唐杰，清华大学计算机系 WeBank 讲席教授，国家级人才，ACM/AAAI/IEEE Fellow。研究兴趣包括人工智能、知识图谱、数据挖掘、社交网络，最近专注大语言模型：ChatGLM（开源下载近 1000 万）等。曾获 ACM SIGKDD Test-of-Time Award（十年最佳论文）、IEEE ICDM 研究贡献奖、国家科技进步二等奖。

卷首语 Foreword

浅谈超级认知智能

唐杰

超大规模预训练模型（也称“基础模型”，英文为 Foundation Model、Big Model 等）快速发展，成为国际人工智能前沿研究和应用的热点。尤其是最近 ChatGPT 的推出引发社会公众的广泛关注，并引起了该技术是否会引发新一轮行业变革的讨论。基础模型则是 ChatGPT 等生成式人工智能技术产品的核心技术基座，推进 ChatGPT 等产品影响产业格局，甚至成为全新的用户交互方式，造成舆论引导、社会治理、信息服务等方面的不对称优势。人工智能基础模型已成为国际科技竞争“必争之地”，美国目前在占据主导权的同时不断加紧对我国进行战略性遏制，实现国产人工智能基础模型自主可控迫在眉睫。我国人工智能基础模型研究、应用与产业化发展正处于从模仿追赶迈向创新引领的关键时期。

从技术层面而言，ChatGPT 的优异表现主要依托超大规模预训练语言模型 GPT-3/4、有监督指令微调以及基于人类反馈的强化学习。从大模型技术本身演进进程来看，大致可分为三个阶段。

2020 大模型元年。自 2017 年谷歌提出 Transformer 机器学习模型架构，其应用迅速席卷了整个人工智能研究领域，成为自然语言处理等相关研究的主要方法，2018 年先后出现了 BERT、GPT-1，尤其是 BERT 在十多个自然语言理解任务上精度大大超过传统算法；2019 年 GPT-2 实现了自然流畅的文本内容生成；2020 年谷歌的 T5 将自然语言的翻译、分类、回归、摘要生成等任务都统一转成 Text-to-Text 任务，同年更具里程碑式的模型是 OpenAI 发布的 1 750 亿参数 GPT-3，大大提高了模型的内容生成和逻辑推理能力，具备较强通用能力，可完成多场景任务，显著降低学习成本、提高学习效率，同时也开启了大模型元年。

GPT-3 开启了基础模型发展的新时代，其在语言生成、上下文学习和知识（常识）理解等方面展现出惊人能力。随后全球范围内掀起了一股基础模型研究的热潮，国外如 Meta、微软、谷歌等，国内如清华大学、北京智源人工智能研究院、百度、华为、阿里、智谱等，都竞相追赶，提出包括 Gopher、PaLM、OPT-175B、GLM-130B、BLOOM-176B 等多个千亿级模型，积累了一定的技术实力。

2020-2023 大模型成熟期。在初代 GPT-3 的基础上，OpenAI 引入代码训练和指令微调等环节，在过去三年里持续学习形成了 InstructGPT、GPT-3.5、GPT-4 等系列模型。基于这些模型，OpenAI 进一步引入人类反馈强化学习建立了对话模型 ChatGPT，具有更强的自然交互与逻辑推理能力，在面临常识性问题、推理性问题、尚未理解和敏感话题时的处理呈现出高度智能化特征。除了 GPT-4 一枝独秀外，很多工业界和学术界的机构也推出了类 ChatGPT 模型。仅 2023 年 3 月 14 日这天就有 OpenAI 的 GPT-4、Anthropic 的 Claude、谷歌的 PaLM API 服务、智谱 AI 的 ChatGLM、斯坦福的 Alpaca、Midjourney 的 V5。这些都是最早一批可用的大模型，3 月 14 日这一天也被称为大模型里程碑日。

随后 2023 年 7 月 18 日，Meta 发布 Llama2，性能逼近 GPT-3.5，并且免费商业开源。类似开源模型大大促进了全球以及国内的大模型产业发展。

2024 超级认知智能元年。OpenAI 极有可能在 2024 年推出下一代模型，其认知能力将带来通用人工智能的再一次变革。其下一代模型不仅在技术上可能解决目前 ChatGPT 中存在的事实性以及推理能力的缺陷，实现更精细的语义理解、多模态（本文、图像、语音、视频等）输入和输出的支持，具备更强的个性化能力。此外人工智能的发展会更加瞄向通用人工智能，实现超过人类水平的超级认知智能（Super Cognitive Intelligence），实现 AI 的自我解释、自我评测（Self-instruct）、自我监督，并且确保模型的表现符合人类的价值观和安全标准。2023 年 7 月 OpenAI 公布了由首席科学家 Ilya Sutskever 和首席强化学习专家 Jan Leike 发起的超级对齐计划（Superalignment），目标就是实现机器自动对齐人类智能和人类价值观，实现模型自我反思和自我监控。相信在 2024 年会有更多研究者加入到通用人工智能和超级认知智能的研究中。