WESTERN SECURITIES 西部证券

证券研究报告

从训练到推理时代的拐点

——2026年计算机行业中期策略

西部证券研发中心 2025年6月17日

分析师：郑宏达 S0800524020001 邮箱地址：zhenghongda@research.xbmail.com.cn

分析师：谢忱 S0800524040005 邮箱地址：xiechen@research.xbmail.com.cn

分析师：李想 S0800525040006 邮箱地址：lixiang@research.xbmail.com.cn

分析师：卢可欣 S0800525080006 邮箱地址：lukexin@research.xbmail.com.cn

分析师：王朗 S0800526040009 邮箱地址：wanglang@research.xbmail.com.cn

核心结论

2026年是AI从训练转向推理主导的拐点。OpenClaw、ClaudeCode等开源与闭源Agent框架以燎原之势迅速普及，算力的应用从训练走向智能体等推理需求主导，从问答模式走向智能体循环，从单轮生成升级为多步规划、持续执行，推理首超训练成为算力需求的重心。

进入推理时代，算力基础设施的核心关注点是Token成本的低延迟、调度算法和缓存管理，通过极致优化实现高效服务。在芯片层，以Groq LPU为代表的专用推理芯片兴起，推动推理能力和效率普惠，并与GPU结合进行异构计算，实现性能互补；在网络互联层，超节点以高速互联、内存池化、高度集成等优势精准适配推理高并发、实时交互和大显存消耗需求，助推系统向万卡集群演进，以太网交换机凭借其普适性和经济性，有望成为AI时代数据中心网络架构的主流选择。

模型层面，Claude Opus 4.8、GPT 5.5、GLM-5.1、DeepSeek V4、Minimax-M3等前沿模型密集发布，模型迭代节奏进一步加快，模型能力提升聚焦agent、coding与多模态，进一步迈向生产力级别的智能。

展望2026年下半年：芯片、互联、算力，技术进步与整体需求依旧蓬勃发展，模型能力的边界有望进一步拓展，助推AI应用的深度和广度持续提升。建议关注方向：

1）国内外模型商业化提速，规模化应用有望带动Tokens消耗高增，继续看好AI算力。 2）国产模型能力持续提升，具备性能和性价比优势，继续看好国产模型厂商。3）模型应用有望在B端、C端持续展开，看好具备垂直行业know-how的AI应用厂商。

风险提示：AI 技术突破不及预期；大模型应用落地节奏不及预期；宏观经济增长不及预期，IT预算不及预期；国际环境发生变化。

行业评级	超配
前次评级	超配
评级变动	维持

近一年行业走势

相对表现	1个月	3个月	12个月
计算机	-11.21	-11.24	1.07
沪深300	1.48	6.34	27.41

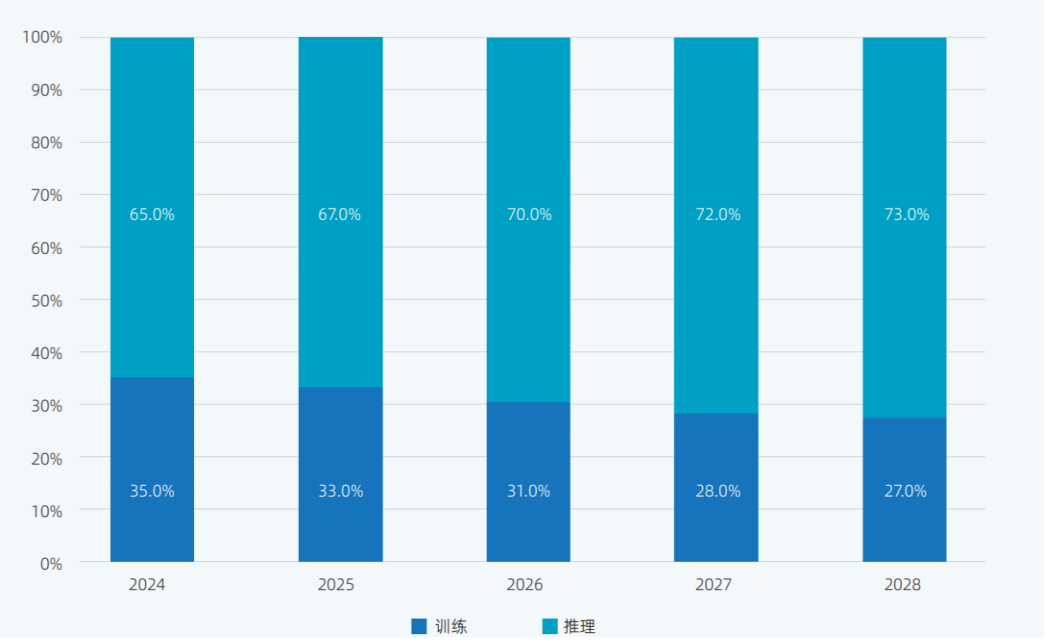
目 录

CONTENTS

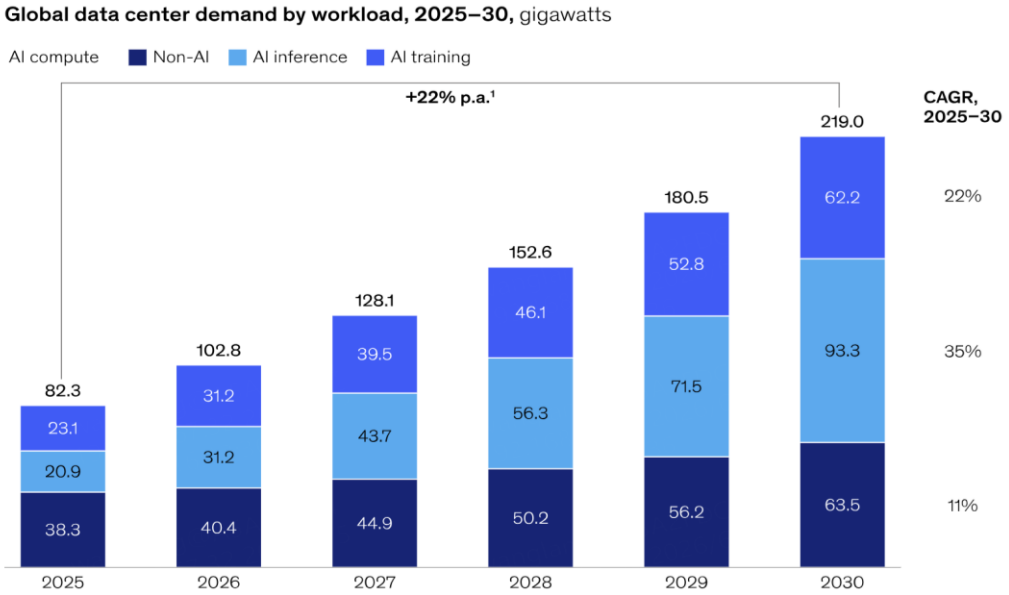
推理算力占比快速提升，2026年首超训练成为算力主体

- **推理首超训练，算力重心发生根本性转移：**AI模型生命周期分训练与推理两段——训练是“一次性建厂”式的集中投入，模型一旦定型、成本即趋于收敛；推理则是“7×24永续调用”，每一次用户请求、每一步Agent任务都在持续消耗算力，需求随用户普及与智能体长任务能力快速增长。算力的应用从训练走向推理主体及智能体主导，算力的架构技术、应用场景、商业模式等发生显著变化：从阅读检索到深度思考从单轮生成升级为多步规划、持续执行；算力架构从注意力经济到生产力经济；智能体驱动从被动问答转向主动任务执行；部署形态从纯云端走向“云—边—端”协同；商业模式从成本中心到价值引擎。
- 根据麦肯锡公司，预计到2030年推理将超越训练，占到全球数据中心AI算力需求的一半以上，推理侧将成为未来几年算力投资的重要增量。
- **中国推理需求倍数更高：**在国产替代与应用落地的双重驱动下，中国推理需求扩张较全球更为剧烈，邬贺铨院士指出中国推理需求已达训练的8倍。同时，推理对单卡算力的要求低于训练，更看重成本、能效与规模化部署，推理有望成为国产算力厂商确定性更高的突围窗口。

