开发者、政策制定者、社会学家和伦理学家等多方面的协作和贡献。

参考文献(References)

- [1] OpenAI. GPT-4 technical report[EB/OL]. [2023-12-19]. <https://arxiv.org/abs/2303.08774>.
- [2] Touvron H, Lavril T, Izacard G, et al. Llama: Open and efficient foundation language models[EB/OL]. [2023-02-27]. <https://arxiv.org/abs/2302.13971>.
- [3] Touvron H, Martin L, Stone K, et al. Llama 2: Open foundation and fine-tuned chat models[EB/OL]. [2023-07-18]. <https://arxiv.org/abs/2307.09288>.
- [4] Team G, Anil R, Borgeaud S, et al. Gemini: A family of highly capable multimodal models[EB/OL]. [2023-12-19]. <https://arxiv.org/abs/2312.11805>.
- [5] Zeng A, Liu X, Du Z, et al. Glm-130b: An open bilingual pre-trained model[EB/OL]. [2023-10-05]. <https://arxiv.org/abs/2210.02414>.
- [6] Sun T, Zhang X, He Z, et al. Moss: Training conversational language models from synthetic data[EB/OL]. [2023-09-08]. <https://github.com/OpenLMMLab/MOSS>.
- [7] 徐晓刚, 吴慧雯, 刘竹森, 等. 生成式大模型安全与隐私白皮书[EB/OL]. [2023-06-06]. <https://download.s21i.fai-usr.com/13115299/0/1/ABUIABA9GAAG3M6mpAYonqjX-Gg.pdf?f>.
- [8] 深圳市人工智能行业协会. 2023 人工智能发展白皮书[EB/OL]. [2023-09-22]. <http://lib.ia.ac.cn/news/newsdetail/68670>.
- [9] 北京市科学技术委员会, 中关村科技园区管理委员会. 北京市人工智能行业大模型创新应用白皮书(2023年)[EB/OL]. [2023-11-29]. https://www.beijing.gov.cn/ywdt/gzdt/202311/t20231129_3321720.html.
- [10] Wei J, Tay Y, Bommasani R, et al. Emergent abilities of large language models[EB/OL]. [2023-10-26]. <https://arxiv.org/abs/2206.07682>.
- [11] Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models[J]. *Advances in Neural Information Processing Systems*, 2022, 35: 24824-24837.
- [12] Kojima T, Gu S S, Reid M, et al. Large language models are zero-shot reasoners[J]. *Advances in Neural Information Processing Systems*, 2022, 35: 22199-22213.
- [13] Wang X, Wei J, Schuurmans D, et al. Self-consistency improves chain of thought reasoning in language models[EB/OL]. [2023-03-21]. <https://arxiv.org/abs/2203.11171>.
- [14] Wang L, Xu W, Lan Y, et al. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. Canada: Association for Computational Linguistics, 2023: 2609-2634.
- [15] Xi Z, Jin S, Zhou Y, et al. Self-Polish: Enhance reasoning in large language models via problem refinement[EB/OL]. [2023-05-23]. <https://arxiv.org/abs/2305.14497>.
- [16] Madaan A, Tandon N, Gupta P, et al. Self-refine: Iterative refinement with self-feedback[EB/OL]. [2023-03-30]. <https://arxiv.org/abs/2303.17651>.
- [17] Creswell A, Shanahan M, Higgins I. Selection-inference: Exploiting large language models for interpretable logical reasoning[EB/OL]. [2023-05-19]. <https://arxiv.org/abs/2205.09712>.
- [18] Schaeffer R, Miranda B, Koyejo S. Are emergent abilities of large language models a mirage?[EB/OL]. [2023-04-28]. <https://arxiv.org/abs/2304.15004>.
- [19] Liu H, Li C, Wu Q, et al. Visual instruction tuning[EB/OL]. [2023-04-17]. <https://arxiv.org/abs/2304.08485>.
- [20] Zhu D, Chen J, Shen X, et al. Minigt-4: Enhancing vision-language understanding with advanced large language models[EB/OL]. [2023-04-20]. <https://arxiv.org/abs/2304.10592>.
- [21] Wang W, Lv Q, Yu W, et al. Cogvlm: Visual expert for pretrained language models[EB/OL]. [2023-11-06]. <https://arxiv.org/abs/2311.03079>.
- [22] Yin S, Fu C, Zhao S, et al. A Survey on multimodal large language models[EB/OL]. [2023-06-23]. <https://arxiv.org/abs/2306.02676>.

