



本报记者 姬晓婷 实习生 卢沫男

大模型Scaling Law(规模定律)将失效？
长期以来,大模型的发展遵循依靠模型参数增加、数据集规模扩充以优化模型性能的规模定律。但随着可用于预训练的开源数据挖掘殆尽、模型训练所需的算力资源紧缺,大模型发展遵循的Scaling Law即将失效。在近日举行的清华基础模型论坛上,清华大学计算机系长聘教授刘知远表示,“密度法则”将接替Scaling Law成为最新大模型迭代规律。根据这一法则,端侧智能将迎来广阔发展前景。

“密度法则”成为大模型发展新定律？

“芯片电路密度(即芯片计算单元数量与芯片面积的比值)每18个月翻1倍。”这是集成电路行业熟悉的摩尔定律。

刘知远提出的“密度法则”与这一规律类似,指大模型知识密度将每100天翻1倍。

基于这一定律,刘知远认为,模型将朝着参数小型化、成本普惠的方向发展,并提出基于该定律的四条推论:

推论一:实现特定AI水平的模型参数呈指数下降,实现相同智能水平的模型参数量每3.3个月下降一半,同时模型推理速度提升一倍。例如,2024年2月发布的2.4B规模的MiniCPM-1,能力与2023年9月发布的7B规模Mistral相当,但参数量仅相当于后者的35%。

推论二:模型推理开销随时间呈指数级下降。例如,在过去20个月内,GPT-3.5级模型的API价格下降至原先的1/266.7,约2.5个月

下降一半,而更低的价格将会带来更多的应用。

推论三:模型训练开销随时间迅速下降。高质量互联网数据规模约为15T,可训练的模型最大参数约为750B,但1年后,仅需58B参数模型便可实现相同能力。

推论四:大模型能力密度呈加速增强趋势。ChatGPT发布之前,大模型能力密度约按每4.8个月增加1倍的速度增长;ChatGPT发布后,大模型密度倍增时间缩短至3.2个月。

模型小型化彰显端侧AI能力

在“密度法则”下,模型将沿着这样的路径发展:同等智能水平模型的训练、推理开销等成本降低,模型规模缩小,能力密度持续增强。

与此同时,在摩尔定律仍在发挥作用的情况下,端侧芯片的承载能力仍在增强。

这意味着,随着时间的推演,将出现能力水平非常高的大模型,

同时其参数规模能够缩小到一定的程度,使其能够装载到消费级计算设备上。届时,端侧AI能力将得以彰显。

但将运行在云端的PB级的知识规模浓缩到GB级的小模型参数中,是一项具备挑战性的工作。如何进行海量互联网数据的治理、如何利用数据背后的知识实现模型高效学习以实现大模型的科学化制备,将成为未来大模型构建的关键课题。

智能算法创新远未收敛

在过去5年中,一项项技术创新推动着大模型的发展:2018年出现的自监督预训练技术,使大模型具备了掌握人类知识的能力。但这一阶段,模型只是实现了学习,还不能实现模型应用;2021年,有监督的微调技术,使模型具备了理解人类意图的能力,可从用户的负责指令中挖掘出人类的真实意图,并调用自己掌握的知识,完成该指令。2025年,模型借助强化学习技

术,具备了深度思考的能力,可以在完成任务之前,像人一样深思熟虑,通过一步一步的思考给出准确的回答。

面向未来,刘知远认为,智能算法创新和演进远没有收敛,未来人工智能迈向AGI(通用人工智能)过程中还要不断拓展能力树,这也是人工智能迈向的核心任务之一。其需要拓展的能力包括自主学习、合作、创新创造等。如果停止算法创新,则将落后于其他发展AI能力的行业参与者。

就像人类个体具备一定智能水平之后,其彼此之间通过协同实现了群体智能涌现一样,刘知远认为,人工智能在结束个体的智能涌现之后,将通过多智能体群聚,迎来群体智能的第二次涌现。就像互联网把全球信息和人连接在一起,群体智能就有机会通过智能体互联网连接在一起,实现智能的二次涌现。制定智能体接入与沟通协议,有利于异质智能体沟通协作,共同完成任务。且不同专业能力的智能体有机结合,将完成个体所不能胜任的更复杂的任务。

近日,OpenAI相关负责人在墨西哥召开的人工智能峰会上表示,备受关注的GPT-5模型即将推出,且性能远超现有模型。

记者了解到,此前,OpenAI首席技术官Mira Murati将GPT与不同水平的人类智力进行了比较,称GPT-5将具备更先进的推理和记忆能力,其智能水平将实现质的飞跃,从GPT-3的儿童智力、GPT-4的高中生智力,到针对特定任务实现博士级的智能。OpenAI研究员Michelle Pokrass表示,GPT-5将是一个“知道何时认真思考,何时闲聊”的模型。

