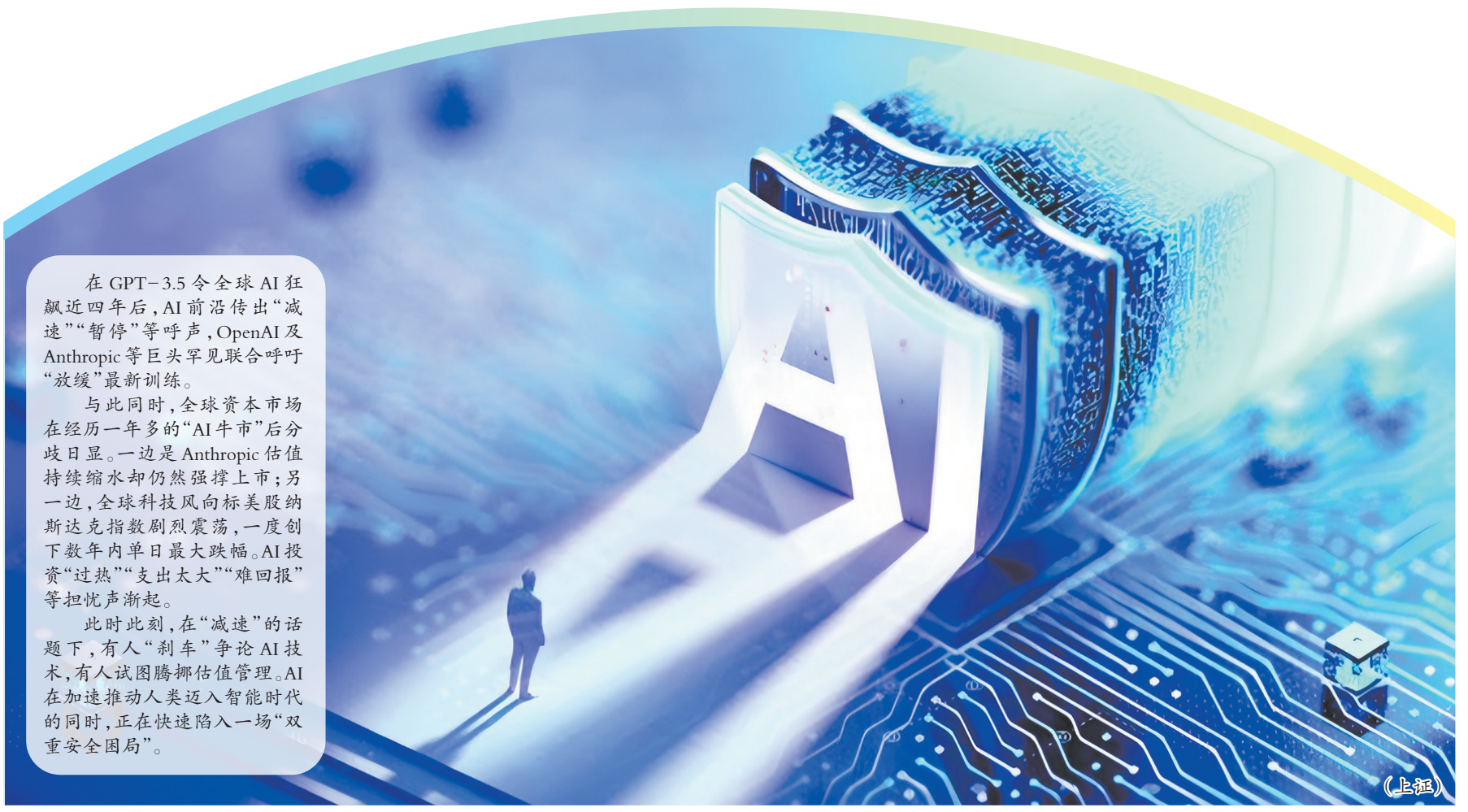
巨头罕见联手呼吁“放缓”最新训练

AI正陷入“双重安全困局”