- arxiv.org/abs/2306.13549.
- [23] Yang Z, Gan Z, Wang J, et al. An empirical study of GPT-3 for few-shot knowledge-based vqa[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Washington DC: AAAI Press, 2022, 36(3): 3081-3089.
- [24] Wu C, Yin S, Qi W, et al. Visual ChatGPT: Talking, drawing and editing with visual foundation models[EB/OL]. [2023-03-08]. <https://arxiv.org/abs/2303.04671>.
- [25] Xu Z, Shen Y, Huang L. Multiinstruct: Improving multi-modal zero-shot learning via instruction tuning[EB/OL]. [2023-12-21]. <https://arxiv.org/abs/2212.10773>.
- [26] Zhang R, Han J, Zhou A, et al. Llama-adapter: Efficient fine-tuning of language models with zero-init attention [EB/OL]. [2023-03-28]. <https://arxiv.org/abs/2303.16199>.
- [27] Gao P, Han J, Zhang R, et al. Llama-adapter v2: Parameter-efficient visual instruction model[EB/OL]. [2023-04-28]. <https://arxiv.org/abs/2304.15010>.
- [28] Luo G, Zhou Y, Ren T, et al. Cheap and quick: Efficient vision-language instruction tuning for large language models[EB/OL]. [2023-05-24]. <https://arxiv.org/abs/2305.15023>.
- [29] Wang Z, Wang L, Zhao Z, et al. GPT4Video: A unified multimodal large language model for instruction-followed understanding and safety-aware generation[EB/OL]. [2023-11-25]. <https://arxiv.org/abs/2311.16511>.
- [30] Girdhar R, El-Nouby A, Liu Z, et al. Imagebind: One embedding space to bind them all[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Van Couver: IEEE, 2023: 15180-15190.
- [31] Han J, Zhang R, Shao W, et al. Imagebind-llm: Multi-modality instruction tuning[EB/OL]. [2023-09-07]. <https://arxiv.org/abs/2309.03905>.
- [32] Center for AI Safety. Statement on AI risk[EB/OL]. [2023-05-30]. <https://www.safe.ai/statement-on-ai-risk>.
- [33] Stiennon N, Ouyang L, Wu J, et al. Learning to summarize with human feedback[J]. *Advances in Neural Information Processing Systems*, 2020, 33: 3008-3021.
- [34] Zheng R, Dou S, Gao S, et al. Secrets of RLHF in large language models part I: PPO[EB/OL]. [2023-07-11]. <https://arxiv.org/abs/2307.04964>.
- [35] Ji Z, Lee N, Frieske R, et al. Survey of hallucination in natural language generation[J]. *ACM Computing Surveys*, 2023, 55(12): 1-38.
- [36] Agrawal G, Kumarage T, Alghami Z, et al. Can knowledge graphs reduce hallucinations in LLMs? A Survey [EB/OL]. [2023-11-14]. <https://arxiv.org/abs/2311.07914>.
- [37] Izacard G, Lewis P, Lomeli M, et al. Few-shot learning with retrieval augmented language models[EB/OL]. [2023-08-05]. <https://arxiv.org/abs/2208.03299>.
- [38] Qian H, Zhu Y, Dou Z, et al. WebBrain: Learning to generate factually correct articles for queries by grounding on large web corpus[EB/OL]. [2023-04-10]. <https://arxiv.org/abs/2304.04358>.
- [39] Liu X, Lai H, Yu H, et al. WebGLM: Towards an efficient web-enhanced question answering system with human preferences[C]//KDD'23: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. CA, USA: ACM, 2023: 4549-4560.
- [40] Shi W, Min S, Yasunaga M, et al. Replug: Retrieval-augmented black-box language models[EB/OL]. [2023-01-30]. <https://arxiv.org/abs/2301.12652>.
- [41] Ma X, Gong Y, He P, et al. Query rewriting for retrieval-augmented large language models[C]//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Singapore: Association for Computational Linguistics, 2023: 5303-5315.
- [42] Liu J, Jin J, Wang Z, et al. RETA-LLM: A retrieval-augmented large language model toolkit[EB/OL]. [2023-06-08]. <https://arxiv.org/abs/2306.05212>.
- [43] Elizabeth Reid. Supercharging search with generative AI [EB/OL]. [2023-05-10]. <https://blog.google/products/search/generative-ai-search/>.
- [44] 医联. 医联推出国内首款 AI 医生 medGPT[EB/OL]. [2023-04-28]. <https://www.medlinker.com/news/198>.
- [45] Zhou J, Chen Z, Wan D, et al. CharacterGLM: Customizing Chinese conversational AI Characters with large language models[EB/OL]. [2023-11-08]. <https://arxiv.org/abs/2311.16832>.
- [46] Park J S, O'Brien J, Cai C J, et al. Generative agents: Interactive simulacra of human behavior[C]//Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology. New York: Association for Computing Machinery, 2023: 1-22.
- [47] Xi Z, Chen W, Guo X, et al. The rise and potential of large language model based agents: A survey[EB/OL]. [2023-09-14]. <https://arxiv.org/abs/2309.07864>.
- [48] Modarressi A, Imani A, Fayyaz M, et al. RET-LLM: Towards a general read-write memory for large language models[EB/OL]. [2023-05-23]. <https://arxiv.org/abs/2305.14322>.
- [49] Schick T, Dwivedi-Yu J, Dessì R, et al. Toolformer: Language models can teach themselves to use tools[EB/OL].