图：中国AI服务器工作负载中训练与推理的比例（2024-2028）



图：全球数据中心负载规模及预测（2025-2030）

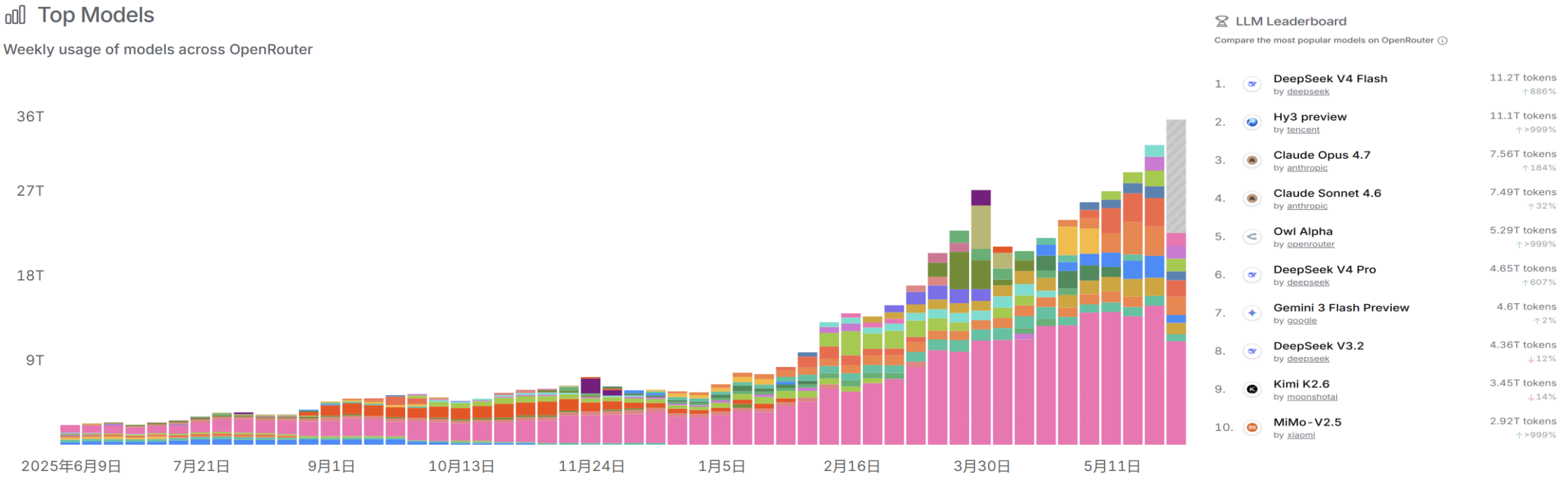


资料来源：IDC《2025年中国人工智能计算力发展评估报告》、麦肯锡官网、西部证券研发中心

Token调用量与日俱增，国产开源模型成为全球推理新主力

- **Token调用量伴随推理需求指数级放大：**AI工作重心从训练转向推理、再转向Agent，频繁的工具调用、多步规划与长程执行不断拉长模型的输出链路，单次任务的Token消耗相比单轮对话是数量级的差距。全球知名的多模型聚合平台OpenRouter为全球超过800万用户，提供涵盖Anthropic、Google、OpenAI、xAI和DeepSeek等领先AI供应商的400多个模型访问，是观察全球真实推理调用的窗口。根据OpenRouter官方，该平台的周Token处理量已从2025年11月的约5万亿激增至2026年5月的25万亿（约100万亿/月），半年增长了5倍。
- **国产模型在agent调用方面性价比凸显：**开源模型与闭源模型形成稳定“双轨”结构——闭源定义性能与可靠性上限，开源以成本、透明与性价比优势承接agent时代的大规模推理负载。PinchBench专门针对OpenClaw任务对大模型进行基准测试：在成功率方面，Qwen3.7达到92.5%，仅次于第一名的ClaudeOpus4.8，Mimo-v2.5、Deepseekv4等国产开源模型也排名靠前。

图：OpenRouter 周度 Token 处理量及月度模型调用量排名



资料来源：OpenRouter 官网、西部证券研发中心

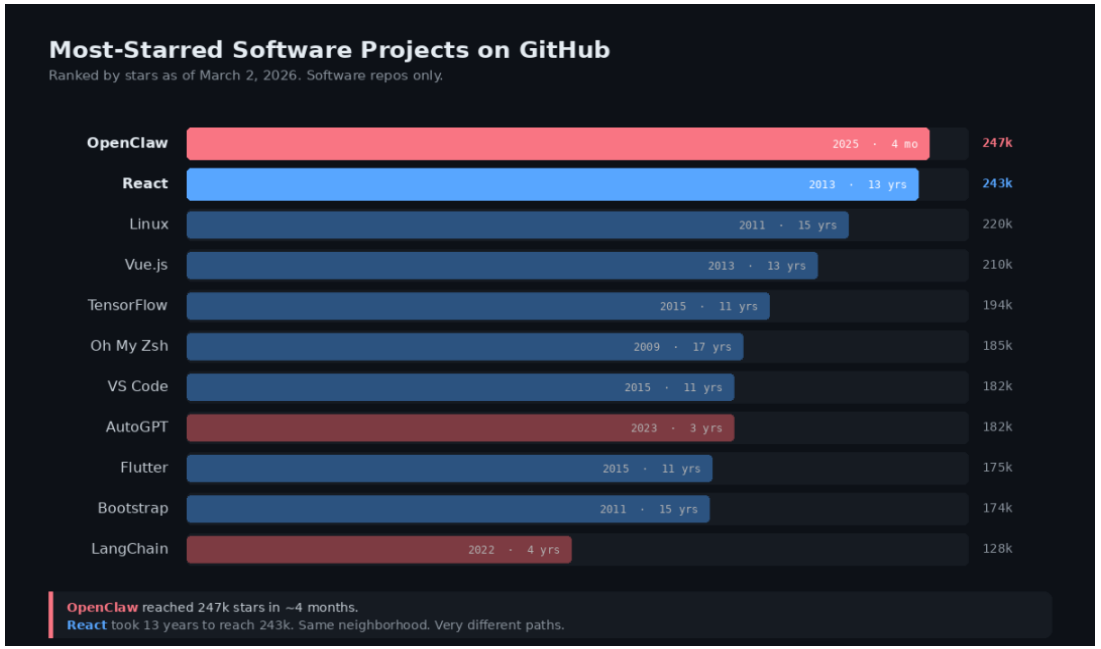
Agent框架广泛流行，智能体循环主导推理需求

- **从单次对话到智能体循环：**AI模型的使用形态正在从“单轮文本补全”转向“多步骤、工具集成且推理密集型的工作流”，LLM请求的不仅仅是简单的问题或孤立指令，而是结构化的、类似智能体的循环的一部分，调用外部工具、对状态进行推理，并在更长的上下文中持续运行。
- **OpenClaw为代表的开源Agent框架崛起：**OpenClaw诞生于2025年底，是一款开源的AI智能体框架，在四个月内登顶Github软件项目星标榜史上第一。它能主动调用工具、访问网络、操作软件、发送消息，可持续在后台运行，像一个永不下班的私人助理OpenClaw的核心能力：底层的操作系统级访问（终端命令、文件读写、进程管理）、中间层的应用程序控制（浏览器自动化、邮件客户端操作、即时通讯接入），以及上层的多步骤任务编排（将复杂目标拆解为子任务并自主执行)。OpenClaw在技术栈中占据了一个前所未有的位置，一个能操控所有其他应用程序的智能代理。

图：OpenRouter的AI Agent流行度排名

1.	Hermes Agent Hermes Agent is an open-sour... Personal Agents CLI Agents	13.6T tokens	11.	GitLawb Collaborate on a live decentralized git network Cloud Agents	579B tokens
2.	OpenClaw OpenClaw is an open-source A... Personal Agents CLI Agents	5.36T tokens	12.	Roo Code Roo Code is a VS Code extensi... IDE Extensions Cloud Agents	545B tokens
3.	Kilo Code Kilo Code is an open-source AI c... CLI Agents IDE Extensions	4.52T tokens	13.	Pioneer (production) Inference API that improves with your traffic	429B tokens
4.	Claude Code Claude Code is Anthropic's agentic coding tool ... CLI Agents	1.85T tokens	14.	Mira is the leading AI agent inside Telegram Mira is a Telegram-native AI assistant. Personal Agents	420B tokens
5.	Descript AI Video & Podcast Editor Video Generation	1.7T tokens	15.	SillyTavern SillyTavern is the LLM frontend for power users, a ... Roleplay	242B tokens
6.	pi There are many coding agents, but this one is y... CLI Agents	1.34T tokens	16.	VidMuse Generate music videos, shorts, and more f... Video Generation	222B tokens
7.	Janitor AI Janitor AI is a chatbot platform where users create... Roleplay	862B tokens	17.	LangChain docs.langchain.com	215B tokens
8.	ISEKAI ZERO AI adventures. Travel with your favorite characters Game	778B tokens	18.	Pieces.app: Ambient Artificial Memory Pieces is a private, on-device, long-term m... Personal Agents	198B tokens
9.	Lemonade The AI tool for Roblox games. Programming App	705B tokens	19.	AA-LCR Benchmark Applied research lab for frontier models	196B tokens
10.	Cline Cline is an open-source AI codin... IDE Extensions CLI Agents	644B tokens	20.	OpenHands AI coding agent that can run commands, browse... CLI Agents	180B tokens

图：OpenClaw登顶Github软件项目星标榜第一



资料来源：OpenRouter官网、新智元公众号、西部证券研发中心

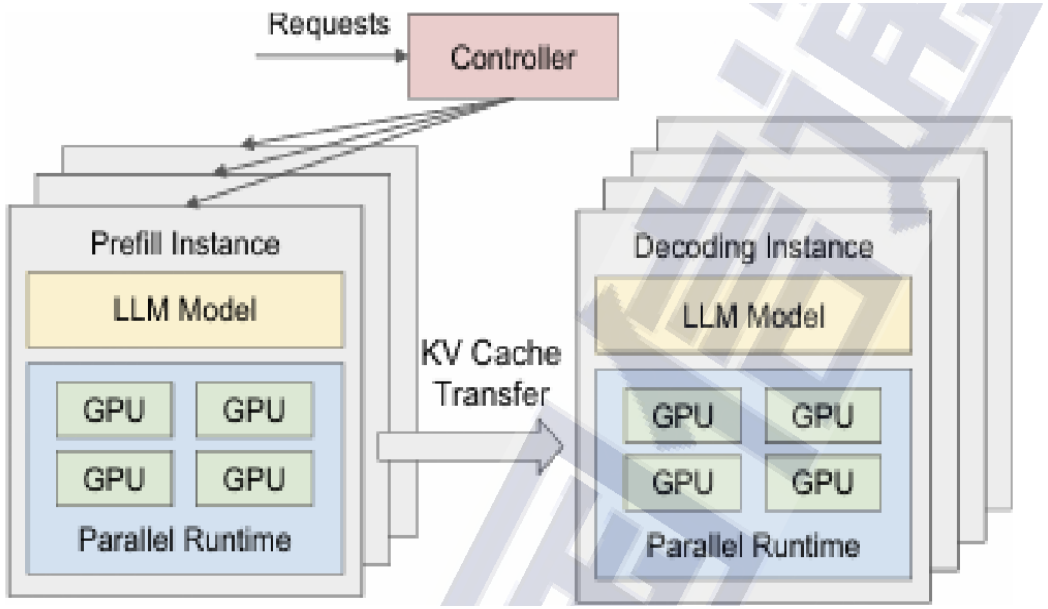
目 录

CONTENTS

训练架构到推理架构：从规模化到精细化

- **训练时代以“规模化”为核心：**为满足大模型对计算资源的高需求，提升单节点的计算性能（Scale-up）变得至关重要，这包括增加单芯片或单个机架的计算能力；通过增加节点数量实现计算能力的横向扩展（Scale-out），高速互联网络和分布式计算框架将有效支持千卡、万卡甚至十万卡的集群建设。
- **推理时代转向“精细化”：**伴随大模型从训练阶段迈向应用阶段，推理工作负载持续增加，面向应用和推理需求对芯片和系统架构进行设计愈加重要；大语言模型推理包含预填充（Prefill）和解码（Decode）两阶段，对计算和存储资源的访问频率与调度需求不同，实操中往往采用 P-D 解耦部署策略，通过构建分离式算力资源池，缩短计算时间、降低计算成本、提高资源利用率。推理阶段核心关注点是Token成本的低延迟、调度算法和缓存管理，通过极致优化实现高效服务。

