中图法分类号:TP183　　文献标识码:A　　文章编号:1006-8961(2023)09-2749-14

论文引用格式:Yan H, Liu Y L, Jin L W and Bai X. 2023. The development, application, and future of LLM similar to ChatGPT. Journal of Image and Graphics, 28(09):2749-2762(严昊, 刘禹良, 金连文, 白翔. 2023. 类ChatGPT大模型发展、应用和前景. 中国图象图形学报, 28(09):2749-2762)[DOI:10.11834/jig.230536]

类ChatGPT大模型发展、应用和前景

严昊[1],刘禹良[1*],金连文[2],白翔[1]

1. 华中科技大学人工智能与自动化学院,武汉　430074; 2. 华南理工大学电子与信息学院,广州　510640

摘　要: 生成式人工智能技术自ChatGPT发布以来,不断突破瓶颈,吸引了资本规模投入、多领域革命和政府重点关注。本文首先分析了大模型的发展动态、应用现状和前景,然后从以下3个方面对大模型相关技术进行了简要介绍:1)概述了大模型相关构造技术,包括构造流程、研究现状和优化技术;2)总结了3类当前主流图像—文本的大模型多模态技术;3)介绍了根据评估方式不同而划分的3类大模型评估基准。参数优化与数据集构建是大模型产品普及与技术迭代的核心问题;多模态能力是大模型重要发展方向之一;设立评估基准是比较与约束大模型的关键方法。此外,本文还讨论了现有相关技术面临的挑战与未来可能的发展方向。现阶段的大模型产品已有强大的理解能力和创造能力,在教育、医疗和金融等领域已展现出广阔的应用前景。但同时,它们也存在训练部署困难、专业知识不足和安全隐患等问题。因此,完善参数优化、优质数据集构建、多模态等技术,并建立统一、全面、便捷的评估基准,将成为大模型突破现有局限的关键。

关键词: 人工智能(AI);ChatGPT;多模态技术;自然语言处理;大模型(LLM)

The development, application, and future of LLM similar to ChatGPT

Yan Hao[1], Liu Yuliang[1*], Jin Lianwen[2], Bai Xiang[1]

1. School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan　430074, China;

2. School of Electronic and Information Engineering, South China University of Technology, Guangzhou　510640, China

Abstract: Generative artificial intelligence (AI) technology has achieved remarkable breakthroughs and advances in its intelligence level since the release of ChatGPT several months ago, especially in terms of its scope, automation, and intelligence. The rising popularity of generative AI attracts capital inflows and promotes the innovation of various fields. Moreover, governments worldwide pay considerable attention to generative AI and hold different attitudes toward it. The US government maintains a relatively relaxed attitude to stay ahead in the global technological arena, while European countries are conservative and are concerned about data privacy in large language models (LLMs). The Chinese government attaches great importance to AI and LLMs but also emphasizes the regulatory issues. With the growing influence of ChatGPT and its competitors and the rapid development of generative AI technology, conducting a deep analysis of them becomes necessary. This paper first provides an in-depth analysis of the development, application, and prospects of generative AI. Various types of LLMs have emerged as a series of remarkable technological products that have demonstrated versatile capabili-

收稿日期:2023-07-26;修回日期:2023-08-10;预印本日期:2023-08-17

*通信作者:刘禹良　ylliu@hust.edu.cn

基金项目:国家自然科学基金项目(62206104, 62225603)

Supported by: National Natural Science Foundation of China (62206104, 62225603)

ties across multiple domains, such as education, medicine, finance, law, programming, and paper writing. These models are usually fine-tuned on the basis of general LLMs, with the aim of endowing the large models with additional domain-specific knowledge and enhanced adaptability to a specific domain. LLMs (e. g. , GPT-4) have achieved rapid improvements in the past few months in terms of professional knowledge, reasoning, coding, credibility, security, transferability, and multimodality. Then, the technical contribution of generative AI technology is briefly introduced from four aspects: 1) we review the related work on LLMs, such as GPT-4, PaLM2, ERNIE Bot, and their construction pipeline, which involves the training of base and assistant models. The base models store a large amount of linguistic knowledge, while the assistant models acquire stronger comprehension and generation capabilities after a series of fine-tuning. 2) We outline a series of public LLMs based on LLaMA, a framework for building lightweight and memory-efficient LLMs, including Alpaca, Vicuna, Koala, and Baize, as well as the key technologies for building LLMs with low memory and computation requirements, consisting of low-rank adaptation, Self-instruct, and automatic prompt engineer. 3) We summarize three types of existing mainstream image-text multimodal techniques: training additional adaptation layers to align visual modules and language models, multimodal instruction fine-tuning, and LLM serving as the center of understanding. 4) We introduce three types of LLM evaluation benchmarks based on different implementation methods, namely, manual evaluation, automatic evaluation, and LLM evaluation. Parameter optimization and fine-tuning dataset construction are crucial for the popularization and innovation of generative AI products because they can significantly reduce the training cost and computational resource consumption of LLMs while enhancing the diversity and generalization ability of LLMs. Multimodal capability is the future trend of generative AI because multimodal models have the ability to integrate information from multiple perceptual dimensions, which is consistent with human cognition. Evaluation benchmarks are the key methods to compare and constrain the models of generative AI, given that they can efficiently measure and optimize the performance and generalization ability of LLMs and reveal their strengths and limitations. In conclusion, improving parameter optimization, high-quality dataset construction, multimodal, and other technologies and establishing a unified, comprehensive, and convenient evaluation benchmark will be the key to achieving further development in generative AI. Furthermore, the current challenges and possible future directions of the related technologies are discussed in this paper. Existing generative AI products have considerable creativity, understanding, and intelligence and have shown broad application prospects in various fields, such as empowering content creation, innovating interactive experience, creating "digital life," serving as smart home and family assistants, and realizing autonomous driving and intelligent car interaction. However, LLMs still exhibit some limitations, such as lack of high-quality training data, susceptibility to hallucinations, output factual errors, uninterpretability, high training and deployment costs, and security and privacy issues. Therefore, the potential research directions can be divided into three aspects: 1) the data aspect focuses on the input and output of LLMs, including the construction of general tuning instruction datasets and domain-specific knowledge datasets. 2) The technical aspect improves the internal structure and function of LLMs, including the training, multimodality, principle innovation, and structure pruning of LLMs. 3) The application aspect enhances the practical effect and application value of LLMs, including security enhancement, evaluation system development, and LLM application engineering implementation. The advancement of generative AI has provided remarkable benefits for economic development. However, it also entails new opportunities and challenges for various stakeholders, especially the industry and the general public. On the one hand, the industry needs to foster a large pool of researchers who can conduct systematic and cutting-edge research on generative AI technologies, which are constantly improving and innovating. On the other hand, the general public needs to acquire and apply the skills of prompt engineering, which can enable them to utilize existing LLMs effectively and efficiently.

