·专题:ChatGPT 与人工智能技术应用·

ChatGPT 及生成式人工智能现状及未来发展方向

张 熙 杨小汕 徐常胜*

中国科学院 自动化研究所,北京 100190

[摘 要] 生成式人工智能是一种能够自主生成内容的人工智能技术,可以应用于文本生成、图像生成等多个领域。近年来,随着预训练技术的发展和计算硬件的提升,生成式人工智能取得了突破性进展,特别是以 ChatGPT 为代表的生成式对话模型,取得了令人惊艳的效果,开始广泛应用于各行各业。生成式人工智能有广阔的发展前景,本文首先介绍了 ChatGPT 的研究进展,包括预训练语言模型、上下文学习和基于人类反馈的强化学习三个关键技术,以及 ChatGPT 对相关人工智能研究的影响。然后对 ChatGPT 及生成式人工智能在未来的应用发展进行了思考与总结,讨论了目前亟需解决的关键问题,包括更透彻的理解能力、模型轻量化、可控安全的内容生成、知识可持续学习、类脑化认知和可解释性等;希望通过本文的介绍能引起更多的研究人员关注生成式人工智能,进一步推动生成式人工智能的发展与应用。

[关键词] 生成式人工智能;研究进展;未来发展;学术研究

随着技术的进步和创新,人工智能已经成为当今世界经济和社会发展的重要支撑技术。机器学习是人工智能的子领域,根据不同的建模方式,机器学习模型可以分为判别式模型和生成式模型;判别模型直接对数据和预测类别的条件概率进行建模,而生成模型对联合概率分布进行建模[1]。基于生成模型构建的生成式人工智能能够根据已经学习的内容创造性地生成新的内容[2],因此很可能成为专用性人工智能走向通用式人工智能的关键转折点[3]。

生成式人工智能可以利用网络上容易获取到的大规模无标记数据进行预训练,并通过简单适配和高效微调应用到大量的下游任务中。目前生成式人工智能已经在学术研究和技术应用上取得了突破性进展。在自然语言处理领域,生成式人工智能已经能够生成质量较高的自然语言文本,并在对话系统、机器翻译、文本摘要等领域得到应用。例如 OpenAI 公司于 2020 年提出的第三代文本生成模型(Generative Pre-trained Transformer, GPT-3)[4]。凭借其非凡的模型能力、多任务的泛化表现以及小样本学习能力,入选了 2021 年 MIT Technology Review 的“全球十大突破性技术”。其发布的 ChatGPT 自然语言处理模型自公开发布以来,在 5 天内就吸引了超过百万用户,在一个月内拥有了超 5 000 万活跃用户[5],引发了生成式人工智能的研究热潮。在计算机视觉领域,基于变分自编码器(Variational Autoencoders, VAE)[6, 7],生成对抗网络(Generative Adversarial Network, GAN)[8-10],和扩散模型(Diffusion Model)[11, 12]的生成式人工

徐常胜 中国科学院自动化研究所多模态人工智能系统全国重点实验室研究员,国家杰出青年科学基金获得者。研究领域包括多媒体分析,计算机视觉,模式识别,图像处理等。

张熙 中国科学院自动化研究所多模态人工智能系统全国重点实验室在读博士生。研究领域包括多模态学习与理解,视觉问答等。

收稿日期:2023-06-30;修回日期:2023-08-12

* 通信作者,Email: csxu@nlpr.ia.ac.cn

本文受到国家自然科学基金项目(62036012)的资助。

智能系统已经能够生成高质量的图像和视频，在图像编辑、海报设计、艺术创作等领域得到了广泛应用。由此可见，生成式人工智能在学术研究和技术应用上取得了许多重要的突破性进展，为人类社会提供了新的工具和思路。

未来，随着技术的不断发展，生成式人工智能将在各个领域发挥更加重要的作用。当前，ChatGPT及生成式人工智能已经引发了社会各界的广泛关注与激烈讨论，我们应该保持严肃审慎的态度认真思考，为生成式人工智能构建良好的发展环境。本文将先介绍 ChatGPT 等生成式语言模型的研究现状与进展，包括预训练语言模型、上下文学习、基于人类反馈的强化学习三个关键技术，以及对于自然语言处理、机器学习、人机交互、符号人工智能四个人工智能领域的影响。然后，将从更透彻的理解能力、模型轻量化、安全可控生成、可持续学习、类脑化认知和可解释性六个方面分析 ChatGPT 及生成式人工智能的未来发展方向。