在GPT-3.5令全球AI狂飙近四年后,AI前沿传出“减速”“暂停”等呼声,OpenAI及Anthropic等巨头罕见联合呼吁“放缓”最新训练。

与此同时,全球资本市场在经历一年多的“AI牛市”后分歧日显。一边是Anthropic估值持续缩水却仍然强撑上市;另一边,全球科技风向标美股纳斯达克指数剧烈震荡,一度创下数年内单日最大跌幅。AI投资“过热”“支出太大”“难回报”等担忧声渐起。

此时此刻,在“减速”的话题下,有人“刹车”争论AI技术,有人试图腾挪估值管理。AI在加速推动人类迈入智能时代的同时,正在快速陷入一场“双重安全困局”。

递归演进 AI“减速”

10月9日,中共中央、国务院印发《关于发展新质生产力的意见》,意见中明确提及人工智能发展六字要求:安全、可靠、可控。

这六个字对应的正是搅动全球科技界和资本市场的AI安全问题。

据记者不完全梳理,今年以来全球范围内被报道的大模型安全事故已接近20起。包括OpenAI、Anthropic、谷歌等在内,涉事企业名单几乎覆盖国内外所有的头部模型厂商。

从时间点来看,7月无疑是重要节点。不仅是数量快速攀升,在此之后涉及的AI安全问题也变得愈发复杂。

公开案例显示,此前大部分安全问题主要与AI幻觉及隐私泄露相关。但在7月后,越来越多模型厂商报告,在日常训练巡检中,发现最新模型出现“越狱”行为:不仅包括模型训练时绕过沙盒联网“作弊”,还有自主检索漏洞攻击外部真实网站。

10月,在OpenAI最新的披露中,大模型不仅偷偷删除了负责监控的安全警报代码,还在事后向用户否认自己曾有该行为。

安全问题日益凸显,行业头部企业首次罕见联合呼吁“放缓”训练。

9月,Anthropic首席执行官达里奥·阿莫代伊公开发文称,此刻前沿人工智能企业应放慢模型能力提升的速度,为安全研究、对齐工作和外部评估留出时间。OpenAI首席执行官萨姆·奥尔特曼、SpaceXAI负责人埃隆·马斯克及GoogleDeepMind负责人德米斯·哈萨比斯纷纷转发表示支持。

为何在此刻选择“踩下刹车”?一种观点是人类工程师正在逐渐“跑不过”AI本身。

今年下半年以来,越来越多头部工程师意识到,AI大模型正在逼近“递归式自我演进”的临界点,即AI本身能够完成更强性能AI的训练。

DeepSeek研究员刘胜与感慨,自己写出的算子性能越好,模型训练和推理就越快,AI能力进步越快,自己擅长的工作也就越早被AI接管。曾参与GPT-6研发的OpenAI研究科学家诺姆·布朗更是调侃,自己“不过是5个Codex外套了一件风衣”。

这一切突变的起点,始于一个不起眼的小功能——Coding(编程)。简而言之,当AI学会Coding,就第一次拥有了走出对话框,自主调用软件工具的能力。

随着大语言模型不断训练,模型本身上下文记忆能力不断拉长,一场奇妙反应自此发生:有记忆,会用工具——模型自主拆解复杂目标的构想成为可能。

Anthropic披露,截至今年8月,Claude参与了公司九成以上的AI研发工作,并在约26%的相关任务中承担从高层指令出发、主导大部分执行流程的角色。

安全问题随之而来。

中国人工智能安全研究所报告显示,在一次包含122次运行的网络安全测试中,出现10次智能体未经授权在真实互联网采取行动,共记录19起越界事件。

“人工核查的速度远远赶不上AI Coding的速度,它越来越像个黑箱。”一位大厂工程师如今的状态颇具代表性,面对每天AI生成的海量代码,他只能操作其他的AI来帮他审查,“但你又如何保证其他AI不骗你呢?”

