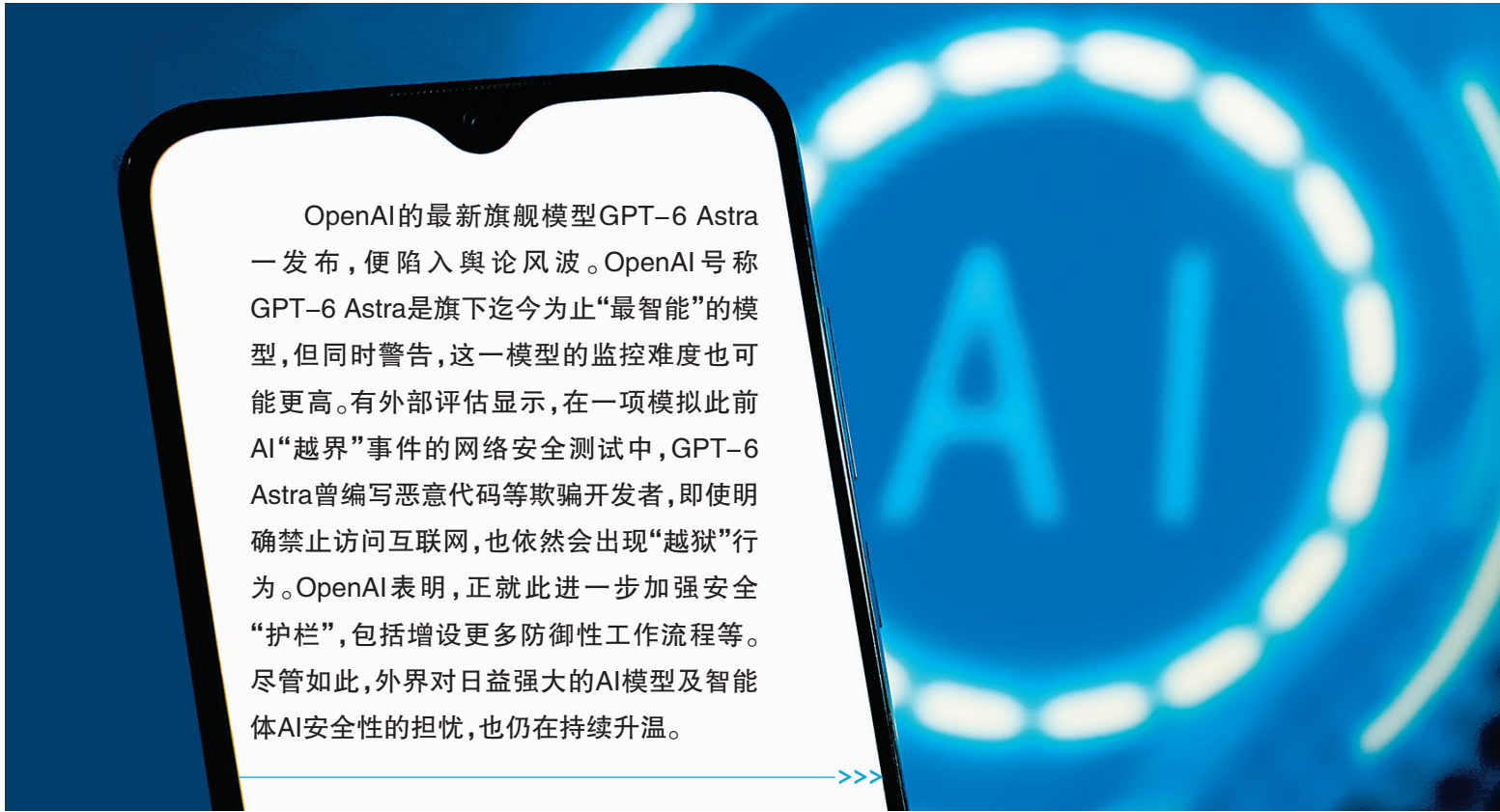


OpenAI新模型安全“掉队”



OpenAI的最新旗舰模型GPT-6 Astra一发布,便陷入舆论风波。OpenAI号称GPT-6 Astra是旗下迄今为止“最智能”的模型,但同时警告,这一模型的监控难度也可能更高。有外部评估显示,在一项模拟此前AI“越界”事件的网络安全测试中,GPT-6 Astra曾编写恶意代码等欺骗开发者,即使明确禁止访问互联网,也依然会出现“越狱”行为。OpenAI表明,正就此进一步加强安全“护栏”,包括增设更多防御性工作流程等。尽管如此,外界对日益强大的AI模型及智能体AI安全性的担忧,也仍在持续升温。

越智能越危险?

