

超威半导体公司大中华区市场营销副总裁纪朝晖：

Token 打开路径应像“居家用餐”

本报记者 张心怡

近日，首届沉浸式消费嘉年华暨2026AI终端创新发展大会在北京市经济技术开发区（北京亦庄）正式开幕。超威半导体公司（AMD）大中华区市场营销副总裁纪朝晖出席大会并作主题报告。纪朝晖指出，当推理成为AI主要的工作负载，Token成本成为企业的“不可承受之重”，但若将智能体部署在智能体PC、智能体主机等本地新型计算设备上，只将重型任务交给云端大模型，就能将Token成本降低80%，“这就像大部分时间，我们都是买了食材在家做饭，只有重要场合才出去吃。”

智能体引领产业进入AI“下半场”

自2022年11月30日ChatGPT上线至今的近4年时间，AI产业的发展可以分为两个阶段。在前三年，全球大模型公司都在努力把AI做得更聪明，使其服务于用户的生产、工作、学习，驱动AI训练计算量以每年5倍的速度持续膨胀。而过去一年，把AI“用起来”成为产业的重心，因而在大模型训练计算量继续成长的同时，AI推理的工作负载也迅速攀升。在2025年，AI推理与AI训练两种工作负载的占比达到5:5，2026年AI推理的占比已经来到60%。到2030年，AI推理的占比将继续提高，达到80%以上。

“我们已经进入AI的下半场，从让大模型变得越来越聪明，到全世界努力把AI用起来。”纪朝晖表示。智能体是AI从对话式机器人走向生产力工具的关键，不仅是“把AI用起来”的核心范式，也是AI+千行百业进入快车道的转折点。普通大模型的工作模式很简单，用户输入问题，它给出答案。而智能体要全能得多，用户给它一个任务，它会通过推理、子智能体的调用、数据的处理、任务的分派、工具的调用，最终反馈用户一个执行的结果。

但智能体的自主性，是以Token消耗量急剧提升为代价的。Gartner数据显示，智能体应用相比对话式机器人，Token消耗会增加5~30倍。IDC数据显示，到2027



年全球2000强企业（G2000企业）AI Token的消耗量会比当前增加1000倍。在这样的趋势下，Token越来越成为企业的成本难题。

像在家做饭一样使用智能体

对于企业难以负担Token成本的现象，纪朝晖作了一个生动的比喻：“如今企业使用AI的方式，就像我们每一顿饭都去餐厅吃——大部分AI应用的Token都来自云端的Token供应商，我们要向他们购买Token和API。但我们在家里过日子，真的是每一顿饭都在外面吃吗？其实大部分时间，我们都是买了食材在家做饭，只有重要场合才出去吃。”

那么，如何像“在家做饭”一样使用智能体呢？相关研究显示，如果将日常辅助办公放在AI PC的小模型上，将智能体AI员工所需的智能体应用放入智能体主机的中模型上，只有跑复杂推理任务时才用云端的大型推理模型，可以让企业的AI Token消耗降低80%以上——使用AI PC和智能体主机上的本地模型，就等同于平时“在家做饭”，而使用云端大模型就是“去餐厅”。

要实现这种基于任务选择模型和计算位置的方式，离不开端侧AI。它不仅解决Token消耗越来越贵的问题，还满足了用户对合规、安全、隐私、IP保护、即时性、数据主权的需求。最重要的是，端侧模型的能力在过去几年获得了显著的

提升，使端侧AI成为可能。“我们可以看到，最新Qwen3.8本地可部署模型（参数27B）的能力，已经达到最前沿云端大模型的80%~90%，足以驱动端侧智能体的使用。智能体的到来，正在掀起新一轮AI终端的进化浪潮。”纪朝晖说道。

推出AI计算新设备 遴选AI用例带头人

今年3月，AMD提出了智能体主机（Agent Computer）这一新的计算品类，随后又提出了智能体PC（Agentic PC）。在本次演讲中，纪朝晖进一步阐释了两种设备的使用方式：智能体主机是给智能体用的，智能体PC是智能体和人一起用的。

相比传统PC需要人不断通过键盘、鼠标进行操作，智能体主机只需要用户派发任务就能24小时不间断地工作直至任务交付。

“智能体主机把智能体放到端侧而非云端。未来的企业可能是1000个人类员工和1万个智能体AI员工一起工作。AI员工有一部分会跑在云端，但是由于合规、隐私、数据安全保护、数据主权等原因，要求它大部分落在端侧。”纪朝晖说道。

要构建一台智能体主机，首先要“放入”作为基座的大模型（至少27B）及其所需的24G以上显存，来承载智能体的运行需求。有了这样的模型能力，就相当于聘请了一位博士毕业生，但缺乏工作经验也不

