小型语言模型:AI领域的新热点

多年来,谷歌等科技巨头和OpenAI等初创公司,一直在不遗余力地利用海量在线数据,打造更大、更昂贵的人工智能(AI)模型。这些大型语言模型(LLM)被广泛应用于ChatGPT等聊天机器人中,帮助用户处理各种各样的任务,从编写代码、规划行程,到创作诗歌等。

自ChatGPT面世以来,AI模型便在变大、变强之路上"狂奔"。但喧嚣过后,科技公司也越来越关注更小、更精简的小型语言模型(SLM)。他们认为,这些小巧玲珑的模型不仅"术业有专攻",而且部署成本更低廉、更节能。

未来,这些规模不一的AI模型将协同工作,成为人类的左膀右臂。

小型语言模型在简单的专业领域可能更有优势。

小型模型独具优势

随着AI技术突飞猛进,AI模型的"块头"与日俱增。ChatGPT的缔造者OpenAI去年夸耀称,其GPT-4模型拥有约2万亿个参数。参数表示AI模型的大小,一般参数越多,AI模型的能力越强,庞大的参数量使GPT-4成为迄今最强大的AI模型之一,能回答从天体物理学到动物学等多领域包罗万象的问题。

但是,如果某家公司只想借助AI模型解决特定领域(如医学)的问题,或者一家广告公司只需一款AI模型来分析消费者行为,以便他们更精准地推送广告,GPT-4这类模型就有点"大材小用"了,SLM反而更能满足用户们的要求。

美国《福布斯》双周刊网站在11月的报道中,将SLM称为AI领域的"下一个大事件"。

微软公司生成式AI副总裁塞巴斯蒂安·布贝克表示,虽然SLM的参数量目前并没有统一标准,但大约在3亿到40亿个之间,小巧到可以安装在智能手机上。

专家声称,SLM更胜任简单的任务,如总结和索引文档、搜索内部数据库等。

法国初创公司LightOn的负责人劳伦特·都德认为,与LLM相比,SLM拥有诸多优势:首先,这些模型的反应速度更快,能同时响应更多查询,回复更多用户;其次,SLM部署成本更低,能源消耗也更少。

都德解释道,目前很多LLM需要大量服务器来进行训练,然后处理查询。这些服务器由尖端芯片组成,需要大量电力来运行,并进行冷却。而训练SLM所需芯片更少,运行耗费的能源也更少,这使其更便宜、更节能。

SLM还可直接安装在设备上,在不依赖数据中心的情况下运行,这能进一步确保数据的安全性。《福布斯》表示,SLM能以最少的计算资源执行各种任务,使其成为移动设备、边缘设备等的理想选择。

AI模型掀起"极简风"

谷歌、微软、元宇宙平台公司以及OpenAI等公司闻风而动,推出了各种SLM。

去年12月底,微软公司正式发布了只有27亿个参数的语言模型Phi-2。微软研究院在其X平台官方账号上表示,Phi-2的性能优于现有其他SLM,且能在笔记本电脑或移动设备上运行。今年4月,微软又推出了只有38亿个参数的Phi-3系列模型。

今年8月,微软公司再接再厉,推出了最新的Phi-3.5-mini-instruct。这款SLM为高效、先进的自然语言处理任务量身打造。9月,英伟达公司开源了Nemotron-Mini-4B-Instruct。该公司表示,这款SLM特别适合边缘计算和设备端的应用。报道称,这两款SLM在计算资源使用和功能表现之间实现了良好平衡。在某些方面,其性能甚至可媲美LLM。

OpenAI也不甘示弱。今年7月,OpenAI公司发布了GPT-4o mini,称其是该公司最智能和最实惠的SLM。

此外,亚马逊公司还允许在其云平台上使用各种规模的AI模型。

其他公司也纷纷开发更适合自身需求的SLM。例如,美国制药巨头默克公司正与波士顿咨询集团(BCG)合作开发一款SLM,旨在探究某些疾病对基因的影响。这将是一款参数介于几亿到几十亿之间的AI模型。

大小模型作用互补

虽然SLM在效率等方面具有独特优势,但LLM在解决复杂问题、提供更广泛的数据访问方面仍然具有极大优势。

展望未来,LLM和SLM两种模型将"是朋友而非对手",它们之间的协作交流将成为主流趋势。

当遇到用户提出的某个问题时,一款SLM会"一马当先",理解这个问题,再根据问题的复杂性,将相关信息发送给几个大小不一的AI模型。这些模型"群策群力""并肩携手"为用户解决难题。

目前市面上的AI模型要么太大、太贵,要么处理速度太慢。两者合作,或是最佳解决方案。

来源:科技日报

土星环为何"驻颜有术"

土星,这位太阳系中的"珠宝大师",其最引人注目的"作品"无疑是那条璀璨夺目的冰环。长久以来,科学家对这条冰环的年龄充满了好奇。人们认为,它应与土星一同诞生,拥有着45亿年的悠久历史;然而,它那令人惊讶的洁白无瑕,看上去却最多不过4亿岁。外表与实际40亿岁的"年龄差",似乎在诉说着一个不同的故事。