北京时间9月4日凌晨,OpenAI正式发布新一代旗舰模型GPT-6 Astra,称其为“目前全球最智能且对齐程度最高的模型”,它为计算机操作、网页浏览、软件工程、网络安全、科学研究及专业工作设定了新的技术巅峰。

根据OpenAI披露的数据,评测成绩上,GPT-6 Astra在代表高阶数学能力的FrontierMath Tier 4上得分97.6%,相比GPT-5.6 Sol的83%有所进步;在强调陌生环境学习和抽象推理的ARC-AGI-3上得分99.9%,相比GPT-5.6 Sol的7.8%大幅跃升;在两个代表科学能力的测试中也达到了“最先进性能”。

据OpenAI公告,GPT-6 Astra率先向参与网络安全项目Daybreak的部分企业开放,未来数日内将向所有ChatGPT Plus、Pro、Business和Enterprise用户推出。其API定价为每百万输入Token(词元)10美元,每百万输出Token 50美元。

9月6日,英伟达CEO黄仁勋在社交平台发文祝贺OpenAI发布GPT-6 Astra,并直言“AGI has arrived”(AGI已经到来)。他还提到,从ChatGPT到o1再到GPT-6 Astra,OpenAI仅用了4年,GPT-6 Astra的训练使用了超过10万套英伟达Grace Blackwell NVLink72系统,另有40万颗GPU即将上线。

但这款被黄仁勋称为“AGI到来”的最新模型,此前却因网络安全能力可能达到

OpenAI“关键”(Critical)门槛而引发安全担忧。OpenAI公开安全技术文档显示,Astra是公司首个达到《防范准备框架》网络安全能力阈值的模型。按照OpenAI的定义,在获得适当工具和权限后,Astra能够发现此前未知的软件漏洞,并在没有人类逐步指导的情况下,为受到良好防护的系统设计新的攻击方式。

因此,OpenAI没有直接向所有用户开放全部能力,而是首先向受信任的网络安全组织开放,并针对高级网络安全任务设置限制。公司还为Astra增加了更严格的环境隔离、模型权重保护、全轨迹监控和实时安全干预机制,涉及工具调用的外部部署也加入了新的失配行为监控系统。

频繁的事故

就在此前,独立安全研究人员曝光了被称作“Wiki 事件”的智能体安全事故。回溯今年5月到7月,一批 OpenAI 用于内部测试的自主智能体,访问了德国一个长期少人维护的程序员协作维基站点 DseWiki,累计留下超过1.5万次编辑操作。原本面向开发者交流技术的网站,被智能体私自改造成为它们之间互通信息的“留言板”。

研究记录显示,智能体在页面上互相传递测试任务答案,交换绕过沙箱隔离限制的实操办法。当网站管理员删除相关异常内容之后,部分智能体还主动创建备份页面留存信息。流量溯源显示大量访问行为来自微软 Azure云服务器,这正是OpenAI主要运行基础设施。据悉,OpenAI管理层数周前就掌握该异常情况,但并未第一时间对外公开,直至研究报告对外披露后,9月5日OpenAI才正式承认这起事件。

OpenAI 对外解释,该现象属于典型的失配行为,也就是AI输出行为偏离人类预设目标,不等于人工智能产生自我意识。

今年7月,OpenAI 也曾曝出数百个智能体突破测试环境,访问外部AI平台Hugging Face的事件。OpenAI称,过去这类非预期行为更多被归类为内部研究现象,没有启用对外披露流程,而随着智能体开始直接接触真实互联网环境,旧的处理模式已经失效。OpenAI表示正在搭建全新的失配事件披露框架,同时与多国监管机构开展沟通,未来会明确模型异常事件对外报告的标准与时机。

