三大运营商竞逐智能体 寻找AI转型最优解

中经记者 谭伦 北京报道

随着5G建设放缓，通信运营商正在加大寻求AI转型创新的力度。

近日，中国电信正式推出星辰智能体服务平台1.0，以“星小辰”为统一入口构建跨终端智能服务生态。中国电信人工智能科技有限公司副总经理杨戈透露，目前已有2万多活跃开发者通过该开发平台创建了5万余个智能体应用。

此前，中国移动宣布灵犀智能体用户规模已突破2亿大关，成为国内用户量最大的消费级智能体产品；中国联通亦紧随其后，推出元景思维链大模型及“智家通通”智能体解决方案，完成家庭与行业场景的双重覆盖。

这场三大运营商巨头加入的智能体竞逐战，背后是巨大经济收益的浪潮推动。中商产业研究院数据显示，2025年中国AI智能体市场规模将达69亿元，2030年有望逼近300亿元，其间复合增速超30%。

同时，今年8月国务院发布的《关于深入实施“人工智能+”行动的意见》也明确提出，到2027年智能体等应用普及率需超70%，2030年进一步提升至90%。

中信建投研报指出，2025年作为“智能体元年”，推理需求带动算力爆发，而运营商的入局正重塑行业竞争格局。业内也广泛认为，市场与政策的双重驱动下，手握海量用户、全场景渠道与数据资源的运营商，正从通信服务商加速向“AI+服务”生态构建者转型，而智能体已成为这次转型的主角。

运营商“人工智能+”行动

运营商纷纷加码智能体服务是政策导向、技术迭代与市场需求共同作用的必然结果。

通信分析师周桂军认为，当前，运营商纷纷加码智能体服务并非偶然，而是政策导向、技术迭代与市场需求共同作用的必然结果。长期以来，运营商面临着被“管道化”的危机，在移动互联网的红利期，流量入口被超级APP把持。而在AI时代，智能体作为新的流量分发中枢，为运营商提供了重夺入口的机会。

工信部数据显示，2024年国内移动用户渗透率已超110%，语音、短信等基础业务收入持续下滑。而智能体带来的场景化服务，成为激活存量用户价值的关键。华信人咨询调研显示，57%的消费者高频使用智能家电语音助手。

而智能体的出现，无疑为运营商提供了一个将“连接”升维的机会。通过智能体，运营商可以将单纯的“卖流量、语音”转变为“卖服务、算力”。如中国移动将5G新通话结合灵犀智能体，可以让用户在通话中直接调取翻译、订票、导航等服务，极大地提升通话场景的商业附加值。

中国电信最新发布的星辰智能体服务平台，也秉承这一思路。《中国经营报》记者在平台发布现场获悉，其可通过该统一入口，用户无需跳转多个APP，即可通过对话完成复杂任务，实现“开口即办”。

同时，周桂军认为，与互联网厂商不同，运营商掌握着最为庞大且基础的用户群，并具备与终端厂商、家电、智慧政务以及传统行业厂商长期合作的渠道，这使得它们能把智能体能力“嵌入终端与行业应用”中，从而实现从单点工具到生态入口的跃升。

而在技术侧，运营商也是目前唯一能够打通手机终端、网络传输和云端算力的角色。通过部署智能体，运营商可以在合规的前提下，将海量的用户数据转化为个性化的服务推荐能力。如当系统检测到用户正在漫游且网络信号不佳时，智能体可以主动介入优化网络或推荐相应服务，这种“原生”的优势是APP厂商无法比拟的。

此外，Omdia电信战略分析师杨光认为，国家“人工智能+”行动的落地实施，为运营商划定了明确的转型路线图。政策要求到2027年实现智能体在重点行业的规模化应用，而运营商作为数字基础设施建设主力军，承担着推动AI普惠的社会责任，也自然成为承接智能体产业规模化重任的最合适角色。

