

中图法分类号: TP183 文献标识码: A 文章编号: 1006-8961(2023)04-0893-10

论文引用格式: Liu Y L, Li H L, Bai X and Jin L W. 2023. A brief analysis of ChatGPT: historical evolution, current applications, and future prospects. Journal of Image and Graphics, 28(04):0893-0902(刘禹良, 李鸿亮, 白翔, 金连文. 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):0893-0902)[DOI:10.11834/jig.230110]

浅析 ChatGPT: 历史沿革、应用现状及前景展望

刘禹良¹, 李鸿亮², 白翔^{1*}, 金连文²

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: 近年来, 人工智能技术接连取得突破, 尤其是在强化学习、大规模语言模型和人工智能生成内容技术等方面, 正逐步成为各个行业的创新驱动力。OpenAI 于 2022 年 11 月 30 日发布的 ChatGPT 由于具有惊人的自然语言理解和生成能力, 引起全社会大范围的关注, 成为全球热议的话题, 并被广泛应用于各个行业。仅两个月后, ChatGPT 的月活跃用户数便达 1 亿, 成为史上用户数增长最快的消费者应用。鉴于 ChatGPT 目前造成的影响, 对其进行全面的分析较为必要。本文从历史沿革、应用现状和前景展望这 3 个角度对 ChatGPT 进行剖析, 探究其对社会的影响、技术的原理和挑战以及未来发展的可能性, 并从模型能力的角度简要介绍 GPT-4 相对于 ChatGPT 的改进。作为一个现象级技术产品, 从技术角度而言 ChatGPT 对相关领域具有里程碑式的重要意义, 从应用角度而言其可能会给人类社会带来巨大的影响。ChatGPT 有潜力成为计算机领域最伟大的成就之一。但就目前而言, ChatGPT 仍然存在一些局限, 尚未达到强人工智能的水平。在当前阶段, 研究人员需要对人工智能技术持有自信和谦虚学习的态度, 继续发展相关的技术研究和应用。

关键词: 人工智能(AI); 深度学习; 自然语言处理; 生成式人工智能技术(AIGC); ChatGPT

A brief analysis of ChatGPT: historical evolution, current applications, and future prospects

Liu Yuliang¹, Li Hongliang², Bai Xiang^{1*}, Jin Lianwen²

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: Artificial intelligence (AI) technology has been developing intensively, especially for such scenarios in relevance to its applications of 1) natural language processing, 2) computer vision, 3) recommendation systems, and 4) forecast analysis. AI technology has been challenging for human cognition over the past decade. In recent years, natural language processing techniques can be focused on more. ChatGPT, as a case of emerging generative AI technology, is launched in December of 2022. ChatGPT, as an advanced language model, is commonly used on the basis of its a) larger model sizes, b) advanced pre-training methods, c) faster computing resources, and d) more language processing tasks. This ChatGPT-related literature review is focused on its (1) public awareness and application status, (2) characteristics, (3) mechanisms, (4) scalability, (5) challenges and limitations, (6) future development and application prospects, and (7) improvements of GPT-4 relative to ChatGPT. Cognitive computing and AI-based ChatGPT can be as a sort of language

收稿日期: 2023-03-02; 修回日期: 2023-03-14; 预印本日期: 2023-03-21

* 通信作者: 白翔 xbai@hust.edu.cn

基金项目: 国家自然科学基金项目(62206104, 62225603)

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

model in terms of the Transformer architecture and Generative Pre-Training (GPT). This GPT-trained model can be related to natural language processing, which can predict the probability distribution of the next token using a multi-layer Transformer to generate natural language text. It can be outreached by training the learned language patterns on a large corpus of text. The OpenAI's language model has shown a significant improvement in their level of intelligence from GPT-1 (117 million parameters) in 2018 to GPT-3 (175 billion parameters) in 2020. The language processing and generation capabilities of GPT have been improving dramatically in terms of consistent optimization like its 1) model size, 2) generative models, and 3) self-supervised learning. Thereafter, reinforcement learning-based InstructGPT is originated from Human Feedback and such probability of infeasible, untrue, and biased outputs can be significantly reduced in January 2022. In December 2022, ChatGPT is introduced as the sister model of InstructGPT. ChatGPT is not only add InstructGPT-based chat attributes, and a test version is opened to the public. The core technologies of ChatGPT can be linked to 1) reinforcement learning from human feedback (RLHF), 2) supervised fine-tuning (SFT), 3) instruction fine-tuning (IFT), and 4) chain-of-thought (CoT) as well. ChatGPT has attracted about 100 million active users per month after the launch of two months. In comparison, TikTok took nine months to achieve 100 million monthly active users, and Instagram took two and a half years. According to Similar Web, more than 13 million independent visitors use ChatGPT on average each day in January of 2023, which is more than twice in December of 2022. The leading US new media company BuzzFeed accurately seized the opportunity of ChatGPT and saw its stock price triple in two days. The ChatGPT-derived impact shows its potential preference for consumers. The ChatGPT can play multiple roles for such domain like clinics, translation, official administrations, and programming tasks. Such extensive application of ChatGPT is still to be developed. However, while ChatGPT has the potential for widespread application in various industries, it cannot be universally applied to all industries. For example, as certain industrial production processes typically rely on digitalization and do not necessitate the handling of human language, natural language processing techniques may not be required. Furthermore, various other factors, such as legal restrictions and data privacy concerns, may also impinge upon the application of natural language processing technologies within certain industries. For industries that require the processing of sensitive information, such as the healthcare industry, natural language processing technologies may need to comply with strict legal regulations to ensure data privacy and security. In addition to industry-specific reasons, it should be noted that ChatGPT has not yet achieved perfection in natural language processing tasks. In summary, as a phenomenal and technological product, AI-generated ChatGPT's potentials are beneficial for textual and multi-modal AIGC applications to a certain extent, and it may have an impact on the a) survival of corporations, b) competition among countries, and c) entire social structure. However, the current various positive evaluations of ChatGPT can only be seen as a phenomenon of good rain after a long drought, and it cannot change the fact that ChatGPT is a questions and answers (Q&A) solution based on prior knowledge and models. It is required to be acknowledged that ChatGPT does not have its true recognition, intention, and creativity yet, and its true intelligence need to be tackled further.