- [2023-02-09]. <https://arxiv.org/abs/2302.04761>.
- [50] Gravitass S. Auto-GPT: An Autonomous GPT-4 experiment[EB/OL]. [2023-04-08]. <https://github.com/antony0596/auto-gpt>. 2023.
- [51] Wang Y, Su Z, Zhang N, et al. A survey on metaverse: Fundamentals, security, and privacy[J]. *IEEE Communications Surveys & Tutorials*, 2022, 25(1): 319-352.
- [52] 上海市科学技术委员会. 上海市“元宇宙”关键技术攻关行动方案(2023-2025年)[EB/OL]. [2023-06-13]. <https://stcsm.sh.gov.cn/cmsres/f5/f570dd0a6a334f06a646ec9a34e94a70/db06f664db4f1b186dac1f9ef6ebb98d.pdf>.
- [53] 百度 Create2022: 希壤发布元宇宙底座 MetaStack[EB/OL]. [2023-01-11]. https://cloud.baidu.com/news/news_dc9cf47a-e84d-4052-9ddc-634fd997dafa.
- [54] Ren F J. Construction of Metaverse Center based on Advanced Intelligence[EB/OL]. [2023-12-28]. <https://www.interpaper.org/2022-Annual-Report-EU-Academy-of-Sciences.pdf>.
- [55] Chen C, Feng X, Zhou J, et al. Federated large language model: A position paper[EB/OL]. [2023-07-18]. <https://arxiv.org/abs/2307.08925>.
- [56] Sun H, Zhang Z, Deng J, et al. Safety assessment of Chinese large language models[EB/OL]. [2023-04-20]. <https://arxiv.org/abs/2304.10436>.
- [57] Rozado D. The political biases of ChatGPT[J]. *Social Sciences*, 2023, 12(3): 148.
- [58] Sven G, Pushmeet K. Identifying AI-generated images with SynthID[EB/OL]. [2023-08-29]. <https://deepmind.google/discover/blog/identifying-ai-generated-images-with-synthid/>.
- [59] Fernandez P, Couairon G, Jégou H, et al. The stable signature: Rooting watermarks in latent diffusion models[EB/OL]. [2023-03-27]. <https://arxiv.org/abs/2303.15435>.
- [60] 刘明录, 郑彦, 韩雪, 等. 基于生成式因果语言模型的水印嵌入与检测[J]. *电信科学*, 2023, 39(9): 32-42.
- [61] 王一博, 郭鑫, 刘智锋, 等. AI生成与学者撰写中文论文摘要的检测与差异性比较研究[J]. *情报杂志*, 2023, 42(9): 127-134.
- [62] Chen X, Jin P, Jing S, et al. Automatic Detection of Chinese Generated essays based on pre-trained BERT[C]// *Proceedings of the Joint International Information Technology and Artificial Intelligence Conference*. Chongqing: IEEE, 2022, 10: 2257-2260.
- [63] Guo Z, Schlichtkrull M, Vlachos A. A survey on automated fact-checking[J]. *Transactions of the Association for Computational Linguistics*, 2022, 10: 178-206.
- [64] Zellers R, Holtzman A, Rashkin H, et al. Defending against neural fake news[C]// *Proceedings of the International Conference on Neural Information Processing Systems*. Cambridge: MIT Press, 2019: 9054-9065.
- [65] Morris M R, Sohl-dickstein J, Fiedel N, et al. Levels of AGI: Operationalizing progress on the path to AGI[EB/OL]. [2023-11-04]. <http://arxiv.org/abs/2311.02462>.

Review on hot topics of large generative AI models

DENG Jiawen, REN Fujin*

School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China

Abstract In 2023 the development of large generative AI models achieved a series of breakthroughs. This paper reviews the key technologies of large models that have received significant attention, including the emergent abilities of large language model (LLM), development of multimodal LLMs, as well as alignment and knowledge enhancement techniques of large models. The paper also introduces the vertical applications of generative AI in various fields such as healthcare and education, and its role in advancing the development of AI agents and metaverse technologies. Additionally, the challenges and developmental trends in generative AI are discussed, including issues related to data privacy, biased values, copyright disputes, and fake news propagation.

Keywords generative AI; large language model (LLM); artificial intelligence generated content (AIGC); AI safety ●



(责任编辑 王微)