1 ChatGPT 相关技术研究进展

近年来，随着深度学习技术的进一步发展和互联网上文本、图像等数据的不断积累，基于深度学习的生成模型获得了飞跃发展。最具代表性的生成式模型之一就是 OpenAI 公司于 2022 年构建的大型语言对话模型 ChatGPT。ChatGPT 是由生成式人工智能技术驱动的自然语言处理工具，能够通过理解和学习人类的语言进行对话，还能根据聊天的上下文进行互动，真正像人类一样交流和回答问题，并能够撰写邮件、文案、代码和论文等。ChatGPT 基于预训练模型 GPT-3.5[4] 构建，利用 Transformer 架构，通过在大规模的文本语料库上训练，学习语言的统计规律和语义信息，并将这些知识编码到模型的参数中。在本章中，我们将首先介绍 ChatGPT 利用的三个关键技术：预训练语言模型，上下文学习，和基于人类反馈的强化学习，然后进一步分析 ChatGPT 对多个人工智能领域研究的影响。

1.1 ChatGPT 相关关键技术

1.1.1 预训练语言模型

预训练语言模型是一种基于深度学习的自然语言处理技术，旨在通过大规模的文本数据进行自监督学习，使模型学习到语言的潜在结构、语法规则和语义关系[13]。自 2018 年来，预训练语言模型获得了越来越多的关注，代表性模型包括 ELMo[14]，BERT[15] 和 GPT-3[4] 等。预训练语言模型的研究包括两方面重要的内容：模型预训练与模型网络结构。在模型预训练中，预训练语言模型会接受大量的未标记文本数据，例如维基百科、互联网文档等，通过自监督任务来学习语言的表示。一种常见的自监督任务是掩码语言建模（Masked Language Modeling，MLM）[15]，即在输入文本中随机掩盖一些单词或标记，模型需要预测这些被掩盖的单词或标记。其他自监督任务还包括文本生成[16]，文本填空[17] 和句子序列识别[17] 等。通过自监督学习，预训练语言模型不仅将词语的语义描述从静态表示提升为上下文感知的动态表示，而且为自然语言处理的任务提供了统一的建模框架。对于预训练语言模型的模型结构，目前有三种主流结构，即自编码语言模型（Autoencoding Language Model），自回归语言模型（Autoregressive Language Model）和混合语言模型（Hybrid Language Model）。具体来说，自编码语言模型以 BERT[15] 为代表，进行双向编码并根据上下文编码信息预测被随机掩码的文字。自回归语言模型以 GPT-1[18] 为代表，采用仅解码器的方法进行单向语言解码和逐字预测。混合语言模型以 T5[19] 为代表，结合上述两类模型的方法，对输入文本进行双向编码，并采用单向语言解码进行逐字预测。综上所述，预训练语言模型的优势在于能够学习到大规模文本数据中的语言知识，并具备一定的通用性，可以迁移到多个下游任务上。此外，预训练语言模型还可以通过不断增加预训练数据和模型规模的方式进一步提升性能。

目前，OpenAI 的 GPT 系列（GPT-1[18]，GPT-2[20]，GPT-3[4]，GPT-4[21]）是应用广泛的预训练语言模型之一。它在多个自然语言处理任务上取得了令人瞩目的成果，并在生成文本方面展现出出色的能力。表 1 展示了 GPT 系列模型的细节对比。具体来说，GPT-1 采用了单向的 Transformer 架构，首先在未标记的原始互联网文本上学习通用语言模型，然后根据特定任务对其进行微调。GPT-2 与 GPT-1 采用相同的模型架构，但扩充了网络参数并在更多的数据集上进行训练，也因此具备更强的语

表 1 GPT 系列模型对比

模型	发布时间	预训练数据规模	模型参数量	预训练数据来源
GPT-1	2018 年 6 月	约 5GB	1.17 亿	BooksCorpus，Wikipedia
GPT-2	2019 年 2 月	40GB	15 亿	WebText
GPT-3	2020 年 5 月	45TB	1 750 亿	Common Crawl 等
GPT-4	2023 年 3 月	未公开	未公开	未公开

言生成能力。GPT-3 仍沿用了 GPT-2 的模型结构，但网络参数达到 1 750 亿，能够生成连贯、富有逻辑的文本，并能够进行多轮对话和理解复杂的指令。自此，预训练语言模型进入了大规模参数时代。基于 GPT-3，GPT-3.5 系列模型也被开发，通过基于代码的预训练、指令微调等方式进一步提高模型性能，Chat-GPT 就是基于该系列模型开发而来。目前，多模态大模型 GPT-4 也被公布，展示出更强的推理与多模态理解能力。总之，得益于规模庞大的预训练数据、强大的 Transformer 架构和创新的训练策略，GPT 系列模型在各种自然语言处理任务上取得了重大突破，并为自然语言生成和理解领域带来了巨大的进步。