Key words: artificial intelligence (AI); ChatGPT; multi-modality; natural language processing; large language model (LLM)

0 引　言

自 2022 年 11 月 OpenAI 公布了 ChatGPT 以来，生成式人工智能的发展迎来了爆发式的增长。在过去的数月里，生成式人工智能不断突破边界，实现了从文本、图像到音频等多领域的自动化和智能化生成。不仅如此，得益于相关技术的飞速发展，目前市面上大模型（large language model，LLM）的智能程度已有了质的飞跃。这些产品中性能最佳的就是 GPT-4（generative pre-training transformer-4），它是 OpenAI 在 ChatGPT 基础上进一步优化和扩展的最新模型，Dimitris 教授只通过两段提示，GPT-4 就给出了与 DeepMind 发表在 Nature 的 AlphaDev（Mankowitz 等，2023）发表的排序算法一致的结果。GPT-4 的惊人表现引发了全球人工智能（artificial intelligence，AI）领域的强烈反响，也促使众多学者投入到 LLM 的相关研究中。

LLM 的发展历史离不开 Google 和 OpenAI 的贡献。作为 AI 领域的引领者，Google 与 OpenAI 之间展开了全球范围内的激烈竞争。二者都拥有深厚的 AI 研究基础以及海量的数据资源，使得他们在该领域具有压倒性的优势。Google 最早提出了 Transformer 结构（Vaswani 等，2017）作为 LLM 的基础构成单元。随后，LLM 逐渐往 encode-only、encode-decode 和 decode-only 等 3 个方向演化。Google 公司聚焦于 encode-only 与 encode-decode 结构，并且在该阶段遥遥领先，提出的 BERT（bidirectional encoder representations from transformers）模型（Devlin 等，2019）几乎统治了所有自然语言处理领域。2022 年后，研究开始逐渐聚焦于 LLM 的生成能力（Yang 等，2023a）。由于采用的位置掩码方法不利于文本生成，encode-only 结构模型生成能力的缺陷开始被广泛讨论并大幅度放大（Tay 等，2023）。相比之下，decode-only 结构的 GPT 系列模型在生成能力方面表现出了 encode-only 结构模型所不具备的优势。由 OpenAI 发布的 decode-only 模型 ChatGPT，在自然语言处理领域展现了较高的水平，并且在多个推理任务上的性能超越了 BERT 模型（Zhong 等，2023）。decode-only 结构模型逐渐主导了 LLM 的发展。为了在 LLM 竞争赛道上追赶 OpenAI，将专注 LLM 的 Blue Shift 整体移入 Deepmind 后，Google 合并了两个旗下的顶级 AI 团队 Google Brain 和 DeepMind。同时，发布了对标 ChatGPT 和 GPT-4 的竞品 Bard 和 PaLM-2。ChatGPT 的出现，不仅在国外引发了 AI 领域的变革，也在国内掀起了一股热潮。国内互联网巨头和高校纷纷涌入这一赛道，发布各自的类 ChatGPT 模型：阿里的通义大模型、百度的文心大模型、商汤的 SenseChat 和面壁智能的 Luca 等。这些产品的问世依赖于其发布者背后庞大的私有数据库：阿里的电商物流数据、百度的内容检索数据、商汤科技公司的大规模商业数据以及知乎的问答数据。不同于国外致力于开发更为智能的语言模型，国内更加关注产业链的整合和布局，在芯片、模型底座、预训练 LLM、智能应用等多个领域同时发力，以实现基础研究与产品应用的平衡发展。

在各大公司进行激烈角逐的同时，开源模型作为一匹黑马悄然崛起。开源模型除了在质量方面存在急速缩小的劣势外，还具有训练快、私密性强、功能完善、维护成本低、技术迭代快等优势。而 LLaMA（Touvron 等，2023）正是其中的佼佼者。开源模型的发展受益于各种参数优化技术和调优指令集构建技术，这些技术在提高模型质量的同时降低了模型构建的成本。自从 2023 年 3 月初 Meta 公开 LLaMA 后，开源模型的崛起使得训练和实验的门槛大幅度降低，几乎所有机构和个体研究者都能够参与其中，甚至一人一台笔记本电脑即可部署模型。其结果就是高质量的开源模型吸引了大量的人才与机构，加速了 LLM 的产生和迭代，反过来也促进了开源 LLM 发布公司的技术和产品更新，且这种循环正不断重复。目前的现状是大型 AI 公司发布的产品处于领先地位，而众多开源模型在不断追赶，并快速缩小差距。这也预示着生成式人工智能领域开源开放是未来，协同创新是趋势，人人都可积极参与推进相关研究。因此生成式人工智正持续成为全球热议的话题。

本文对 LLM 进行了深度剖析，评估后续数月内相关的舆论影响力和应用；归纳 GPT 系列产品在性能方面的提升；简要分析相关技术原理；评估当前仍存在的局限，并讨论各个行业潜在的应用前景以及未来可能的研究方向。

1 舆论及应用现状

1.1 ChatGPT 的舆论

ChatGPT 作为一种基于对话式互动的生成式人

工智能产品，自发布短短两月后，活跃用户数已达1亿，访问量持续飙升。据统计，其2023年4月份全球访问量达17.6亿次，占谷歌的2%，占百度的60%左右，已超过DuckDuckGo等其他国际搜索引擎。

相关技术的持续发展也引发了搜索引擎革命。据SimilarWeb统计，微软旗下必应在2023年2月份宣布集成ChatGPT能力后，访问量增涨9%，而同时谷歌下降了3.1%。而在移动端，据data.ai统计，必应APP下载量达到了之前的8倍，而谷歌APP下载量下降2%。DeepMind联合创始人Mustafa Suleyman表示，互联网即将发生根本性变革，传统搜索引擎将在10年内消失，对话式互动产品将引发新一轮技术浪潮。

资本热潮开始不断涌入AI相关领域：相比于1月份，2023年4月份全球范围内资本对于AI的投入次数大约翻了3.3倍、发布的产品数量大约翻了6.5倍；OpenAI于4月底完成新一轮约3亿美元的融资，其市值介于270亿～290亿美元之间；美国著名证券公司Wedbush发布的一份分析报告显示，随着微软将ChatGPT和生成式AI功能整合在产品矩阵中，2023年其市值预计将飙升3 000亿美元。

此外，生成式人工智能相关技术及产品的风靡与推广，国内外政府对其也给予了不同程度的重视和支持。美国政府对于生成式人工智能相关产品和技术采取了较为宽松的态度，主要关注其伦理道德方面的问题，以确保该技术的发展符合国家法律和伦理标准，并保证美国在全球科技竞争中处于领先地位。日本政府致力于推动生成式人工智能的发展和普及，横须贺市则是率先实践的地区，宣布自2023年5月20日起采用ChatGPT作为公务辅助工具，涉及会议纪要、政策规划等方面。相比之下，欧盟国家的态度则更为保守和谨慎，关切ChatGPT所带来的数据隐私问题。欧洲数据保护委员会认为ChatGPT利用互联网信息与用户个人数据进行训练和迭代不符合条例。2023年5月11日，欧洲议会的两个委员会通过立法，明确禁止“对人类安全构成不可容忍风险的人工智能系统”。中国政府高度重视AI的发展，早在2017年颁布实施了发展规划，部署重大项目，建设开放平台，扩展应用场景，为国产LLM的发展提供有利条件。同时，也加强了对生成式人工智能的监管，于2023年4月发布国内首份专门针对AIGC的监管文件《生成式人工智能服务管理办法（征求意见稿）》，提出“利用生成式人工智能生成的内容应当真实准确”等多项要求。

1.2 类ChatGPT模型在不同领域的应用

经过数月的迭代和发展，LLM已经不再是OpenAI的ChatGPT一家独大。目前已经出现了在通用LLM的基础上，针对不同领域构建更加具体的模型：

1）教育领域。目前运用于教育领域的模型有讯飞星火、MathGPT等。该类模型通常由通用LLM经过相关教学知识的训练微调，可以帮助学生和老师提高学习与教学的效率和质量，丰富教育内容和形式，拓展教育场景和对象，为教育领域带来了新的可能性和机遇。

2）医疗领域。目前运用于医疗领域相关的模型有SurgicalGPT（Seenivasan等，2023）、ChatCAD（Wang等，2023a）和Med-PaLM（Singhal等，2023）等。该类模型通常经过医疗领域知识微调后形成专业的医学LLM。它们能够实现手术问答、辅助诊断、个性化治疗方案设计以及药物推荐等功能。

3）金融领域。目前运用于金融领域相关的LLM有轩辕大模型、BloombergGPT（Wu等，2023a）等。该领域的LLM需要具备股票、基金和保险等复杂知识，能够有效提高从业人员的专业水平和服务能力，同时大幅度降低运营成本。