图：PD 分离架构示意图



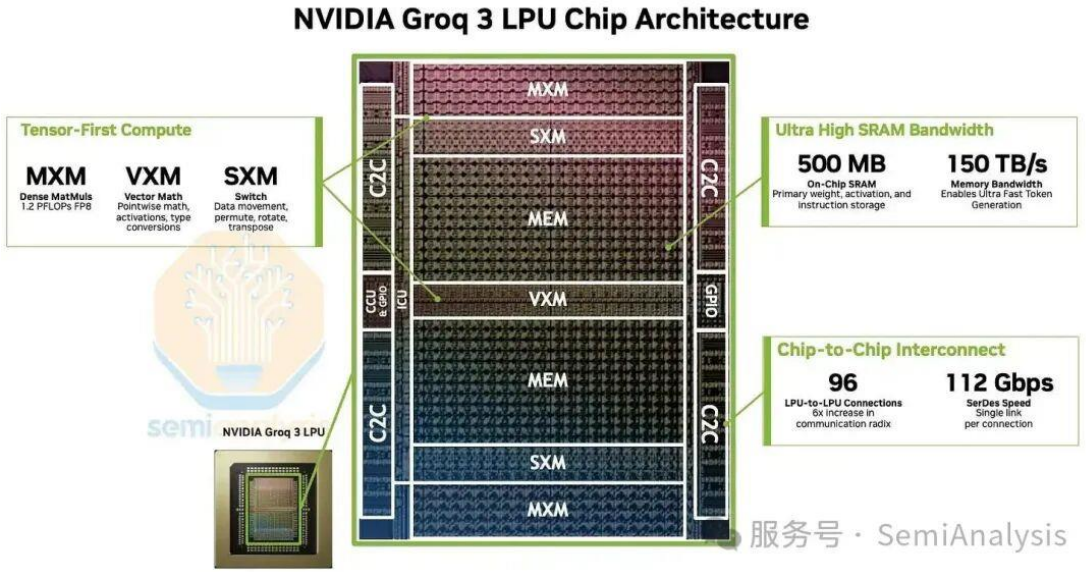
图：大模型推理核心目标

资料来源：中国信通院《大模型推理优化关键技术及应用实践研究报告（2026年）》、西部证券研发中心

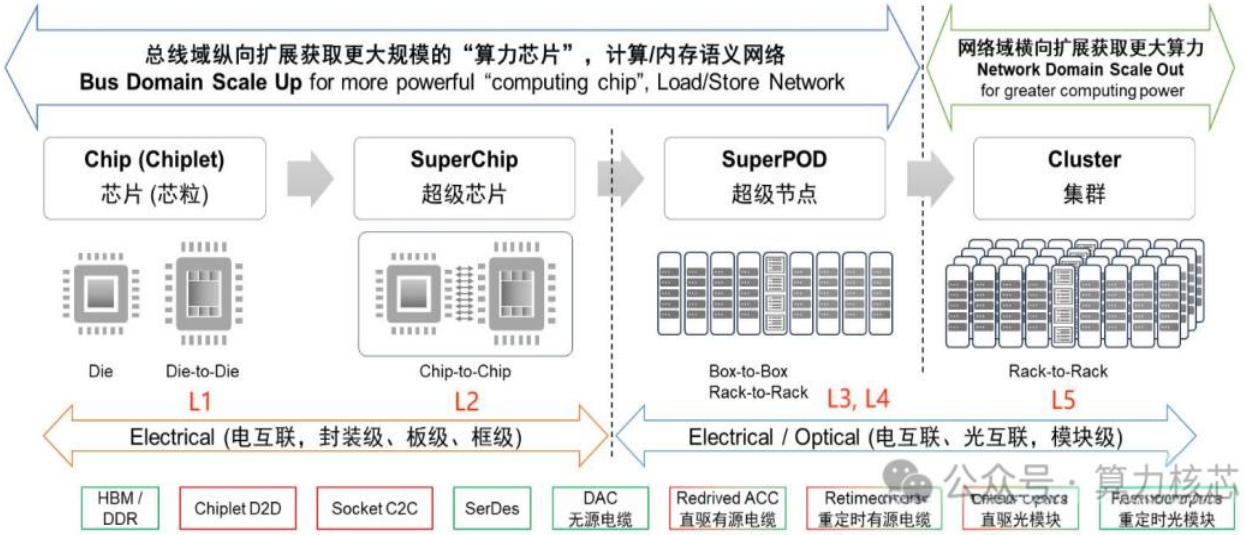
推理时代算力底座三大特征：资源池化·低延迟互联·推理专用芯片

- **算力资源池化，打破物理边界：**推理时代需要协同异构基础设施，将整个数据中心作为协同工作的有机体，整合多种计算资源，优化数据处理流程和模型训练效率，通过灵活的计算任务调度，高效执行人工智能任务；通过资源池化、动态分配和智能调度等技术手段，突破传统算力供给模式的局限性，提高资源利用率和灵活性。
- **低延迟传输与互联，构建面向 AI 的运力体系：**高带宽能够显著提升数据传输速度，目前网络速率已经可达到 400G/800G，1.6T 是超大规模数据传输和高效能需求的下一步计划，未来行业目光将投向 3.2T 乃至更高速率；RoCEv2 的出现，使集群可以基于 RDMA 技术扩展到超大规模，较传统以太网的通信方式大大降低了延迟。
- **推理硬件创新 — 以 Groq LPU 为代表的专用芯片：**专为语言处理任务设计的专用处理器（Language Processing Unit），聚焦推理阶段的极致速度与低延迟，是 AI 产业从“训练竞赛”转向“推理普惠”的核心产物。遵循“软件优先、可编程流水线、确定性计算、片上存储”四大原则，摒弃传统 GPU 的复杂缓存层级，聚焦推理任务的数据流优化。