众说纷纭GPT-5

本报记者 宋婧 实习生 陈存

数月前,OpenAI公司在推出GPT-4.5时再次透露,这会是OpenAI最后一个非思维链接模型,未来问世的GPT-5将能集合o系列模型(如o3推理模型)与GPT系列模型;还将融合此前推出的各项多模态功能,如语音、canvas、搜索、DeepResearch等。免费用户也可以无限制访问标准智能级别的GPT-5,而Plus和Pro用户则能以更高的智能水平运行。

目前,OpenAI公司仍在对其进行开发,或将在7月正式发布,以应对Gemini 2.5 Pro和Claude 4等竞品带来的挑战。

业内人士判断,GPT-5的开发成本虽尚不明朗,但预计价格不低。早在发布GPT-4.5之时,这个“史上最贵”的大模型就已带来一些争议。相比GPT-4o,GPT-4.5的输入成本达到75美元/百万token,定价贵了30倍,输出成本为150美元/百万token,贵了15倍。

然而,“模型越贵越好用”的认知此前已经被以DeepSeek为代表的国产开源大模型打破。昂贵的GPT-5或将为性价比更高的中国大模型产品开拓海外市场提供新机遇。在多个基准测试成绩持平的基础上,DeepSeek-R1模型低谷期价格是GPT-4.5的1/270;百度发布的文心大模型4.5,API调用价格仅为GPT-4.5的1%。5月,DeepSeek、豆包、通义千

问、文心一言等大模型更是纷纷降价,甚至推出多款免费大模型。

相关专家表示,AI大模型赛道已呈现显著分化趋势。

一条路径是“大力出奇迹”,通过卷算力、卷数据来训练效果更好的模型。OpenAI称,为研发GPT4.5耗尽了GPU资源,不得不推迟新模型的全面发布。此次“千呼万唤始出来”的GPT-5能否再次引领大模型技术的新一轮能力跃迁,OpenAI又能否凭借新模型再创辉煌,还有待观察。

另一条路径是“四两拨千斤”,通过有限资源的高效利用,“花小钱办大事”。在DeepSeek开源生态和普惠定价策略的影响下,大模型行业迎来颠覆性变革,百度、阿里、字节跳动等国内大模型巨头企业纷纷跟进,提速加码,在低价领域展开激烈角逐。

快思慢想研究院院长田丰指出,全球AI行业已进入免费竞争与成本驱动的关键阶段,开源模型和性能高低决定企业生死。

中国科学院科技战略咨询研究院研究员周城雄强调:“国内外的通用大模型在性能上趋同,难以通过技术差异来建立产品的壁垒,要靠价格来竞争。免费策略能够满足大模型企业快速积累用户规模、通过流量入口抢占生态位的市场扩张需求。”

IBM收购人工智能平台Seek AI

本报讯 近日,IBM宣布收购人工智能平台Seek AI,收购金额尚未公开。Seek AI允许用户使用自然语言就企业数据提问。

Seek AI于2021年创立,总部位于纽约。该公司提供一系列工具,帮助企业更好地利用数据。其软件通过类似聊天机器人的界面,将自然语言问题自动转化为数据库查询,并进行高级数据分析与总结。Seek AI首席执行官兼创始人Sarah Nagy表示:“我们将扩大平台规模,为IBM客户部署关键任务解决方案,赋能下一代AI开发者,并大幅扩充我们的团队。”

此次收购正值IBM加大在人工智能领域的投资之际,尤其是在企业级AI应用方面。IBM的这一战略已经取得了显著成效,其第一季度财报显示,软件业务增长强劲,AI需求旺盛,推动公司业绩表现超出预期。

收购完成后,Seek AI将成为IBM新成立的Watsonx AI Labs的重要组成部分。Watsonx AI Labs是IBM位于纽约市的人工智能加速器,旨在汇聚IBM的研究人员和工程师,打造一个“协作中心”,为企业开发AI解决方案。IBM数据与AI总经理Ritika Gunnar表示:“Watsonx AI Labs将为顶尖AI开发者提供世界级工程师和资源,助力他们构建新的业务和应用。选择在纽约市落地这一项目,旨在投资于多样化、世界级的人才库,并融入一个充满活力的社区,这个社区的创新成果一直推动着科技行业的发展。”

此外,Watsonx AI Labs还将寻求与当地高校和研究机构合作。成功在该加速器推出产品的初创企业将获得IBM的资源支持,并有机会获得IBM风险投资部门IBM Ventures及其5亿美元企业AI基金的投资。(艾文)

