2025 年 08 月 18 日

WAIC 2025 聚焦多种大模型、AI 算力芯片和服务器

——人工智能月度跟踪

行业及产业

电子

强于大市

一年内行业指数与沪深 300 指数对比走势：

资料来源：聚源数据，爱建证券研究所

相关研究

《电子行业专题报告：全品类科技硬件产品的贸易风险分析》2025-08-15

《电子行业点评：AI 时代半导体的变与不变》2025-08-13

《电子行业周报：国产 GaN 芯片进入 NVIDIA 供应链》2025-08-11

《电子行业周报：WAIC 2025，华为发布昇腾 384 超节点》2025-08-06

《电子行业周报：雅江水电工程带动 HVDC 加速发展》2025-07-29

证券分析师

许亮

S0820525010002

0755-83562506

xuliang@ajzq.com

联系人

朱俊宇

S0820125040021

021-32229888-25520

zhujunyu@ajzq.com

投资要点：

- 2025 年 7 月 26 日至 29 日，WAIC 2025（世界人工智能大会）在上海举行。本次大会以“智能时代，同球共济”为主题，汇聚全球顶尖科技企业与创新公司，集中展示前沿产品、技术及行业发展趋势，聚焦“学术突破、软硬结合、全球治理”等亮点。会议涵盖大模型、AI 算力芯片、服务器等人工智能核心领域，重点展示了多种大模型及智能体、AI 算力芯片和服务器。

- 在 WAIC 2025 大会上，阶跃星辰、商汤科技、腾讯、阿里云等企业展示了国产 AI 大模型与智能体的最新进展，各产品在技术特性与表现上各有侧重。阶跃星辰的 Step-3 采用 MoE 架构，在 MMMU、MathVision 等榜单获开源最佳成绩，同时 Step-3 在国产芯片上的推理效率、兼容性及成本控制表现突出，计划全球开源并联合厂商构建生态联盟；商汤科技日日新 V6.5 强调图文交错思维，多模态交互得分超过 Gemini 2.5 Flash 和 GPT-4o，性价比较前代提升 3 倍，依托升级的“商汤小浣熊”智能体综合测试得分与 Claude-4-Opus 相当；腾讯混元 3D 世界模型 1.0 开源后下载量超 230 万次，支持通过文本或图像快速生成可编辑虚拟世界，同步开源的四款端侧小模型具备 Agent 能力和长上下文处理能力，其中 7B 模型在特定测试中优于同类产品，体现轻量化趋势；阿里云 Qwen3-Coder 在编程测试中成绩接近 ClaudeSonnet4，输入输出成本均有优势，同步发布的千问 3 系列开源模型位居全球开源榜前列。这些进展反映出国产 AI 在多模态、特定领域能力及开源生态建设上的推进。

- 沐曦股份在本次 WAIC2025 发布基于国产供应链的旗舰 GPU——曦云 C600，这一成果标志着国产高性能 GPU 实现了历史性突破。曦云 C600 基于沐曦自主知识产权核心 GPU IP 架构，构建了从设计、制造到封装测试全流程的国产供应链闭环，实现核心技术自主可控。该芯片集成大容量存储与多精度混合算力，支持 MetaXLink 超节点扩展技术，并内置 ECC/RAS 多重安全防护模块。此外，沐曦还推出锚定云端智算推理的曦思 N 系列、覆盖智算与数据分析的曦云 C 系列通用 GPU 以及专攻图形渲染加速、服务云游戏与元宇宙的曦彩 G 系列，均依托自主技术与完整软件栈，适配不同场景需求。

- 华为于本次 WAIC 2025 展出 2025 年 4 月发布的 AI 算力集群解决方案 Cloud Matrix 384（即 Atlas 900 A3 SuperPoD）。Cloud Matrix 384 基于超节点架构，采用全对等（Peer-to-Peer）UB 总线，将 384 颗 NPU 与 192 颗鲲鹏 CPU 紧密互联；Atlas 900 A3 SuperPoD 还搭载昇腾 910C 芯片，算力达 300 PFLOPs，点到点访问时延不足 1 微秒，适用于大模型推理、MOE 训练等场景。与英伟达 GB 200 NVL72 相比，其芯片封装层性能稍逊（如 BF16 算力为对方 0.3 倍），但系统层级优势明显（BF16 算力、HBM 容量分别为对方 1.7 倍、3.6 倍）。

- **投资建议：**从本次 WAIC 的主要新技术产品展示情况来看，国产 AI 产业正在大模型，算力芯片和服务器设备多方面齐头并进发展。2022 年美国《芯片和科学法案》对先进算力芯片与半导体设备设出口限制，但是国内企业依然在持续升级大模型能力与硬件设备指标。我们认为国内 AI 产业链的相关投资机会值得长期关注。