图：Groq 3 LPU架构



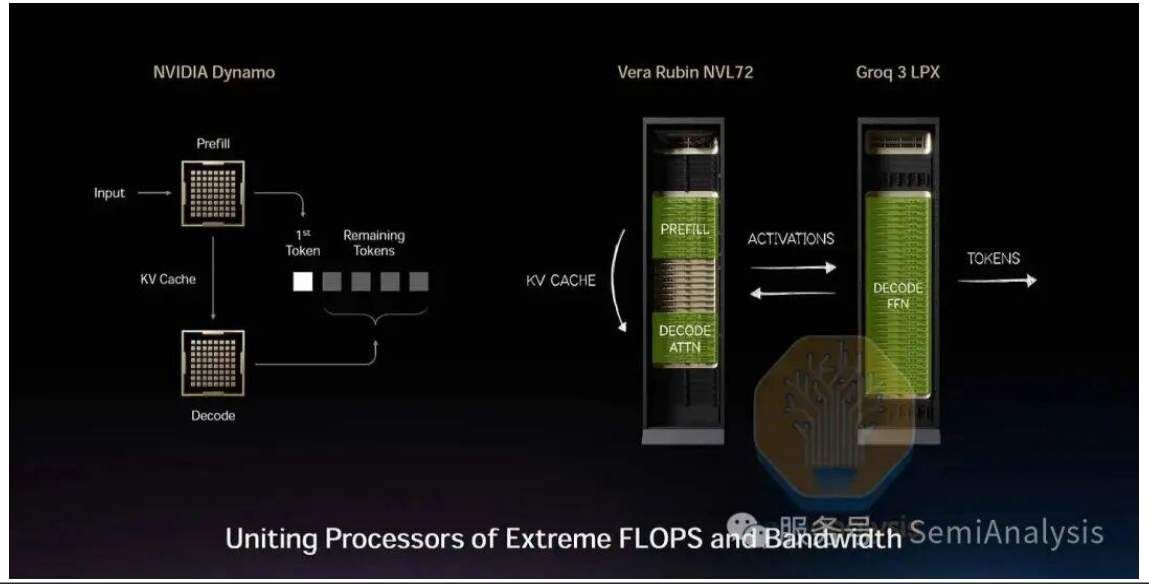
图：L0-L5的高速互联体系



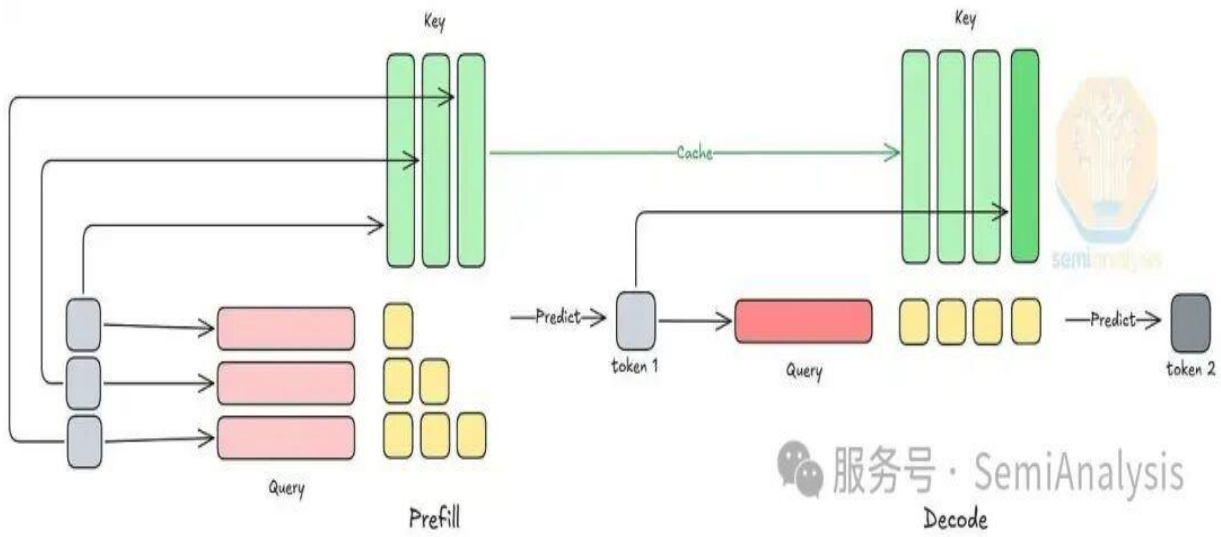
LPU融入英伟达芯片体系：异构计算再加速

- **LPU与GPU结合进行异构计算**：大语言模型（LLM）推理包含两个阶段：预填充（prefill）和解码（decode）。预填充阶段处理完整的输入上下文：它是计算密集型的，非常适合 GPU 处理。另一方面，解码阶段负责预测新 token，属于内存受限型（memory-bounded）任务。由于模型需要逐个预测新 token，解码阶段对延迟极其敏感，而 LPU 的高 SRAM 带宽和低延迟特性有助于加速这一迭代过程。
- **相比GPU和TPU，LPU用确定性来提高计算效率**：LPU通过编译器预先确定每一纳秒的数据流向，消除了TPU和GPU因动态调度产生的硬件冗余与延迟，从而实现了算力资源的更高效使用（去掉了用来处理不确定性的预测器、缓存器）。
- **异构推理架构正在逐步形成行业共识**：亚马逊与Cerebras宣布合作，将AWS的Trainium-3加速器与Cerebras的晶圆级加速器结合部署，逻辑与英伟达的GPU-LPU系统如出一辙。

图：英伟达引入LPU改善延迟性能



图：大模型推理两阶段：预填充（prefill）和解码（decode）

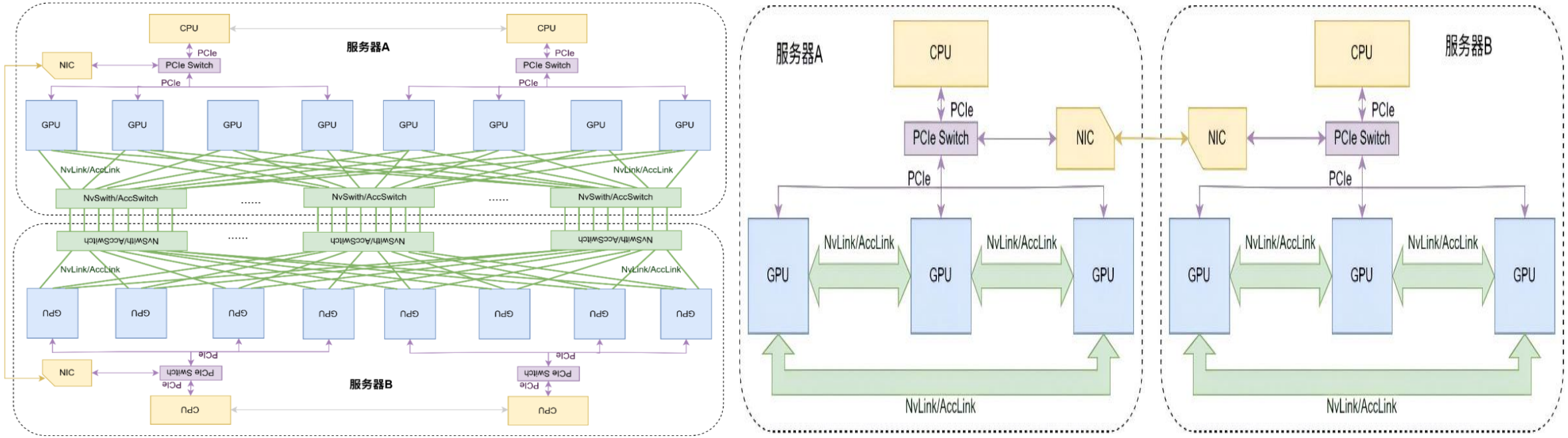


资料来源：SemiAnalysis公众号、西部证券研发中心

推理架构落地形式：超节点

- **超节点的技术特征包括，**1）大量GPU互联系统。由于业界已经有成熟的8卡产品方案，因此超节点超过8卡是最基本要求；2）统一内存地址空间：为互联的GPU提供统一寻址和内存一致性，如英伟达UVA（unified virtual address）、UM（unified memory），系统内任一GPU可以像访问本地HBM一样访问任意互联的HBM。3）超高带宽、超低时延互联：除PCIe/CXL外提供额外GPU显存高速互联，如英伟达NVLink，为系统内GPU提供高速访问的物理通道，带宽达数百GB到数百TB，纳秒级时延满足GPU间内存语义通信的同步操作要求。4）原生可扩展性：互联协议层面预留相应的bit位数满足未来可能的GPU扩展规模，拓扑层面支持一级和二级交换实现更大集群规模的扩展。
- **超节点以高速互联、内存池化、高度集成等优势，能够精准适配推理高并发、实时交互和大显存消耗需求。**此外，Agentic AI等场景下推理中产生的KV缓存随上下文长度线性增长，超节点通过全局内存池化实现KV缓存的共享与复用，可支持处理数十万甚至上百万Token的上下文，极大增强其理解复杂问题、执行长期任务的能力。

图：超节点集群（左）与传统集群架构对比



资料来源：TWCL探微芯联公众号、西部证券研发中心

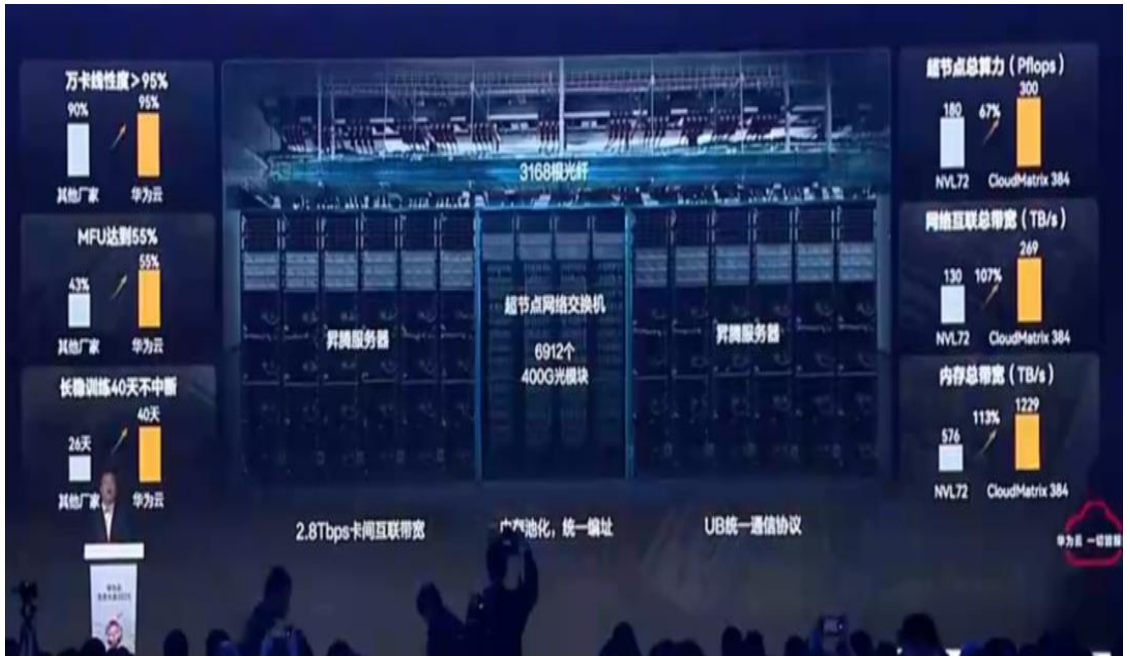
推理架构落地形式：超节点

- 1）互联方案主要有两类：封闭生态：英伟达（NVLink）、华为（UB）等，内部沟通成本低、效率高，但需全套自研（GPU、交换机、网卡、机柜等）。开放生态：中国移动OISA、腾讯ETH-X、阿里ALS等，基于以太网协议，通用开放，产业内企业可灵活加入。
- 2）华为是国内较早将 AI 集群从“八卡服务器”推进到“超节点”形态的厂商之一，CloudMatrix 384将互联范围从单节点扩展到多柜一体化超节点，把 Scale-Up 域从传统的 8 卡或 16 卡级别提升到 384 NPU 级别。华为超节点路线并不以单卡性能领先为主要特征，而是通过昇腾 NPU、鲲鹏 CPU、自研 HCCS/UB 灵衢互联以及云侧调度软件，把更多强耦合通信尽量保留在高带宽域内。CloudMatrix 384 是这一路线第一次较完整落地到超节点形态的代表性产品，其后续方向则是继续沿着更大 Scale-Up 域和更高层级互联能力演进，向千卡甚至万卡级别的系统扩展。

