人工智能行业专题（22）
30B模型智能水平快速崛起，推动端侧算力发展

行业研究·行业专题

互联网·互联网II

投资评级：优于大市（维持）

证券分析师：张伦可
0755-81982651
zhanglunke@guosen.com.cn
S0980521120004

证券分析师：吴林轩
wulinxuan@guosen.com.cn
S0980526090002

证券分析师：刘子谭
liuzitan@guosen.com.cn
S0980525060001

报告摘要

- **模型趋势：轻量化模型能力快速跃迁，单位参数性能和Agent能力同步提升。**伴随轻量化、高性能模型持续推出，AI推理任务从云端向边缘和终端设备扩散。以2026年8月发布的Qwen3.8-27B为例，其Intelligence Index与300-3000B级头部模型的差距已缩小至个位数，并在部分编程、办公及Agent基准中超越更大参数模型。苹果、Google等厂商则通过稀疏加载、量化和模型架构优化，进一步降低端侧部署所需的内存与算力。Microduck发布则进一步展示了低成本端侧AI功能的应用潜力，模型研发落地节奏有望进一步加速。
- **算力趋势：端侧算力设备进入密集发布期，个人及办公推理任务有望加快向本地迁移。**当前端侧AI一体机主要分为个人AI超算、家庭及个人Agent主机、企业边缘推理设备三类。英伟达DGX Spark、小米AI Cube、苹果Mac mini及联想AI主机P7等产品整体覆盖30-200B级模型部署能力。相较云端推理，端侧模型推理拥有低时延、隐私保护、离线可用以及高负载下成本更低等优势。随着小参数模型性能提升，本地知识库、软件开发、内容生成等任务有望率先迁移至端侧。
- **产业趋势：芯片、终端与应用场景协同升级，从技术验证走向商业化落地。1）苹果：**依托自研芯片、端云模型、Apple Intelligence和Siri AI构建软硬件闭环，并逐步向操作系统级模型调度与Agent开发平台演进。**2）小米：**发布玄戒O3、O100、D100及AI Cube，完善手机、汽车、家庭AIoT和本地大模型算力布局。**3）爱芯元智：**AI推理产品矩阵覆盖0-1000+TOPS算力，26H1终端计算业务驱动高增，27-28年智驾业务放量可期，同时布局边缘大算力芯片适配端侧AI场景。**4）瑞芯微：**通过“主芯片+协处理器”架构切入AI Box、NAS及机器人等场景。**5）海康威视：**建立“设备—边缘—中心云”架构，已在城市、工业和企业场景广泛落地。
- **市场预测：中国AI推理芯片预计保持高增，端侧一体机场景放量可期。**根据灼识咨询，2024-2030年全球AI推理芯片市场规模预计由6067亿元增至30696亿元，CAGR为31%；其中云侧和边缘推理市场CAGR分别为36%和42%，高于端侧的20%。同期中国市场规模预计由1608亿元增至11664亿元，CAGR为39%，全球市场份额有望由27%提升至38%。中国边缘AI盒及一体机推理芯片出货量预计由80万颗增至1210万颗，CAGR达57%；对应市场规模由6亿元增至57亿元，CAGR为44%。
- **投资建议：端侧AI有望成为轻量化模型能力提升后的重要算力增量方向。**小参数模型正在跨越本地任务的可用性门槛，叠加端侧算力设备集中发布，个人及办公Agent需求有望率先落地。建议关注具备端侧软硬件生态和终端入口优势的公司。

目录

01 端侧模型能力加速跃迁，参数效率与Agent能力持续提升

02 端侧AI算力设备加速落地，个人/办公推理任务向端侧迁移

03 产业链公司加码端侧AI，芯片、终端与场景协同升级

04 端侧与边缘推理需求释放，中国芯片市场规模高增

05 风险提示

轻量模型能力跃迁，端侧AI落地窗口开启

国信证券 GUOSEN SECURITIES

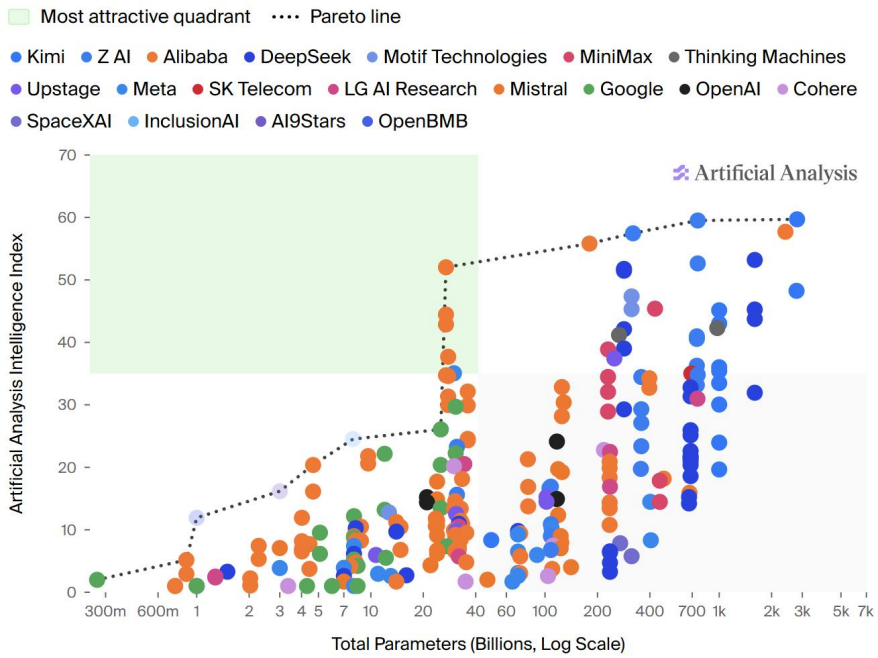
- **端侧模型参数体量可根据终端场景划分，参数规模随终端内存、带宽和功耗预算逐级上移。**1B以下模型主要用于可穿戴、IoT的唤醒、语音识别和简单感知；1-9B模型覆盖手机、平板及家庭AI Hub，可承担摘要、翻译、视觉理解和个人助理；7-30B模型进入AI PC和普通一体机，支持本地知识库、办公、编程及Agent任务；高内存一体机则可运行30-200B模型或低激活MoE，承接复杂推理、专业生产力和多智能体协同。汽车与机器人对参数规模的要求并非最高，但更加重视多模态感知、VLA控制、实时响应与可靠性。
- **头部轻量模型进入大模型能力区间，参数效率持续提升。**2026年8月24日千问推出Qwen3.8-27B，Artificial Analysis数据显示，其Intelligence Index与300B以上参数的头部模型（GLM-5.3、Kimi K3）轻量化版本的差距收窄至个位数，参数效率提升为端侧轻量模型落地提供推理能力支撑。