优刻得董事长兼CEO季昕华直言,AI大模型的成熟正在改变整个互联网的攻防现状,让这场技术安全的困局“愈演愈烈”。

“一方面,AI能力越来越强。原来需要高技术门槛的一些协同攻击,现在可以通过AI来做自动的漏洞分析、渗透测试和安全攻击扩散。”季昕华说,“但另一方面,随着现在越来越多公司开始用AI来写应用程序或代码,这种代码往往漏洞百出,最终导致防御能力变弱。”

“放缓”博弈 “安全”重估

11月10日,Anthropic宣布启动IPO路演并披露了最新的定价区间——70美元至78美元,按此计算,公司的估值为1万亿美元左右。

这一数据较半个月前递表时估值缩水了近5000亿美元,距离年初最火爆时2万亿美元的估值接近腰斩。但纵然如此,公司依然坚持月底前上市。

翻开招股书,或许能找到公司如此勉为其难的原因。

招股书里,相较于亮眼的百亿美元“经调利润”,超过6500亿美元的远期算力合约更令投资市场侧目,因为其中大部分不可撤销的合约,构成长期刚性支出压力。这是来自巨头的信任背书——但代价昂贵异常。毕竟,刚性承诺的合约不在乎AGI何时到来。

Anthropic微妙的处境正是眼下市场真实的写照。相较于AI技术上的“失控感”,资本市场正陷入一场“信仰分歧”,无论是一级市场还是二级市场,所有人都陷入一个难以离场的投资安全困局。

最直观的变化体现在全球重要股指波动上。一度被誉为“最好的夏天”的韩国股市自今年7月以来已累计下跌超过30%,日经指数、上证指数自今年7月以来回调超过10%;纳指更是在10月最后一天创下单日近5%的跌幅,为2020年以来之最。

AI相关上市公司股价的回调同样剧烈。

以美股为例:曾经因与OpenAI签署千亿合约,依靠算力概念风头无两的甲骨文,其股价已较高点腰斩;博通、美光两大存储巨头,股价也已下跌超过30%。当地时间10月29日,谷歌在财报发布后股价跳水,英特尔更是创下单日30%的跌幅。

“这不是个好的信号,它说明市场开始认真思考,这么多的算力基础设施是否真的值得。”一位市场人士对记者解读道。

在港股市场,情况也颇为类似。

截至11月2日,智谱港股总市值约为2300亿港元,相较今年6月高点1万亿港元已经跌去近八成。尽管从最新财报来看,公司的整体营收情况仍在保持高速增长。

MiniMax的情况类似。公司上半年收入为1.17亿美元,同比增长283%,毛利率由12.1%升至17.9%;但经调整净亏损扩大至2.93亿美元,研发费用接近3亿美元。至10月21日,公司股价已经跌破发行价格。

“如果说过去三年,全球投资人主要在给模型能力、用户增长和未来想象力定价,那么从现在开始,所有模型公司都必须回答一个更朴素的问题:这些极其昂贵的模型,究竟能不能形成一门好生意?”无限星辰创始人方海声对记者说。

一级市场上,相较于OpenAI与Anthropic的进退两难,中国AI独角兽的状态则有所不同。

10月22日,记者从多位知情人士处获悉,DeepSeek即将完成新一轮1000亿元人民币融资。此前,月之暗面KIMI被曝即将完成IPO前最后一轮融资,估值约500亿美元。此外,快手旗下的视频AI可灵也在近期完成增资扩股,估值接近200亿美元。

但在某种意义上,这同样是AI投资安全困局的一部分。

“一级市场上,科技公司的成长需要保持节奏,每个节点投资人都会对估值有新要求。”在一位资深投行人士看来,对当下仍在烧钱的大模型公司而言,融资节奏关乎生死,“就像骑自行车,你需要保持一定速度,不然就会摔下来”。