图：超节点的标准

组织/项目	类型	牵头单位	参加单位
ETH-X	开放标准	信通院/腾讯	信通院、腾讯、中国移动、快手、京东、Intel、博通、锐捷、新华三、联想、中兴、盛科、光讯、华勤、云和智网、云豹智能、立讯
ALS	开放标准	阿里	信通院、AMD、博通、华勤、新华三、Intel、浪潮、立讯、沐曦、楠菲、天数智芯、芯潮流、中兴、澜起、奇异摩尔
OISA	开放标准	中国移动	数十家单位
UA LINK	开放标准	AMD、Intel、谷歌、微软、博通、思科、Meta、HPE	
NVLINK	私有协议	英伟达	无
UB	私有协议	华为	无

图：华为CloudMatrix 384超节点

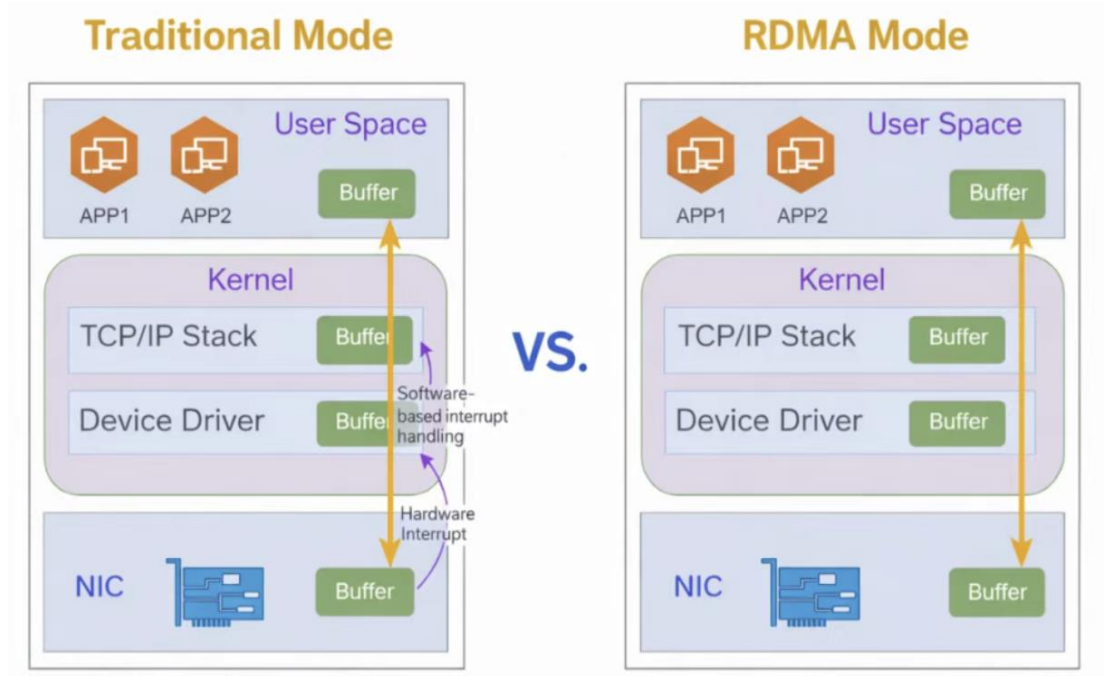


资料来源：鲜枣课堂公众号、西部证券研发中心

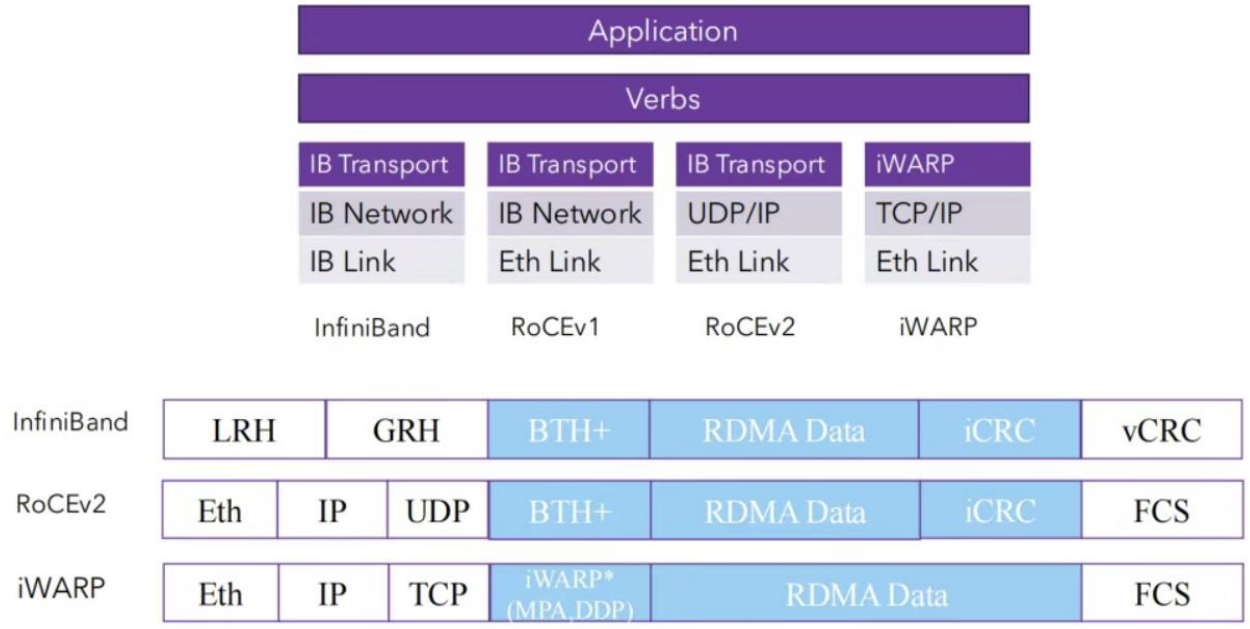
以太网交换机凭借其普适性和经济性，有望成为AI时代数据中心网络架构的主流选择

- **未来增长驱动：以太网交换机凭借其普适性和经济性，有望成为AI时代数据中心网络架构的主流选择**
- **RDMA（远程直接内存访问）是一种用于提高传统以太网数据传输效率的技术。**RDMA技术依赖于支持RDMA的网络接口卡 (NIC) 和交换设备，广泛应用于大规模科学计算、模型训练和推理等高性能场景。
- RDMA允许高吞吐、低延迟的网络通信，InfiniBand和RoCE为AIDC主流方案。1）InfiniBand通过基于通道的点对点消息队列转发模型，允许应用程序直接访问远程内存，而无需内核和 CPU 干预，但需要专用的联网硬件，导致成本较高。2）RoCE是一种在传统以太网上实现RDMA的技术，分为 RoCEv1 和 RoCEv2，其中RoCE v2 基于以太网 UDP/IP 协议实现，可在三层网络上运行，支持跨网段通信，并具有更好的网络可扩展性，因此广泛应用于智能计算数据中心。