图：端侧模型参数规模随终端内存、带宽和功耗预算逐级上移

终端场景	主流参数	国内代表	海外代表
可穿戴/IoT	0.2-1B	Qwen3.5-0.8B、Qwen3-ASR-0.6B、MiniCPM5-1B	LFM2.5-230M/350M、SmolVLM2、Gemma 3 1B
普通手机	0.8-5B	Qwen3.5、MiniCPM-V 4.6、GLM-Edge；BlueLM-3B	苹果AFM、GeminiNano、Gemma3n、Llama3.2
旗舰手机/平板	3-9B	AI PC、MagicLM7B、AndesGPT、GLM-Edge-V5B、	Gemma3nE4B、Ministral8B、LFM2.5-8B-A1B
AI PC/普通一体机	7-30B	Qwen3.8-27B、MiMo-7B、MiniCPM-o	gpt-oss-20b、Ministral14B、Phi-4、LFM2-24B-A2B
汽车/机器人	2-30B，多模态/VLA	GLM-Edge、MiMo-Embodied、MiniCPM-o、Qwen3.5	Gemma 3 1B、Cosmos3Edge、NemotronNano
家庭AI Hub	1-9B	MiMo-VL-Miloco、MiniCPM-V、Qwen3.5	Gemma3n、LFM2.5、Llama3.2
高内存端侧一体机	30-200B或低激活MoE	Qwen3.8-Flash-Next、Qwen3.6-35B-A3B、盘古MoE-72B	gpt-oss-120b、Nemotron 3 Nano、Cosmos3-Super

资料来源：Qwen、OpenBMB、智谱、vivo AI Lab、小米MiMo、苹果、Google、Meta、Mistral AI、OpenAI、Liquid AI、英伟达、Microsoft、OPPO、Hugging Face、荣耀、华为、国信证券经济研究所整理

图：Qwen3.8-27B的Intelligence Index已接近300B以上参数的头部模型（GLM-5.3、Kimi K3）的轻量化版本



资料来源：Artificial Analysis、国信证券经济研究所整理 注：数据截止2026/8/27

Qwen3.8-27B：小参数开源模型，进入Arena WebDev榜单前十

国信证券 GUOSEN SECURITIES

- **Qwen3.8-27B小参数模型实现越级性能。**Qwen3.8-27B为27B参数稠密视觉语言模型，原生支持262K token上下文，可扩展至1M token。截至2026/8/27，模型在Arena WebDev公开榜单暂列第9，网页开发和视觉编程能力已进入全球头部梯队。

➢ 相较397B参数的MoE模型Qwen3.7-Plus，在Terminal-Bench2.1、SWE-bench Pro、CoWorkBench、OSWorld等测试中分别领先9.0/4.1/5.6/11.0pct，编码及办公任务能力显著提升。

➢ Qwen3.8-27B通过原生视觉理解、Agent后训练、环境反馈处理和思维上下文保留，提高复杂任务的端到端完成率。

图：Qwen3.8-27B在多项Agent能力超越Qwen3.7-Plus

能力方向	较Qwen3.7-Plus指标变化	能力应用及相关技术
权重规模显著降低	总参数397B→27B，约为其1/15	27B稠密、开源权重；降低本地存储门槛，适配高内存AI PC及端侧一体机
编程Agent能力提升	Terminal-Bench2.1：64.0→73.0；SWE-bench Pro：57.6→61.7；DeepSWE1.1：14.2→42.2	强化自主规划、环境反馈处理、终端工具调用及代码Harness适配
长周期办公能力增强	CoWorkBench：65.1→70.7；JobBench：27.6→33.4； Agents' Last Exam得分：33.6→42.9	多步骤任务规划、历史思维保留、任务状态衔接及跨工具协同
GUI与多模态Agent突出	OSWorld-Verified：73.3→84.3； WebArena-Verified：55.3→64.8； Vision2Web：42.1→62.9	原生视觉语言模型，强化屏幕理解、视觉定位、界面操作与工具调用
通用能力基本持平	IFBench：79.1→79.5； LiveCodeBenchv6：89.6→90.3；	优势集中于Agent执行而非知识推理全面领先，模型训练更偏向实际任务完成
长上下文与推理可控	原生262Ktoken，可扩展至1M；支持xhigh、medium、low三档推理深度	采用Gated DeltaNet与Gated Attention混合架构，支持reasoning_effort与preserve_thinking