- **风险提示：**1）先进算力芯片限制加强 2）下游应用需求不及预期 3）国产模型迭代升级迟缓

目录

图表目录

1. WAIC 2025 聚焦多种大模型、AI 算力芯片和服务器

2025 年 7 月 26 日至 29 日，WAIC 2025（世界人工智能大会）在上海举行。本次大会以“智能时代，同球共济”为主题，汇聚全球顶尖科技企业与创新公司，集中展示前沿产品、技术及行业发展趋势，聚焦“学术突破、软硬结合、全球治理”等亮点。会议涵盖大模型、AI 算力芯片、服务器等人工智能核心领域，重点展示了多种大模型及智能体、AI 算力芯片和服务器。

在 WAIC 2025 大会上，阶跃星辰、商汤科技、腾讯、阿里云等企业展示了国产 AI 大模型与智能体的最新进展，各产品在技术特性与表现上各有侧重。阶跃星辰的 Step-3 大模型采用 MoE 架构，在 MMMU、MathVision 等榜单获开源最佳成绩，同时 Step-3 大模型在国产芯片上的推理效率、兼容性及成本控制表现突出，计划全球开源并联合厂商构建生态联盟；商汤科技日日新 V6.5 强调图文交错思维，多模态交互得分超过 Gemini 2.5 Flash 和 GPT-4o，性价比较前代提升 3 倍，依托升级的“商汤小浣熊”智能体综合测试得分与 Claude-4-Opus 相当；腾讯混元 3D 世界模型 1.0 开源后下载量超 230 万次，支持通过文本或图像快速生成可编辑虚拟世界，同步开源的四款端侧小模型具备 Agent 能力和长上下文处理能力，其中 7B 模型在特定测试中优于同类产品，体现轻量化趋势；阿里云 Qwen3-Coder 在编程测试中成绩接近 ClaudeSonnet4，输入输出成本均有优势，同步发布的千问 3 系列开源模型位居全球开源榜前列。这些进展反映出国产 AI 在多模态、特定领域能力及开源生态建设上的推进。

沐曦股份在本次 WAIC2025 发布基于国产供应链的旗舰 GPU——曦云 C600，这一成果标志着国产高性能 GPU 实现了历史性突破。曦云 C600 基于沐曦自主知识产权核心 GPU IP 架构，构建了从设计、制造到封装测试全流程的国产供应链闭环，实现核心技术自主可控。该芯片集成大容量存储与多精度混合算力，支持 MetaXLink 超节点扩展技术，并内置 ECC/RAS 多重安全防护模块。此外，沐曦还推出锚定云端智算推理的曦思 N 系列、覆盖智算与数据分析的曦云 C 系列通用 GPU 以及专攻图形渲染加速、服务云游戏与元宇宙的曦彩 G 系列，均依托自主技术与完整软件栈，适配不同场景需求。

华为于本次 WAIC 2025 展出 2025 年 4 月发布的 AI 算力集群解决方案 Cloud Matrix384 （即 Atlas 900 A3 SuperPoD）。Cloud Matrix 384 基于超节点架构，采用全对等（Peer-to-Peer）UB 总线，将 384 颗 NPU 与 192 颗鲲鹏 CPU 紧密互联；Atlas 900 A3 SuperPoD 还搭载昇腾 910C 芯片，算力达 300 PFLOPs，点到点访问时延不足 1 微秒，适用于大模型推理、MOE 训练等场景。与英伟达 GB 200 NVL72 相比，

其芯片封装层性能稍逊（如 BF16 算力为对方 0.3 倍），但系统层级优势明显（BF16 算力、HBM 容量分别为对方 1.7 倍、3.6 倍）。

2. 多种国产大模型及智能体齐聚展会

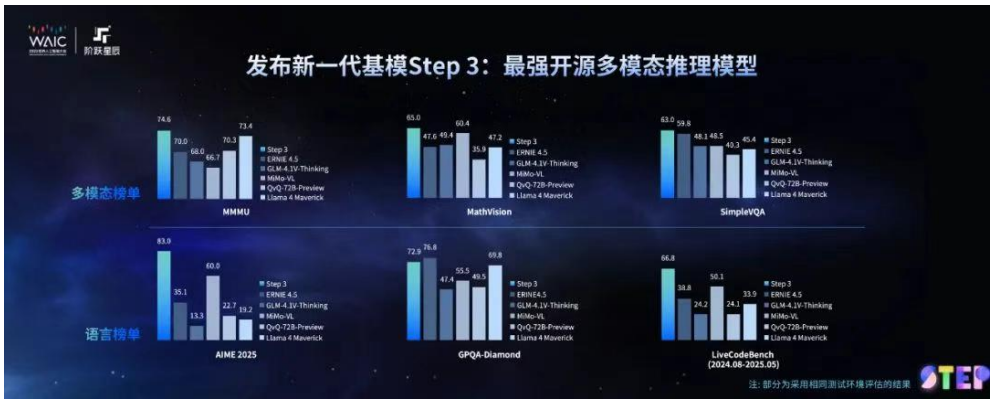
国产大模型与智能体持续加速迭代。阶跃星辰、商汤科技、阿里云、腾讯等企业纷纷推出相关大模型与智能体，助力国产 AI 加速发展。在本次大会上，亮相的大模型包括阶跃星辰的多模态推理大模型 Step-3、商汤科技的日日新 V6.5、腾讯的混元 3D 世界模型 1.0，以及阿里云的通义千问 Qwen3-Coder。这些产品不仅涵盖其他大模型的核心功能，并且在技术与应用层面展现出各自独特的创新。