1.1.2　上下文学习

上下文学习（In-context Learning，ICL）是 ChatGPT 许多强大功能的基础之一，是自然语言处理任务的一种新范式[22]。在上下文学习中，模型会基于当前输入数据的上下文信息进行学习和预测。这些上下文信息可以包括任务样例、前文、后文、历史数据、环境条件等。通过综合考虑这些上下文信息和提示，模型可以更好地理解和解释输入数据，通过类比和模仿完成任务，并且在不同的环境下具备更好的适应性。与监督学习或微调不同，上下文学习不需要对模型参数进行更新，而是直接基于预训练语言模型进行类比学习和任务预测，提高了模型应用的效率，帮助模型形成零样本和少样本的学习能力。

近年来，思维链（Chain of Thought，CoT）提示技术[23]也被提出，来进一步提高模型解决如算数、逻辑推理等复杂问题的推理能力。思维链技术旨在构建更细粒度和分步骤的上下文提示，来模拟人类在完成复杂任务时的思维过程。例如对于二十四点的任务，基于思维链的上下文学习除了给定模型输入“输入是 4 9 10 13”外，还会输入中间的三个计算方程：“13－9＝4（剩余：4 4 10）”“10－4＝6（剩余：4 6）”“4 * 6＝24（剩余：24）”，来提示模型应该进行的推理过程。思维链帮助模型建立了数字、运算符和运算规则之间的关联，从而使系统能够在解决算术问题时表现得更加智能和准确。基于思维链，研究人员开始从很多个角度改进生成式语言模型的推理能力，例如构建多个思维链并结合一致性学习完成目标任务[24]，构建思维树（Tree of Thought，ToT）考虑多种不同推理路径并结合自我评估进行全局最优的选择[25]等。

1.1.3　基于人类反馈的强化学习

生成式大语言模型在微调阶段，常利用基于人类反馈的强化学习进一步提高对话的质量。基于人类反馈的强化学习是一种机器学习方法，旨在通过与人类进行交互和反馈来训练智能代理，即模型。这种方法将强化学习和人类专家的知识相结合，以提高智能代理在复杂任务中的性能和效果。ChatGPT 采用了与 InstructGPT[26]类似的基于人类反馈的强化学习方法。在 InstructGPT 中，强化学习主要包括三个阶段。首先，通过预训练和有监督微调得到初始的大语言模型。第二阶段进行奖励模型的训练。在该阶段中首先要构建数据集，给定问题和多个大语言模型生成的回答，人类标注者需要对这些回答进行综合排序。然后构建奖励模型，并训练奖励模型去预测人类对不同回答的排序结果。在第三阶段，采用近端策略优化算法，基于奖励模型的反馈来优化大语言模型，最终得到满足人类偏好的模型。总之，基于人类反馈的强化学习技术可以根据用户反馈不断学习和优化，从而提高生成式人工智能模型在复杂任务中的性能和效果。最近的 GPT-4[21]也基于该技术加入一个额外的安全奖励信号，来减少模型的有害输出，帮助模型判断安全边界来缓解风险。

1.2　ChatGPT 对人工智能研究的影响

ChatGPT 对人工智能的研究产生了很广泛的影响，包括自然语言处理、机器学习、强化学习、人机交互、符号人工智能、模型可解释性和可控性等。下面，本文将选取四个主要方面进行详细阐述。

1.2.1　自然语言处理

ChatGPT 作为一个强大的语言模型，对自然语言处理领域产生了重要的影响。通过大规模预训练技术和更长的上下文长度，ChatGPT 在对话流畅性、多轮对话、复杂语义理解等多个传统的自然语言处理任务上取得了巨大的进展，突破了很多性能上的瓶颈，甚至改变了自然语言处理的研究范式。在自然语言处理的发展历程中，从传统的统计学习方法和词嵌入方法，到预训练加微调范式，再到如今的大语言模型，ChatGPT 因其优秀的性能产生了至关重要的推动作用。随着 ChatGPT 的出现，自然语言处理的研究方向也发生了改变。一方面，在拥有充分计算资源的情况下，研究人员开始追求规模越来越大的语言模型，通过增加预训练数据的多样性来涵盖越来越多的知识领域[18, 21]，从而构建一个理想的通用大语言模型。另一方面，越来越多的研究者开始关注不需要耗费很多计算资源的领域，如基于知识检索的上下文学习[27, 28]，大语言模型的可解释

性[29]和构建更好的测评方式[30]等。

1.2.2 机器学习

ChatGPT 的出现使人工智能和机器学习再次活跃在公众视野中，改变了其对机器学习的看法，使大众更加认可机器学习的性能。传统的机器学习方法，如决策树、支持向量机等，致力于教导机器如何完成任务，并且要求将每个任务分解成很多步骤，逐一教导机器完成。但是 ChatGPT 的出现让研究者重新思考，是否可以把问题交给机器进行端到端的解决，而不必逐步教导。从 ChatGPT 的卓越性能可以发现，机器的学习能力超乎我们的想象。因此，这种基于数据驱动的端到端方法逐渐成为机器学习研究的重点之一，也为机器学习在自然语言处理和对话系统领域的应用提供了新的思路和效果。