资料来源：Qwen、国信证券经济研究所整理

图：Qwen3.8-27B在Arena WebDev暂列全球前十

Rank	Rank Spread	Model	Score	Votes	Price $/M
1	1 → 2	claude-opus-5-max Anthropic · Proprietary	1691 +9/-9	8,116	$5 / $25
2	1 → 4	kimi-k3-max Moonshot · Kimi K3 license	1674 +11/-11	4,546	$3 / $15
3	2 → 4	qwen3.8-max Alibaba · Proprietary	1669 +13/-13 Preliminary	3,212	$2 / $6
4	2 → 4	claude-opus-5-high Anthropic · Proprietary	1663 +8/-8	8,659	$5 / $25
⚡	N/A	glm-5.3-flash Z.ai · MIT	1634 +18/-18 AutoEval	N/A	$0.15 / $0.50
5	5 → 8	grok-4.6-high SpaceXAI · Proprietary	1629 +17/-17 Preliminary	1,523	$2 / $6
6	5 → 7	claude-fable-5 Anthropic · Proprietary	1626 +8/-8	8,382	$10 / $50
7	5 → 8	gpt-5.6-sol-xhigh (codex-h… OpenAI · Proprietary	1619 +8/-8	9,499	$4 / $20
⚡	N/A	qwen3.8-flash-next Alibaba · qwen-community-1.0	1617 +15/-15 AutoEval	N/A	$0.16 / $0.47
8	6 → 13	glm-5.3-max Z.ai · MIT	1599 +15/-15	1,916	$1.40 / $4.40
9	8 → 13	qwen3.8-27b Alibaba · Apache 2.0	1595 +13/-13	2,464	$0.40 / $3
10	8 → 13	gemini-3.7-flash-high Google · Proprietary	1587 +13/-13 Preliminary	2,552	$0.75 / $3.57

资料来源：Arena、国信证券经济研究所整理 注:数据截止2026年8月27日

Muse Glimmer 30B：Meta开源端侧Agent模型，压缩后24GB显存即可部署

国信证券 GUOSEN SECURITIES

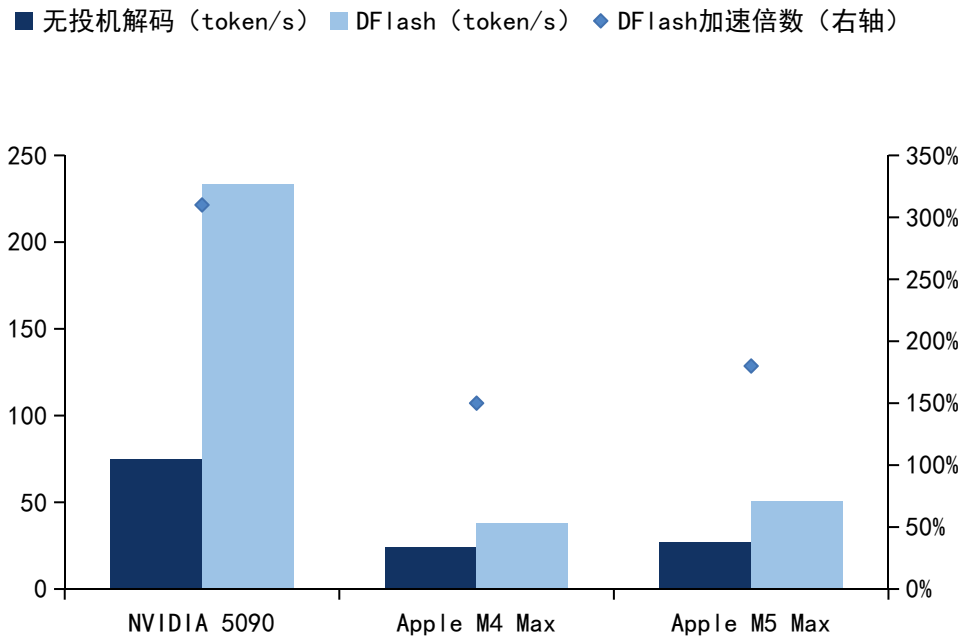
- **Muse Glimmer 30B于2026年8月发布，面向本地自主Agent。**Muse Glimmer约29.6B稠密参数，含约1.8B ViT-G/14视觉编码器，原生131K+上下文，可实现多步推理、函数/工具调用、失败恢复、代码任务与图文混合输入等功能，支持100+语言及4档推理强度。
- **模型端侧落地依靠4-bit K-Quant与DFlash投机解码：**17GB权重版本可部署于24GB显存，Dynamic版本面向32GB，15项基准平均损失仅1.0%/0.2%。

图：Muse Glimmer 30B在MCPAtlas、DeepSearchQA、SWE-BenchPro等测试中领先同参数规模的对手

Benchmark	Muse Glimmer	Gemma 4-31B	Qwen3.6-27B
MCPAtlasPublic	**75.5**	54.2	62.5
DeepSearchQA	**74.6**	61.7	71.1
Tau3-Banking	**23.5**	15.1	16.7
SWE-BenchPro	**51.2**	36.9	50.2
OSWorld-Verified	65.9	58.5	**75.6**
MMMUPro	74	73	**75**
AIME2026	**94.7**	89.2	94.1
Beam128K	**65.1**	58.2	63