■ 阶跃星辰发布多模态推理大模型 Step-3

在 WAIC 2025 大会上，阶跃星辰推出首个全尺寸、原生多模态推理模型 Step-3。该公司在确保模型效果的同时，兼顾了推理成本。目前，目前 Step-3 已授权多家芯片公司并完成适配，于 7 月 31 日面向全球开源。

Step-3 模型采用 MoE 架构，总参数量、激活参数量分别是 321B、38B。与 ERNIE 4.5 和 GLM-4.1V-Thinking 相比，该模型在 MMMU、MathVision、SimpleVQA、AIME2025、LiveCodeBench 等榜单上均取得了开源多模态推理模型的 SOTA（当前最佳）成绩；不过在 GPQA-Diamond 榜单上，Step-3（72.9）则略低于 ERNIE 4.5（76.8）。

图表 1：阶跃星辰 Step 3 测试数据



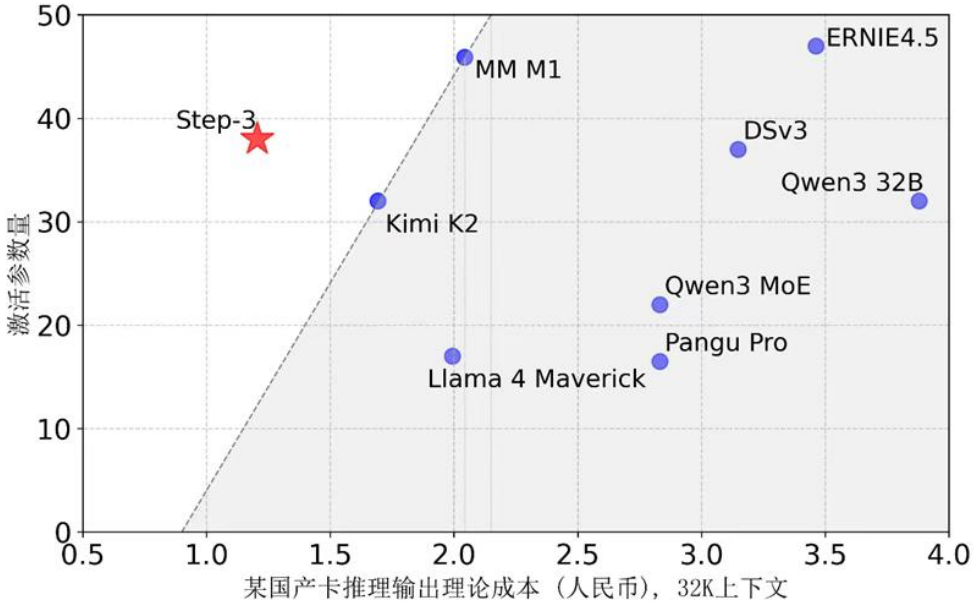
资料来源：阶跃星辰公众号，爱建证券研究所

Step-3 在国产芯片上的推理效率最高可达 DeepSeek-R1 的 300%，同时对所有芯片均具有良好的兼容性。基于英伟达 Hopper 架构芯片的分布式推理测试，Step-3 相比 DeepSeek-R1 吞吐量提升超过 70%。

在同等激活参数条件下，Step-3 的推理成本相对同类模型更低。Step-3 在激活参数的成本对比中表现突出——不仅激活参数量领先于 Kimi K2、Qwen3 MoE 等模型，还保持着较低的推理成本（每 32K 上

下文约 1.2 元）。

图表 2：阶跃星辰 Step-3 激活参数成本对比图



资料来源：阶跃星辰，爱建证券研究所

大会上，阶跃星辰联合近 10 家芯片及基础设施厂商共同发起“模芯生态创新联盟”。其中，华为昇腾芯片已率先完成对 Step-3 的适配与运行，其它联盟厂商的适配工作也正在有序推进。

■ 商汤科技发布日日新 V6.5 大模型

商汤科技在 WAIC 2025 大会上发布了日日新 V6.5（SenseNova V6.5）大模型。该模型是国内首个实现图文交错思维的商业级大模型，其全面升级标志着 AI 逐渐从”工具”向”人类”演进。

日日新 V6.5 具备推理能力强、效率高及支持智能体等优势。该模型采用图文交错的多模态思维链机制，整体性能可媲美 Gemini 2.5 Pro、Claude 4-Sonnet；在多项测试中，其多模态推理能力已超越这两款模型，多模态交互平均得分 77.97，领先于 Gemini 2.5 Flash（76.04）和 GPT-4o（75.40）。