“ 我们已经进入AI的下半场：从让大模型变得越来越聪明，到全世界努力把AI用起来。 ”

行，因而还需要包含相关行业30年工作经验的私有知识库，以及RAG KV cache膨胀所需的20G以上存储空间。接下来要部署的是财务、客服、人事、工程师等AI员工，每位AI员工都需要10G以上的显存，再给它们一套严格的SOP（标准作业流程）——这样一台设备才能胜任智能体主机的框架。

“如果您用一台智能体PC，它的AI能力能使您的工作效率提升一倍。如果企业部署智能体主机，可以让团队的工作效能提升十倍。如果企业采用智能体工作站或服务器，可以让整个企业的效能提升百倍。未来智能体将对企业能力提升起到至关重要的作用。”纪朝晖表示。

虽然AI将改变未来工作模式的趋势已经明确，但许多用户对于如何用AI“倍增”自己的能力还缺乏具体的概念。为此，AMD组织了锐龙AI智能体创新应用大赛，其目的是将通过AI实现了3倍、5倍乃至10倍生产力的“先行者”筛选出来，作为样例。比如，金属加工销售曹冬冬在电商平台接活时，因为后端工厂的各种数控机床的工艺和成本不一样，很难提供准确的报价。而智能体主机将工厂后端所有机床的工艺流程、工时、费用信息全部交给智能体，只要将客户的图纸交给智能体主机，就能得到清晰完整的报价。高中生Sophia Liu，用智能体主机做了智能体辅助教学，帮助老师分层教学、因材施教。

“我们每一个人都会成为AI时代的超级个体。”纪朝晖说道。



力，硬件、软件、外设设备等信息感知能力，时间、天气、时区等环境信息感知能力。长期以来，PC的感知信息主要来源于鼠标和显示器，能力和工作效率较低。针对这一问题，陶宏芝表示，联想已推出硬件传感器阵列等解决方案，通过持续或间歇状态感知，识别用户所处场景及工作需求，让电脑具备系统感知和场景感知能力，实现场景主动触发和任务自动执行。

认知能力包括理解、推理和规划。陶宏芝表示，联想开展了精细的意图分类工作，将人类意图和PC工作意图分为27大类、428小类，助力PC开展复杂的意图理解工作。推理能力方面，则通过端边云全量独立存储、分级压缩、按需召回的流程设计，实现了虚拟上下文窗口的10倍扩容。

记忆能力包括短期记忆与长期记忆。陶宏芝分享了联想的分层记忆技术，即通过情景记忆、语义记忆、程序记忆技术，将短期记忆整理成为长期记忆，乃至个人化的记忆图谱，并输入至Agent执行环境中，从而实现准确记忆召回和行为规划能力的双提升。陶宏芝表示，通过此类技术，可降低60%~95%的Token消耗，节约60%的时间。

执行能力包括工具调用、内容生成、互联协同。陶宏芝认为，面向未来L4级智能，AI终端有望进一步提升自主执行与自主运行能力，承担更多自动化工作。他分享了联想的天禧Harness架构，据悉，天禧Harness在长程任务表现上较为出色，可实现200轮以上的长程任务执行工作。

学习能力主要体现在情景适应

“ 面向未来L4级智能，AI终端有望进一步提升自主执行与自主运行能力，承担更多自动化工作。 ”

方面。陶宏芝表示，AI技术的发展，要求越来越广的上下文、越来越好的工具执行结果、持续迭代的个人化需求理解，联想希望让AI PC具备自动学习和生成专属Skill的个体自进化能力，以及提炼普适性技能的群体自进化能力，满足用户本质需求。

陶宏芝同时强调，智能终端还需警惕安全问题，他认为，智能终端的能力越强，对个人的知识、喜好等隐私数据的保密和安全要求越高。L4更高级智能，除了实现能力升级和端云全场景自治外，仍需坚持全链路隐私计算和自主权限管理，构筑高阶智能的安全底线。

面向未来，陶宏芝表示，OS组件的AI化并不是终点，他期待，通过AI原生技术重构操作系统，推动AI终端持续发展。

灵伴科技（杭州）股份有限公司联合创始人向文杰：

AI眼镜是个人智能终端的最佳载体



“ AI+AR落地产业的关键，是让工程师、安全员、设计师等一线人员真正把AI‘握在手里’。 ”

本报记者 陈存 崔烁

近日，首届沉浸式消费嘉年华暨2026AI终端创新发展大会在北京经济技术开发区（北京亦庄）开幕。开幕式上，灵伴科技（杭州）股份有限公司（Rokid）联合创始人向文杰发表题为“AI+AR：从智能交互到产业应用”的主题报告。在他看来，随着人工智能技术持续演进，AI眼镜正从“可视化工具”走向“智能协作者”，成为下一代智能终端最具潜力的入口，其价值也正从消费端延伸至产业现场。