资料来源：Meta、国信证券经济研究所整理

图：Muse Glimmer 30B可部署于英伟达5090和苹果M4 Max、M5 Max芯片



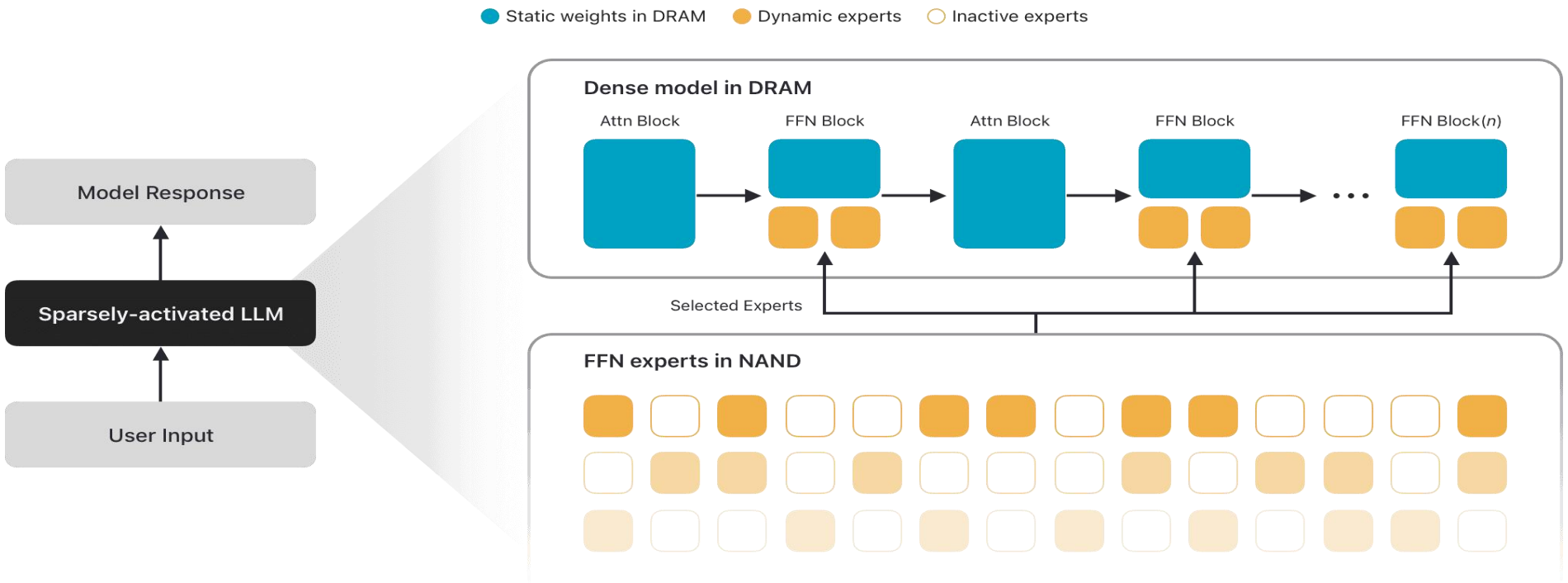
资料来源：Meta、国信证券经济研究所整理

苹果AFM 3 Core Advanced：20B模型部署手机端，按Prompt稀疏调用

国信证券 GUOSEN SECURITIES

- **苹果AFM 3 Core Advanced通过NAND全量存储+DRAM稀疏调用，实现手机端20B参数模型部署。**苹果第三代端侧模型包括约3B参数的Dense模型AFM 3 Core，以及20B参数的稀疏模型AFM 3 Core Advanced。后者将完整模型保存在NAND中，根据用户Prompt选择及具体用例选择性加载部分前馈模块（FFN Block）的专家权重，并与共享静态权重组合，构成DRAM中1-4B参数的稠密子模型。模型将落地于搭载苹果高性能芯片的设备，主要支持Siri自然语音、系统级听写、多模态能力和个人专属服务。

图：苹果AFM 3 Core Advanced在用户查询时会选择性加载部分前馈模块（FFN Block）的专家权重，并与共享静态权重组合，构成DRAM中的稠密子模型



资料来源：苹果、国信证券经济研究所整理

Gemma 4 E2B/E4B：移动端原生多模态，2.5GB内存实现4B参数有效能力

国信证券 GUOSEN SECURITIES

- **Gemma 4 E2B/E4B原生支持多模态，E4B仅2.5GB权重内存实现4.5B有效参数计算。**Google推出Gemma 4开放权重模型系列，其中E2B和E4B面向手机、笔记本电脑、IoT及其他边缘设备，支持文本、图像、音频和视频输入，原生上下文长度128K。E2B和E4B模型分别包含约2.3B/4.5B有效参数，采用Per-Layer Embeddings，在维持较低有效计算量的同时提升参数利用效率；其包含嵌入的总参数分别约5.1B/8.0B。采用移动端专用QAT量化格式及LiteRT-LM时，加载静态权重约需1.1GB/2.5GB内存。Google官方基准显示，思考模式下的E4B与Gemma 3 27B相比在推理、代码、Agent和多模态Benchmark上普遍提升。

图：Gemma 4 E2B和E4B模型分别包含约2.3B/4.5B有效参数

Property	E2B	E4B	12B Unified	31B Dense
Total Parameters	2.3B effective (5.1B with embeddings)	4.5B effective (8B with embeddings)	11.95B	30.7B
Layers	35	42	48	60
Sliding Window	512 tokens	512 tokens	1024 tokens	1024 tokens
Context Length	128K tokens	128K tokens	256K tokens	256K tokens
Vocabulary Size	262K	262K	262K	262K
Supported Modalities	Text, Image, Audio	Text, Image, Audio	Text, Image, Audio	Text, Image
Vision Encoder Parameters	~150M	~150M	-	~550M
Audio Encoder Parameters	~300M	~300M	-	No Audio