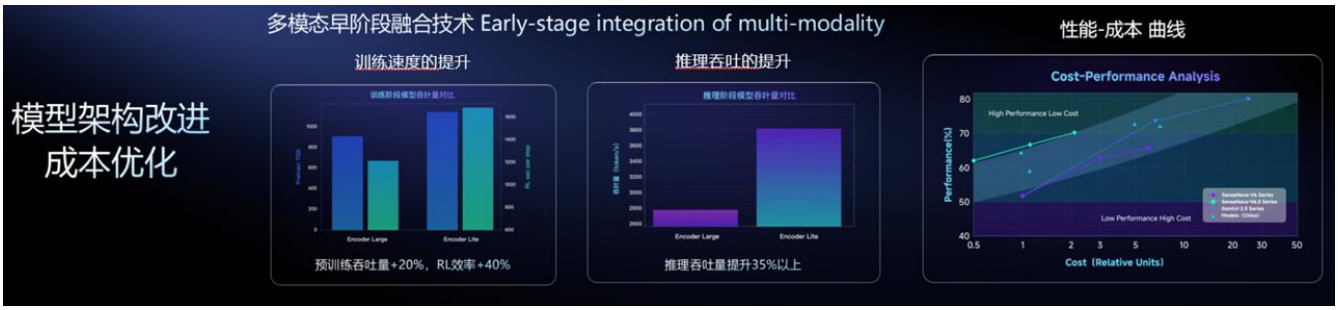
图表 3：日日新 V6.5 多模态推理与交互能力大幅提升

资料来源：商汤科技公司官网，爱建证券研究所

日日新 V6.5 通过优化多模态架构，整体性价比较前代 V6.0 提升 3

倍。依托多模态模型融合架构的改进，日日新 V6.5 实现了跨模态早期融合，有效提升模型预训练吞吐量（+20%）、强化学习效率（+40%），推理吞吐量提升更超 35%，较前代日日新 V6.0 实现了显著的性价比提升。

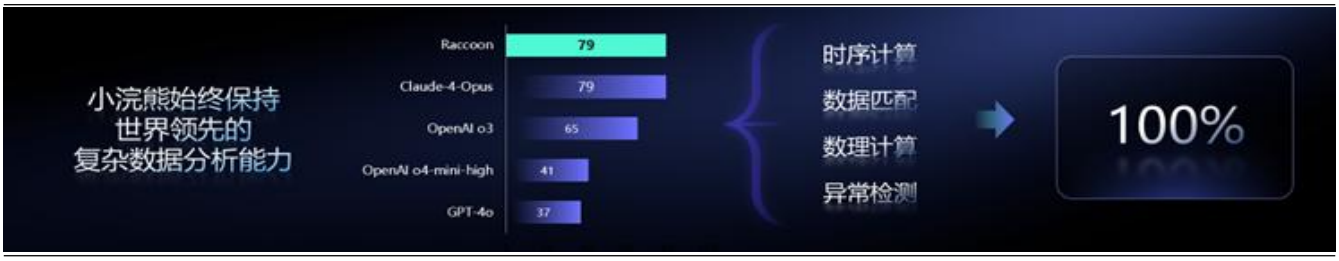
图表 4：日日新 V6.5 模型架构成本优化



资料来源：商汤科技公司官网，爱建证券研究所

智能体“商汤小浣熊”依托日日新 V6.5 的多模态数据分析能力实现全面升级。在客户场景综合测试中，其得分达 79 分，与世界标杆 Claude-4-Opus 水平相当，整体性能领先于 OpenAI o3 等模型；同时，“商汤小浣熊”在多项任务中的准确率接近 100%，交互范式更显高效智能。

图表 5：商汤科技小浣熊智能体在复杂数据分析领域保持世界领先水平



资料来源：商汤科技公司官网，爱建证券研究所

■ 腾讯混元 3D 世界模型 1.0 发布并开源

腾讯元宝在 WAIC 2025 大会上正式发布并开源混元 3D 世界模型 1.0。该模型支持用户通过输入一句话（文生）或一张图（图生），仅需几分钟即可生成可 360°漫游、可编辑的虚拟世界，同时输出标准化 3D 资产，且兼容主流引擎，能显著缩短内容生产周期。

图表 6：腾讯混元 3D 世界模型全景图

资料来源：腾讯云，爱建证券研究所

混元大模型持续迭代发展。2023 年 9 月，腾讯推出混元大模型，其发展方向从文本、图像、视频延伸至 3D 领域，并向轻量化演进。2024 年 5 月，开源文生图模型（混元 DiT）以强化开源生态；2024 年 9 月，推出新一代模型“混元 Turbo”，聚焦性能升级；2025 年 7 月，腾讯推出混元 3D 世界模型 1.0，该模型作为全球最受欢迎的 3D 生成开源模型，下载量超 230 万次；2025 年 8 月，开源四款端侧小模型（0.5B/1.8B/4B/7B），支持手机、车载等低功耗设备，推动模型轻量化发展。