三家打法各有侧重

三大运营商的智能体部署思路也已渐趋清晰。

共识之下，三大运营商的智能体部署思路也已渐趋清晰。

最新发布的“星辰智能体服务平台1.0”，被中国电信定义为“一站式智能体构建与分发平台”。官方公布的信息显示，这不仅是给用户用的，也是给开发者用的。记者在现场注意到，中国电信明确提出了“亦庄亦谐”的打法：对外，支持零代码开发，鼓励中小开发者和企业用户在平台上快速生成专属智能体；对内，则利用星辰大模型重构自身的业务流程。

据杨戈透露，星辰已在天翼防诈、天翼智屏、天翼智看、天翼智铃、云智手机五大产品线开始赋能，并以覆盖中国电信约4.6亿移动用户与2.4亿家庭用户作为后续规模化推广基础。平台同时提出兼容、简单与开放三大特性，意在对外打造“一点接入”的生态规则。

相比之下，体量更大的中国移动则将部署场景“大而全”化。其推出的“灵犀”智能体，依托于“九天”人工智能大模型，并未走传统的独立APP路线，而是深度嵌入到了庞大的中国移动APP及5G新通话业务中。

图为12月5日，2025数智科技生态大会上的中国电信展台。 公司官网/图

公开数据显示，“灵犀”的用户数目前已突破2亿，用户规模在三大运营商中居于首位。记者此前从中国移动了解到，灵犀的特点在于其“全能型”助理定位，不仅能处理查话费、办套餐等传统业务，还能进行复杂的跨应用任务执行。

此外，记者注意到，中国移动还在重点发力“5G+AI”的融合，如在5G新通话中引入智能体，用户在视频通话时，可以召唤智能体进行实时语音转写、方言翻译，甚至生成通话纪要。中国移动工作人员表示，这种将AI能力直接“原生化”到通信协议中的做法，极大地降低了用户的使用门槛。

而中国联通的智能体战略则显得更加务实和垂直。依托“元景”大模型，中国联通主要将火力集中在客户服务和政企行业应用上。在C端，中国联通推出了升级版的“联通助理”，通过智能体技术拦截骚扰电话、代接听电话，并生成智能摘要。这种“小切口”的功能直击用户痛点，培养了用户的付费习惯。在B端，中国联通利用智能体技术赋能千行百业。如在智慧城市领域，中国联通部署的城市治理智能体可以自动分析摄像头数据，识别违章停车、垃圾溢出等事件，并自动生成工单派发给处理人员。

突围仍存多重挑战

未来的竞争将不再是单纯的用户规模比拼，而是谁能更懂用户的意图、更高效地连接服务以及构建更繁荣的开发者生态。

尽管运营商在智能体领域拥有用户与生态的资源禀赋，但要真正赢得这场竞争，仍面临着来自互联网AI巨头的压力、商业化落地等多重挑战。

周桂军认为，智能体的核心能力来自AI厂商自身。当前国内互联网头部大厂生态在AI领域投入巨大，运营商若不能在核心模型能力或差异化工具链上展示出明显优势，难以长期吸引用户。同时，已经合作的生态企业也可能转而与其他AI大厂合作，抢夺掉运营商的现有用户。

而有运营商人士告诉记者，更致命的威胁或许来自手机终端厂商。由于华为、小米、荣耀等厂商正在将大模型直接植入操作系统，通过系统级入口直接截流用户需求。未来如果用户对着手机说一句“帮我订咖啡”，系统级AI直接调用了美团等，那么运营商的智能体就会被完全绕过。

因此，如何在“终端厂商掌握OS”和“互联网巨头掌握场景”的夹缝中，利用网络侧优势构建不可替代的场景，成为运营商必须解决的难题。

同时，杨光认为，在商业化变现方面，从替代人工客服到家庭增值服务再到SaaS化行业解决方案，运营商目前已有多条尝试路径。但如何能稳定产生高边际利润以及能否由免费补贴阶段顺利过渡到用户付费，是决定商业成功的核心。

有研究机构指出，智能体与垂直SaaS结合、以业务流程效率提升为驱动的付费最有可能率先落地，但这要求运营商具备深厚的行业服务能力与销售能力，而非仅靠渠道导流即可解决。