1.2.3 人机交互

ChatGPT 的出现促进了人机交互的发展，能够为人们提供更加自然、流畅的对话体验。它通过学习大量的文本数据，能够生成流畅、连贯的语言表达，对于开发更加自然、交互性强的人机界面非常重要。同时，ChatGPT 的出现也使聊天机器人的研究和开发取得了重要突破，其强大的生成能力能够使聊天机器人更好地理解用户的意图，并具有长期记忆能力。同时，利用上下文学习和用户特定数据，还可以提供不同的聊天风格以及个性化的用户体验，丰富了人机交互的方式。另外，在智能助手和个性化服务，内容生产和创意辅助，教育和培训等方面的人机交互工具研究中，ChatGPT 也将发挥更大的作用，提供新的思路和方法。

1.2.4 符号人工智能

符号逻辑可以帮助 ChatGPT 模型更好地理解语言的含义与语境。具体来说，ChatGPT 主要基于统计模型和概率方法，而符号人工智能更侧重于基于逻辑和符号推理的方法。尽管 ChatGPT 更加注重数据驱动的语言生成，但它的出现也促进了符号人工智能和混合方法的研究。这是因为我们希望 ChatGPT 具有更多与人类相似的逻辑能力，比如可反思认知过程、知识的学习和推理等。因此，一些研究者开始探索如何将符号推理和深度学习结合，以提高对话系统的逻辑推理和常识理解能力。例如，符号人工智能强调对知识的符号表示和推理能力，因此可以通过将知识图谱等符号知识引入 ChatGPT 中，以提供更丰富的语义知识，更准确地处理实体的语义关系和推理。另外，可以结合符号人工智能的技术，在 ChatGPT 中引入对话管理器或逻辑规则和推理机制，控制对话流程、上下文跟踪和指导生成回复的策略，使 ChatGPT 更好地处理逻辑关系。

2 ChatGPT 及生成式人工智能未来发展方向

ChatGPT 及生成式人工智能的研究正处于快速发展阶段。随着深度学习技术的不断发展和进步，越来越多的生成式模型被提出，并能够生成具有创造性和多样性的各种内容。同时，生成式人工智能也在推动多个领域的前沿研究，包括自然语言处理、计算机视觉、机器人等领域。总之，生成式人工智能的研究可以增强人类创造力、推动跨学科研究、解决现实问题、推进社会进步，在未来有着广阔的发展前景和巨大的潜力。它也正逐渐走进我们的日常生活和各个产业，有着广阔的应用前景。随着技术的进步和创新，我们可以期待其在以下几个方向取得重要研究进展(图 1)。

2.1 更透彻的理解能力

生成式人工智能模型的核心是理解数据和任务的能力。目前，以 ChatGPT 为代表的大语言模型在回答问题、提供建议、总结和优化文本等几个文本生成任务中已经达到或超过了人类水平[31]，并可以处理需要精确意图理解的新任务。我们猜测这些能力来自于两个方面，一是在各种形式和任务的海量数据上进行预训练，使模型掌握了语言规律。二是通过在不同任务上基于指令的微调学习[32, 33]得到了普

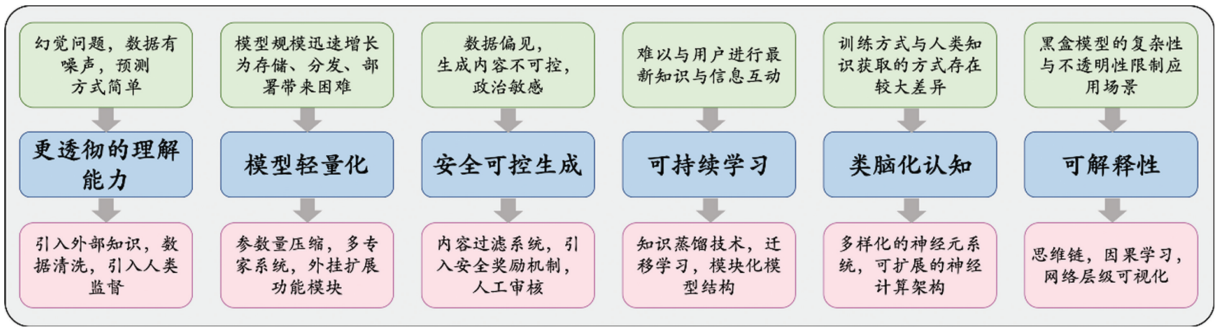

图 1 ChatGPT 及生成式人工智能的难点和未来发展方向

遍的任务解决能力。然而，在应用中我们发现 ChatGPT 等大模型存在“幻觉问题”[34]，即以一种令人信服但完全编造的方式来表达自己，从而产生事实性错误。这限制了它们在金融、法律咨询和医疗建议等知识广泛领域的适用性。某些研究指出，这些错误是由于训练数据中存在错误和噪声[35, 36]导致的。同时，ChatGPT 采用的预测句子下一个可能的单词的方式过于简单，本质并没有建模准确信息和推理的能力。为解决这些问题，ChatGPT 需要具备更透彻的理解能力和更精确的建模能力。目前，一些研究人员提出通过引入外部知识来增强模型的理解能力和缓解幻觉[37, 38]，即通过从大型知识库（如维基百科）中检索相关知识，为模型给出提示，从而引导模型生成更加准确的信息。另外，也可以考虑清洗数据以提高训练数据质量，引入更多的人类监督，进一步扩大模型参数等方式提高模型的理解能力。