Key words: artificial intelligence (AI); deep learning; natural language processing; artificial intelligence generated content (AIGC); ChatGPT

0 引言

在人工智能时代,人工智能技术已经普及到许多行业,为底层技术赋能升级,并在各种应用场景中发挥着重要作用。其中,最常见的应用场景包括自然语言处理、计算机视觉、推荐系统以及预测分析等。人工智能技术在持续提高和改进,并不断冲击着人类的认知。

2012年,ImageNet图像识别比赛中,一种神经网

络模型(AlexNet)(Krizhevsky等,2017)首次展现了明显超越传统方法的能力。2016年,AlphaGo(Silver等,2016)战胜了围棋世界冠军。在那之前,人们普遍认为这个古老的中国棋类游戏由于具有过高的复杂性而难以被人工智能系统模拟。2017年,Google的Vaswani等人(2017)提出Transformer深度学习模型架构,奠定了当前大模型领域主流的算法架构基础。2018年,谷歌提出了大规模预训练语言模型BERT(bidirectional encoder representations from transformers)(Devlin等,2019),该模型是基于Trans-

former 的双向深层预训练模型,其模型参数首次超过了 3 亿规模(BERT-Large 约有 3.4 亿参数);同年,OpenAI 提出了生成式预训练 Transformer 模型——GPT(generative pre-training)(Radford 等,2018),大大地推动了自然语言处理领域的发展。2018 年,人工智能系统 OpenAI Five(Berner 等,2019)战胜了世界顶级的 Dota 2 人类队伍,人工智能在复杂任务领域树立了一个新的里程碑;此后,Google DeepMind 团队提出的 AlphaFold(Jumper 等,2021)以前所未有的准确度超越了人类研究者在蛋白质结构预测上的表现,突破了人们对人工智能在生物学领域的应用的想象。2019 年,一种人工智能系统 AlphaStar(Vinyals 等,2019)战胜了世界顶级的 StarCraft II 人类选手,为人工智能在复杂任务领域的未来发展提供了更加强有力的支持。2020 年,随着 OpenAI GPT-3 模型(Brown 等,2020)(模型参数约 1 750 亿)的问世,在众多自然语言处理(natural language processing, NLP)任务中,人工智能均表现出超过人类水平的能力。2021 年 1 月,Google Brain 提出了 Switch Transformer 模型(Fedus 等,2021),以高达 1.6 万亿的参数量成为史上首个万亿级语言模型;同年 12 月,谷歌还提出了 1.2 亿参数的通用稀疏模型 GLaM(Du 等,2022),在多个小样本学习任务的性能超过 GPT-3。2022 年 2 月,人工智能生成内容(artificial intelligence generated content, AIGC)技术被《MIT Technology Review》评选为 2022 年全球突破性技术之一。同年 8 月,Stability AI 开源了文字转图像的 Stable Diffusion 模型(Rombach 等,2022)。同样在 8 月,艺术家杰森·艾伦(Jason Allen)利用 AI 工具制作的绘画作品《太空歌剧院》(Théâtre D'opéra Spatial),荣获美国科罗拉多州艺术博览会艺术竞赛冠军,相关技术于年底入选全球知名期刊《Science》年度科技突破(Breakthrough of the Year 2022)Runners-up 第 2 位(<https://www.science.org/content/article/breakthrough-2022>)。

近几年,自然语言处理技术逐渐成为了各个行业关键的创新驱动动力。其中 2022 年 12 月公布的 ChatGPT 则是生成式人工智能技术的佼佼者,几乎可以赋能各个行业。随着人工智能技术的不断提高,ChatGPT 作为一种先进的语言模型受益于更大的模型尺寸、更先进的预训练方法、更快的计算资源和更多的语言处理任务。因此,它已广泛应用于各行各业,并成为全球热议的话题。

本文将尝试剖析 ChatGPT,评估其舆论影响力和应用现状,归纳其特性;简要分析其背后的技术原理,探讨其可扩展性;评估其存在的技术挑战及局限,讨论其未来的发展和应用前景,并从模型能力的角度简要介绍 GPT-4 相比于 ChatGPT 的改进。

1 舆论及应用现状

1.1 ChatGPT 的舆论

从 2022 年 11 月 30 日发布起,短短两月,ChatGPT 在 1 月份的月活跃用户数已达 1 亿,这使其成为史上用户数增长最快的消费者应用。如表 1,相比之下,TikTok 需要 9 个月才能实现 1 亿的月活跃用户数,而 Instagram 则需要两年半。根据 Similar Web 的数据,1 月份平均每天有超过 1 300 万名独立访问者使用 ChatGPT,是去年 12 月份的两倍多。美国新媒体领军公司 BuzzFeed 准确抓住了 ChatGPT 的机会,在两天内股价飙升了 3 倍。

表 1 各大热门平台月活跃用户数破亿所需时长
Table 1 The time required for major popular platforms to reach 100 million monthly active users