图：RDMA与传统数据传输模式对比



图：RDMA的传输协议

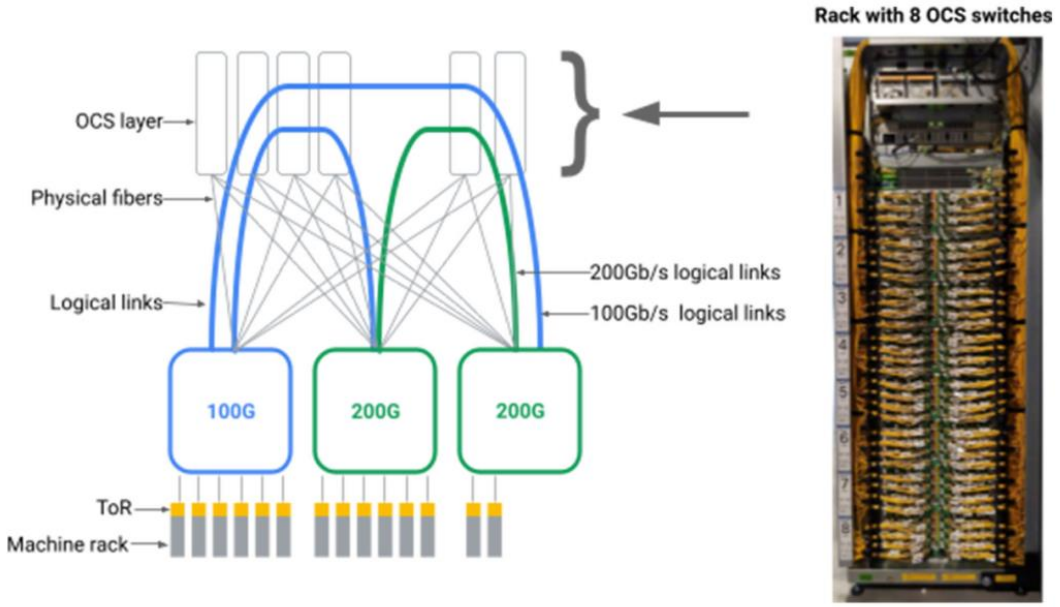


资料来源：SDNLAB公众号、智猩猩芯算公众号、西部证券研发中心

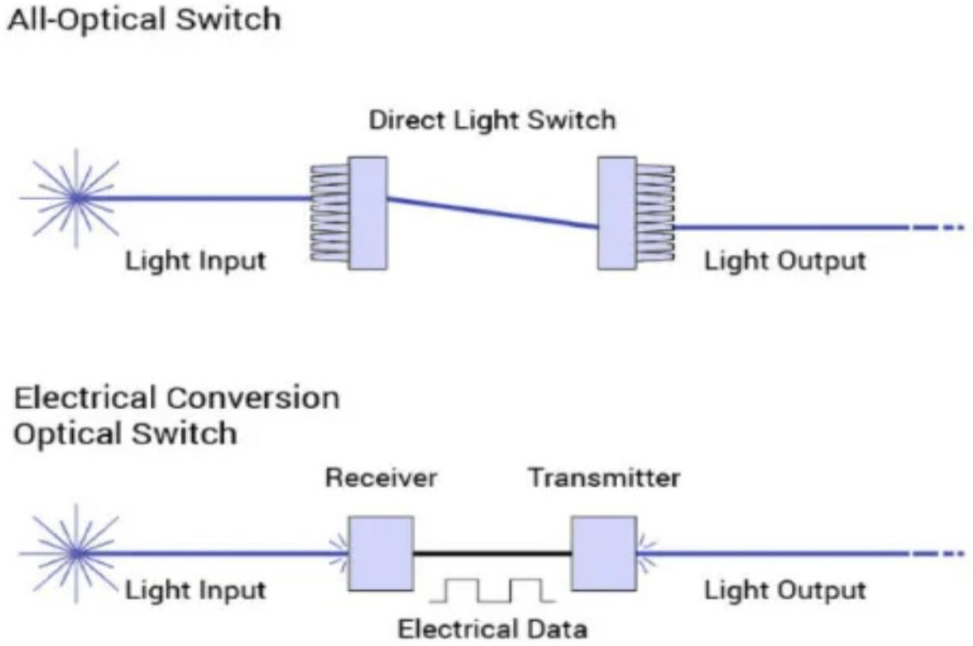
OCS：谷歌主推，行业从0到1

- **谷歌已在TPU集群中大规模部署OCS，用于替代传统电交换机。**
- **OCS（光路交换机）是一种无需光电/电光（O/E/O）转换，直接实现光信号在光纤端口间切换的技术，应用于AI算力集群、超大规模数据中心的叶脊架构互连、超节点集群高速通信等场景。**OCS通过直接在光域完成数据信号的路由与切换，无需进行光电/电光转换，大幅缩短了信号传输的时延，助力AI算力集群、数据中心光互连系统整体功耗降低30%以上。
- 根据观研天下，OCS市场规模呈现爆发式增长态势。2020-2024年全球光交换机（OCS）市场规模由7278万美元增长到3.66亿美元，期间年复合增长率达49.8%。预计2031年全球光交换机（OCS）市场规模将达到20.22亿美元，2024-2031年年复合增长率达27.7%。

图：谷歌的OCS光路交换机内部结构



图：OCS 建立全光直连路径



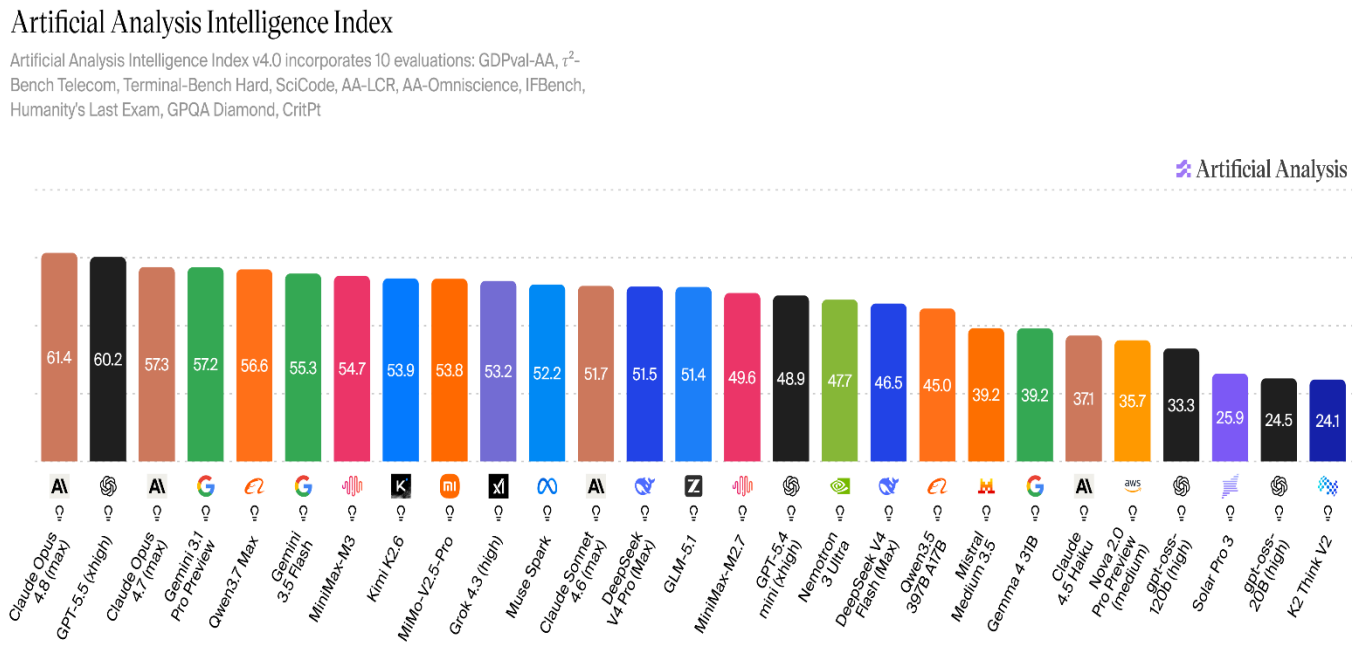
目 录

CONTENTS

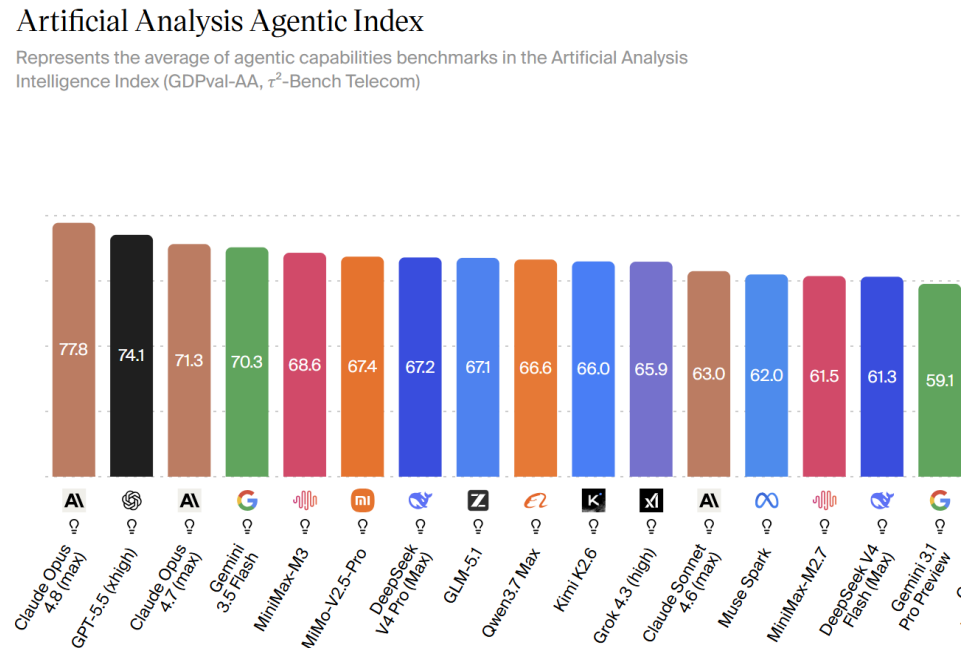
前沿模型密集发布，小步快跑提升性能

- **今年上半年，全球大模型迭代节奏进一步加快，模型更新更倾向于小步快跑。**例如Anthropic的旗舰模型迭代节奏从两个月缩短到六周——Claude 4.6（2 月）、4.7（4 月）、4.8（5 月底），几乎是行业最激进的迭代速度。
- 根据Artificial Analysis发布的模型智能指数基准（涵盖Humanity's Last Exam等多项模型评测基准），Claude Opus 4.8 位居首位、GPT-5.5紧随其后，Gemini 3.1 Pro Preview位列第四；国产模型方面，Qwen 3.7 Max、Minimax-M3、Kimi K2.6、MIMO-V2.5-Pro、GLM-5.1、Deepseek V4-Pro等多款模型表现靠前。
- **大模型能力迭代聚焦Coding、Agentic两项关键能力，本质上是模型全面接管工作任务以及生产力水平的高低。**
- **DeepSeek为代表的国产模型厂商推进模型普惠。**5月，DeepSeek官宣 V4-Pro 模型 API 永久锁定 2.5 折定价，创下全球大模型调用价格新低，即每百万tokens输入（缓存命中）0.025元，输入（缓存未命中）3元，输出6元。