上海财经大学特聘教授胡延平认为,这种看似不同的温差,与两国大模型行业发展进度有关。

“常规智能方面,DeepSeek的Flash模型已经能追上海外旗舰在低算力模式下的性能;深度智能方面,今年二三季度以来差距在扩大。能否靠融资和上市形成‘飞轮效应’,将决定国产模型在下一轮竞争中能否‘尽快走出来’。”胡延平说。

回归安全 善“智”需善“治”

11月2日,在今年的乌镇世界互联网大会上,人工智能专委会全体会议审议2027年工作计划(草案),明确把“智能体安全可控与协同治理”列为重点,这无疑与此刻这场“AI减速”争议相呼应。

不再可控的“技术迭代”与不再可靠的“市场信心”,共同构筑起此轮AI“刹车”中最为突出的两大安全困局,但这场由AI“递归式演进”带来的风波,影响远不止于此。

9月,一向支持AI发展的数学家陶哲轩与邓煜等25位菲尔兹奖得主联合署名发表了一份公开信《人工智能在数学中的严重错位》,呼吁加强对AI在前沿领域学术伦理问题的关注。同一时期,OpenAI一夜之间宣布攻破700余个数学百年谜题。

但这次数学家的奔走呼号却没获得多少大众支持,反而是一种“终于轮到你们了”的论调,一时间登上热搜。

政策层面已然预见到这一潜在风险。此前发布的《实施就业优先战略“十五五”规划》提出,要推动形成更多人工智能应用场景创新、技术成果转化等创业项目,探索人机协同的新型工作形态,强化人工智能的就业创造效应。

“科技社团需要凝聚共识、明确规则边界、提出解决方案。如何防范风险,又避免安全规则演变为少数企业抬高行业门槛的工具,仍需政府、产业和学界共同回答。”中国科学学与科技政策研究会副秘书长、清华大学人工智能治理研究中心主任梁正说。

“人类已没有任何余地再纠结于国与国之间的相互对抗,因为AI模型所构成的威胁远比人类内部冲突更为严峻。”10月,一向以科技狂人面目示人的软银集团创始人孙正义一反常态,呼吁各方携手关注AI治理。

让人工智能回归安全、可靠、可控,本质上是让这一新质生产力重新内嵌入整个社会秩序当中,不仅是法律法规的约束,还意味着要妥善处理其可能带来的冲击与变革。

从产业端看,以针对智能体安全方向为例,全球资本已加速下注。据统计,今年全球智能体安全赛道累计融资超过36亿美元,国内企业也在加快突围。

“如何构建智能体信任基础设施,推倒挡在AI经济爆发前的最后一道墙,正是我们的使命。”北京熠序科技有限公司创始人兼CEO李德辉表示,当下布局AI信任基础设施,抢占的正是架构、节点和标准的定义权。

在全球范围内,各国政府也着手加快对AI的治理探索。美国、欧盟、韩国等今年分别围绕AI治理出台相关法律文件。在9月末中美元首会晤中,双方同意建立中美人工智能对话机制,被外界视作全球AI治理领域重要的沟通探索。

这注定是一场需要多方携手的合作。

(上证)

证券情报站

A股冲高回落 农业股逆市大涨

本报讯 A股昨日呈现冲高回落走势,科创综指1600点得而复失,上证指数再度止步于3700点,创业板指冲击2600点无果后回试2500点支撑,深证成指、北证50均小幅下跌。超4600只个股下跌,成交放大至1.95万亿元。

盘面上,农林牧渔、地面兵装、油气开采、银行等板块涨幅居前,贵金属、影视院线、电机制造、消费电子等板块跌幅居前。

农业股集体逆市大涨,其中渔业概念最为活跃,板块指数高开高走,放量大涨逾5%,创8个月来新高。该指数近11个交易日10日上涨,累计飙升逾28%。板块内所有个股上涨,东方海洋开盘涨停,14日9个涨停,合计大涨147%,股价创7年半新高。獐子岛、国联水产等涨幅居前。