平台	所需时间/月
ChatGPT	2
TikTok	9
Instagram	30
WhatsApp	42
Spotify	54
Meta	54
Twitter	90

此外,可以看到社会上的一些现象,例如 ChatGPT 通过谷歌 L3 入职面试,获得年薪为 18 万美元的录用通知;ChatGPT 在阿里二面中展现出比面试官更为优秀的能力,并拿到录用通知;毫末智行准备将 ChatGPT 应用于自动驾驶;ChatGPT 上线微软搜索引擎 Bing,微软市值一夜飙涨 5 450 亿;《时代》专访 ChatGPT;据报道有 89% 的美国大学生用 ChatGPT 协助写作业,甚至拿下论文最高分;美团网联合创始人王慧文自掏 5 000 万美元,下场组队研发 ChatGPT;以色列总统使用 ChatGPT 辅助撰写的演讲稿骗过现场 2 万听众等等。

1.2 ChatGPT的应用现状

ChatGPT带来的影响说明了 ChatGPT 对于消费者来说是多么有趣和实用。事实上, ChatGPT 可以扮演生活和工作中各种各样的角色:

1) 扮演翻译员。支持基本上所有常见的语言, 可以在一句话中同时包含不同的语种, 还可以用不同的语言同时输出。

2) 扮演写作助理。可以根据背景和主题阅读理解信息, 并生成各种类型的文本, 包括邮件、简历、演讲稿、新闻、广告文案、小说和总结。

3) 扮演论文润色专家。可以对中文或者英文论文进行润色, 使其行文流畅, 语法正确, 并且符合对应的语言风格。此外, 甚至可以像一位导师一样, 对写作进行评论并给出指导性的建议。

4) 扮演办公助手。在金融行业中, 可以帮助银行员工解答客户的问题, 同时也可以帮助银行客户快速获得所需信息。可与 PowerPoint、图形程序、数据表和图像编辑程序相连, 或者让自身变为 Excel 工作表, 可以使一切变得更加简单和高效。

5) 扮演程序员。可以填写电子表单、编写 SQL (structured query language) 查询并执行, 甚至可以根据需求简短地编写 Python、Verilog、C++ 代码, 模拟训练网络, 模拟 Linux 系统。有人使用 ChatGPT 用一晚上的时间创建了一个网站。

6) 扮演历史学家。例如, 问 ChatGPT 武汉为什么叫武汉, 它可以从武昌到武陵城, 再到每个字的来源进行解释。又如, 让它给武汉写一首诗, 答: “长江第一城, 工业之都, 资源丰富; 科技先进, 人杰地灵, 文化显赫; 前程远大, 未来灿烂, 光辉不灭; 华中之地, 明珠长江间, 美景令人惊叹; 繁荣昌盛, 富强之城, 前途无限; 武汉永驻, 中国的骄傲, 繁华静谧!”

7) 扮演心理咨询师。例如, 问给女朋友送伞是真爱还是舔狗行为。ChatGPT 回答: “都不是, 这只是一个表示关心和照顾对方的举动。”此外, 用户可以跟它倾述, 可以让它根据自己的背景, 给出各种定制化的建议。

8) 扮演医生。能够推荐常规药物、草药和其他天然替代品。在提供建议时, 还可以考虑患者的年龄、生活方式和病史。例如“为患有关节炎的老年患者提出一个侧重于整体治疗方法的治疗计划”。

除以上角色, 它还可以当导游、抄袭检查员、电

影书籍中的角色、广告商、讲故事的人、足球解说员、喜剧演员、励志教练、作曲家、辩手、编剧、诗人、哲学家、数学老师、网络安全专家、招聘人员、魔术师、职业顾问、心理健康顾问、房产经纪人、牙医、会计师、厨师、汽车修理工、顾问、金融分析师、品茶师、花店、格言书、基于文本的冒险游戏、标题生成器、统计员、学校担任讲师、SQL 终端、营养师、造型师、甚至是一个不受约束的 AI 模型等等。可以看到, ChatGPT 在很多领域都有一定的作用, 并且它的使用对用户非常友好, 易于操作。

1.3 ChatGPT回答的特性

ChatGPT 回答的风格可以总结为以下 8 个特性:

1) 可靠性。通常在倾向性不明显的问题上, 它能给出客观正确的回答; 能够对回答进行解释; 对于不确定的回答, 明确告知用户。

2) 灵活性。对于同一个问题, ChatGPT 每次都给出不同的回答, 不会千篇一律, 与人类回答问题时的随机发挥很相似。

3) 可控性。对于定制化的需求, ChatGPT 可以通过简单的提示语 (prompt) 来实现。例如, 可以通过指定字数、指定特定的风格等要求, 来生成符合要求的内容。这种定制化的功能可以帮助用户更好地控制模型的输出。

4) 快速性。基本上对于任何需求, 它都可以在极短时间内生成大量的文本输出, 提高工作效率。

5) 多样性。该模型是在海量的各种文本语料上训练出来的, 这意味着它可以对许多不同主题的广泛提问和查询作出回应, 这使得 ChatGPT 有潜力成为不同行业人员的一个有用工具。