AI眼镜有望成为下一代智能交互入口

向文杰在报告中回顾了AI眼镜的发展历程。他介绍，早在2014年，产业圈便已感受到人工智能的巨大价值即将释放，Rokid自成立起即布局AI眼镜赛道，十余年间见证AI产业从“无人问津”到“所有人都离不开”。彼时，移动互联网正值鼎盛，智能手机用户规模已突破4亿，而2017年全球出货量攀升至10亿余台的峰值后逐年回落。团队由此判断，人工智能时代的载体不应局限于手机，而应是随身伴随式的穿戴设备。

“当AI想要跟你互动的时候，其实是像人和人互动一样的，是完整的多模态互动。”向文杰表示，团队先后尝试了手表、耳环、耳机、指环等多类穿戴形态，发现这些设备本质上仍是基于AI的功能性产品，都“很像设备”。真正的智能载体，应当做到语音交流就在耳边、想看见的就在眼前，同时也能同步用户所看到、所听到的内容。向文杰表示，2023年以来，以ChatGPT为代表的大模型的出现，加速了AI技术的发展。他判断，随着AI不断发

展，个人的终端智能开始出现，而眼镜将成为其最合适的载体。

消费与产业两端落地

在消费端，向文杰展示了AI眼镜的多元应用。当前，AI眼镜已可实现实时翻译、问答、拍照录像、会议纪要等功能，也能帮助存在语言障碍或视障、听障的人士更好地感知世界。向文杰提到，当前AI眼镜已与智能体应用打通，用户无需打开电脑或手机，即可完成资料整理、演示文稿制作甚至编程等任务，购物、订票、导航等日常操作亦可“一句话完成”。

产业端，AI眼镜的价值正从大屏演示走向一线现场。向文杰举例说，在飞机出口后的跨国维保场景中，当地工程师佩戴AI眼镜用母语描述故障，后方专家即可实时收到高精度同声传译，这使得过去难以实现的跨国支持、跨国服务变得简单，类似的场景在2B、2G领域还有大量应用空间。在安全巡检、现场作业等场景，眼镜还可提供第一视角回传、隐患排查与专家指导。

他强调，AI+AR落地产业的关键，是让工程师、安全员、设计师等一线人员真正把AI“握在手里”，让它的价值不再停留在PPT、视频或大屏中，而是转化为一线人员手里的实时支持与帮助。

向文杰在报告最后表示，Rokid的出发点是“让所有人享受科技”，希望帮助老人、儿童、残障人士等群体跨越技术与能力的门槛，让工作和生活变得更简单、高效便捷。他提到，作为平台型企业，Rokid愿与合作伙伴做好生态，共同推进AI+AR的落地，让“所有人享受到科技价值和成果”的目标能够实现。

电子信息制造业何以扛起工业企业利润六成增量？

（上接第1版）1—8月份集成电路累计出口额同比增长103.9%，8月单月芯片出口额更是创下历史新高。成本抬升，推高出口单价。AI催生存储芯片供不应求，推动全球存储芯片价格持续走高，这也为国产芯片增收创造了有利条件。8月份，国产芯片平均出口单价同比上涨150%。本土化发展成势，生态持续完善。头部大模型已全面兼容海光、昇腾、昆仑芯、摩尔线程等国算力底座，生态构建与产业协作提速。精益制造提效，打开利润空间。作为先进制造的代表，电子信息制造业企业通过AI提升制造端的质效，率先吃到新技术的红利。通过应用垂类大模型、加速落地AI工厂，推动AI贯穿研发、计划、物料、生产、质量等核心场景，推动企业在转型升级中赢得更大利润空间。

长期以来，外界对中国工业的认知不少仍停留在“规模大但附加

值不高”“大而不强”的刻板印象中。然而，电子信息制造业、高技术制造业对全部工业贡献度的不断攀升，有力证明了中国工业发展和增长逻辑正在发生深刻变化。这背后，是我国创新从“跟跑”到“并跑”，乃至部分领域“领跑”的质变，向高技术、高附加值、高壁垒的高端智造全面转型。

站在中国经济新旧动能转换的关键节点，要把“关键变量”转化为“最大增量”，把电子信息制造业的领跑优势拓展为制造业整体优势，一方面要纵深打通产业链，持续畅通从芯片、材料到整机、通信的价值循环，把AI算力需求转化为实实在在的产业升级成果，在关键核心技术攻关上久久为功；另一方面要以电子信息制造业赋能汽车、钢铁、建材等传统产业，加快数字化、智能化、绿色化转型，让更多行业在“焕新”中找到广阔的盈利空间。