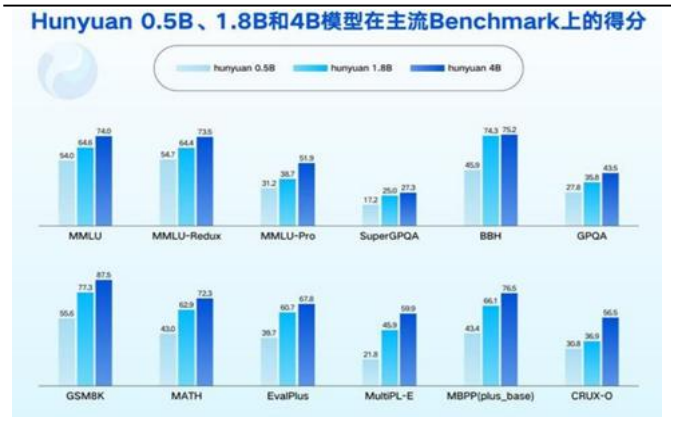
图表 7：腾讯混元模型发展

时间	发展事件	意义影响
2023 年 9 月	腾讯推出混元大模型，上线微信小程序“腾讯混元助手”	在指令理解、会中问答、会议摘要、会议待办项等多个方面，混元大模型均获得较高的用户采纳率。
2024 年 5 月	腾讯发布“腾讯元宝”，开源文生图模型（混元 DiT）	-
2024 年 9 月	腾讯推出新一代模型“混元 Turbo”	相较于前代模型推理效率提升 100%，解码速度提升 20%
2025 年 7 月	腾讯开源混元 3D 世界模型 1.0	全球最受欢迎的 3D 开源模型
2025 年 8 月	腾讯开源四款端侧小模型（0.5B/1.8B/4B/7B）	支持手机/车载等低功耗设备，提供 256K 上下文、双模式推理

资料来源：AINEWS，腾讯云，爱建证券研究所

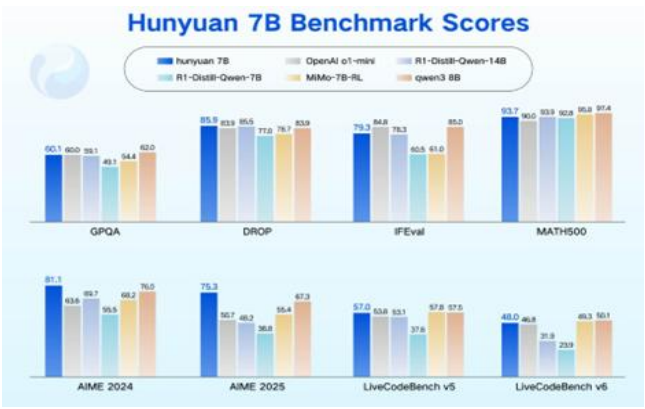
腾讯混元 8 月开源的四款模型（0.5B/1.8B/4B/7B），亮点在于具备 Agent 能力与长文处理能力，其上下文窗口达 256k；同时支持单卡部署，部分手机、平板等设备可直接接入，实现轻量化发展。其中，混元 7B 模型在 AIME 2024 和 AIME 2025 测试中的得分明显优于同类模型，在语言理解、数学、推理等领域表现出色。

图表 8：腾讯混元各模型于主流 Benchmark 上得分表现



资料来源：Github:Hunyuan-1.8B，爱建证券研究所

图表 9：腾讯混元 7B 模型于主流 Benchmark 上得分表现



资料来源：INCPAK，爱建证券研究所

■ 阿里云主要展出通义千问 Qwen3-Coder

在 WAIC2025 大会上，阿里通义千问重点展出最新开源的 AI 编程大模型 Qwen3-Coder。该模型具备 480B 参数（激活 35B 参数），原生支持 256K 上下文，还可通过 YaRN 扩展至 1M 长度。在 OpenAI 发布的 Agent 测试基准 SWE-bench Verified 中，Qwen3-Coder 500 轮交互测试得分达 69.6%，常规测试分数为 67%，可与最强闭源代码模型之一 Claude Sonnet4 的 70.4%、68%分数表现相媲美。

图表 10：阿里云通义千问 Qwen3-Coder

资料来源：阿里云官网，爱建证券研究所

在能力评测中，Qwen3-Coder 在浏览器调用（WebArena）、工具调用（BFCL）等智能体能力相关评测里，刷新了开源模型的纪录，成绩成功超越 DeepSeek-V3、GPT4.1。而在用于考察模型自主规划解决编程任务的 SWE-Bench 评测中，Qwen3-Coder 同样取得了开源模型中的最佳成绩，达到了可与 Claude4 媲美的水平。

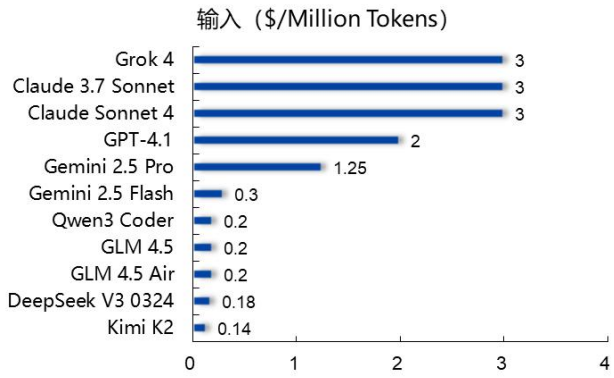
图表 11：Qwen3-Coder 与 Kimi K2、DeepSeek-V3 性能比较