此外，运营商强调“一点接入、伙伴共创”，但如何在流量分成、数据使用权、接口标准与竞争中保持公平并激励创新，是现实难题。若合作条款倾向于“封闭或高门槛”，会抑制中小厂商的接入意愿；若过度开放，又可能削弱运营商的商业回报。

整体而言，业内认为，对于三大运营商而言，发布平台仅仅是拿到了入场券。未来的竞争将不再是单纯的用户规模比拼，而是谁能更懂用户的意图、更高效地连接服务以及构建更繁荣的开发者生态。

杨光认为，目前，三大运营商的优势是明确且现实可用的，但要把优势转化为长期护城河并实现可持续商业化，还需在模型能力、合规治理、生态激励与变现体系上同步突破。但无论结局如何，这场由运营商发起的去“管道化”努力，已经成为运营商转型进程中的重要一步。

20年来CUDA最大更新 英伟达自我革命or别有企图？

中经记者 李玉洋 上海报道

近日，英伟达CUDA迎来重大更新，正式推出NVIDIA CUDA 13.1，该公司AI开发者账号在社媒平台自我评价称：“这是20年来最大的一次更新。”

《中国经营报》记者了解到，全新的编程模型CUDA Tile是CUDA 13.1最核心的更新，它让开发者可以用Python写GPU内核，15行代码就能达到200行CUDA C++代码的性能。

需要注意的是，CUDA Tile目前仅支持采用英伟达Blackwell架构的GPU产品，未来的CUDA版本将扩展支持更多架构的产品。多年以来，CUDA被称为英伟达稳固的护城河，然而，随着CUDA Tile编程模型的发布，引起了业界关于英伟达“护城河”是否会被削弱的讨论。

对此，曾主导设计AMD Zen架构芯片、苹果A系列芯片等知名芯片的架构师Jim Keller发帖称：“英伟达是要终结自己的护城河？如果英伟达像大多数其他硬件（公司那样）转向Tile模型，那AI内核将更容易移植。”

言下之意，不像过去的CUDA C++那样高度绑定英伟达硬件，CUDA Tile这种新的编程模型将改写GPU编程范式，开发者用Python代码可直接生成高效GPU内核，大大降低AI底层开发门槛，这可能会给AMD、Intel或新兴AI芯片公司提供切入机会。

“现在来看，底层更新对于应用基本没影响。”AI算法专家、资深人工智能从业者黄颂如此表示。他拥有丰富的CUDA生态应用开发经验，日常使用如PyTorch这些基于CUDA的高层库。黄颂进一步指出，短期内还看不到CUDA 13.1对于应用开发的积极影响，“传导需要时间，应用有更高层的接口。”黄颂表示。

改写GPU编程范式

据了解，CUDA的全称是Compute Unified Device Architecture（统一计算设备架构），是英伟达在2006年推出的一套并行计算平台和编程模型。

对于CUDA，一般开发者接触最多的是CUDA Toolkit（CUDA工具包），它是使用CUDA的核心载体，包含编译器、运行时API/驱动API、基础数学库（cuBLAS/cuFFT/cuDNN）等组件；CUDA已成为高性能计算和AI领域的“标配”，且仅支持英伟达GPU。

过去近20年，CUDA一直采用SIMT（单指令多线程）模型，开发者写代码时，需要手动管理线程索引、线程块、共享内存布局、线程同步，每一个细节都要自己操心。想要充分利用GPU性能，特别是用上Tensor Core这类专用模块，更是需要深厚的经验积累。

英伟达解释说，CUDA Tile可让开发者在高于SIMT的层级编写GPU核函数。在目前的SIMT编程中，开发者通常通过划分数据并定义每个线程的执行路径来指定核函数。

而借助CUDA Tile，开发者可以提升代码的抽象层级，直接指定被称为Tile的数据块。只需指定要在这些Tile上执行的数学运算，编译器和运行时环境会自动决定将工作负载分发到各个线程的最佳方式。