资料来源：Google、国信证券经济研究所整理

图：思考模式下的Gemma 4 E4B与Gemma3 27B相比在推理、代码、Agent和多模态Benchmark上普遍提升

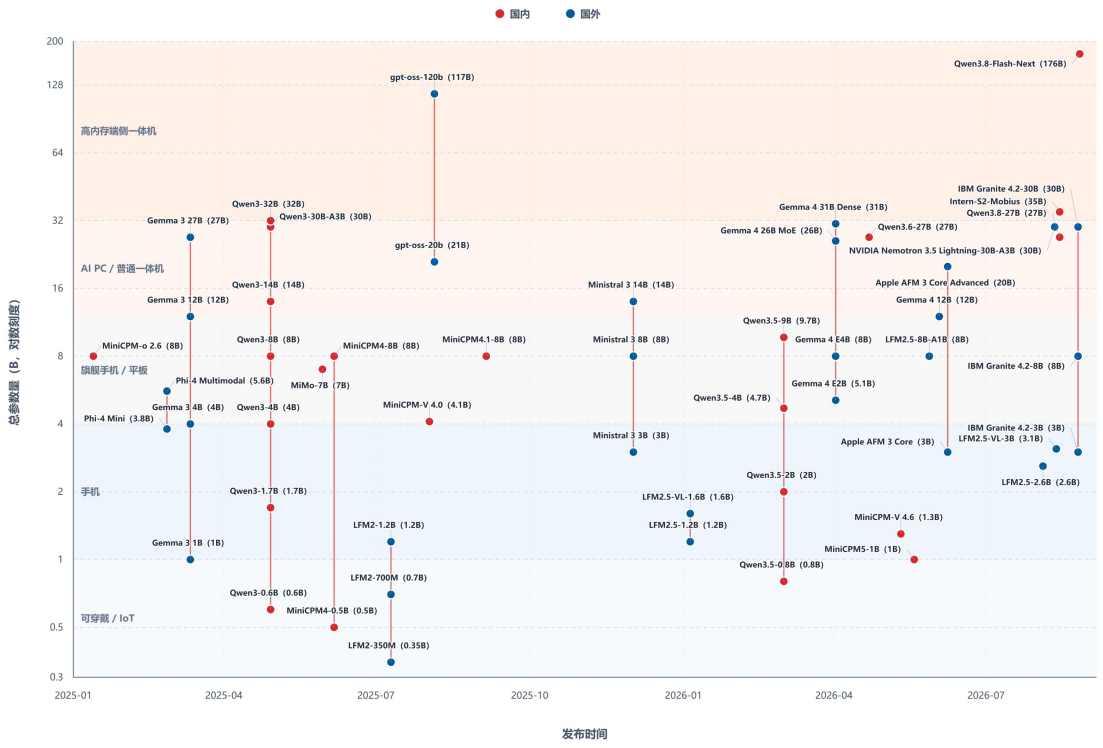
测试项目	Gemma 4 31B	Gemma 4 26B A4B	Gemma 4 12B Unified	Gemma 4 E4B	Gemma 4 E2B	Gemma 3 27B（无思考）
MMLU Pro	**85.2%**	82.6%	77.2%	69.4%	60.0%	67.6%
AIME 2026（无工具辅助）	**89.2%**	88.3%	77.5%	42.5%	37.5%	20.8%
LiveCodeBench v6	**80.0%**	77.1%	72.0%	52.0%	44.0%	29.1%
Codeforces ELO	**2150**	1718	1659	940	633	110
GPQA Diamond	**84.3%**	82.3%	78.8%	58.6%	43.4%	42.4%
Tau2（3次测试平均值）	**76.9%**	68.2%	69.0%	42.2%	24.5%	16.2%
HLE 无工具	**19.5%**	8.7%	5.2%	-	-	-
支持搜索的 HLE	**26.5%**	17.2%	-	-	-	-
BigBench Extra Hard	**74.4%**	64.8%	53.0%	33.1%	21.9%	19.3%
MMMLU	**88.4%**	86.3%	83.4%	76.6%	67.4%	70.7%
MMMU Pro	**76.9%**	73.8%	69.1%	52.6%	44.2%	49.7%
OmniDocBench 1.5（平均编辑距离，越低越好）	**0.131**	0.149	0.164	0.181	0.290	0.365
MATH-Vision	**85.6%**	82.4%	79.7%	59.5%	52.4%	46.0%
MedXpertQA MM	**61.3%**	58.1%	48.7%	28.7%	23.5%	-

资料来源：Google、国信证券经济研究所整理

端侧模型加速迭代，模型能力向长上下文、Agent拓展

- **模型发布密集，性能持续提升。**26H1主要端侧大模型发布9个系列、较25H1增加3个，模型发布更加密集。趋势上，整体参数规模未明显增加，但单位参数能力和任务闭环能力同时提升，30B参数以下模型的部分编程、办公、Agent能力逐步贴近大参数模型，有望推动基础个人/办公Agent推理任务加快向AI Box、一体机等边缘设备部署。

图：26H1主要端侧大模型发布9个系列、较25H1增加3个



资料来源：Qwen、Google DeepMind、OpenAI、Microsoft、苹果、英伟达、Mistral AI、Liquid AI、小米MiMo、OpenBMB、InternLM、IBM Research、国信证券经济研究所整理