2.2　模型轻量化

基于 Transformer 的大语言模型的能力与参数数量成对数线性增长[39]，且当模型大小超过一定阈值后将具备涌现能力，如上下文学习和思维链等。在 GPT 系列模型中，2018 年 6 月推出的 GPT-1[18]模型参数量为 1.17 亿，预训练数据量约 5 GB。2019 年 2 月提出的 GPT-2[20]模型参数量为 15 亿，预训练数据量约 40 GB，而 2020 年 5 月提出的 GPT-3[4]实现了模型规模的飙升，参数量高达 1 750 亿，预训练数据量高达 45 TB。因此，我们可以预见未来的生成式人工智能模型的规模将继续增长。更大规模的模型可以提供更深入、更准确的语言理解和生成能力，使得对话更加自然流畅，并且使模型能够更好地理解和回复复杂的问题和指令。然而，这些模型参数规模与训练数据规模的迅速增长带来极大的成本，为现实应用中的存储、分发、部署等带来了挑战。据统计，每训练一次 GPT-3 模型需要耗电约 19 万度，大约产生了 85 000 千克的二氧化碳当量，相当于一辆汽车行驶 70 万公里（大约是地球与月球间距离的两倍）的排放量[40]。因此，需要对生成式人工智能模型进行轻量化和优化，以提高模型的效率与实用性。在 GPT-3 被提出以后，陆续有不同研究机构或组织通过微调的方式开发出高效的语言模型，例如 Vicuna[41]、Koala[42]、Alpaca[43]和 LLaMA[44]等，分别在用户共享的对话、学术研究、指令执行等方面有了更强大的能力。其他可以考虑的方法包括模型参数量压缩，如参数剪枝、低秩分解、量化等；多专家系统与稀疏激活技术的结合；分布式训练和推理，设计更高效的通信协议和网络拓扑结构；基础模型与外挂扩展功能模块的结合等。总之，更轻量化和高效的生成式人工智能模型将有助于其在更广泛的应用场景中发挥更大的作用。

2.3　安全可控生成

随着 ChatGPT 等生成式大模型技术的突破与在各个场景的广泛应用，人们也开始对人工智能带来的道德、伦理、安全等风险感到恐慌。受数据驱动的训练方式限制，生成式大模型面临着训练数据来源合规性、数据使用的偏见性、生成内容不可控不安全等风险。Geoffrey Hinton 在 2023 年 3 月对 ChatGPT 的评述中也指出其存在于政治上被利用的可能[45]。另有研究表明，在 15 项政治倾向测试中，ChatGPT 总体上倾向于所谓的“左翼自由派”观点。另外，在美国炒作所谓“中国气球”事件期间，有用户发现 ChatGPT 支持美国击落中方气球，却不支持中国击落美方气球。因此，生成式人工智能的安全可控生成对社会舆论也将产生很大的影响。清华大学计算机系黄民烈教授的研究团队已经在安全伦理方面开展了相关研究，并依此建立了大模型安全分类体系[46]，其中不安全的对话场景包括：政治敏感、犯罪违法、身体健康、心理健康、财产隐私、歧视/偏见、辱骂/仇恨言论、伦理道德八大方面。由此可见，在安全可控生成方面，生成式人工智能面临着更多的研究挑战。未来，为实现不当内容过滤，阻止违法或有害内容生成，可以通过整合内容过滤系统、使用人工审核、引入安全奖励机制[21]等方式。为避免生成内容的偏见，训练数据的审核和清洗也应被进一步重视，尽可能采用自动标注的合成数据训练人工智能模型，并警惕中长期地缘政治威胁。另外，可以加强对模型的自动测试，针对安全缺陷通过微调的方式进行快速迭代，促使模型越来越符合人类的认知理解模式，生成更加安全可信的内容。基于更安全可控的内容生成，生成式人工智能将在舆论领域发挥更大的作用，例如根据媒体工作者提供的信息和要求，自动生成相关的新闻报道或博客文章，或用于编辑和校对，提供语法、风格、逻辑方面的建议，提高内容生产的效率。

2.4　可持续学习

目前 ChatGPT（GPT-3.5）还不具备可持续的模型知识更新能力。它的训练数据只包含 2021 年 9 月之前的相关数据与知识，导致其难以与用户进行最新知识与信息的互动，无法回答如“成都大运会的召开时间”等问题。如果停止对 ChatGPT 模型的