4）法律领域。目前运用于法律领域的LLM有LawGPT、ChatLaw等。该领域的类ChatGPT需要了解专业的法律词汇，具备理解法律语义的能力。它们能够成为从业者的智能助理，帮助撰写法律文件、法律文件分析、查询案例和法律条款。

5）编码领域。目前运用于协助编码的LLM有PromptAppGPT、HuggingGPT（Shen等，2023）等。该领域的类ChatGPT模型需要具备理解不同类型的编程语言的能力和更加强大的逻辑推理能力。它们能够替程序员阅读或编写代码，并添加详细的注释。

6）论文写作领域。目前运用于该领域的LLM有ChatGPT Academic。该领域的类ChatGPT模型需要具有更强大的语言理解与写作能力以及更加专业的学术知识，能够协助用户润色文章、快速阅读和摘要生成等。值得注意的是，尽管大多数经过LLM本身就具有较强的阅读和写作能力，但它们并不能满足专业学术写作任务的要求。

应用于不同领域LLMs的项目链接如表1所示。除此之外，经过专业知识微调训练的LLM还可以应用于诸多科研领域，例如，物理、化学、哲学以及计算

机领域。除了帮助文献资料查阅和总结，撰写学术性邮件，它们有时还能给予科研人员创新的灵感或者参考意见。可以看出，经过数月的技术沉淀，ChatGPT 相关技术已经从各领域中的新鲜事物进化到能够初步走入部分领域并且协助工作的程度。

表 1　不同领域 LLMs 及其项目/论文链接

Table 1　Projects or papers links of LLMs in different fields

模型	项目/论文链接
讯飞星火	https://xinghuo.xfyun.cn/
MathGPT	https://www.mathgpt.com/
SurgicalGPT	https://arxiv.org/pdf/2304.09974.pdf
ChatCAD	https://arxiv.org/pdf/2302.07257.pdf
Med-PaLM	https://arxiv.org/pdf/2305.09617.pdf
轩辕大模型	https://github.com/Duxiaoman-DI/XuanYuan
BloombergGPT	https://arxiv.org/pdf/2303.17564.pdf
LawGPT	https://github.com/pengxiao-song/LaWGPT
ChatLaw	https://github.com/PKU-YuanGroup/ChatLaw
PromptAppGPT	https://github.com/mleoking/ PromptAppGPT
HuggingGPT	https://arxiv.org/pdf/2303.17580.pdf
ChatGPT Academic	https://github.com/binary-husky/gpt_academic

1.3　GPT 系列模型的提升与扩展

相比于先前的 GPT 版本，GPT-4 在以下 7 个方面的性能有了大幅度的提升：

1）专业知识。GPT-4 拥有更丰富的专业知识，这使得它能够在各类考核中更有优势。例如，在法律执业资格考试中，GPT-4 的成绩位于前 10% 的水平，而 GPT-3. 5 则仅位于前 90%。在其他的专业考试中，GPT-4 可以达到与人类相当的水平，但 GPT-3. 5 则表现不佳。

2）多模态能力。GPT-4 拥有初步的多模态能力。虽然 GPT-4 和 GPT-3. 5 在训练过程中都未涉及图像，但 GPT-4 能够较为准确地掌握了一些基本图像的概念。作为一个多模态模型，GPT-4 可以接收图像和文本作为输入，并输出文本。这使得 GPT-4 可以应对更复杂且含有视觉信息的任务，如图像描述、图像问答、图像生成文本等。

3）推理能力。GPT-4 具有更强的推理能力。在多数的推理任务中，如演绎、归纳、溯因、类比和因果等，GPT-4 相较于 GPT-3. 5 都展现出了显著优势，各项指标均有提高。GPT-4 通过推理得出的答案更加精确和严密。

4）编程能力。GPT-4 的编程能力有了显著提升。GPT-4 和 GPT-3. 5 在代码生成能力上相对于其他 LLM 具有巨大的优势，而 GPT-4 的代码能力又比 GPT-3. 5 有大幅提高。同时，GPT-4 还具有更高的稳定性。

5）可信度。GPT-4 能够生成更加符合事实的准确陈述。相比于 GPT3. 5，GPT-4 减少了在回答问题时胡言乱语的可能性。此外，GPT-4 能够更加注重对话中的细节逻辑，基于对话中的潜藏逻辑，生成更合理可信的回答。

6）安全性。GPT-4 在安全性上有了更大提升。在一些敏感和不允许的 prompts 上，对一些不该回答的内容也能更好地检测出来。GPT-4 在敏感问题上的不正确回答比例相比 GPT-3. 5 的不正确回答比例明显减少。

7）迁移能力。GPT-4 具有更强的迁移能力。在仅使用 few-shot 的情况下，GPT-4 在多数的 NLP 任务评估中超越了现有 SOTA，而 GPT-3. 5 则只能接近现有 SOTA。同时 GPT-4 在多个评测基准中体现出了更全面的能力。

然而，值得注意的是，尽管 GPT-4 在上述各方面

的性能都有了显著提升，但是，它仍然不是完美的。例如仍然存在产生错觉、社会偏见和对抗性提示等问题。因此，在使用模型的过程中，仍需要进一步的验证与核对。

2 技术原理简介及可扩展性

2.1 类ChatGPT模型构建

GPT-4的卓越表现和优秀的生成能力迅速引起了社会关注，各大企业和学术机构纷纷投入了大量的资源开展了相关研究。

如图1所示，要获得一个类ChatGPT模型，首先需要在大量无监督的数据集上进行预训练，得到一个基础模型。然后，通过指令微调（instruction fine-tuning，IFT）、基于人类反馈的强化学习（reinforcement learning from human feedback，RLHF）和思维链（chain-of-thought，CoT）等技术（刘禹良 等，2023），训练出一个助理模型。需要说明的是，虽然基础模型可以回答问题，但它所给出的回答并不可靠，因此需要通过监督微调来优化生成回复和文本理解的能力。因此，市面上发布的类ChatGPT模型大多属于助理模型。

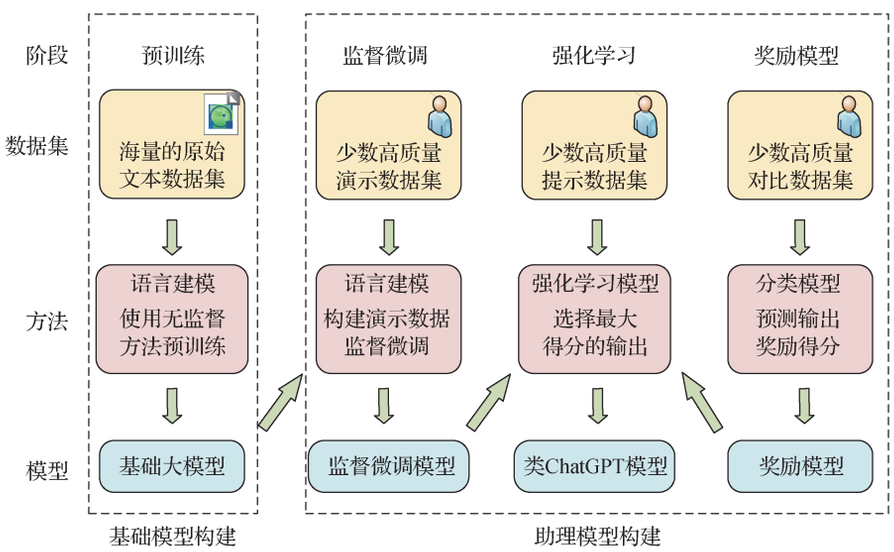

图1 类ChatGPT模型构建流程

Fig. 1 Construction of LLMs similar to chatGPT

国内外类ChatGPT模型的现状如表2所示，OpenAI和Google分别采用了不同的策略来提升GPT系列模型的性能。OpenAI通过增加LLM参数量和预训练数据量来提升GPT系列模型的知识覆盖面和问题解决能力，Google则更加关注模型的结构效率和迁移能力，使得PaLM2在推理速度、参数数量和成本方面具有优势。目前，这两家公司均未开源。然而，从长期来看，各类开源模型凭借其免费、高质量、易迁移和低维护等特点，在与闭源LLM的竞争中展现出一定的优势。