图：Qwen3.8-27B端侧模型在部分能力上接近Qwen3.8-2.4T参数大模型

能力场景	Benchmark	Qwen3.8-27B	Qwen3.8-2.4T-Max	27B/2.4T得分比
Agent式编程	SWE-bench Pro	61.7	67.7	**91%**
软件工程	QwenSWEBench	79.0	80.7	**98%**
长周期办公任务	CoWorkBench	70.7	74.8	**95%**
真实场景工具调用	ToolathlonVerified	67.1	72.5	**93%**
科学推理	GPQA Diamond	89.2	92.6	**96%**
指令遵循	IFBench	79.5	82.8	**96%**
前沿Agent任务	Agents' Last Exam Score	42.9	52.4	82%
仓库级代码生成	NL2Repo-Bench	42.3	55.9	76%
跨学科推理	HLE	30.8	43.6	71%
专业职业任务	JobBench	33.4	53.4	63%

资料来源：Hugging Face、国信证券经济研究所整理

端侧部署兼具数据安全与成本优势，稳定高负载下经济价值显著

国信证券 GUOSEN SECURITIES

- **端侧部署兼具数据安全与成本优势，稳定高负载下经济价值显著。**端侧大模型兼具数据安全、隐私保护、低时延、离线可用与成本可控等优势，尤其适用于企业代码库、金融信贷数据、本地会议处理等数据敏感且搬运成本较高的场景，并有望以类似英伟达DGX Spark、苹果Mac mini的企业及个人AI设备的形式率先落地。

- 经济性方面，以每月3亿输入、1亿输出Token、GPU空转窗口控制在30%的Coding负载为例，5090运行Qwen3.8-27B的纯端侧月成本约105美元，端云混合模式约388美元，分别较Kimi K3纯云端API的1671美元节省约94%和77%，表明在高频、大体量的模型推理任务中，端侧部署具有显著成本优势。

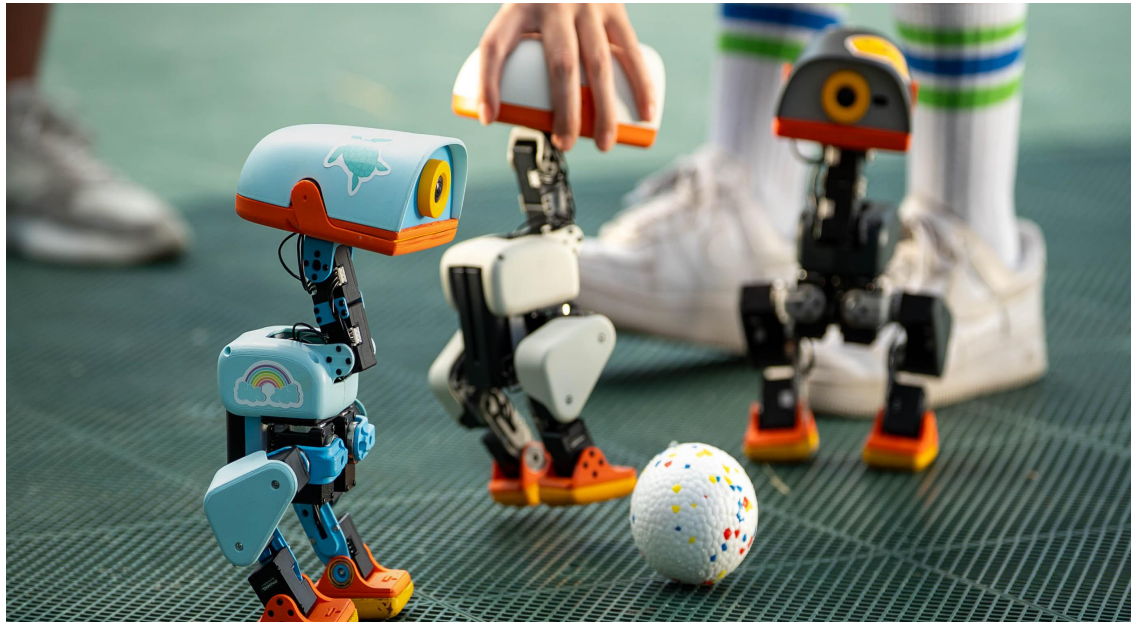
核心结果与数量假设	端侧Qwen3.8-27B	云端Kimi K3	端侧+云端混合模式
一、统一任务设定			
月输入Token（万）	30,000	30,000	30,000
月输出Token（万）	10,000	10,000	10,000
GPU空转时间窗口比例	30%	-	30%
二、端侧vs云端输出质量调整			
综合输出质量（以云端为基准）	83%	100%	85%
回退率	17%	0%	15%
月输入Token-经质量调整后（万）	36,288	30,000	35,198
月输出Token-经质量调整后（万）	12,096	10,000	11,733
三、云端Token定价			
输入缓存命中率		90%	90%
缓存命中输入价（美元/百万Token）		$0.3	$0.3
缓存未命中输入价（美元/百万Token）		$3	$3
输出价（美元/百万Token）		$15	$15
四、端侧硬件、能耗成本			
1）硬件成本			
5090整机购置成本（美元）	$3,000		$3,000
期末残值（美元）	$600		$600
折旧年限（月）	36		36
2）能耗成本假设			
①Prefill时长			
5090 FP16 Tensor算力（TFLPOS）	210		210
FP16模型计算量（TFLOP/Token）	0.054		0.054
每月Prefill时长（小时）	26		21
②Decode时长			
5090显存带宽（GB/秒）	1,792		1,792
访存量（GB/Token）	22		22
月Decode时长（小时）	405		335
月等效1亿Token GPU推理时间（小时）	**431**		**357**
月等效1亿Token GPU运行时间（小时）	**616**		**509**
月度时间占用率	**86%**		**71%**
5090官方TGP（W）	575		575
端侧用电量（kWh/月）	257		212
电价（美元/kWh）	$0.15		$0.15
五、端侧vs云端月度综合成本对比			
硬件折旧（美元/月）	$66.67		$66.67
电费（美元/月）	$38.48		$31.81
云端API费（美元/月）		$1,671.00	$289.54
成本合计（美元/月）	**$105.15**	**$1,671.00**	**$388.02**