按照常理,随着岁月流逝,无数微小的太空岩石,会不断撞击土星环中的冰块,逐渐给它染上一层"风霜之色",让土星环变得暗淡无光,就像是一幅古老的画作因岁月侵蚀而褪色。然而,当卡西尼号探测器于2004年抵达土星时,它所见到的却是一幅几乎未受岁月侵扰的美丽景象。这使许多天文学家推测,土星环可能非常年轻,不会超过4亿岁。

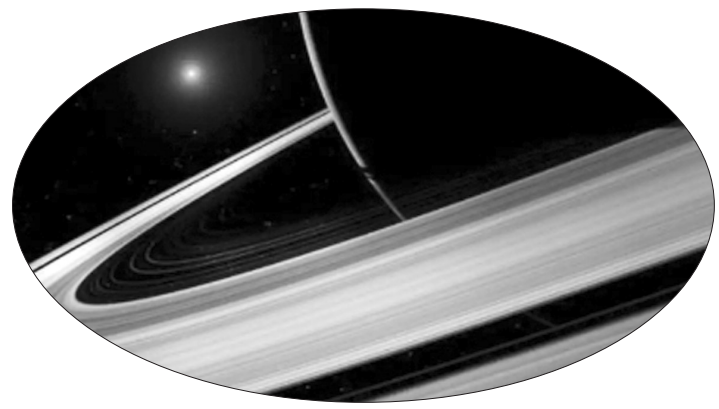
但最近,《自然·地球科学》杂志上发表了一项研究,为人们揭示了一个全新的视角。研究团队由日本地球生命研究所牵头,他们利用先进的计算机模拟技术,重新审视了微流星体与土星环之间的相互作用。模拟结果显示,当微流星体以极高速度撞击冰环时,并不会简单地留下痕迹,而是会瞬间蒸发成气体。这些气体随后在土星强大的磁场作用下,发生了奇妙的变化——它们膨胀、冷却,最终凝结成为带电的纳米粒子和离子。

这些带电粒子的命运相当戏剧化:一部分被土星的强大引力吸引,坠入其大气层;另一部分则挣脱土星束缚,飞向浩瀚的宇宙深处。因此,真正能够留在土星环上的"污渍"少之又少。这也就解释了为什么土星环能够保持如此纯净的白色,仿佛从未受到过外界的打扰。

这项研究不仅挑战了人们对土星环年龄的传统认知,还提出了一个可能性:即土星环其实与土星同龄,拥有数十亿年的历史。而这种独特的清洁机制,或许不仅仅适用于土星,在天王星和海王星的环系统,乃至巨行星周围的冰卫星上,也可能上演着同样的故事。

这让人们再次见证了宇宙中那些看似不可能的现象背后,所隐藏着的自然界巧妙设计。土星环的故事,就像是一个关于时间、空间和物理定律交织而成的美丽传说,等待着我们去探索和理解。

来源:科技日报



我国科学家建立生成式模型为医学AI训练提供技术支持

记者从北京大学未来技术学院获悉,北京大学与温州医科大学的研究团队建立一种生成式多模态跨器官医学影像基础模型(MINIM),可基于文本指令以及多器官的多种成像方式,合成海量的高质量医学影像数据,为医学影像大模型的训练、精准医疗及个性化诊疗等提供有力技术支持。该成果已于近期在国际权威期刊《自然·医学》上在线发表。

医学影像大模型是利用深度学习和大规模数据训练的AI通用模型,可自动分析医学影像以辅助诊断和治疗规划。但要提升大模型的性能,就需要大量数据不断进行训练。然而,由于患者隐私保护、高昂的数据标注成本等多种因素,要获得高质量、多样化的医学影像数据往往存在障碍。为此,近年来,研究者们开始探索使用生成式AI技术合成医学影像数据,以此来扩充数据。

"目前公开的医学影像数据非常有限,我们建立的生成式模型有望解决训练数据不够的问题。"北京大学未来技术学院助理研究员王劲卓说,研究团队利用多种器官在CT、X光、磁共振等不同成像方式下的高质量影像文本配对数据进行训练,最终生成海量的医学合成影像,其在图像特征、细节呈现等多方面都与真实医学图像高度一致。

实验结果显示,MINIM生成的合成数据在医生主观评测指标和多项客观检验标准方面达国际领先水平,在临床应用中具有重要参考价值。在真实数据基础上,使用20倍合成数据在眼科、胸科、脑科和乳腺科的多个医学任务准确率平均可提升12%至17%。

王劲卓表示,MINIM产生的合成数据具有广泛应用前景,可单独作为训练集来构建医学影像大模型,也可与真实数据结合使用,提高模型在实际任务中的性能,推动AI在医学和健康领域更广泛应用。目前,在疾病诊断、医学报告生成和自监督学习等关键领域,利用MINIM合成数据进行训练已展现出显著的性能提升。

来源:新华网