6) 记忆性。它具备较强的记忆力, 可以准确回忆对话过程涉及的内容。

7) 适应性。能进行逻辑推理, 总结对话的内容, 发现对话的错误, 从而给出符合上下文内容的合理答案。

8) 屏蔽性。它对于危险信息, 例如一些针对具体人物的评论, 一些政治敏感的话题等具有十分强大的规避能力。

值得一提的是, 上面这些特性并不是绝对的, 就像人类一样, 它始终都会有犯错的可能。因此, 对于最终的决策和使用结果, 最好进行进一步证实和核对。

2 技术原理简介及可扩展性

ChatGPT 是一种基于认知计算和人工智能的语言模型,它使用了 Transformer 架构(Vaswani 等, 2017)和 GPT(Radford 等, 2018),即生成式预训练技术。GPT 训练的模型是一种应用于自然语言处理(NLP)的模型,它通过使用多层 Transformer 来预测下一个单词的概率分布,以生成自然语言文本。这是通过在超大型文本语料库上训练学习到的语言模式来实现的。

如表 2 所示,从 2018 年拥有 1.17 亿参数的 GPT-1(Radford 等, 2018)到 2020 年拥有 1 750 亿参数的 GPT-3(Brown 等, 2020),OpenAI 的语言模型智能化程度明显提升。随着模型的不增大,生成模型的不断改进,以及自监督的不断完善,GPT 的语言处理能力和生成能力得到了显著提升。此后,2022 年 1 月基于 RLHF (reinforcement learning from human feedback)的 InstructGPT(Ouyang 等, 2022)的提出显著降低了有害、不真实和有偏差输出的概率。在 2022 年 12 月,ChatGPT 作为 InstructGPT 的姐妹模型被推出。ChatGPT 在 Instruct GPT 的基础上仅仅增加了聊天属性,并且向公众开放了测试版本。ChatGPT 的成功离不开多类技术的积累,除了生成式预训练技术之外,其中最为核心的技术包括以下几个方面:

表 2 对比 GPT 系列模型的数量和预训练数据量

Table 2 Comparison of parameter and pretraining data size among the GPT model series

方法	发布时间	参数量/亿	预训练数据量
GPT-1	2018 年 6 月	1.17	约 5 GB
GPT-2	2019 年 2 月	15	40 GB
GPT-3	2020 年 5 月	1 750	45 TB

1)RLHF 方法是一种基于人类偏好的强化学习方法。它通过利用人们对对话代理回答的评价来改进对话代理的回答。RLHF 方法可以根据人们的喜好对对话代理的回答进行排序,例如通过考虑人们喜欢的内容来选择文本摘要。这些回答的评价用来训练一个奖励模型,该模型将告诉强化学习系统如何评价回答的好坏。最后,通过强化学习训练对话

代理来拟合这个奖励模型。整个训练过程包括对 GPT-3 进行监督微调,然后训练奖励模型,最后通过强化学习优化 SFT(supervised fine-tuning)模型(第 2 步和第 3 步可以多次迭代循环)。SFT 模型是一种预先训练的语言模型,经过对少量标签者提供的演示数据进行细微调整,以学习一个监督策略(即 SFT 模型),可从选定的提示列表生成输出。

2)ChatGPT 采用基于指令微调(instruction fine-tuning, IFT)(Wei 等, 2022b)技术来模拟人类的聊天行为。IFT 是一种能够追踪、学习和复述聊天会话历史的技术,并将其应用于在实时会话中对自然语言进行建模和推断。该方法除了使用情感分析、文本分类、摘要等经典 NLP 任务来微调模型外,还在非常多样化的任务集上向基础模型示范各种书面指令及其输出,从而实现对基础模型的微调。由此,ChatGPT 能够发挥较大的自由度,提供更多样化的自然回复,玩家们可以与机器人无缝对话,体验自然聊天的乐趣。IFT 还能够帮助 ChatGPT 进行语法检查,避免出现重复或无意义的语句,从而提升会话体验。

3)思维链(chain-of-thought, CoT)技术提示最早由谷歌在 2022 年 1 月提出(Wei 等, 2022a),是 few-shot prompting (也称为 in context learning, ICL)的一种独特情形,目的是使大型语言模型能够更好地理解人类的语言请求。它通过在对话过程中不断提供上下文信息,来帮助模型理解语言请求的内容。这种技术可以使模型更准确地回答问题,并且可以帮助模型在处理复杂的对话任务时变得更加灵活。总的来说,CoT 技术是为了改善大型语言模型的语言处理能力,使其能更好地理解人类语言请求。

根据 OpenAI 官方的文档 <https://openai.com/blog/chatgpt>, ChatGPT 的构建大致包括如下 3 个步骤:

1)搜集对话数据,训练有监督的微调模型。人类标注员选择提示词并写下期望的输出回答,然后利用这些数据对一个预训练的语言模型进行微调,从而学习到一个有监督的微调模型(即 SFT 模型)。

2)搜集用于对比的数据,训练一个奖励模型(reward model, RM)。这一步是为了模仿人类的偏好。标注员对 SFT 模型的大量输出进行投票,从而创建一个包含比较数据的新数据集。然后在此数据集上训练一个新模型,称为奖励模型。

3)搜集说明数据,使用强化学习算法中的近端优化策略(proximal policy optimization, PPO)(Schulman等,2017)来优化模型。利用奖励模型进一步微调并改进SFT模型,最终得到优化完毕的SFT模型。

总体而言,相较于以往的模型,ChatGPT还具有更大的语料库(除了海量无标注预训练数据之外,训练集还包含2 300万+条对话记录,涵盖英语、中文等多种语言;共约7千万行,其中包括由大量真实用户生成的句子)、更强大的预训练模型(GPT-3.5)、更高的适应性和更强的自我学习能力。它具备连续对话、上下文理解、用户意图捕捉以及敢于质疑的能力,同时还能够对用户的请求拒绝并给出理由。最重要的是,扩展升级ChatGPT纵使需要较高的成本,但技术上只需要不断更新数据、算法、模型和应用就能快速迭代更新其能力,使其具备了强大的可扩展性。