Benchmarks	Open Models			Proprietary Models	
	Qwen3-Coder 480A35-Instruct	Kimi-K2 Instruct	DeepSeek-V3 0324	Claude Sonnet-4	OpenAI GPT-4.1
Agentic Coding					
Terminal-Bench	37.5	30.0	2.5	35.5	25.3
SWE-bench Verified					
w/ OpenHands, 500 turns	69.6	-	-	70.4	-
w/ OpenHands, 100 turns	67.0	65.4	38.8	68.0	48.6
w/ Private Scaffolding	-	65.8	-	72.7	63.8
SWE-bench Live	26.3	22.3	13.0	27.7	-
SWE-bench Multilingual	54.7	47.3	13.0	53.3	31.5
Multi-SWE-bench mini	25.8	19.8	7.5	24.8	-
Multi-SWE-bench flash	27.0	20.7	-	25.0	-
Aider-Polyglot	61.8	60.0	56.9	56.4	52.4
Spider2	31.1	25.2	12.8	31.1	16.5
Agentic Browser Use					
WebArena	49.9	47.4	40.0	51.1	44.3
Mind2Web	55.8	42.7	36.0	47.4	49.6
Agentic Tool Use					
BFCL-v3	68.7	65.2	56.9	73.3	62.9
TAU-Bench Retail	77.5	70.7	59.1	80.5	-
TAU-Bench Airline	60.0	53.5	40.0	60.0	-

资料来源：51CTO，爱建证券研究所

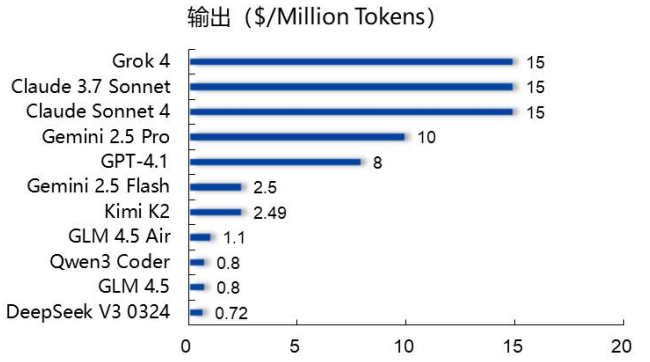
使用成本上，Qwen3-Coder 相较于 Grok 4、Claude Sonnet 4、Kimi K2 等模型具有价格低的优势。目前 Qwen3-Coder 输入、输出价格分别为$0.2/Million Tokens、$0.8/Million Tokens。

图表 12：Qwen3-Coder 输入价格优势明显



资料来源：openrouter，爱建证券研究所

图表 13：Qwen3-Coder 输出价格优势明显



资料来源：openrouter，爱建证券研究所

此外，阿里云还发布并开源了千问 3 最新版基础模型（非思考版）与千问 3 推理模型，两者均在全球开源榜中位居冠军。同时，阿里云还推出了首款专为 AI Agents 打造的“超级大脑”——无影 AgentBay。该产品能够实现跨系统无缝切换，并且可以随时调用算力、存储与工具链，为用户提供即开即用的智能开发环境，极大地提升了 AI Agents 开发的便捷性与效率。

3. 沐曦发布旗舰 GPU 曦云 C600

沐曦于 WAIC 2025 正式发布了基于国产供应链的旗舰 GPU 曦云 C600，标志着国产高性能 GPU 实现突破。

沐曦正式发布了基于国产供应链的旗舰 GPU 曦云 C600。该芯片基于沐曦自主知识产权核心 GPU IP 架构，构建从设计、制造到封装测试的全流程的国产供应链闭环，核心技术自主可控。曦云 C600 同时集成大容量存储与多精度混合算力，支持 MetaXLink 超节点扩展技术，并内置 ECC/RAS 多重安全防护模块，为金融、政务等关键领域提供高可靠算力基座。

图表 13：沐曦旗舰 GPU 曦云 C600

资料来源：沐曦公司公众号，爱建证券研究所

从公司产品性能来看：

曦思 N 系列深度锚定云端智算推理场景，依托高带宽内存与领先视频编解码能力，以高速显存配置、澎湃算力输出，支撑大规模数据推理与超高清视频流处理，搭配完整软件栈，实现智算任务高效部署。

曦云 C 系列作为通用 GPU 芯片，基于自主知识产权架构，具备超高精度算力与片间互联 MetaXLink 技术，支持多 GPU 系统无缝协同，借自主软件栈 MXMACA 构建全生态方案，覆盖智算研发、数据分析等复杂场景。