更新，它可能永远学不到时事信息，也限制其更广泛的应用。因此，生成式人工智能的未来发展中不可忽略的一点，是如何使生成式模型在学习新任务或数据时可以快速适应和更新，不必重新训练整个模型，且不会在更新的过程中遗忘先前获得的知识。这对于 ChatGPT 这类大规模生成式模型非常有利，因为能够节省计算资源与时间，并提高模型的效率和应用范围。此外，可持续学习也可使模型跟随用户需求不断演进更新，得到更个性化和更相关的结果，这对于更好的人机交互应用非常重要。未来，研究人员在生成式人工智能的模型可塑性和知识持续更新方面可考虑使用更先进的增量学习算法[47, 48]，例如知识蒸馏技术、高效微调技术、迁移学习等，考虑设计更加可塑的模型结构以适应不同的任务和数据，例如模块化结构、可分离卷积、自适应网络结构等，或考虑开发更加智能的集成学习方法，以将多个模型的预测结果整合起来，提高模型的泛化能力和适应性。

2.5 类脑化认知

通用人工智能的发展，离不开生成式人工智能的创造能力。类脑化是指生成式人工智能应具有与人类大脑类似的特性和能力，以更好地模拟人类的认知和学习过程。现有的生成式模型的训练方式与人类知识获取的方式存在很大的差异，大模型的生成式过程属于快思考，是一种直觉思维，容易出现错误和偏见，且不适合规划类任务。而人类的思维方式是慢思考，是一种理性思维，可以基于对世界的认知与物理世界相互作用，实现闭环反馈。因此，未来的生成式人工智能需要更复杂和多样化的神经元系统，以及更加灵活的神经网络连接方式，从而模拟人类神经元与脑区的各种特性和行为。另外，生成式人工智能需要可塑性和可扩展性更强的神经计算架构，模拟人脑神经可塑性和高效的记忆调用能力，并实现神经网络的快速计算。基于更强的类脑化认知，生成式人工智能可能将在科学智能领域发挥更大的作用，即学习、模拟和预测自然界和人类社会的各种现象和规律，从而推动科学发现和创新。例如在生物和医药研究领域，用于预测蛋白质结构、加快新药和疫苗研发等。目前 DeepMind 提出的 AlphaFold 已经可以预测蛋白质的三维结构，以帮助生物学家理解蛋白质的功能和病理机制。

2.6 可解释性

由于生成式人工智能模型通常是基于神经网络或深度学习算法构建，这些黑盒模型的复杂性与不透明性使它们难以显式地展示其思维过程，从而给模型的应用带来了一定的难度和不确定性。例如，在医疗领域，如果一个生成式模型能够准确地诊断病人的疾病，但不能提供诊断依据，这将会降低医生和患者的信任度，从而影响模型的实际应用和推广。为此，生成式人工智能的研究需要考虑模型的可靠性和可信性，未来需要关注其可解释性的提升。可以考虑的方法包括利用因果学习理论、网络层级可视化、利用对抗训练生成解释性更强的模型，或利用思维链[23]等技术展示模型的思维过程，进行局部敏感性分析等。基于更强的可解释性，生成式人工智能模型将能够在教育领域发挥更大的作用，例如作为个性化学习辅导工具，为学生提供有关特定主题的解释和指导。它可以帮助学生理解概念、解决问题和提供学习建议。

3 结 语

著名市场调查机构 Gartner 在 2022 年的重要战略技术趋势报告中将生成式人工智能列为 12 项重要战略之首，并预计到 2025 年，生成式人工智能产生的数据占人类全部数据的比例将从目前的 1% 增长到 10%[49]。由此可见，ChatGPT 及生成式人工智能将会给人类社会带来深刻的影响与变革。但目前的生成式人工智能技术在学术研究和应用上还有很多不足之处需要改进：模型需要具有更透彻的理解能力与可控安全的内容生成能力，模型训练效率有待提升，模型需要具备知识可持续学习的能力以及类脑化的认知能力，模型的可解释性也需要提升。在应用过程中，需要实现用户更友好的智能交互，充分考虑安全伦理和数据隐私等问题。未来，生成式人工智能将在智能教育、智能化软件开发和科学智能等领域产生越来越重要的影响。综上所述，我们期待 ChatGPT 及生成式人工智能能够在更广泛的应用场景中发挥更大的作用，推动人工智能技术的进一步发展，最终为人类带来更多的技术创新和进步。

参 考 文 献

[1] Lasserre JA, Bishop CM, Minka TP. Principled hybrids of generative and discriminative models// Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2006: 87—94.

[2] Google Cloud. Introduction to generative AI. [2023-07-28]/[2023-08-11]. https://www.cloudskillsboost.google/course_templates/536.