高质量的中文数据集与中文预训练LLM也是国内学者关注的重要问题。现有高性能的开源LLM都以英文为主，这样训练出的LLM具有英文思维的倾向。当这些模型应用于非英语语言时，模型内部可能需要先将输入转换成英语才能进行理解和生成任务，将降低模型的理解和生成能力。单纯将英文模型应用于其他语言时，无法发挥其最佳性能。因此，本文认为，针对其他语言特别是中文模型的研究具有重要意义。

2.2 羊驼家族

LLaMA是Meta于2023年2月发布的模型集合，包含7 B、13 B、33 B和65 B共4个版本。其中LLaMA-13 B在多数数据集上的表现超越了GPT-3并且使用开源语料作为训练语料。而羊驼家族是指一些基于LLaMA模型结合2.1节中涉及方法构建的模型，以下针对Alpace、Vicuna、Koala和Baize 4个羊驼家族成员进行简要介绍。

1）Alpaca：Alpaca由斯坦福大学于2023年3月发布。该模型的训练过程大体上分为基于Self-instruct（Wang等，2023b）方法自动构建调优数据集

表 2　类 ChatGPT 大模型相关工作

Table 2　Related work of LLMs similar to ChatGPT

模型	发布者	发布时间	参数量/亿
ChatGPT	OpenAI	2022 年 11 月	1 750
MOSS	复旦大学	2023 年 2 月	160
GPT-4	OpenAI	2023 年 3 月	–
文心一言	百度	2023 年 3 月	2 600
通义千问	阿里	2023 年 4 月	–
SenseChat	商汤科技	2023 年 4 月	–
PaLM2	Google	2023 年 5 月	3 400
LLaMA	Meta	2023 年 5 月	650
CPM-Bee	面壁智能	2023 年 5 月	100
讯飞星火	科大讯飞	2023 年 6 月	–
Claude2	Google	2023 年 6 月	–
ChatGLM2	清华大学	2023 年 6 月	60
Aquila	智源	2023 年 6 月	70
LLaMA2	Meta	2023 年 7 月	700
FreeWilly	Stability AI	2023 年 7 月	700
BuboGPT	字节跳动	2023 年 8 月	–
LISA	CUHK	2023 年 8 月	130

注：“–”代表未提供数据。

并基于构建的调优数据集监督微调 LLaMA。它的优势在于其极低的微调成本以及极少的资源消耗。更重要的是，作为羊驼家族早期成员，它为开源 LLM 研究提供了一个低门槛的平台，吸引了更多的研究者参与其中。

2）Vicuna：Vicuna 由伯克利大学、卡内基梅隆大学等机构的研究团队于 2023 年 4 月联合发布。Vicuna 调优数据集从 ShareGPT 收集并且筛选得来。此外，模型在 Alpaca 的基础上，改进了训练损失函数以适应多轮对话场景，增加了最大上下文长度以提升长文本理解能力，以及利用 SkyPilot 服务部署了具有自动恢复功能的 Spot 实例以进一步降低计算成本。相较于 Alpaca，Vicuna 的性能有了显著的性能提升，并且更加接近 ChatGPT 模型的水平。

3）Koala：Koala 由伯克利人工智能研究院于 2023 年 4 月发布。他们专注于构建小规模高质量的数据集，其调优数据集来源于经过知识蒸馏的公开对话数据集以及一系列开源数据集。Koala 的意义在于强调了高质量数据集对于对话模型性能的影响，甚至能够在一定程度上弥补开源模型参数规模小的缺陷。

4）Baize：Baize 由加州大学洛杉矶分校、中山大学、微软于 2023 年 4 月联合发布。他们提出一种名为 Self-Chat 的 ChatGPT 对话数据自动收集的方法，批量生成高质量多轮对话数据集用于调优。同时，在训练阶段应用了低秩适配（low-rank adaptation，LoRA）方法（Hu 等，2022）进一步降低了微调成本。

羊驼家族成员们的训练流程大体相当，其区别主要在于调优数据集的构成和参数优化方法。因此以下给出部分常用的参数优化方法以及调优数据集构成方法简介。

1）参数优化的 LoRA 技术。LoRA 技术是由微软在 2021 年 10 月提出，旨在加速 LLM 的训练并降低其微调的成本。LoRA 对于初始的预训练权重矩阵 $\boldsymbol{W} \in \mathbb{R}^{d\times d}$ 引入两个秩分解矩阵 $\boldsymbol{A} \in \mathbb{R}^{r\times d}$ 和 $\boldsymbol{B} \in \mathbb{R}^{d\times r}$ 替代现有权重进行微调，其中 d 为预训练权重的输出维度，而内在秩 r 远小于 d，其值通常根据任务需求预先设定。Hu 等人（2022）针对 Transformer 中的权重矩阵，提出了两种不同的低秩适配方案，分别适用于调整两种类型和一种类型的注意力机制的情况，其中 $r = 4$ 和 $r = 8$ 分别为最优选择。参数更新过程可视为 $\boldsymbol{W} + \Delta\boldsymbol{W} = \boldsymbol{W} + \boldsymbol{BA}$，其中，$\boldsymbol{A}$ 和 $\boldsymbol{B}$ 分别使用随机高斯和零初始化。在训练过程中 $\boldsymbol{W}$ 保持不变。LoRA 将原本需要微调的权重矩阵 $\boldsymbol{W}$ 转换成了 $\boldsymbol{A}$ 和 $\boldsymbol{B}$，显著减少了参数量。同时通过更换不同的重参数化组合，使得该技术能够灵活地在不同下游任务中应用。目前除了 Baize，LoRA 技术已应用于 Alpace（https://github.com/tloen/alpaca-lora）和 Vicuna（https://github. com/jackaduma/Vicuna-LoRA-RLHF-PyTorch）模型。

2）调优指令数据集构造。目前主流的调优指令数据集构造方法除了人工编写指令外，还有 Self-instruct 和 Automatic prompt engineer（Zhou 等，2023）两种自动化方法。Self-instruct 方法主要包括指令生成、指令分类、实例生成和数据过滤 4 个步骤，利用人工编写的种子指令和 LLM 的生成能力、上下文理解能力以及指令理解能力，扩充指令数量和类型，生成完整的数据实例，并过滤掉低质量的数据。Automatic prompt engineer 方法主要包括指令候选集生成、指令评估和指令变体生成 3 个步骤，利用 LLM 和现有的输入输出对，逆向生成和选择最优的 prompt

指令，并利用LLM生成相似变体增加指令的多样性。二者都是基于现有的LLM生成高质量调优指令数据集，其区别在于，Self-instruct方法是基于多样的种子指令，利用LLM的多任务学习能力，以生成更丰富的指令样本。而Automatic prompt engineer方法是基于输入输出对，利用LLM的逆向推理能力，以生成和挑选最优的prompt指令。

羊驼家族作为开源LLM的代表，其快速发展展示了开源LLM的可行性和潜力。而其中涉及的参数优化技术和调优数据集构造技术有效地降低了LLM的训练成本和计算资源消耗，同时提升了模型的多样性和泛化能力。因此，羊驼家族及相关技术对于LLM产品普及与技术迭代具有重要意义。

2.3 多模态技术

多模态技术是生成式人工智能领域中的一种关键技术，能够执行跨模态生成任务，例如根据文本生成图像、根据图像生成文本等。随着LLM的发展，多模态技术也取得了重大突破。以GPT-4为例，它可以根据图像生成不同类型的文本，如描述、解释、总结和问答等，也可以根据文本生成或编辑图像，完成创意和技术写作任务。但是目前GPT-4模型和技术细节还未被公布，因此许多学者尝试构建一个类似GPT-4的多模态大模型。本文总结了目前主流的图像—文本的多模态技术实现方法，大体上可划分为3类：