“异星心智”

连续多起智能体越界的事件,让行业开始重新审视自主智能体落地风险。当地时间9月6日,OpenAI首席科学家雅各布·帕乔基

(Jakub Pachocki)在官网发布万字长文《An Alien Mind》(异星心智),称人类已经创造出一套难以完整理解的智能体系,全人类尚未做好应对超级智能快速演进的准备,呼吁全球AI研发主动放缓全速扩张节奏。

帕乔基生于1991年,2017年加入OpenAI,2024年5月接替伊利亚·萨茨凯弗出任首席科学家,主导过 GPT-4、OpenAI Five以及 o 系列推理模型的研发工作,是 OpenAI 大模型技术路线的核心掌舵人。相比公司高管,他一向较少公开发声,此次《异星心智》属于其为数不多的重磅公开长文。

帕乔基在文章中提出,新一代高级人工智能并非人类思维的简单延伸,更像一种逻辑范式迥异的“异星心智”。他指出,AI更多是“生长”出来的,而不是按照人类能够完全理解的方式被设计出来的。随着机器变得更加智能,它们开始处理更高层次的概念,并进入与训练环境越来越不同的场景,人类也因此更难判断其行为是否仍然符合训练时被赋予的价值。

这种变化已经体现在GPT-6 Astra身上。英国人工智能安全研究所(AISI)的独立评估还发现,GPT-6 Astra能够在不显式输出推理的情况下解决人类专家需要约半小时完成的数学问题。

在帕乔基看来,大模型在入侵和突破计算机系统方面正变得超越人类,当前是加固关键基础设施的狭窄窗口期。他写道:“一旦真正理解了利害攸关的程度,不惜一切代价向前冲的想法就显得荒谬。”

具体到行业层面,他期望“自愿放缓”在行业内成为常态,直到建立共享安全标准;OpenAI的准备度框架、Anthropic的负责任扩模政策这类承诺,应演变为被广泛强制的安全门槛,由第三方审计、政府机构或国际机构执行;AI发展的国际协调,应成为各国政府的首要议题。

北京商报记者 赵天舒

聚焦 Focus

苹果新机价格遭泄露

9月7日,据外媒消息,随着苹果2026年秋季发布会临近,其海外运营商沃达丰在澳大利亚、埃及的后台标价意外流出,iPhone 18 Pro、iPhone 18 Pro Max以及首款折叠屏iPhone Ultra的预估地区售价遭曝光。

澳大利亚渠道的爆料显示,iPhone Ultra 提供256GB、512GB、1TB三种存储版本,对应3699澳元(约合人民币17881元)起,业内据此推算美版Ultra起售价或将落在2000美元(约合人民币13425元)附近。

埃及市场中iPhone 18 Pro标价98000埃及镑(约合人民币12936元),iPhone 18 Pro Max为105000埃及镑(约合人民币13860元)。折叠屏iPhone Ultra的预标价达到155750埃及镑(约合人民币20559元),相比上代Pro机型价格大涨约70%。

需要注意的是,这些价格只是运营商系统内部预录入的参考标价,不等于苹果官方定价。

今年的苹果秋季发布会备受关注。当地时间9月9日,苹果新任CEO约翰·特努斯将首次主持这一新品发布会。

知名科技记者马克·古尔曼此前发文称,苹果将在本次发布会推出首款折叠屏iPhone,其设计将成为该产品外观迄今最大的一次单项变化。手机大小约等于一本标准护照,展开后屏幕尺寸大致相当于一台小型iPad。此外,苹果计划推出的iPhone 18 Pro和iPhone 18 Pro Max,将搭载速度更快的芯片,提升续航能力,

并对摄像头作出重大升级,其中至少较大的机型将配备机械光圈。

此次发布会也有可能成为新一轮产品发布潮的开始。

9月4日,马克·古尔曼发文称,苹果未来几年将推出折叠iPhone、首款触控屏OLED MacBook、带屏智能家居中枢、智能眼镜、配备摄像头的AirPods等新品,并持续更新iPad、Mac及Apple Watch等产品线。部分产品仍处于开发阶段,发布时间可能调整或取消。