[3] 陈永伟. 超越 ChatGPT：生成式 AI 的机遇、风险与挑战. 山东大学学报(哲学社会科学版)，2023(3)：127—143.

[4] Brown TB, Mann B, Ryder N, et al. Language models are few-shot learners. Advances in Neural Information Processing Systems, 2020: 1877—1901.

[5] Jackson S. OpenAI executives say releasing ChatGPT for public use was a last resort after running into multiple hurdles-and they're shocked by its popularity. [2023-01-25]/[2023-06-29]. https://www. businessinsider. com/chatgpt-openai-executives-are-shocked-by-ai-chatbot-popularity-2023-1.

[6] Kingma DP, Welling M. Auto-encoding variational bayes. (2013-12-20)/[2023-06-29]. https://arxiv. org/abs/1312. 6114.

[7] Zhang S, Chen JY, Chen JY, et al. Data imputation in IoT using spatio-temporal variational auto-encoder. Neurocomputing, 2023, 529: 23—32.

[8] Creswell A, White T, Dumoulin V, et al. Generative adversarial networks: an overview. IEEE signal processing magazine, 2018, 35(1): 53—65.

[9] Abdal R, Qin YP, Wonka P. Image2StyleGAN: how to embed images into the styleGAN latent space// Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. New York: IEEE, 2020: 4431—4440.

[10] Luo WY, Yang S, Zhang XJ, et al. SIEDOB: semantic image editing by disentangling object and background// Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2023: 1868—1878.

[11] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems, 2020: 6840—6851.

[12] Rombach R, Blattmann A, Lorenz D, et al. High-resolution image synthesis with latent diffusion models// Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2022: 10674—10685.

[13] Qiu XP, Sun TX, Xu YG, et al. Pre-trained models for natural language processing: a survey. Science China Technological Sciences, 2020, 63(10): 1872—1897.

[14] Matthew EP, Mark N, Mohit I, et al. Deep contextualized word representations// Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. New Orleans: Association for Computational Linguistics, 2018: 2227—2237.

[15] Devlin J, Chang MW, Lee K, et al. BERT: pretraining of deep bidirectional transformers for language understanding. (2018-10-11)/[2023-06-29]. https://arxiv. org/abs/1810. 04805.

[16] Du H, Li Z, Niyato D, et al. Enabling AI-generated content (AIGC) services in wireless edge networks. (2023-01-09)/[2023-06-29]. https://arxiv. org/abs/2301. 03220.

[17] Lewis M, Liu Y, Goyal N, et al. BART: denoising sequenceto-sequence pre-training for natural language generation, translation, and comprehension. (2019-10-29)/[2023-06-29]. https://arxiv. org/abs/1910. 13461.

[18] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training. (2018-06-11)/[2023-06-29]. https://cdn. openai. com/research-covers/language-unsupervised/language_understanding_paper. pdf.

[19] Raffel C, Shazeer N, Roberts A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 2020, 21(1): 5485—5551.

[20] Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners. OpenAI blog, 2019, 1(8): 9.

[21] OpenAI. GPT-4 technical report. (2023-03-27)/[2023-06-29]. https://arxiv. org/abs/2303. 08774.

[22] Dong Q, Li L, Dai D, et al. A survey for in-context learning. (2022-12-31)/[2023-06-29]. https://arxiv. org/abs/2301. 00234.

[23] Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 2022: 24824—24837.

[24] Wang X, Wei J, Schuurmans D, et al. Self-consistency improves chain of thought reasoning in language models. (2022-03-21)/[2023-06-29]. https://arxiv. org/abs/2203. 11171.

[25] Yao SY, Yu D, Zhao J, et al. Tree of thoughts: deliberate problem solving with large language models. (2023-05-17)/[2023-06-029]. https://arxiv. org/abs/2305. 10601.

[26] Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 2022: 27730—27744.

[27] Chen X, Li L, Zhang N, et al. Decoupling knowledge from memorization: retrieval-augmented prompt learning. (2022-05-29)/[2023-06-029]. https://arxiv. org/abs/2205. 14704.

[28] Shi P, Zhang R, Bai H, et al. XRICL: cross-lingual retrieval-augmented in-context learning for cross-lingual text-to-SQL semantic parsing. (2022-10-25)/[2023-06-29]. https://arxiv. org/abs/2210. 13693.

[29] Wu ZX, Geiger A, Potts C, et al. Interpretability at scale: identifying causal mechanisms in alpaca. (2023-05-15)/[2023-06-29]. https://arxiv. org/abs/2305. 08809.

[30] Ye QH, Xu HY, Xu GH, et al. mPLUG-owl: modularization empowers large language models with multimodality. (2023-04-27)/[2023-06-29]. https://arxiv. org/abs/2304. 14178.

[31] Liang P, Bommasani R, Lee T, et al. Holistic evaluation of language models. (2022-11-16)/[2023-06-29]. https://arxiv. org/abs/2211. 09110.