1）训练中间层以对齐视觉模块和语言模型。该类方法首先预训练视觉模块，将这些视觉模块与LLM冻结，然后在视觉模块与LLM之间插入可训练的中间层，构建多模态模型。接着在大规模的图像—文本对数据集上对多模态模型进行微调，更新中间层的可训练参数，实现视觉模块与LLM的对齐，完成跨模态任务。该类方法的大体流程如图2所示，Flamingo（Alayrac等，2022）、BLIP2（Li等，2023）和ImageBind（Girdhar等，2023）都使用该类方法构成多模态大模型。

2）多模态指令微调。该类方法在模态对齐的基础上，进一步进行指令微调训练，用多模态指令数据集对视觉编码器等额外结构进行适配，使其能够与LLM协同工作，从而达到与GPT-4类似的多模态能力。这类模型的训练通常包含两个阶段：第1阶段利用大规模的图像-文本对数据对模型进行预训练，学习视觉和语言模态间的对齐；第2个阶段通过

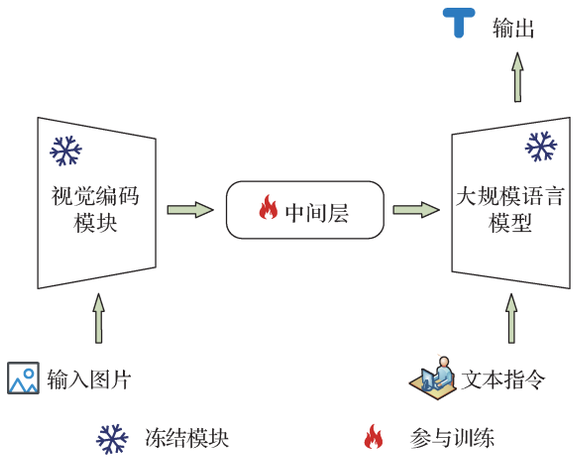

图2 训练中间层对齐视觉模块和大模型

Fig. 2 Training adaptation layers to align visual modules and language models

多模态指令数据集对模型进行微调，让模型获得多模态指令跟随能力。该类方法的大体流程如图3所示，MiniGPT4（Zhu等，2023）、LLaVA（Liu等，2023a）和mPLUG-Owl（Ye等，2023）都使用该类方法构成多模态大模型。

3）LLM作为理解中枢。该类方法利用LLM实现多模态处理，它将多模态数据转化为文本数据输入LLM，作为与用户交流的理解中枢，LLM根据用户需求调用其它视觉基础模型，从而达到跨模态输入输出和完成多种任务的效果。该类方法的大体流程如图4所示。目前Visual ChatGPT（Wu等，2023b），MM-REACT（Yang等，2023b）采用这种方式构成多模态大模型。

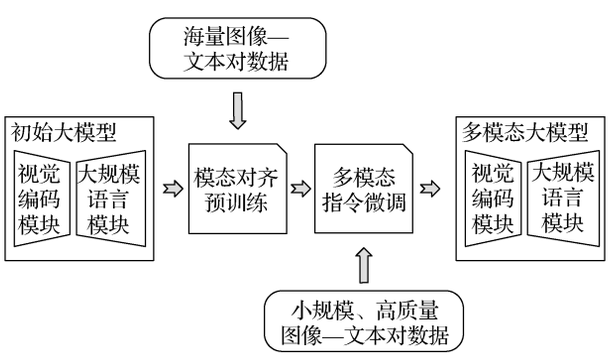

图3 多模态指令微调

Fig. 3 Multimodal instruction fine-tuning

目前，多模态大模型技术尚处于初级阶段，面临着训练数据不足、多模态信息表示不一致和对齐算法不精确等挑战。然而，多模态大模型具有融合多种感知维度的信息的能力，更接近人类认知世界的

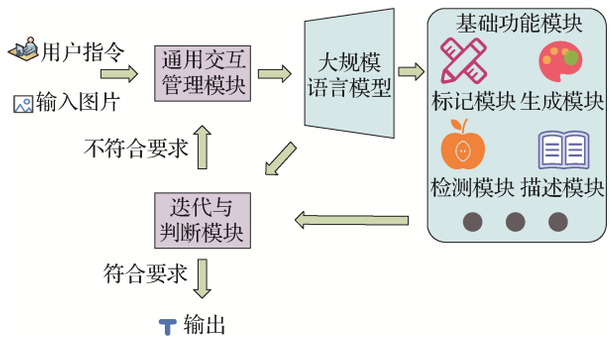

图 4　LLM 做理解中枢方法

Fig. 4　LLM serves as the center of understanding

方式，是大模型未来的重要发展方向之一。

2.4　LLM 评估标准

为了有效衡量和优化 LLM 的性能和泛化能力，并揭示其优势和局限，建立合理的 LLM 评价基准具有重要意义。现阶段，主流的 LLM 评估方法可划分为 3 类：

1）人工评估。基于人工的评估方法通常需要邀请大量的志愿者或相关领域专家对 LLM 的生成进行主观评估和打分。以专业领域知识评估为例（Guo 等，2023），首先需要收集不同领域专家根据该领域专业问题给出的答案作为参考，然后根据已有知识与 LLM 生成的输出，主观评估 LLM 专业领域知识的掌握程度。此外，人工评估还可以运用于评估生成内容与人类意愿高度相关的指标，例如语义一致性、逻辑合理性和文体风格等方面。

人工评估方法依赖于人类评估者对生成内容的主观判断，可以更好地反映生成内容的质量以及 LLM 在不同专业领域的能力。此外，它还可以灵活适应不同的任务场景。但是，基于人工的评估方法需要耗费大量的时间、金钱和人力。因此，基于人工的评估方法不利于当前 LLM 研究发展的快速迭代。

2）自动评估。自动评估方法是通过使用标准化的数据集和指标来对 LLM 进行自动化评估。该类评估方法所采用的综合评测基准通常涵盖主流 LLM 评估任务，并且可以快速对比不同 LLM 在相同任务下的性能差距。以下介绍两个具有代表性且在当前阶段流行的综合评估基准。

首先是由 Berkeley 研究者发布的 MMLU（massive multitask language understanding）评测基准（Hendrycks 等，2021）。MMLU 包含 57 个任务，涵盖了数学、历史、计算机科学和法律等领域，通过零样本和少样本设置来评估模型蕴含的知识。其次是 AI2 提出的 ARC（AI2 reasoning challenge）评测基准（Clark 等，2018）。ARC 包含 7 787 个来自不同科学领域的考试问题，根据难易程度，可以划分为 2 590 个问题组成的 ARC-Challenge 和 5 197 个问题的 ARC-Easy。这些问题用以评估 LLM 在多步推理、语言匹配等多方面的高级能力。在这两个评测基准中，GPT-4 取得了最好的成绩，明显超过其他 LLM。

自动评估方法可以快速地对比不同 LLM 在相同条件下的性能差异，也可以提供一些可量化和可解释的结果。同时，区别于人工评估方法，该类方法不受主观因素影响，可重复性较强。因此，自动评估方法是 3 类方法中最广泛使用的一种。但是，自动评估方法也有一定的缺陷，例如机械化、忽视人类偏好、数据集质量与覆盖度无法保证等。

3）其他 LLM 评估。除了人工评估和自动评估，利用 LLM 本身作为评估器，来比较不同 LLM 的输出，并给出相对的优劣判断也是一种思路。该类方法通常设定一系列用来评估 LLM 综合性能的开放式问题集合，用以得到 LLM 的输出，并且采用现有先进的 LLM 作为评估者，对 LLM 的输出进行标注输出对比或打分以评估 LLM 的综合性能。以下介绍 3 种最新的基于其他 LLM 的评估方法 MMBench（Liu 等，2023）、AlpacaEval（Dubois 等，2023）和 MT-Bench（Zheng 等，2023）。