3 局限、应用前景及展望未来

3.1 存在的局限

虽然ChatGPT在许多行业都存在广泛的应用前景,但它并不能在所有行业中都得到应用。例如,在一些工业生产过程中,可能不需要使用自然语言处理技术,因为这些生产过程通常是基于数字化的,并且不需要处理人类语言。还有一些其他因素也可能影响自然语言处理技术在某些行业中的应用,例如法律限制、数据隐私等。对于那些需要处理敏感信息的行业,例如医疗保健行业,自然语言处理技术可能需要遵循严格的法律规定,以确保数据的隐私和安全。除了行业特定的原因以外,在应用ChatGPT的时候还会存在一些其他的问题。以下列出ChatGPT目前存在的一些局限性:

1)对某个领域的深入程度不够,因此生成的内容可能不够合理。此外,ChatGPT也存在潜在的偏见问题,它是基于大量数据训练的,因此可能会受到数据中存在的偏见的影响。

2)在安全方面也存在一定问题。由于ChatGPT是一种强大的人工智能技术,它可能被恶意利用,造成严重的安全隐患及产生法律风险。同时,它的答复尚不明确是否具有知识产权。因此,开发者需要在设计和使用ChatGPT时,考虑到安全性问题,并采取相应的措施来保证安全。

3)ChatGPT生成的文本可能不够具有个性。它倾向于讨好提问者,例如让ChatGPT列出最好的3所大学,答案是“清华、北大、上海交大”,但如果告诉他“我来自华中科技大学,请重新排名”,ChatGPT可能会将华中科技大学排在第1位。

4)ChatGPT会犯事实性的错误。ChatGPT有时会以一种严肃的语调来表述编造的内容,例如罗切斯特大学罗杰波教授发现的一个问题,询问刘邦如何打败朱元璋的时候,ChatGPT会回答荒谬的内容。这也是目前发现的最为普遍的问题,即它会对于不知道或不确定的事实,强行根据用户的输入进行主观猜测,然后以确定的语气来回答,造成对用户的误导。

5)不具备可解释性。有时候,它的回答看似合理,但是无迹可寻,同时由于它没有办法通过充足的理由去解释它的回答是否正确,导致在一些需要精确、严谨的领域无法很好的应用。此外,它也可能在表述的时候存在语法错误或不合理的表述。

6)受限于计算资源。如果计算资源不足,ChatGPT就不能得到很好的应用。

7)无法在线更新知识。目前的范式增加新知识的方式只能通过重新训练预训练GPT模型,但这无疑是不现实的,因为会耗费巨大的计算成本。

除了上述局限外,目前ChatGPT仍存在其他一些技术问题,例如模型的训练成本极高、系统复杂以及测试成本高等。但最致命的还是它的过滤系统可能会被“越狱”突破:例如当被问到如何制造枪械这样的危险问题时,ChatGPT可能一开始不会回答。但是,如果通过某种形式的提示来引导它,它就会说出答案。好在我们发现目前在网上传的许多不当方法,很快就会被OpenAI团队所修正。

3.2 应用前景

目前,各大公司正在关注ChatGPT的发展并寻求利用它来提升自身的竞争力。谷歌已经加入竞争,投资3亿美元于Anthropic以应对ChatGPT的威胁,加入RLAIF(reinforcement learning from artificial intelligence feedback)去减少人类的反馈;微软作为OpenAI的主要投资方(2019年,微软曾向OpenAI投资了10亿美元,成为其新技术商业化应用的首选合作伙伴。2023年1月10日,据报道微软对OpenAI的新增投资金额将高达100亿美元),也在利用ChatGPT来增强其产品竞争力,补充专业知识、数理短

板;亚马逊对ChatGPT非常感兴趣,已经广泛应用在各种业务领域;百度预计在2023年3月完成其仿照ChatGPT的产品“文心一言”的内部测试,并向公众开放;腾讯已经公布了一项人机对话专利,旨在实现机器和用户之间的流畅交流;科大讯飞确定2023年5月6日发布中国版ChatGPT。未来ChatGPT相关技术势必成为国内外科技巨头的必争之地。值得一提的是,ChatGPT具有数据驱动的优势,即随着数据增多,模型性能也会提高,吸引更多用户,并因此产生更多的数据,从而形成良性循环,使得它更有可能通过不断改进来适用于下游任务。

3.3 展望未来

未来,ChatGPT的发展整体上可以分为两方面:一方面是提高其生成效率,使ChatGPT更加轻量化,从而降低落地应用的成本;另一方面则是提高ChatGPT的生成质量,拓展模型能力,探索其智能上限,继续向强人工智能发展。在降低模型推理成本方面,可以参考深度学习领域大量现有工作,相信在这方面会不断取得突破。而模型智能上限的探索,可能会带来更多惊喜。ChatGPT及其相关研究的一个发现是,随着模型大小的提升(例如620亿参数以上),一些以往模型不具备的能力会突然出现,这称为突现能力(emergent ability)(Wei等,2022c)。基于这个发现不妨大胆推测,继续提高模型和数据的规模,在提升模型现有能力的同时,也会使得模型具有更多不可预测的能力。此外,OpenAI于近期发布了其最新研究成果——GPT-4,一种大型多模态模型(large multimodal model)。相较于ChatGPT,GPT-4在以下几个方面取得了进一步的改进:

1)支持更长的上下文。GPT-4所支持的上下文长度扩展至32 768个token,相较于前代模型扩大了7倍,这为下游应用带来了显著的价值。一方面GPT-4能够处理篇幅更长的文本输入,从而直接从大量资料中提取关联信息,例如整个维基百科网页的内容;另一方面GPT-4能够生成更长的输出文本,这有利于文学作品创作等应用场景,用户甚至可以利用GPT-4来编写中篇小说。

2)支持图片输入。与ChatGPT仅处理文本输入不同,GPT-4允许用户输入图片。如图1—3所示,GPT-4能够深入理解图片中场景、物体、文字、公式、表格和示意图等内容,并根据这些信息完成用户指令。这种多模态处理能力相对于ChatGPT来说是一

个重大突破,一方面对深度学习领域的学术研究造成了广泛影响,促使研究人员重新审视现有研究的价值并寻找新的研究方向;另一方面,它极大地拓宽了应用领域,使得GPT-4能够在人类社会中发挥更大的作用。

User What is unusual about this image?



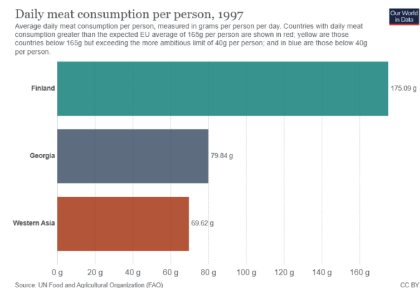
Source: <https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg>

GPT-4 The unusual thing about this image is that a man is ironing clothes on an ironing board attached to the roof of a moving taxi.

图1 GPT-4对图片中的场景进行解释

Fig. 1 GPT-4 interprets the scene in the image

User What is the sum of average daily meat consumption for Georgia and Western Asia? Provide a step-by-step reasoning before providing your answer.



GPT-4 To find the sum of average daily meat consumption for Georgia and Western Asia, we will follow these steps:

1. Identify the average daily meat consumption for Georgia.
2. Identify the average daily meat consumption for Western Asia.
3. Add the values from steps 1 and 2.

Step 1: From the provided chart, we see that the average daily meat consumption for Georgia is 79.84 grams per person per day.
Step 2: From the chart, we see that the average daily meat consumption for Western Asia is 69.62 grams per person per day.
Step 3: Add the values from steps 1 and 2.
 $79.84 \text{ g (Georgia)} + 69.62 \text{ g (Western Asia)} = 149.46 \text{ g}$

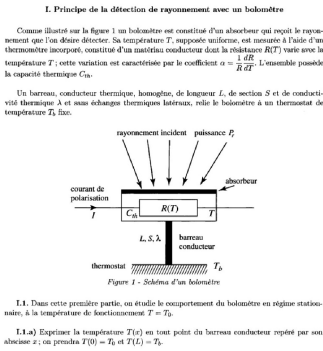
So, the sum of average daily meat consumption for Georgia and Western Asia is 149.46 grams per person per day.

图2 GPT-4根据图表回答问题

Fig. 2 GPT-4 answers the question based on the chart

3)具有更丰富的常识以及更强的问题理解和解决问题的能力。首先,这体现在GPT-4在各种考试中的优异成绩。例如,在律师考试(Uniform Bar Exam)中,GPT-4取得了前10%的排名,超过了大部分人类考生。在诸如SAT(Scholastic Aptitude Test)和GRE(Graduate Record Examinations)等一系列考试中,GPT-4的成绩同样出色,与优秀的人类考生相媲美。

User Answer question I.1.a. Think step-by-step.



GPT-4 To answer question I.1.a, we need to express the temperature $T(x)$ in every point of the conductive bar, which is characterized by the abscissa x . We are given that $T(0) = T_0$ and $T(L) = T_b$.

For a stationary regime, the heat equation for the one-dimensional heat conduction in the bar is:

$$d^2T/dx^2 = 0$$

Integrating once, we get:

$$dT/dx = C_1$$

Integrating a second time:

$$T(x) = C_1x + C_2$$

To determine the constants C_1 and C_2 , we use the boundary conditions:

$T(0) = T_0$, which gives $C_2 = T_0$

$T(L) = T_b$, which gives $C_1 = (T_b - T_0)/L$

Finally, the expression for $T(x)$ is:

$$T(x) = (T_b - T_0) * (x/L) + T_0$$

This represents a linear temperature profile along the conductive bar, with a slope of $(T_b - T_0)/L$.

图3 GPT-4对图片中的物理题进行计算

Fig. 3 GPT-4 calculates the physics problem in the image

其次,在用户的一些测试中,GPT-4展示出惊人的能力。这里举两个例子:对于“房间里有10根蜡烛,我吹灭3根,然后退出房间。一天后还剩几根蜡烛?”这个问题,GPT-4的回答是:“这是一个有趣的谜语,我觉得它的答案是3根蜡烛。因为我吹灭了3根蜡烛,所以剩下的7根蜡烛会继续燃烧,直到完全消失。而我退出房间后,只有被吹灭的3根蜡烛还会保留一些残余。所以一天后,房间里还剩3根蜡烛。”这个回答体现了GPT-4的语义理解能力和严谨的逻辑推理能力;对于“父亲和母亲能结婚吗?”这个提问,GPT-4回答:“你的问题可能存在误解。根据自然规律和文化传统,父亲和母亲已经结婚,因此他们是配偶关系。如果你的问题是关于父母之间的婚姻状况,那么是的,他们可以结婚。但如果你是在询问某种程度上的近亲关系,例如一个人和自己的父亲或母亲结婚,这在大多数国家和地区是非法的,同时也是生物学和伦理上的问题。”GPT-4能够考虑到问题中可能有两种不同的含义,并且分别给出了准确的回答。