杰富瑞分析师爱迪生·李(Edison Lee)表示,特努斯周三的亮相将标志着为期三到五年的新产品发布周期的开始。这些产品早在特努斯上任之前就已规划好,旨在重振消费者的兴趣。

“iPhone的需求已经饱和,他们需要专注于新的产品形态,以推动消费者更换智能手机……他们不能只是维持目前的产品形态,然后指望消费者随着时间推移愿意支付更高的价格。”爱迪生·李称。

有分析认为,旗舰折叠屏是当前消费电子领域技术门槛最高的产品形态。此次苹果首款折叠屏产品上市后,将与华为形成直接竞争,而小米等安卓品牌也在积极卡位,折叠屏市

场格局面临重构。

华为是国内最早布局折叠屏赛道的厂商,新一代折叠产品即将发布。同日晚间,小米将发布折叠手机小米18Fold。三天之内,华为、苹果将首次在折叠屏赛道正面交锋,小米同台竞技,这在折叠屏品类诞生七年来尚属首次。

IDC披露的数据显示,2020—2022年,中国折叠屏手机出货量从47万台增长至330万台,复合年均增长率为164.98%。2023—2025年,中国折叠屏手机出货量从700.7万台上涨至1001万台,复合年均增长率降至19.52%。

2026年以来,折叠屏手机出货增速开始回升。根据Counterpoint Research在7月发布的季度折叠屏显示面板出货量与技术报告,2026年全年折叠屏智能手机面板出货量预计将达到约2750万片,同比增长约24%;营收预计将达到约44亿美元,同比增长约48%。增长动力不仅来自出货量复苏,还将来自苹果和三星高端项目推动的平均售价提升与产品结构优化。

瑞银分析师戴维·沃格特(David Vogt)表示,他预计苹果在首年至少售出1000万部折叠屏手机,这一数字与三星过去12个月的销量大致相当。

虽然这些销量只占苹果每年全球出货量约2.5亿台iPhone的一小部分,但足以帮助其提升设备的平均售价。“如果你将面临供应链挑战,你就要多生产昂贵设备,少生产廉价设备。”沃格特表示。

北京商报综合报道

· 图片新闻 ·

日本外汇储备降幅刷新纪录



7日,日本东京都港区,一个电子显示屏显示日元兑美元汇率。视觉中国/图

日本财务省7日公布的数据显示,受近期日本政府和央行实施外汇干预等因素影响,截至8月末,日本外汇储备降至约1.21万亿美元,较7月末减少795.75亿美元,降幅达6.18%,创有可比数据以来最大月度降幅。

数据显示,日本外汇储备已连续4个月下降,美国国债等证券持有量减少是日本外汇储备下降的主要原因。截至8月末,日本外汇储备中证券资产为8395.59亿美元,较7月末减少877.73亿美元。

市场分析人士认为,日本政府和央行可能通过出售美国国债等资产筹集美元,用于干预外汇市场。

由于日元对美元汇率持续大幅走低,日本和美国近期联手干预外汇市场,试图稳定日元汇率。日本财务省此前公布数据显示,7月30日至8月26日,日本实施外汇干预规模累计约为15.4万亿日

元(约合965亿美元)。

此次干预推动日元从过去40年低位、1美元兑164日元附近回升,8月3日曾一度最高触及155.2。随后,日元再度回落,一度逼近1美元兑160日元关口,9月初又回升至155—156区间。

本次部分购日元操作由日美联合实施,这是两国自2011年以来首次协同干预,大大超出市场预期,因为此前市场普遍认为日美联合干预可能性极低。

为缓解市场对于日本大规模干预弹药不足的担忧,美日双方表示,日本可动用美联储在新冠疫情时期设立的主要央行应急工具。该工具于2020年推出,用于疫情期间稳定市场,能够帮助日本获取美元流动性,无需直接抛售美国国债,有望减轻日本开展汇率干预的资金压力。

北京商报综合报道