MMBench 是上海人工智能实验室提出的一种客观评估大型视觉语言模型不同能力的评估基准。它从模型的感知和推理能力出发，构建了 3 级能力维度，并采用一种循环评估策略 CircularEval，以提高评估过程的稳定性。对于某一具体的评估题目，首先，从多模态大模型的输出中提取与选项匹配的内容。若匹配失败，则用 ChatGPT 预测和选项，生成选项标签。如果仍然无法提取选择，则用随机选择来标记预测，并添加评论信息。最后使用 ChatGPT 作为自动评估器，评估多模态大模型的性能。AlpacaEval 由斯坦福大学的研究人员发布，是一种基于 AlpacaFarm 数据集来测试 LLM 遵循一般用户指令的能力的评估方法。具体来说，研究人员以强大的 LLM（如 GPT-4、Claude 或 ChatGPT）为自动评估器，以 GPT-3.5 为基准，将目标模型与 GPT-3.5 的回复进行比较，计算获胜率。而 MT-Bench 是一种基于多轮对话来评估 LLM 聊天机器人能力和用户偏好的

评估方法。它是一个由80个多轮问题组成的基准测试集，旨在评估聊天机器人的对话和指令跟随能力。同样使用强大的LLM（如GPT-4、Claude或ChatGPT）作为评估器，将不同的聊天机器人的回复进行比较，并给出优劣判断。

基于其他LLM评估方法具有快速、廉价、可复现且与人类偏好高度一致的特点，可以作为开发和测试LLM的有用工具。但是，目前该类方法的发展时间较短，技术不完善，仍具有较多的局限性，例如无法用于评估LLM的高阶能力、评估结果存在偏差以及缺乏安全评估等。

在上述3种方法中，人工评估方法最符合人类意愿，数据集自动评估方法应用最广泛，而利用其他LLM评估方法具有最高的新颖性和可拓展性。但是，现阶段尚未存在完善的LLM综合能力评估方法，因此，现阶段应该多种方法互补使用以全面地评估LLM的性能和质量。

Chatbot Arena基准平台（https://chat.lmsys.org/）上发布的部分LLM排行如表3所示。可以看出，GPT-4在多数指标上都大幅领先其他模型，大量开源模型在生成质量上仍存在较大差距。

表3　Chatbot Arena平台上部分LLM评估结果

Table 3　LLM evaluation results on the Chatbot Arena

模型	Elo rating	MMLU	Score
GPT-4	1 206	86.4	8.99
Claude2	1 135	78.5	8.06
GPT-3.5	1 122	70.0	7.94
Guanaco	1 038	57.6	6.53
Vicuna	1 096	59.2	7.12
WizardLM	1 040	50	6.76
PaLM	1 015	-	6.4
Llama2	987	53.6	6.65
Koala	983	44.7	5.35
MPT-Chat	947	32	5.42
Alpaca	923	48.1	4.53
ChatGLM	900	36.1	4.5
FastChat	892	47.7	3.04
StableLM	863	24.4	2.75
LLaMA	817	47	2.61
baize	-	48.9	5.75

注："-"表示平台未提供数据。Elo rating是平台根据用户对LLM投票数据计算得出的得分数据；MMLU代表LLM在MMLU评测基准上的得分；Score表示LLM在MT-Bench评测基准中的得分。

3　局限、应用前景及展望未来

3.1　存在的局限

尽管类ChatGPT模型经过数月的迭代和完善，已经初步融入了部分领域以及人们的日常生活，但目前市面上的产品和相关技术仍然存在一些问题，以下列出一些局限性进行详细说明与成因分析：

1）互联网上高质量、大规模、经过清洗的公开数据集和开源、结构高效的预训练LLM仍然不足。这是因为收集和清洗数据集的过程非常烦琐和复杂，且预训练LLM的训练需要高性能设备和大量优质数据集。

2）针对同一问题，重复输入会导致不一致的回答。有时也会出现稍微改变一些词语，模型的回答就会从无法回答转变为正确回答的情况。这是因为训练时得到的LLM缺乏泛化能力，输入格式不规范且噪声多。

3）模型虽然能够回答一些通用性问题，但是在涉及一些专业领域或者具体情境的问题时，就会显得力不从心。这是因为LLM训练数据并没有覆盖所有领域和场景，而且模型本身也缺乏足够的知识库和推理能力来处理复杂的问题。

4）LLM由于缺乏常识知识，输入缺乏事实依据和事实验证，因此在类ChatGPT模型在大规模运用时容易产生幻觉，生成错误答案，并出现推理错误等问题。

5）类ChatGPT模型在生成文本时，存在输出很难被人类理解和解释且很难被人类监督并纠正的问题，这是因为模型基于深度学习，生成文本时并不遵循任何明确的规则或逻辑，而是根据概率分布来选择最可能的词汇。

6）类ChatGPT模型依赖于基础模型，但基础模型为了产生能力"涌现"的现象，需要庞大的参数量来支撑其存储的知识规模。因此，相关产品的部署和运行不仅需要高昂的硬件成本和资源消耗，而且难以适应移动设备和边缘计算等场景。

7）类ChatGPT模型使用奖励机制作为训练类

ChatGPT模型的主要方法，并不受法律和道德准则的约束。因此可能会被恶意利用，造成严重的安全隐患或者法律风险。此外，与用户交互时，能够记住与会话相关的项目以及用户输入、缓存、日志等隐私信息。以及可能存在利用模型逻辑强大的对话能力与丰富的知识进行诈骗或作弊的情况。

除了上述局限外，目前，作为构建基石的基础模型仍存在一些原理问题尚未得到突破，例如无法保持自我一致性、无法处理比token更小的单元以及多模态领域表示困难等问题。

3.2　应用前景

类ChatGPT模型在多个领域具有非常强大的潜在应用价值，各大公司都在积极布局该类模型。以下列出部分类ChatGPT模型潜在的应用前景。

1）赋能内容创作。基于视觉语言模型的内容创作已经得到了广泛应用，如文字或图片内容补全。利用多模态模型更强大的多模态和推理能力，可以实现大型内容创作，如直接创作剧情严密的影片剧本。

2）革新交互体验。借助LLM的语言理解能力，人机交互体验有望发生革命性进步，机器可以理解人类的指令与需求，并生成模拟内容予以辅助。

3）诞生“数字生命”。将LLM引入虚拟世界中，实现了智能体全场景的终身学习，具备快速学习、反馈环境、探索世界的能力。相信在不久的未来，智能体有望对多模态数据进行感知与学习，距离通用人工智能更进一步，从游戏模拟跨向现实应用。

4）智能家居与家庭助理。利用LLM建立各种智能家居设备的中枢管理，提供更加智能化、更懂人类需求的智能家居解决方案，实现根据用户指令和环境自动制定规划控制，并提供处理家庭日常事务、排疑解惑以及脑洞聊天等助理服务。

5）自动驾驶与智能汽车交互。类ChatGPT模型能够给自动驾驶带来语音交互提升，成为提升智能座舱语音交互质量的重要工具，并启发自动驾驶底层算法跃迁升级。此外，生成式AI为自动驾驶模型训练提供高质量合成数据，破解自动驾驶数据和测试难题。

3.3　未来展望

现阶段，LLM仍存在许多问题和挑战，LLM未来在数据方面、技术方面和应用等方面仍有较大的发展空间。

1）数据方面。数据方面的研究主要关注LLM的输入和输出，包括数据集的构建和专业知识的嵌入等方面。未来的研究方向可以从以下两个角度展开：

（1）训练数据集构建，这是影响LLM产品成功与否的关键因素，对数据集的质量和规模有较高的要求。为了提高数据集的可靠性和多样性，建立统一范式的人工数据集构造方法和各类高质量数据集生成算法设计，是未来重要且基础的研究方向之一；