图表 15：沐曦股份主要产品分类

芯片名	介绍	产品特点	应用场景
曦思 N 系列	曦思 N 系列是面向云端应用的智算推理产品，采用高带宽内存，提供强大的算力和领先的视频编解码能力。	高速显存； 澎湃算力； 领先的视频处理能力； 完整的软件栈	智算
曦云 C 系列	曦云 C 系列通用 GPU(GPGPU)芯片是针对智算及通用计算的完美解决方案	自主知识产权 GPUPU; 超强高精度及混合精度算力 片间互联 MetaXLink 无缝连接多 GPU 系统； 自主软件栈 MXMACA 提供全面生态解决方案	智算； 数据分析
曦彩 G 系列	曦彩 G 系列 GPU 是针对图形渲染加速的解决方案，沐曦自主知识产权架构提供卓越的图形图像渲染与视频处理能力	卓越的图形图像渲染与视频处理能； 国产全功能显卡； 采取沐曦自主知识产权； 兼容主流 GPU 生态的完整软件栈	云游戏与元宇宙

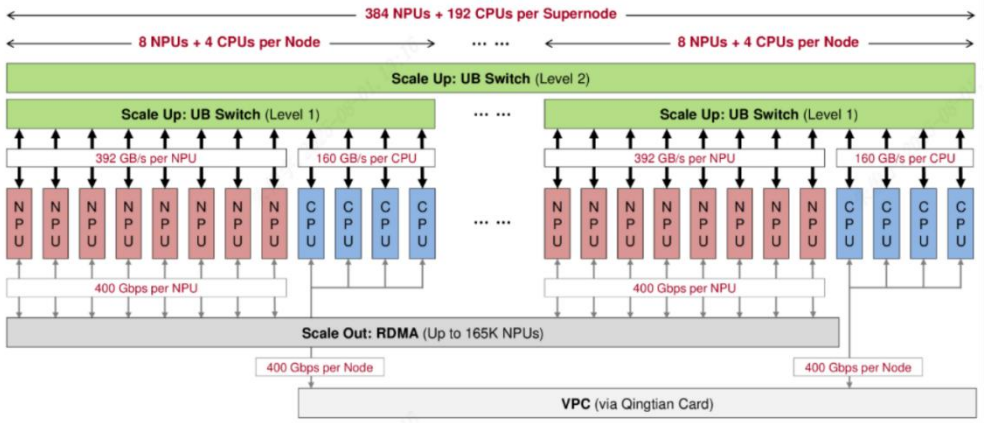
资料来源：沐曦股份官网，爱建证券研究所

曦彩 G 系列专攻图形渲染并加速，凭自主架构输出卓越图形图像渲染与视频处理性能，以国产全功能显卡身份，兼容主流 GPU 生态，为云游戏、元宇宙等场景提供高画质、低延迟的算力支撑。

4. 华为展示昇腾 384 超节点

2025 年 4 月华为发布的 AI 算力集群解决方案 Cloud Matrix 384，通过全互连拓扑架构实现芯片间的高效协同，并在本次 WAIC 2025 大会上展出。本次 WAIC 2025 上，华为线下展出的昇腾 384 超节点基于超节点架构，采用全对等（Peer-to-Peer）UB 总线，将 384 颗 NPU 与 192 颗鲲鹏 CPU 进行互联。

图表 16：华为 Atlas 900 A3 SuperPoD 架构分析



资料来源：Serving Large Language Models on Huawei CloudMatrix384，爱建证券研究所

Atlas 900 基于昇腾 910 芯片构建，总算力规模达 98PFLOPs，主要应用于大模型训练、科学研究及商业创新等场景，系统功耗为 736KW。

Atlas 900 A3 SuperPoD 搭载昇腾 910C 芯片，具备超大带宽、

超低延迟、超强性能优势。Atlas 900-A3-SuperPoD 算力达 300PFLOPs；在昇腾超节点集群上，LLaMA3 等千亿稠密模型训练性能为传统集群的 2.5 倍以上，Qwen、DS 及 MOE 模型性能提升达 3 倍；点到点访问时延不足 1 微秒，是业界唯一突破 15ms Decode 时延的方案；超节点内任意两服务器单向带宽达 392GB/s。该产品功耗 559KW，适用于大模型推理、MOE 模型训练、轨道交通、智能制造等场景。

图表 17：Atlas 900 与 Atlas 900-A3-SuperPoD 参数对比

参数	Atlas 900	Atlas 900-A3-SuperPoD
芯片构成	昇腾 910	昇腾 910C
算力规模	256-1024 PFLOPs @FP16	300 PFLOPs @BF16
通信时延	-	0.2 微秒
系统功耗	736KW	559KW
应用场景	大模型训练，科学研究与商业创新	大模型推理、MOE 模型训练、轨道交通、智能制造

资料来源：各公司官网，新浪财经，爱建证券研究所

华为昇腾 910C Cloud Matrix384（改名为 Atlas 900 A3 SuperPoD）与英伟达 GB200 NVL72 性能差异显著。芯片封装层，GB200 BF16 dense TFLOPS 达 2500 TFLOPS，是昇腾 910C（780 TFLOPS）的 3.2 倍；HBM 容量 192GB，为昇腾 910C（128GB）的 1.5 倍，带宽 8.0TB/s，是其 2.5 倍。