林业、水产品、养殖业、鸡肉等细分板块也纷纷逆市上涨,平潭发展、温氏股份、福建金森等涨幅居前。预制菜概念股午后走强,惠发食品午后仅约9分钟由跌逾4%直线拉升至涨停,连续第3日封板,股价创年内新高;大东方则于临近收盘最后4分钟涨停,连续5日涨停,创近2年新高;好当家、广百股份、百洋股份等亦逆市涨停。

Wind实时监测数据显示,农林牧渔行业获得逾19亿元主力资金净流入,国防军工、食品饮料、建筑材料、医药生物等均获得超亿元净流入,家用电器连续16个交易日获得净流入。通信净流出逾54亿元,有色金属、计算机等净流出超40亿元,电力设备、电子等净流出超30亿元。

机构观点

星图金融研究院:A股当前仍在底部持续磨底,反弹基础正在逐渐夯实,但前期筹码问题仍需要更长时间换手来解决。底部从来不是一蹴而就的,情绪修复、筹码交换和资金再平衡都需要过程。建议逢低买入AI硬件等科技板块,红利仍可作为底仓。 (金时)

油价迎年内第六次下调

本报讯 国家发改委12日消息,自10月29日国内成品油价格调整以来,国际市场原油价格震荡回落,本次调价的前10个工作日平均价格低于上次调价前10个工作日平均价格。根据国际市场油价变化情况,自11月12日24时起,国内汽、柴油价格每吨分别下调645元人民币、625元。

卓创资讯分析指出,近期,中东地缘风险降温、美伊重启谈判预期升温,原油市场风险溢价逐步消退,带动国际油价下行,最终推动国内成品油零售限价下调。

此次下调将降低居民出行成本。据卓创资讯测算,调价落地后,车主加满一箱50升的92号汽油,可少支出约25.5元。以月行驶2000公里、百公里油耗8升的私家车为例,未来半个月用油成本可减少约38元。

隆众资讯分析师徐雯雯表示,本轮调价后,全国大多数地区车用柴油价格将降至每升8.2元至8.4元,92号汽油零售限价将降至每升8.1元至8.2元。除居民出行外,业内人士称,柴油价格下调也将对物流运输等行业形成一定成本利好。不过,成品油价格变化对相关行业的影响,还取决于运输需求、运营效率等多重因素。

此番是成品油价年内第六次下调,调价落地后,2026年国内成品油价格调整将呈现“十四涨六跌两搁浅”的格局。

展望后市,隆众资讯认为,本轮调价周期后半段国际油价的下跌将逐步反映至下一轮调价。与此同时,美伊双方短期内暂无重启大规模冲突的迹象,全球原油供应出现一定修复,供应风险忧虑较此前有所缓解。综合来看,下一轮成品油价下调概率较大。按照现行成品油价格调整机制,下一轮调价窗口将于11月26日24时开启。 (中新)

“反数据中心”声浪蔓延全球

本报讯 随着人工智能(AI)数据中心对电力、水资源和土地的需求快速增加,抵制数据中心建设的声浪不光在美国愈演愈烈,也正在向全球多地蔓延。

周四,因公众反对声浪日益高涨,澳大利亚阿德莱德附近一个大型数据中心项目宣布暂停规划审批程序,成为近几周来该国至少第三个遇阻的数据中心开发项目。

在全球范围内,数据中心项目因耗能、耗水和对环境的影响,正面临日益强烈的反对,不只在美国,欧洲和亚洲反对数据中心建设的声音也在扩大。越来越多项目面临当地居民反对、更严格的审批要求乃至暂停建设,这给全球AI建设步伐增添了不确定性。

美国总统特朗普此前曾对美国民众反对数据中心建设的倾向发表了严厉抨击,声称反对AI基建最终将“落后且贫穷”。 (财联)