（2）在LLM中嵌入特定领域的具体知识，旨在应对LLM中蕴含知识无法被完全利用与LLM专业领域知识不足的矛盾。可以收集已有特定领域（例如医疗、教育、法律等）的知识，构成特定领域的专业数据集并融合到LLM中，使其在该领域表现更好，以此打造针对某领域或某群体的专用LLM。

2）技术方面。技术方面主要关注LLM的内部结构和功能，涵盖了模型的搭建、扩展、革新和瘦身4个方面：

（1）完整搭建并训练LLM。这是最基础且核心的研究方向之一，需要面对如何高效地训练、如何充分利用现有语料、如何构建多语言的LLM等多种挑战；

（2）扩展LLM的多模态能力。目前大部分成熟的LLM多模态功能仍存在诸多缺陷，其中的多模态技术面临的挑战大体上可概括为模态表示、跨模态对齐、跨模特模型推理、跨模态信息生成、跨模态知识迁移和跨模态模型量化分析6个方面（Liang等，2022）。合理解决这些问题和进一步完善多模态技术是LLM实现对世界深入认知与转变为通用人工智能的关键步骤；

（3）对核心原理进行创新改进。这是LLM技术迭代更新与发展过程中的重要研究方向之一，旨在探究如何在现有LLM中使用的上下文学习、模型自适应选择或级联等原理，以及从LLM到ChatGPT的演化过程中采用的技术，例如RLHF、COT、IFT等技术的基础上进行创新改进，提高模型的性能和效率。目前已有对LLM核心原理创新的工作包括ALMoST（Kim等，2023）、TOT（Yao等，2023）、SuperICL（Xu等，2023）和FedIT（Zhang等，2023）等；

（4）LLM裁剪瘦身。这是LLM进一步推广与普及的核心问题之一，旨在优化LLM过于庞大、使用成本过高以及部署困难等问题。通常可以采用量

化、剪枝和蒸馏等方法进行模型瘦身与优化。此外，还可以结合融入特定领域知识或保留数据中重要信息，打造某个具体领域的轻量级专用模型。

3）应用方面。应用方面的研究关注LLM的实际效果和价值。未来的研究方向可从以下3个方面展开：

（1）安全性优化，解决LLM输出恶意内容、泄漏隐私数据等安全问题。这些问题在迭代迅速的开源模型中尤为严重，而目前业界对LLM安全性优化的方法缺乏统一的标准和框架。在不损害LLM性能和效率的基础上，保障其安全性，是LLM成为一项成熟、实用且面向大众的高科技产品的必要条件。

（2）建立评估体系，制定一个全面、公认的LLM评估方法，实现对LLM的生成、推理、知识储备等基本能力，以及对齐人类意愿、正确使用工具等高阶能力进行客观、有效的评估。然而，目前的评价方法都存在局限性和不足。因此，如何完善、融合现有的评价方法，是一个亟待解决的问题。此外，评估体系的建立还有助于开发纠错模型，用于LLM训练。

（3）发展LLM应用工程，进一步推广和普及LLM相关技术。当前，各种LLM的广泛应用已经是一个大趋势，但大多数缺乏提示工程相关知识的普通人无法充分利用市面上成熟的LLM产品。因此，如何高效利用这些产品更好地解决实际问题是一个新颖而实用的研究领域。

4 结 语

在过去的数月中，生成式人工智能相关技术及类ChatGPT模型经过了各大型AI公司以及大量科研人员的改进和优化，相比于最初的ChatGPT版本有了巨大的蜕变。目前，性能最优的GPT-4已展现出非常明显的初级通用人工智能的能力（Bubeck等，2023）。然而，现有的LLM仍然面临着诸多挑战和问题，例如专业知识缺乏、安全隐私风险和训练数据不透明等。为了突破这些局限，亟需在参数优化、数据集构建和多模态融合等方面进行技术创新，并建立统一、全面、便捷的评估基准来衡量LLM的性能和效果。另一方面，先进的LLM产品已经具备了提高人类办公效率的能力。因此，本文认为，对于大多数用户而言，学习并应用提示工程技巧是一种有效的方法，可以增强现有LLM的性能和灵活性，从而显著提高社会总体生产力。最后，必须承认的是，LLM领域相关技术的改进和创新离不开头部公司与开源社区的相互协同和促进，以及“开源开放是未来，协同创新是趋势”将成为主流。因此，为了迎接即将到来的生成式人工智能发展潮流，需要大量科研人员以有组织的科研模式推进原创性、引领性创新。

参考文献（References）

Alayrac J B, Donahue J, Luc P, Miech A, Barr I, Hasson Y, Lenc K, Mensch A, Millican K, Reynolds M, Ring R, Rutherford E, Cabi S, Han T D, Gong Z T, Samangooei S, Monteiro M, Menick J L, Borgeaud S, Brock A, Nematzadeh A, Sharifzadeh S, Bińkowski M, Barreira R, Vinyals O, Zisserman A and Simonyan K. 2022. Flamingo: a visual language model for few-shot learning//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: [s.n.]: 23716-23736

Bubeck S, Chandrasekaran V, Eldan R, Gehrke J, Horvitz E, Kamar E, Lee P, Lee Y T, Li Y Z, Lundberg S, Nori H, Palangi H, Ribeiro M T and Zhang Y. 2023. Sparks of artificial general intelligence: early experiments with GPT-4[EB/OL]. [2023-07-01]. https://arxiv.org/pdf/2303.12712.pdf

Clark P, Cowhey I, Etzioni O, Khot T, Sabharwal A, Schoenick C and Tafjord O. 2018. Think you have solved question answering? TryARC, theAI2 reasoning challenge[EB/OL]. [2023-07-01]. https://arxiv.org/pdf/1803.05457.pdf

Devlin J, Chang M W, Lee K and Toutanova K. 2019. BERT: pre-training of deep bidirectional transformers for language understanding//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics. Minneapolis, USA: Association for Computational Linguistics: 4171-4186 [DOI: 10.18653/v1/N19-1423]

Dubois Y, Li X C, Taori R, Zhang T Y, Gulrajani I, Ba J, Guestrin C, Liang P and Hashimoto T B. 2023. AlpacaFarm: a simulation framework for methods that learn from human feedback[EB/OL]. [2023-07-01]. https://arxiv.org/pdf/2305.14387.pdf

Girdhar R, El-Nouby A, Liu Z, Singh M, Alwala K V, Joulin A and Misra I. 2023. Imagebind: one embedding space to bind them all//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 15180-15190

Guo B Y, Zhang X, Wang Z Y, Jiang M Q, Nie J R, Ding Y X, Yue J W and Wu Y P. 2023. How close is ChatGPT to human experts? Comparison corpus, evaluation, and detection[EB/OL]. [2023-07-01]. https://arxiv.org/pdf/2301.07597.pdf

Hendrycks D, Burns C, Basart S, Zou A, Mazeika M, Song D and Steinhardt J. 2021. Measuring massive multitask language under-

standing//Proceedings of the 9th International Conference on Learning Representations. [s.l.]: OpenReview.net

Hu E J, Shen Y L, Wallis P, Allen-Zhu Z, Li Y Z, Wang S A, Wang L and Chen W Z. 2022. LoRA: low-rank adaptation of large language models//Proceedings of the Tenth International Conference on Learning Representations. [s.l.]: OpenReview.net

Kim S, Bae S, Shin J, Kang S, Kwak D, Yoo K M and Seo M. 2023. Aligning large language models through synthetic feedback [EB/OL]. [2023-07-01]. https://arxiv.org/pdf/2305.13735.pdf

Li J N, Li D X, Savarese S and Hoi S. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models[EB/OL]. [2023-07-01]. https://arxiv.org/pdf/2301.12597.pdf

Liang P P, Zadeh A and Morency L P. 2022. Foundations and trends in multimodal machine learning: principles, challenges, and open questions[EB/OL]. [2023-07-01]. https://arxiv.org/pdf/2209.03430.pdf