图表 18：华为昇腾 910 Cloud Matrix 384 与英伟达 GB200 NVL72 性能比较

芯片及封测层面				
	单位	GB200	Ascend 910C	华为与英伟达对比
BF16 密集型 TFLOPS	TFLOPS	2,500	780	0.3x
高带宽内存（HBM）容量	GB	192	128	0.7x
HBM 带宽	TB/s	8	3.2	0.4x
纵向扩展带宽	Gb/s uni-di	7,200	2,800	0.4x
横向扩展带宽	Gb/s uni-di	400	400	1.0x
系统层面				
	单位	NVL72	CM 384	华为与英伟达对比
BF16 密集型 PFLOPS	PFLOTS	180	300	1.7x
HBM 容量	TB	13.8	49.2	3.6x
HBM 带宽	TB/s	576	1229	2.1x
纵向扩展带宽	Gb/s uni-di	518400	1075200	2.1x
纵向扩展域规模	GPUs	72	384	5.3x
横向扩展带宽	Gb/s uni-di	28800	153600	5.3x
系统总功耗	W	145000	599821	4.1x
每 BF16 密集型 FLOP 的总功耗	W/TFLOP	0.81	2	2.5x
每内存带宽的总功耗	W per TB/s	251.7	488.1	1.9x
每内存容量的总功耗	kW/TB	10.5	12.2	1.2x

资料来源：Semi analysis，爱建证券研究所

在系统层级，华为 CM384 BF16 dense PFLOPS 达 300 PFLOPS，是英伟达 GB200 NVL72 的 1.7 倍；HBM 容量为 49.2TB，是 GB200 NVL72（13.8TB）的 3.6 倍。但在全系统功耗方面，昇腾 CM384 达 599.82KW，约为 GB200NVL72 的 4.1 倍。

5. 风险提示

1）先进算力芯片限制加强
2）下游应用需求不及预期
3）国产模型迭代升级迟缓

爱建证券
AIJIAN SECURITIES

爱建证券有限责任公司

上海市浦东新区前滩大道 199 弄 5 号
电话：021-32229888　　传真：021-68728700　　服务热线：956021
邮政编码：200124　　邮箱：ajzq@ajzq.com　　网址：http://www.ajzq.com

评级说明

投资建议的评级标准

报告中投资建议所涉及的评级分为股票评级和行业评级（另有说明的除外）。评级标准为报告发布日后 6 个月内的相对市场表现，也即：以报告发布日后的 6 个月内的公司股价（或行业指数）相对同期相关证券市场代表性指数的涨跌幅作为基准。其中：A 股市场：沪深 300 指数（000300.SH）；新三板市场：三板成指（899001.CSI）（针对协议转让标的）或三板做市指数（899002.CSI）（针对做市转让标的）；北交所市场：北证 50 指数（899050. BJ）；香港市场：恒生指数（HIS.HI）；美国市场：标普 500 指数（SPX.GI）或纳斯达克指数（IXIC.GI）。

股票评级

买入	相对同期相关证券市场代表性指数涨幅大于 15%
增持	相对同期相关证券市场代表性指数涨幅在 5%～15%之间
持有	相对同期相关证券市场代表性指数涨幅在-5%～5%之间
卖出	相对同期相关证券市场代表性指数涨幅小于-5%

行业评级

强于大市	相对表现优于同期相关证券市场代表性指数
中性	相对表现与同期相关证券市场代表性指数持平
弱于大市	相对表现弱于同期相关证券市场代表性指数

分析师声明

本报告署名分析师在此声明：我们具有中国证券业协会授予的证券投资咨询执业资格或相当的专业胜任能力，本报告采用信息和数据来自公开、合规渠道，所表述的观点均准确地反映了我们对标的证券和发行人的独立看法。研究报告对所涉及的证券或发行人的评价是分析师本人通过财务分析预测、数量化方法、或行业比较分析所得出的结论，但使用以上信息和分析方法可能存在局限性，请谨慎参考。

法律主体声明

本报告由爱建证券有限责任公司（以下统称为“爱建证券”）证券研究所制作，爱建证券具备中国证监会批复的证券投资咨询业务资格，接受中国证监会监管。

本报告是机密的，仅供我们的签约客户使用，爱建证券不因收件人收到本报告而视其为爱建证券的签约客户。本报告中的信息均来源于我们认为可靠的已公开资料，但爱建证券对这些信息的准确性及完整性不作任何保证。本报告中的信息、意见等均仅供签约客户参考，不构成所述证券买卖的出价或征价邀请或要约。该等信息、意见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任何人的个人推荐。客户应当对本报告中的信息和意见进行独立评估，并应同时考量各自的投资目的、财务状况和特定需求，必要时就法律、商业、财务、税收等方面咨询专家的意见。对依据或者使用本报告所造成的一切后果，爱建证券及其关联人员均不承担任何法律责任。

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、评估及预测后续可随时更改。过往的表现亦不应作为日后表现的预示和担保。在不同时期，爱建证券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。

版权声明