百度推出千帆慧金金融大模型

本报讯 近日,百度正式推出千帆慧金金融大模型。该模型基于海量金融语料深度训练,构建金融专用合成数据管线,优化算法策略,并提供知识增强大模型和推理增强大模型两类模型,每类模型分别提供8B和70B两个版本,支持最长32K上下文输入,覆盖金融行业多数场景。

在金融领域的Benchmark评测中,千帆慧金金融大模型综合表现领先,百亿参数模型得分超过千亿参数的通用模型。在金融销售赋能场景中,相比通用模型,千帆慧金金融大模型能更完整地列出贷款材料清单、拆解工作流程,并明确风险管控要点,展现出深厚的行业知识与推理能力。(文编)

中国电科发布大模型“小可3.0”与智能体“极智”

本报讯 记者宋婧报道:6月5日,中国电科正式发布大模型“小可3.0”和智能体“极智”,并启动集团公司“人工智能+”专项行动计划。

记者从发布会现场了解到,中国电科自研的大模型“小可3.0”构建了完整的智能底座金字塔,包括Lite-3B、Std-13B和Max-30B三个版本,采用创新架构,确保内存消耗线性增长,训练效率大幅提升,真正实现无限上下文处理,适配从基层办事窗口到国家数据中心的不同场景需求。

“极智”智能体平台则通过“超级智能体+超级总线”的创新架构,新增百余项专业功能单元,构建“数字员工”,形成了一系列行业定制化解决方案,并在政务、司法、农

业、医疗等多个民生领域实现深度应用。

在AI+政务领域,中国电科通过部署“小可”政务大模型,支撑深圳市福田区打造了城市管理、民生诉求、绩效管理 etc 20个智能化应用场景。在城市管理方面,提升环卫处置效率50%,节约人力成本40%;在市容环卫方面,街道市容环境综合指数保持全市街道第一名;在公务员绩效管理方面,福保街道作为全省唯一入选中组部公务员绩效管理试点单位,创新开发“总结生成”和“辅助评分”绩效智能体,30秒解决“总结烦”,高度改善“打分难-凭印象”的情况。

在AI+司法领域,中国电科助力中国司法大数据研究院基于司法专业模型形成“云上综治”解决

方案,包含30多个智能化应用场景,已在武汉光谷、江西九江、北京石景山、河南汝州、广东佛山等600多个市县区部署使用。在武汉光谷,系统上线以来,共解决14900条法律问题,矛盾调处2343次,普法内容传播729.8万次,矛盾纠纷立案数较前一年下降28%。

在AI+农业领域,基于“小可”多智能体协同架构研发“穗问”农业大模型2.0版,整合海南气象数据、作物生长数据库、千种病虫害特征图谱,通过智能传感器和物联网设备,智能灌溉硬件联动,为农户提供精准种植指导。农户上传的病虫害照片,可实现秒级识别,准确率高达95%,同步生成用药方案,一键直连电商平台,决策效率大幅提升。

采用。OpenGMSL以此坚实基础为依托,将推动自动驾驶、ADAS及新一代信息娱乐技术加速创新,助力现有蓬勃发展的生态系统迈向开放协作的未来。

OpenGMSL标准专注于高清视频和高速数据传输,为智能汽车的关键应用提供稳定、低延迟的连接方案。联盟作为一个非营利性组织,将由独立董事会管理,并欢迎全球行业伙伴参与。为确保不同供应商产品的兼容性,采

用OpenGMSL标准的产品需通过强制性合规测试。

ADI公司作为联盟发起方之一,强调OpenGMSL将帮助汽车制造商利用多供应商资源,打造更具成本效益、灵活性和创新性的智能汽车解决方案。

随着汽车智能化进程加速,OpenGMSL联盟的成立有望为行业提供统一的技术框架,推动自动驾驶和智能座舱技术的进一步发展。(姬晓婷 卢沫男)

OpenGMSL联盟提升多供应商生态系统互操作性

本报讯 近日,多家全球领先的汽车原始设备制造商(OEM)、一级供应商、半导体制造商及生态系统合作伙伴共同宣布成立OpenGMSL联盟,旨在推动千兆多媒体串行链路(GMSL)技术成为全球汽车生态系统的开放标准,以促进自动驾驶、高级驾驶辅助系统(ADAS)及信息娱乐等领域的创新。

该联盟将基于亚德诺半导体公司(ADI)GMSL SerDes技术,制

定统一的开放标准,使OEM和供应商能够采用高效协同的解决方案,缩短产品开发周期、降低成本,并提升多供应商生态系统的互操作性。

OpenGMSL联盟主席Paul Fernando表示,GMSL是汽车行业最成熟且经过路测验证的高速视频链路技术之一。基于GMSL SerDes技术的芯片的出货量已超过10亿颗,获得全球超过25家OEM车厂和50家一级供应商的