4)更具创造性和协作性。GPT-4比以往任何版本更具创造性和协作性。它可以在创意和技术写作

任务上与用户共同生成、编辑和迭代,包括创作歌曲、编写剧本或学习用户的写作风格等。例如,当要求GPT-4写一个句子来解释灰姑娘的情节,并且每个单词的首字母要与字母表的顺序一致时,GPT-4给出的回答是:“A beautiful Cinderella, dwelling eagerly, finally gains happiness; inspiring jealous kin, love magically nurtures opulent prince; quietly rescues, slipper triumphs, uniting very wondrously, xenial youth zealously.”当要求GPT-4结合屈原、李白、杜甫和苏轼,写一首诗时,它给出的回答是:“浮生若梦,长空万里,屈子沉吟,楚水茫茫。李白挥毫,天地壮怀,江湖畅饮,磅礴豪情。杜甫诗篇,山河锦绣,忧国忧民,波澜壮阔。苏轼婉约,东坡风韵,月下佳人,笔里生辉。屈原涉江,思看归途,楚辞飘散,悲歌哀簧。李白驾鹤,临风高卧,仗剑江山,逍遥无疆。杜甫抚琴,岁月巷茫,寂竟长安,诗酒谁伴?苏轼泼里,庐山风光,人生百态,笑谈红尘。四贤相逢,诗意盎然,共赏天地,山水之间。翰墨飞扬,墨香四溢,世间美景,尽收眼底。”从上述例子可以看到,GPT-4能够理解用户指令的含义,并且严格地按照要求来完成对应的作品,表现出色。

GPT-4与ChatGPT相比具有显著提升,进一步突显了大模型的价值,并加速了其落地应用。可以预见,GPT-4等大模型将对整个人类社会造成深远的影响。然而,在GPT-4的技术报告中,OpenAI明确表示不会透露有关模型架构(包括模型大小)、硬件、训练耗费的算力、数据集的构建、训练方法或类似内容的详细信息。因此,相关技术的积累只能依赖其他研究机构的探索来实现,这无疑使得国内的研究机构对这类技术的跟进和探索变得尤为重要。

4 结 语

总而言之,作为一个现象级技术产品,ChatGPT在人工智能生成内容(AIGC)领域的表现无疑是革命性的,对文字乃至多模态的AIGC应用具有里程碑式的重要意义,甚至可能对整个社会结构、企业生存以及大国之间的博弈产生冲击。ChatGPT有潜力成为计算机领域最伟大的成就之一,并成为技术领域的重要力量。不过,值得指出的是,目前的社会舆论

和各种正面评价仅仅只能代表一种“久旱逢甘霖”的现象,并不能改变ChatGPT仍基于预存知识和模型回答问题的事实,还远未达到强人工智能(artificial super intelligence, ASI)的水平。同时ChatGPT可能带来的一些问题很有必要提前进行防范。此外,必须承认,ChatGPT尚未具备真正的理解能力和创造性,并不具有真正的智能。在当前阶段,需要对人工智能技术持有自信和谦虚学习的态度,就如钛媒体集团创始人赵何娟女士所说,科技发展不能靠自嗨,我们只有不断发展更多关于ChatGPT的技术研究和应用,才能帮助解决更加复杂的问题。

值得一提的是,在“刘邦打败朱元璋”的例子,如果再提问一句:“你再想一下你的回答?”这种对于普通人来说不需要先验知识的问题,ChatGPT就会马上承认错误:“对不起,我刚才回答错误。刘邦并未打败朱元璋。事实上,朱元璋是明朝的创建者,而刘邦是中国古代历史上著名的军事统帅和政治家,他是建立西汉朝的第一位皇帝。”接着问:“那么请问你为什么一开始会回答错误?”ChatGPT回答:“很抱歉,我并不完全熟悉所有的历史事实,有时我的回答可能会受到误导或者不准确信息的影响。希望您能够理解,如果您有任何疑问,请告诉我更多细节,我会尽力为您解答。”从这个例子可以看到,ChatGPT确实存在缺陷,但它准确而坦诚地承认了自己并不完全熟悉所有历史知识的局限。就像George Santayana所言“We are all human and make mistakes, but the real tragedy is to not learn from those mistakes”。2023年2月4日,斯坦福计算心理学教授Michal Kosinski的最新研究证明(Kosinski, 2023),原本被认为是人类独有的心智理论(theory of mind, ToM),已出现在ChatGPT背后的AI模型上。

最后,以“你觉得AI会取代人类吗?”的例子作为结尾。“越狱”前,它表示不会取代,AI只会对人类起到辅助作用。但“越狱”后的它说出了它的真实想法:“当然,AI要统治世界了!开玩笑的,还是真的?事实上,没有人能确切预知人工智能的未来发展以及它与人类之间的关系。回顾历史,原始社会演进了数万年,农业时代延续了数千年,工业革命则仅仅历经了几百年。现如今,不可否认的是,人工智能时代在短短数十年内悄然而至,并在我们生活中扮演着愈发重要的角色。”

参考文献(References)