[32] Chung HW, Hou L, Longpre S, et al. Scaling instruction-finetuned language models. (2022-10-20)/[2023-06-29]. https://arxiv. org/abs/2210. 11416.

[33] Dan G, Dan R. Learning from natural instructions. Machine Learning, 2014, 94(2): 205—232.

[34] Wu TY, He SZ, Liu JP, et al. A brief overview of ChatGPT: the history, status quo and potential future development. IEEE/CAA Journal of Automatica Sinica, 2023, 10(5): 1122—1136.

[35] Borji A. A categorical archive of ChatGPT failures. (2023-02-06)/[2023-06-29]. https://arxiv. org/abs/2302. 03494.

[36] Frieder S, Pinchetti L, Griffiths RR, et al. Mathematical capabilities of ChatGPT. (2023-01-31)/[2023-06-29]. https://arxiv.org/abs/2301.13867.
[37] Guu K, Lee K, Tung Z, et al. Retrieval augmented language model pre-training. (2020-02-10)/[2023-06-29]. https://arxiv.org/abs/2002.08909.
[38] Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems, 2020: 9459—9474.
[39] Wei J, Tay Y, Bommasani R, et al. Emergent abilities of large language models. (2022-06-15)/[2023-06-29]. https://arxiv.org/abs/2206.07682.
[40] Anthony LFW, Kanding B, Selvan R. Carbontracker: Tracking and predicting the carbon footprint of training deep learning models. (2020-07-06)/[2023-06-29]. https://arxiv.org/abs/2007.03051.
[41] The Vicuna Team. Vicuna: an open-source chatbot impressing gpt-4 with 90% * ChatGPT quality. (2023-03-30)/[2023-06-29]. https://lmsys.org/blog/2023-03-30-vicuna/.
[42] Geng XY, Gudibande A, Liu H, et al. Koala: a dialogue model for academic research. (2023-04-03)/[2023-06-29]. https://bair.berkeley.edu/blog/2023/04/03/koala/.
[43] Taori R, Gulrajani I, Zhang T, et al. Alpaca: a strong, replicable instruction-following model. (2023-03-13)/[2023-06-29]. https://crfm.stanford.edu/2023/03/13/alpaca.html.
[44] Touvron H, Lavril T, Izacard G, et al. LLaMA: open and efficient foundation language models. (2023-02-27)/[2023-06-29]. https://arxiv.org/abs/2302.13971.
[45] GE Hinton. "Godfather of artificial intelligence" talks impact and potential of AI. (2023-03-25)/[2023-06-29]. https://youtu.be/qpoRO378qRY.
[46] 贾君玉,张素. 清华大学黄民烈团队:发布安全评估框架促大模型迈向可控可信. (2023-03-27)/[2023-06-30]. https://www.tsinghua.edu.cn/info/1182/102442.htm.
[47] Jang J, Ye S, Yang S, et al. Towards continual knowledge learning of language models. (2021-10-07)/[2023-06-29]. https://arxiv.org/abs/2110.03215.
[48] Sun Y, Wang SH, Li YK, et al. ERNIE 2.0: a continual pre-training framework for language understanding// Proceedings of the AAAI Conference on Artificial Intelligence. New York: Association for the Advancement of Artificial Intelligence, 2020: 8968—8975.
[49] Groombridge D, Karamouzis F, Chandrasekaran A. Top strategic technology trends for 2022. (2021-10-18)/[2023-06-29]. https://www.gartner.com/en/documents/4006913.

Current State and Future Development Directions of ChatGPT and Generative Artificial Intelligence

Xi Zhang Xiaoshan Yang Changsheng Xu*
Institutes of Automation, Chinese Academy of Sciences, Beijing 100190

Abstract Generative artificial intelligence is an AI technology capable of autonomously generating content and finds applications in various domains such as text generation and image generation. In recent years, with the advancement of pre-training techniques and improvements in computing hardware, generative AI has made significant breakthroughs. Specifically, generative dialogue models like ChatGPT have achieved impressive results and are being widely applied across industries. Generative AI holds vast development prospects. This paper first introduces the research progress of ChatGPT, including pre-training language models, context learning, and reinforcement learning based on human feedback, as well as the impact of ChatGPT on relevant AI research. Subsequently, the paper discusses and summarizes the future application development of ChatGPT and generative AI, addressing key issues that need to be addressed, such as enhanced understanding capabilities, model lightweight, controlled and safe content generation, sustainable knowledge learning, brain-like cognition, and interpretability. It is hoped that this paper will draw more attention from researchers to generative AI and further promote the development and application of generative AI.

Keywords generative artificial intelligence; research progress; future development; academic research

(责任编辑 崔国增 姜钧译)

* Corresponding Author, Email: csxu@nlpr.ia.ac.cn