图：Artificial Analysis Intelligence Index模型排名



图：Artificial Analysis Agentic Index模型排名

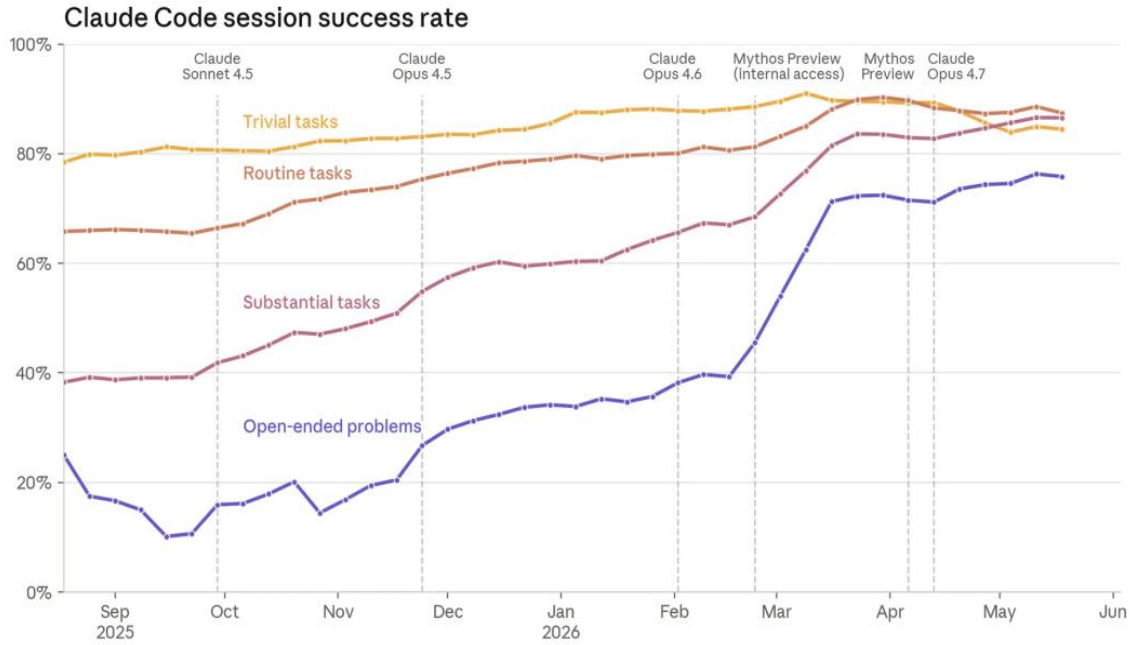


资料来源：Artificial Analysis 官网、西部证券研发中心

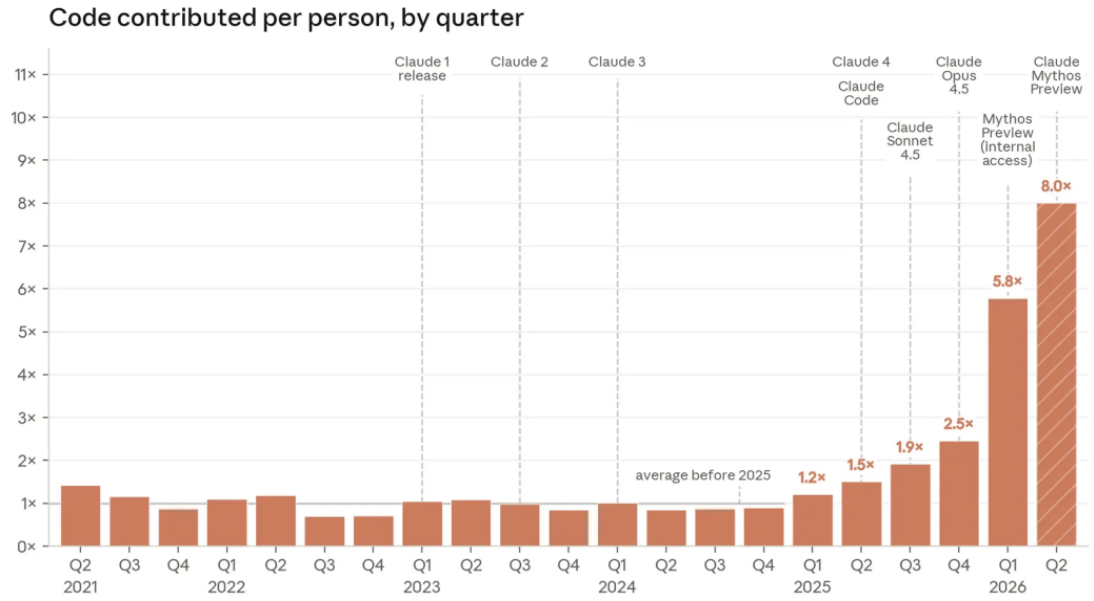
从Copilot到Autopilot，模型能力提升形成飞轮效应

- AI的角色正从Copilot迈向Autopilot，从辅助角色到自主角色，以编程类agent为例：
- 第一阶段Copilot（2022-2024）—— 代码补全：GitHub Copilot等作为“AI副驾驶”，核心交互是按Tab键接受建议，需要人类全程掌控方向盘；
- 第二阶段Agent Mode（2025）—— 多步骤任务：AI可实现跨文件修改、运行终端、管理数据库迁移；
- 第三阶段 Autopilot（2026）—— 自主完成复杂工程，正如OpenAI官方所述，以往你需要步步为营地引导 AI，现在，你只需将一个繁杂的多阶段任务交给 GPT-5.5；它具备极强的自主性，能够自行制定计划、调用工具、核查结果并在模糊的边界中寻找最优路径，始终保持高效推进。
- **AI模型能力用于模型开发迭代，形成模型能力提升的飞轮效应。**在Anthropic内部，截至2026年5月，公司正式并入代码库的代码中，超80%由Claude编写，2026年二季度，工程师日均合入代码量达到2024年基准水平的8倍，原因是绝大多数代码由Claude编写，工程师仅负责任务统筹与代码审核，不再手动敲码。

图：Claude Code在各类开发任务的成功率持续提升



图：Claude模型加速公司AI开发效率

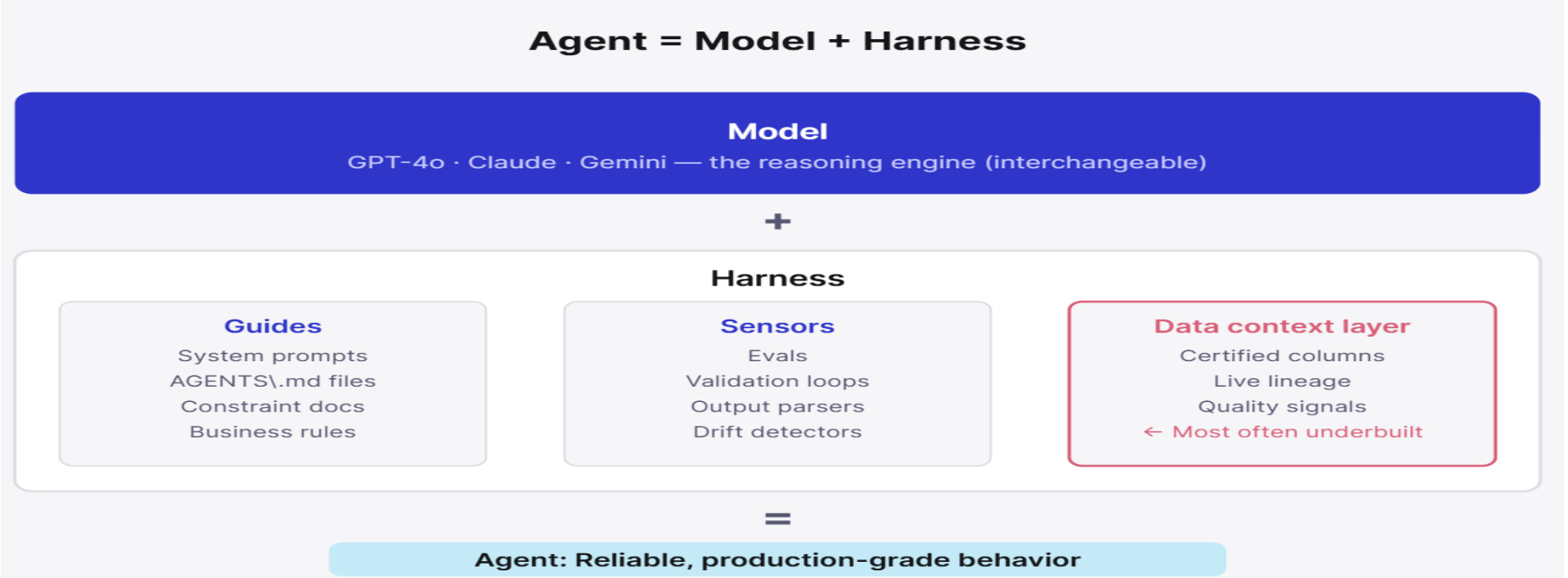


资料来源：头部科技公众号、西部证券研发中心

Agent新范式——Harness Engineering，实现可控的智能

- **Harness Engineer是AI时代软件工程范式转变下的新角色，旨在通过工程化的手段将AI的智能转化为可靠、可控的生产力。**术语源自"harness"（马具）的隐喻——就像骑师通过缰绳和马鞍来引导马的力量走正确的方向，而非增强马本身的体能，Harness Engineering 强调的是引导和约束 AI 模型的能力，而非提升模型本身。
- **AI Agent = SOTA 的模型（野马）+ Harness（驾驭系统）= 千里马**
- Harness的核心理念是：每当发现AI智能体（Agent）犯错时，应当构建一套工程化方案，确保它在未来不会重蹈覆辙。
- Harness 的核心要素可以归结为三个支柱：1）通过上下文工程，管理长短期记忆，进行知识检索注入，渐进式披露和上下文防火墙；2）架构约束：进行Agent编排模式，状态机设计，降级策略，做好安全边界；3）持续管理：做好质量门禁、知识生命周期管理、持续进化等。

图： Harness Engineer核心理念与架构



海内外AI需求加速，模型商业化拐点已现

- 海外：ChatGPT 与 Gemini 均突破 9 亿用户，今年2月ChatGPT 月活用户超过9亿，个人订阅用户超过5000 万，付费企业用户 900 万 ，ARR 已超 250 亿美元；Anthropic 走“企业 + 编码”路线，年化收入ARR 跃升至440亿美元、反超 OpenAI；Claude Code是近期增长加速的关键变量，自2025年5月公开推出，2026年2月年化收入已达25亿美元。AI厂商的收入增速远超传统软件时代的企业，横向对比来看，AWS用了13年才达到350亿美元年收入，Salesforce从1999年成立到2021年才跨过200亿美元收入线，ServiceNow用了约20年超过90亿美元。
- 国内 ：根据QuestMobile 报告，截至2026年3月，国内 AI 原生 App 月活规模达 4.4 亿，豆包 3.45 亿 / 千问 1.66 亿 / DeepSeek 1.27 亿位列前三，单季度新增超 1亿用户；行业从“用户量飙涨”转向“用户量与粘性齐涨”，进入用户量 + 商业化的下半场。