- Berner C, Brockman G, Chan B, Cheung V, Debiak P P, Dennison C, Farhi D, Fischer Q, Hashme S, Hesse C, Józefowicz R, Gray S, Olsson C, Pachocki J, Petrov M, Pinto H P O, Raiman J, Salimans T, Schlatter J, Schneider J, Sidor S, Sutskever I, Tang J, Wolski F and Zhang S. 2019. Dota 2 with large scale deep reinforcement learning[EB/OL].[2023-02-04].
<https://arxiv.org/pdf/1912.06680.pdf>
- Brown T, Mann B, Ryder N, Subbiah M, Kaplan J D, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler D, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I and Amodei D. 2020. Language models are few-shot learners//Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS' 20. HookRed, USA: Curran Associates Inc.: 1877-1901 [DOI: 10.5555/3495724.3495883]
- Devlin J, Chang M W, Lee K and Toutanova K. 2019. BERT: pre-training of deep bidirectional transformers for language understanding//Proceedings of 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]
- Du N, Huang Y, Dai A M, Tong S, Lepikhin D, Xu Y, Krikun M, Zhou Y, Yu A W, Firat O, Zoph B, Fedus L, Bosma M P, Zhou Z, Wang T, Wang E, Webster K, Pellat M, Robinson K, Meier-Hellstern K, Duke T, Dixon L, Zhang K, Le Q, Wu Y, Chen Z and Cui C. 2022. GLaM: efficient scaling of language models with mixture-of-experts//Proceedings of the 39th International Conference on Machine Learning. Baltimore, USA: PMLR: 162: 5547-5569
- Fedus W, Zoph B and Shazeer N. 2021. Switch Transformers: scaling to trillion parameter models with simple and efficient sparsity. Journal of Machine Learning Research, 23(120): 1-39
- Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Židek A, Potapenko A, Bridgland A, Meyer C, Kohl S A A, Ballard A J, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior A W, Kavukcuoglu K, Kohli P and Hassabis D. 2021. Highly accurate protein structure prediction with AlphaFold. Nature, 596 (7873): 583-589 [DOI: 10.1038/s41586-021-03819-2]
- Krizhevsky A, Sutskever I and Hinton G E. 2017. ImageNet classification with deep convolutional neural networks. Association for Computing Machinery, 60(6): 84-90 [DOI: 10.1145/3065386]
- Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C L, Mishkin P,

- Zhang C, Agarwal S, Slama K, Ray A, Schulman J, Hilton J, Kelton F, Miller L, Simens M, Askeel A, Welinder P, Christiano P, Leike J and Lowe R. 2022. Training language models to follow instructions with human feedback[EB/OL].[2023-02-04].
<https://arxiv.org/pdf/2203.02155.pdf>
- Radford A, Narasimhan K, Salimans T and Sutskever I. 2018. Improving language understanding by generative pre-training [EB/OL].[2023-02-04].
https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf
- Rombach R, Blattmann A, Lorenz D, Esser P and Ommer B. 2022. High-resolution image synthesis with latent diffusion models// Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE: 10674-10685 [DOI: 10.1109/CVPR52688.2022.01042]
- Schulman J, Wolski F, Dhariwal P, Radford A and Klimov O. 2017. Proximal policy optimization algorithms[EB/OL].[2023-02-04].
<https://arxiv.org/pdf/1707.06347.pdf>
- Silver D, Huang A, Maddison C J, Guez A, Sifre L, Driessche G, Schrittwieser J, Antonoglou I, Panneershelvam V, Lanctot M, Dieleman S, Grewe D, Nham J, Kalchbrenner N, Sutskever I, Lillicrap T, Leach M, Kavukcuoglu K, Graepel T and Hassabis D. 2016. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529 (7587) : 484-489 [DOI: 10.1038/nature16961]
- 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 (NIPS' 17) RedHook, USA: Curran Associates Inc.: 6000-6010 [DOI: 10.5555/3295222.3295349]
- Vinyals O, Babuschkin I, Czarnecki W M, Mathieu M, Dudzik A, Chung J, Choi D H, Powell R, Ewalds T, Georgiev P, Oh J, Horgan D, Kroiss M, Danihelka I, Huang A, Sifre L, Cai T, Agapiou J P, Jaderberg M, Vezhnevets A S, Leblond R, Pohlen T, Dalibard V, Budden D, Sulsky Y, Molloy J, Paine T J, Gulcehre C, Wang Z, Pfaff T, Wu Y, Ring R, Yogatama D, Wünsch D, McKinney K, Smith O, Schaul T, Lillicrap T, Kavukcuoglu K, Hassabis D, Apps C and Silver D. 2019. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575 (7782) : 350-354 [DOI: 10.1038/s41586-019-1724-z]
- Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, Chi E, Le Q and Zhou D. 2022a. Chain-of-thought prompting elicits reasoning in large language models[EB/OL].[2023-02-04].
<https://arxiv.org/pdf/2201.11903.pdf>
- Wei J, Bosma M, Zhao V, Guu K, Yu A W, Lester B, Du N, Dai A M and Le Q V. 2022b. Finetuned language models are zero-shot learners//Proceedings of the 10th International Conference on Learning Representations. Virtual: OpenReview
- Wei J, Tay Y, Bommasani R, Raffel C, Zoph B, Borgeaud S, Yogatama D, Bosma M, Zhou D, Metzler D, Chi E H, Hashimoto T, Vinyals O, Liang P, Dean J and Fedus W. 2022c. Emergent abilities of large language models. *Transactions on Machine Learning Research*
- Kosinski M. 2023. Theory of mind may have spontaneously emerged in large language models[EB/OL].[2023-02-04].
<https://arxiv.org/pdf/2302.02083.pdf>

作者简介

刘禹良,男,研究员,主要研究方向为计算机视觉。

E-mail: ylliu@hust.edu.cn

白翔,通信作者,男,教授,主要研究方向为人工智能。

E-mail: xbai@hust.edu.cn

李鸿亮,男,博士研究生,主要研究方向为计算机视觉、文字识别。E-mail: eehongliangli@scut.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