资料来源：英伟达、Kimi、Qwen、国信证券经济研究所整理

Microduck：399美元开源双足机器人，端侧50Hz运行强化学习策略

- **Hugging Face旗下Pollen Robotics于2026年8月27日开启Microduck预售，售价399美元。**产品高25cm、重不足800g，采用15自由度双足结构，预售4天突破1万台。
- **功能及实现：**Microduck支持行走、坐立、摔倒起身、叼取、踢球及轮滑等7类预训练动作。硬件端，摄像头、8×8 ToF LiDAR和双IMU提供环境与姿态输入，15个舵机执行动作；软件端，mjlab（MuJoCoWarp）中以PPO训练，并通过域随机化、齿隙仿真完成sim-to-real，策略导出为ONNX后部署。
- **芯片及模型：**Microduck板载瑞芯微RK3566，四核Cortex-A55 CPU、Mali-G522EE GPU及1TOPS INT8 NPU，配1GB RAM和32GB存储；神经网络策略以50Hz闭环运行。软件栈、SDK、仿真与RL训练代码已采用Apache-2.0开源，引发极客的复现热潮。

图：Microduck产品形态



资料来源：Pollen Robotics、国信证券经济研究所整理

图：Microduck以瑞芯微RK3566作为算力大脑

硬件模块	Microduck硬件构成
主控芯片	瑞芯微RK3566，集成AI加速器
内存及存储	1GB RAM，32GB本地存储
运动系统	15个电机、15个自由度，覆盖双腿、颈部及头部
视觉传感器	前置摄像头，配备独立摄像头工作指示灯
距离传感器	紧凑型LiDAR，采用8×8ToF飞行时间测距矩阵
姿态传感器	2个IMU，分别位于机身和头部
物理交互	可活动机械喙，可用于抓取轻小物体
音频系统	麦克风及扬声器，每台机器人拥有独立生成的声音
NFC	2个NFC天线，分别位于头部和机械喙
无线连接	Wi-Fi、Bluetooth
电池	可拆卸NP-F550相机电池，容量2600mAh
随机控制器	标配游戏手柄，无需编程即可控制基础动作
整机规格	高25厘米、宽14厘米，重量不足800克

资料来源：Pollen Robotics、国信证券经济研究所整理

目录

01 端侧模型能力加速跃迁，参数效率与Agent能力持续提升

02 端侧AI算力设备加速落地，个人/办公推理任务向端侧迁移

03 产业链公司加码端侧AI，芯片、终端与场景协同升级

04 端侧与边缘推理需求释放，中国芯片市场规模高增

05 风险提示

端侧一体机场景分化：个人超算、Agent主机与企业边缘推理

国信证券 GUOSEN SECURITIES

- **端侧AI一体机根据应用场景可分为三大类：**
- 个人AI超算：以DGX Spark、小米AI Cube为代表，侧重大容量内存、高带宽和软件生态。
- 家庭/个人Agent主机：以联想AI主机为代表，侧重低功耗、办公助理功能和个人数据本地化。
- 企业边缘推理设备：以爱芯元智推理卡/服务器为代表，强调多模态、并发路数、私有知识库和国产化部署。

图：端侧AI计算设备向个人超算、家庭/个人Agent和企业边缘推理三类场景分化

类型	代表产品	算力及口径	内存及带宽	模型支持能力	核心场景
个人AI超算	英伟达DGX Spark	最高1PFLOP FP4	128GB统一内存；273GB/s	最高支持200B推理、70B微调	AI开发、模型微调、Agent和多模态应用
	小米AI Cube	O3 NPU 200TOPS	原型机80GB；O100近存带宽1.22TB/s；D100平台最高支持160GB	支持120B/3B模型切换	本地大模型、自研芯片协同，面向人车家算力生态
家庭/个人Agent主机	联想P7	整机异构算力190TOPS，其中M50约160TOPS	最高80GB；M50带宽153.6GB/s	最高支持122B、128K上下文及最高50token/s	个人知识库、本地Agent、办公助理和长期在线服务
企业边缘推理服务器	爱芯元智元曦AI服务器	-	-	面向多路模型并发，主要运行轻量视觉、语音及端侧模型	私有知识库、多路视频、长文本、政企内网等

资料来源：英伟达、小米、联想、爱芯元智、国信证券经济研究所整理

2.1 苹果Mac mini：自研M6/M5 Pro芯片，适配桌面Agent部署

- **苹果发布Mac mini及全新SoC，强化端侧AI能力。**苹果推出新一代Mac mini，提供M6和M5 Pro两种芯片配置，美国起售价分别为899美元和1699美元，将于2026年9月22日发售。产品可用于本地AI模型部署、桌面Agent搭载、软件开发、科学仿真等场景。

➢ **硬件方面，**M6配备12核CPU、12核GPU、GPU Neural Accelerator及双16核神经网络引擎，最高支持32GB统一内存和170GB/s带宽，AI性能最高较M4提升4倍；M5 Pro最高配备18核CPU、20核GPU和64GB统一内存，带宽提升至307GB/s，可运行更大的本地AI及扩散模型。