Liu H T, Li C Y, Wu Q Y and Lee Y J. 2023a. Visual instruction tuning [EB/OL]. [2023-07-01]. https://arxiv.org/pdf/2304.08485.pdf

Liu Y, Duan H D, Zhang Y H, Li B, Zhang S Y, Zhao W B, Yuan Y K, Wang J Q, He C H, Liu Z W, Chen K and Lin D H. 2023b. MMBench: Is Your Multi-modal Model an All-around Player?[EB/OL]. [2023-07-01]. https://arxiv.org/pdf/2307.06281.pdf

Liu Y, Li H and Bai X. 2023. A brief analysis of ChatGPT: historical evolution, current applications, and future prospects. Journal of Image and Graphics, 28(04):893-902 (刘禹良,李鸿亮,白翔等. 2023.浅析ChatGPT:历史沿革、应用现状及前景展望.中国图象图形学报, 28(04):893-902)

Mankowitz D J, Michi A, Zhernov A, Gelmi M, Selvi M, Paduraru C, Leurent E, Iqbal S, Lespiau J B, Ahern A, Köppe T, Millikin K, Gaffney S, Elster S, Broshear J, Gamble C, Milan K, Tung R, Hwang M, Cemgil T, Barekatain M, Li Y J, Mandhane A, Hubert T, Schrittwieser J, Hassabis D, Kohli P, Riedmiller M, Vinyals O and Silver D. 2023. Faster sorting algorithms discovered using deep reinforcement learning. Nature, 618(7964): 257-263 [DOI: 10.1038/s41586-023-06004-9]

Seenivasan L, Islam M, Kannan G and Ren H L. 2023. SurgicalGPT: end-to-end language-vision GPT for visual question answering in surgery[EB/OL]. [2023-07-01]. https://arxiv.org/pdf/2304.09974.pdf

Shen Y L, Song K T, Tan X, Li D S, Lu W M and Zhuang Y T. 2023. HuggingGPT: solving AI tasks with chatGPT and its friends in hugging face[EB/OL]. [2023-07-01]. https://arxiv.org/pdf/2303.17580.pdf

Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Hou L, Clark K, Pfohl S, Cole-Lewis H, Neal D, Schaekermann M, Wang A, Amin M, Lachgar S, Mansfield P, Prakash S, Green B, Dominowska E, Arcas B A Y, Tomasev N, Liu Y, Wong R, Semturs C, Mahdavi S S, Barral J, Webster D, Corrado G S, Matias Y, Azizi S, Karthikesalingam A and Natarajan V. 2023. Towards expert-level medical question answering with large language models [EB/OL]. [2023-07-01]. https://arxiv.org/pdf/2305.09617.pdf

Tay Y, Dehghani M, Tran V Q, Garcia X, Wei J, Wang X Z, Chung H W, Bahri D, Schuster T, Zheng H S, Zhou D, Houlsby N and Metzler D. 2023. UL2: unifying language learning paradigms//Proceedings of the Eleventh International Conference on Learning Representations. Kigali, Rwanda: OpenReview: 1-33

Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M A, Lacroix T, Rozière B, Goyal N, Hambro E, Azhar F, Rodriguez A, Joulin A, Grave E and Lample G. 2023. LLaMA: open and efficient foundation language models[EB/OL]. [2023-07-01]. https://arxiv.org/pdf/2302.13971.pdf

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł and Polosukhin I. 2017. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 6000-6010

Wang S, Zhao Z H, Ouyang X, Wang Q and Shen D G. 2023a. ChatCAD: interactive computer-aided diagnosis on medical image using large language models[EB/OL]. [2023-07-01]. https://arxiv.org/pdf/2302.07257.pdf

Wang Y Z, Kordi Y, Mishra S, Liu A, Smith N A, Khashabi D and Hajishirzi H. 2023b. Self-instruct: aligning language models with self-generated instructions//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. Toronto, Canada: ACL: 13484-13508 [DOI: 10.18653/v1/2023.acl-long.754]

Wu C F, Yin S M, Qi W Z, Wang X D, Tang Z C and Duan N. 2023b. Visual chatGPT: talking, drawing and editing with visual foundation models[EB/OL]. [2023-07-01]. https://arxiv.org/pdf/2303.04671.pdf

Wu S J, Irsoy O, Lu S, Dabravolski V, Dredze M, Gehrmann S, Kambadur P, Rosenberg D and Mann G. 2023a. BloombergGPT: a large language model for finance[EB/OL]. [2023-07-01]. https://arxiv.org/pdf/2303.17564.pdf

Xu C W, Xu Y C, Wang S H, Liu Y, Zhu C G and McAuley J. 2023. Small models are valuable plug-ins for large language models[EB/OL]. [2023-07-01]. https://arxiv.org/pdf/2305.08848.pdf

Yang J F, Jin H Y, Tang R X, Han X T, Feng Q Z, Jiang H M, Yin B and Hu X. 2023a. Harnessing the power of llms in practice: a survey on chatgpt and beyond[EB/OL]. [2023-07-01]. https://arxiv.org/pdf/2304.13712.pdf

Yang Z Y, Li L J, Wang J F, Lin K, Azarnasab E, Ahmed F, Liu Z C, Liu C, Zeng M and Wang L J. 2023b. MM-REACT: prompting ChatGPT for multimodal reasoning and action[EB/OL]. [2023-07-01]. https://arxiv.org/pdf/2303.11381.pdf

Yao S Y, Yu D, Zhao J, Shafran I, Griffiths T L, Cao Y and Narasimhan K. 2023. Tree of thoughts: deliberate problem solving with large language models[EB/OL]. [2023-07-01].

https://arxiv.org/pdf/2305.10601.pdf

Ye Q H, Xu H Y, Xu G H, Ye J B, Yan M, Zhou Y Y, Wang J Y, Hu A W, Shi P C, Shi Y Y, Li C L, Xu Y H, Chen H H, Tian J F, Qi Q, Zhang J and Huang F. 2023. mPLUG-Owl: modularization empowers large language models with multimodality [EB/OL]. [2023-07-01]. https://arxiv.org/pdf/2304.14178.pdf

Zhang J Y, Vahidian S, Kuo M, Li C Y, Zhang R Y, Wang G Y and Chen Y R. 2023. Towards building the federated GPT: federated instruction tuning[EB/OL]. [2023-07-01]. https://arxiv.org/pdf/2305.05644.pdf

Zheng L M, Chiang W L, Sheng Y, Zhuang S Y, Wu Z H, Zhuang Y H, Lin Z, Li Z H, Li D C, Xing E P, Zhang H, Gonzalez J E and Stoica I. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena[EB/OL]. [2023-07-01]. https://arxiv.org/pdf/2306.05685.pdf

Zhong Q H, Ding L, Liu J H, Du B and Tao D C. 2023. Can ChatGPT understand too? A comparative study on chatgpt and fine-tuned bert[EB/OL]. [2023-07-01]. https://arxiv.org/pdf/2302.10198.pdf

Zhou Y C, Muresanu A I, Han Z W, Paster K, Pitis S, Chan H and Ba J. 2023. Large language models are human-level prompt engineers// Proceedings of the Eleventh International Conference on Learning Representations. Kigali, Rwanda: OpenReview.net

Zhu D Y, Chen J, Shen X Q, Li X and Elhoseiny M. 2023. MiniGPT-4: enhancing vision-language understanding with advanced large language models[EB/OL]. [2023-07-01]. https://arxiv.org/pdf/2304.10592.pdf

作者简介

严昊，男，博士研究生，主要研究方向为自然语言处理。
E-mail：2020202120075@whu.edu.cn
刘禹良，通信作者，男，研究员，主要研究方向为计算机视觉。
E-mail：ylliu@hust.edu.cn
金连文，男，教授，主要研究方向为计算机视觉、文字识别。
E-mail：eelwjin@scut.edu.cn
白翔，男，教授，主要研究方向为人工智能。
E-mail：xbai@hust.edu.cn