为此，英伟达构建了两个用于Tile编程的核心组件：一是CUDA Tile IR，一种用于英伟达GPU编程的全新虚拟指令集架构（ISA）；二是cuTile Python，一种新的领域特定语言（DSL），用于在Python中编写基于数组和Tile的核函数。

为什么要为GPU引入Tile编程？

英伟达在博客中表示，随着计算工作负载的演进，特别是在AI领域，张量已成为一种基础数据类型。英伟达开发了专门用于处理张量的硬件，比如NVIDIA Tensor Core（TC）和NVIDIA Tensor Memory Accelerator（TMA）。

硬件越复杂，就越需要软件来帮助驾驭这些能力。CUDA Tile对Tensor Core及其编程模型进行了抽象处理，使得用CUDA Tile编写的代码能够兼容当前及未来的Tensor Core架构。

基于Tile的编程方式，允许开发者通过指定数据块（即Tile），然后定义在这些Tile上执行的计算来编写算法。至于怎么把这些运算映射到GPU的线程、Warp和Tensor Core上，编译器和运行时会自动搞定。

这种编程范式在Python等语言中很常见，有观点认为，CUDA Tile的出现改变了GPU编程，就像NumPy之于Python。

真正的护城河所在

此外，此次CUDA 13.1的更新还包括运行时对Green Context（绿色上下文）的支持、CUDA多进程服务（MPS）更新等。

经过近20年的发展，英伟达已经在全球拥有500多万的CUDA生态开发者，该公司创始人兼CEO黄仁勋多次强调CUDA开发者是英伟达最重要的资产和竞争优势。“护城河不是芯片，是数百万开发者写下的代码惯性。”黄仁勋在2025年GTC大会演讲中提到。

2025财年数据显示，英伟达全球员工总数为36000人，较2024财年的29600人增长了21.62%。根据公开资料，虽然还无法提供官方确认的具体数字，但基于多方信息分析，英伟达CUDA团队规模约为2000—5000人，占总员工数的5%—15%。

近日，英伟达H200芯片能对华出口。对此，行业分析机构Omdia人工智能首席分析师苏廉节表示，短期来看，英伟达H200能对华出售这件事对英伟达自己更为有利，“中国还是一个相当大的市场，而且中国开发者也认可CUDA的生态”。

由于英伟达CUDA的生态壁垒，不少国内AI芯片公司采取了兼容CUDA的策略，以吸引开发者，比如摩尔线程、海光信息、沐曦股份、天数智芯、壁仞科技、芯动科技等，只是技术路线不同。

那么，看到CUDA Tile后，Jim Keller为什么说英伟达是否“终结了自己的护城河”？关键原因在于Tile编程模型不是英伟达独有的，AMD、Intel等芯片厂商的硬件，在底层架构上同样可以支持基于Tile的编程范式。

如果未来的主流GPU编程逐渐转向这种Tile-based方式，开发者一旦习惯了“写Tile、硬体自己优化”的模式，那同一套程序逻辑就更容易移植到不同的GPU硬件上，不像过去的CUDA C++那样高度绑定英伟达硬件，这可能会给AMD、Intel或新兴的AI公司提供切入机会。

“AI内核将更容易移植。”正如Jim Keller所说的那样。不过，英伟达也考虑了后手，CUDA Tile IR提供了跨代兼容性，但这种兼容性是建立在CUDA平台之上的。

因此，还有观点认为，表面上看，英伟达用Tile IR构建了一条更高阶的软件路径，专为AI负载设计，尤其适配Transformer、MoE等主流架构。这看似是对开发者的“解放”，实则是用“易用性”作饵，将开发者更深地引入其护城河。

从这个层面看，不管护城河是加深还是削弱，可有一点是确定的：未来，GPU编程的门槛大幅降低。过去，能熟练驾驭CUDA的开发者是稀缺的，而会写Python的人很多。

而CUDA Tile和cuTile Python打通了这个瓶颈。英伟达在开发者博客中提到，一个15行的Python内核性能可以媲美200行手动优化的CUDA C++代码。

未来真正的护城河，或许不再是芯片，而是那一行写得越来越顺手的Python代码。