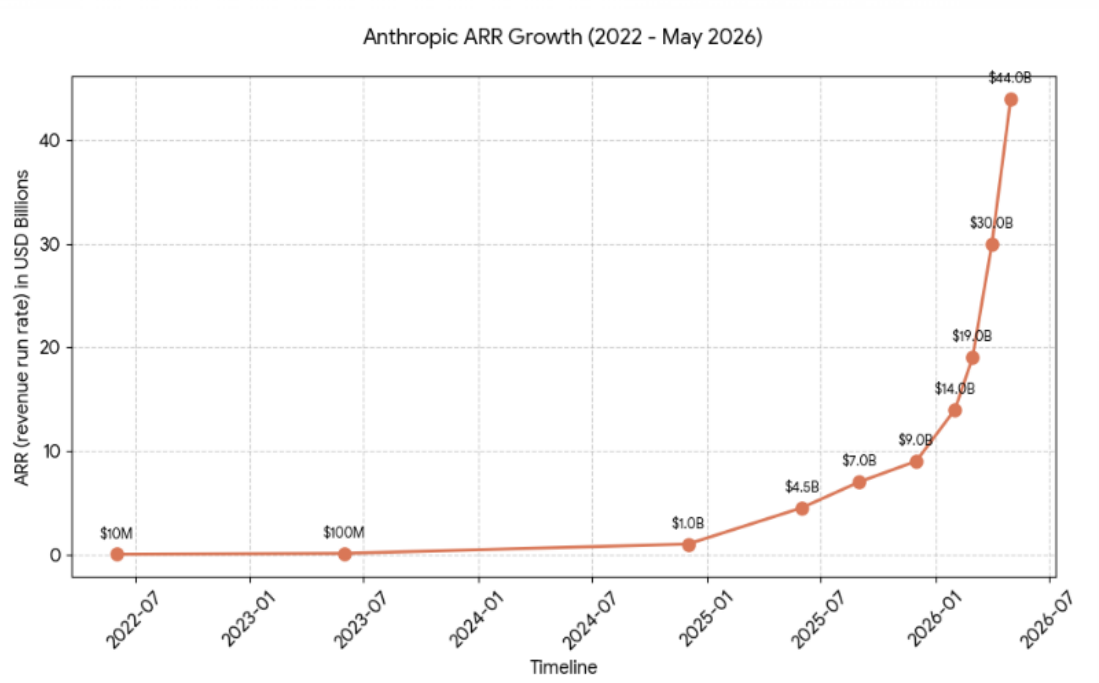
图：2026年3月国内TOP10 AI原生APP

应用层 | 系统层

2026年3月 活跃用户规模TOP10 AI原生App

序号	应用名称	AI赛道	活跃用户规模（单位：万）	规模增量（对比1月）	排名变化（对比1月）
1	豆包	AI综合助手	34,493	10,107	--
2	千问	AI综合助手	16,567	12,633	2位
3	DeepSeek	AI搜索引擎	12,717	63	1位
4	元宝	AI综合助手	5,735	820	1位
5	蚂蚁阿福	AI专业顾问	2,715	140	--
6	豆包爱学	AI学科教育	1,434	144	--
7	即梦AI	AI创作设计	1,352	507	--
8	Lovekey	AI社交互动	1,034	478	2位
9	Kimi	AI综合助手	834	57	1位
10	快对AI	AI学科教育	636	-2	1位

图：Anthropic ARR增长（2022-2026.5）



资料来源：华尔街见闻官网、QuestMobile公众号、西部证券研发中心

目 录

CONTENTS

建议关注

展望2026年下半年：芯片、互联、算力，技术进步与整体需求依旧蓬勃发展，模型能力的边界有望进一步拓展，助推AI应用的深度和广度持续提升。建议关注方向：

- AI芯片：寒武纪（已覆盖）、海光信息（已覆盖）、天数智芯
- 交换芯片：盛科通信（已覆盖）、曦智科技
- 互联芯片：澜起科技、裕太微、万通发展
- 光互联：中际旭创（通信组已覆盖）、新易盛、天孚通信（通信组已覆盖）
- 制造：中芯国际（电子组已覆盖）、华虹宏力（电子组已覆盖）
- 算力租赁：协创数据、宏景科技
- AIDC：商汤-W（已覆盖）、东阳光、润泽科技
- 大模型：智谱、MiniMax、商汤-W（已覆盖）
- 网络层：网宿科技、恒为科技（已覆盖）
- 应用：腾讯控股、阿里巴巴（已覆盖）、快手-W（已覆盖）、中控技术（已覆盖）、鼎捷数智（已覆盖）

目 录

CONTENTS

风险提示

- **AI 技术突破不及预期。**AI 发展受到许多因素的影响，包括数据的质量和可用性、算法的效率和准确性以及计算资源的限制等。
- **大模型应用落地节奏不及预期。**无论是通用大模型还是垂类大模型应用都在加速落地，虽然大模型推动了部分企业营收快速增长，但由于算力成本及研发投入较高，若大模型对企业盈利能力没有产生明显的拉动作用，那么整体的落地节奏可能会放缓。
- **宏观经济增长不及预期，IT预算不及预期。**下游需求主要来源于企业、政府机构等客户IT开支，与经济景气度、产业政策密切相关，若实际经济增长或IT开支较弱，将带来板块盈利和估值冲击。
- **国际环境发生变化。**若中美贸易摩擦加剧、地缘政治摩擦升级等，对于产业发展可能带来不利影响，不论是数字化、国产化趋势都将受到阻碍，此外国际政治经济环境变化也可能带来市场流动性、风险偏好等系统性变化，我们提示相关风险。

投资评级说明

西部证券—投资评级说明

行业评级	超配：	行业预期未来6-12个月内的涨幅超过市场基准指数10%以上
	中配：	行业预期未来6-12个月内的波动幅度介于市场基准指数-10%到10%之间
	低配：	行业预期未来6-12个月内的跌幅超过市场基准指数10%以上
公司评级	买入：	公司未来 6-12个月的投资收益率领先市场基准指数20%以上
	增持：	公司未来 6-12个月的投资收益率领先市场基准指数5%到20%之间
	中性：	公司未来 6-12个月的投资收益率与市场基准指数变动幅度相差-5%到5%
	卖出：	公司未来 6-12个月的投资收益率落后市场基准指数大于5%
报告中所涉及的投资评级采用相对评级体系，基于报告发布日后6-12个月内公司股价（或行业指数）相对同期当地市场基准指数的市场表现预期。其中，A股市场以沪深300指数为基准；香港市场以恒生指数为基准；美国市场以标普500指数为基准。		

分析师声明

本人具有中国证券业协会授予的证券投资咨询执业资格并注册为证券分析师，以勤勉的职业态度、专业审慎的研究方法，使用合法合规的信息，独立客观地出具本报告。本报告清晰准确地反映了本人的研究观点。本人不曾因，不因，也将不会因本报告中的具体推荐意见或观点而直接或间接收到任何形式的补偿。

联系地址

联系地址： 上海市浦东新区耀体路276号12层
北京市西城区月坛南街59号新华大厦303
深圳市福田区深南大道6008号深圳特区报业大厦10C

联系电话： 021-38584209

免责声明

本报告由西部证券股份有限公司（已具备中国证监会批复的证券投资咨询业务资格）制作。本报告仅供西部证券股份有限公司（以下简称“本公司”）机构客户使用。本报告在未经本公司公开披露或者同意披露前，系本公司机密材料，如非收件人（或收到的电子邮件含错误信息），请立即通知发件人，及时删除该邮件及所附报告并予以保密。发送本报告的电子邮件可能含有保密信息、版权专有信息或私人信息，未经授权者请勿针对邮件内容进行任何更改或以任何方式传播、复制、转发或以其他任何形式使用，发件人保留与该邮件相关的一切权利。同时本公司无法保证互联网传送本报告的及时、安全、无遗漏、无错误或无病毒，敬请谅解。

本报告基于已公开的信息编制，但本公司对该等信息的真实性、准确性及完整性不作任何保证。本报告所载的意见、评估及预测仅为本报告出具日的观点和判断，该等意见、评估及预测在出具日外无需通知即可随时更改。在不同时期，本公司可能会发出与本报告所载意见、评估及预测不一致的研究报告。同时，本报告所指的证券或投资标的的价格、价值及投资收入可能会波动。本公司不保证本报告所含信息保持在最新状态。对于本公司其他专业人士（包括但不限于销售人员、交易人员）根据不同假设、研究方法、即时动态信息及市场表现，发表的与本报告不一致的分析评论或交易观点，本公司没有义务向本报告所有接收者进行更新。本公司对本报告所含信息可在不发出通知的情形下做出修改，投资者应当自行关注相应的更新或修改。

本公司力求报告内容客观、公正，但本报告所载的观点、结论和建议仅供投资者参考之用，并非作为购买或出售证券或其他投资标的的邀请或保证。客户不应以本报告取代其独立判断或根据本报告做出决策。该等观点、建议并未考虑到获取本报告人员的具体投资目的、财务状况以及特定需求，在任何时候均不构成对客户私人投资建议。投资者应当充分考虑自身特定状况，并完整理解和使用本报告内容，不应视本报告为做出投资决策的唯一因素，必要时应就法律、商业、财务、税收等方面咨询专业财务顾问的意见。本公司以往相关研究报告预测与分析的准确，不预示与担保本报告及本公司今后相关研究报告的表现。对依据或者使用本报告及本公司其他相关研究报告所造成的一切后果，本公司及作者不承担任何法律责任。

在法律许可的情况下，本公司可能与本报告中提及公司正在建立或争取建立业务关系或服务关系。因此，投资者应当考虑到本公司及/或其相关人员可能存在影响本报告观点客观性的潜在利益冲突。对于本报告可能附带的其它网站地址或超级链接，本公司不对其内容负责，链接内容不构成本报告的任何部分，仅为方便客户查阅所用，浏览这些网站可能产生的费用和风险由使用者自行承担。

本公司关于本报告的提示（包括但不限于本公司工作人员通过电话、短信、邮件、微信、微博、博客、QQ、视频网站、百度官方贴吧、论坛、BBS）仅为研究观点的简要沟通，投资者对本报告的参考使用须以本报告的完整版本为准。

本公司具有中国证监会核准的“证券投资咨询”业务资格，经营许可证编号为：91610000719782242D。