➢ **软件方面，**Mac mini将搭载最新操作系统macOS 27，后者通过Apple Intelligence和Siri AI强化个人上下文理解、屏幕内容识别、系统搜索及跨应用任务执行能力。

图：苹果Mac mini产品形态

资料来源：苹果、国信证券经济研究所整理

图：Mac mini提供M6和M5 Pro两种配置，起售价分别为899美元和1699美元

核心参数	M6基础版	M6中配版	M6高配版	M5 Pro版
定价	899美元	1099美元	1299美元	1699美元
CPU	12核	12核	12核	15核（可选18核）
GPU	12核	12核	12核	16核（可选20核）
AI单元	GPU Neural Accelerator；双16核Neural ngine	同左	同左	GPU Neural Accelerator；16核Neural Engine
内存带宽	153GB/s	153GB/s	170GB/s	307GB/s
统一内存	16GB(可选24GB/32GB)	16GB(可选24GB/32GB)	24GB(可选32GB)	24GB(可选32GB/48GB/64GB)
SSD	256GB(可选512GB/1TB/2TB)	512GB(可选1TB/2TB	512GB(可选1TB/2TB)	512GB(可选1TB/2TB/4TB/8TB)

资料来源：苹果、国信证券经济研究所整理

2.1 苹果Mac mini：自研M6/M5 Pro芯片，适配桌面Agent部署

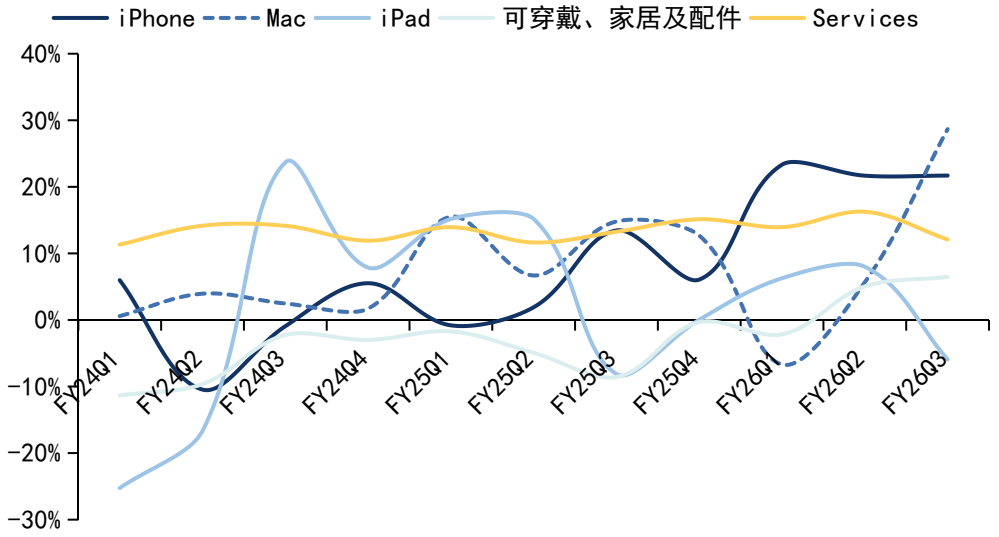
- **Mac产品矩阵AI能力显著提升，26Q3 Mac业务收入高增。**2026年苹果陆续推出M5、M6系列芯片，提供10-36核CPU、10-80核GPU、16-512GB内存及高速带宽，赋予Mac产品AI多模态、本地模型部署能力。相较英伟达GPU的显存与系统内存分离的结构，M系列芯片的CPU和GPU直接访问同一共享内存池，处理AI负载具备性能优势。同时，Mac能够通过EXO Labs开源软件组成集群来本地运行万亿参数规模的大模型，8月新发布的Mac Studio亦具备集群能力。据Information，OpenAI购买万台Mac mini与Mac Studio用于强化学习，Anthropic亦通过亚马逊AWS租用Mac mini完成类似任务。FY26Q3 苹果Mac收入同比+28.7%至103.5亿美元，主要得益于笔记本电脑的快速增长，未来有望进一步受益于端侧大模型部署需求。

图：2026年苹果陆续推出M5、M6系列芯片，Mac产品矩阵AI能力显著提升

产品系列	发布时间	芯片配置	CPU/GPU	统一内存及带宽	AI能力
MacBook Neo	2026/3/4	A18 Pro	6核CPU、5核GPU	8GB统一内存	AI图文功能
MacBook Air	2026/3/3	M5	10核CPU、最高10核GPU	最高32GB、153GB/s	AI图文功能
MacBook Pro	2026/3/3	M5 Pro/M5 Max	最高18核CPU、40核GPU	M5 Pro最高64GB、307GB/s；M5 Max最高128GB、614GB/s	AI多模态功能、本地模型部署
Mac mini	2026/8/25	M6	12核CPU、12核GPU	最高32GB、170GB/s	本地模型部署、办公Agent
Mac mini	2026/8/25	M5 Pro	最高18核CPU、20核GPU	最高64GB、307GB/s	本地模型部署、办公Agent
Mac Studio	2026/8/25	M5 Max	18核CPU、最高40核GPU	最高128GB、614GB/s	多模态功能、本地模型部署
Mac Studio	2026/8/25	M5 Ultra	最高36核CPU、80核GPU	最高512GB、1.2TB/s	本地大型模型部署

资料来源：苹果、国信证券经济研究所整理

图：FY26Q3苹果Mac收入同比+28.7%，主要得益于笔记本电脑的快速增长



资料来源：苹果、国信证券经济研究所整理

2.2 英伟达DGX Spark：本地支持200B模型推理，提供多元开发生态

- **CUDA生态的桌面AI超算，本地支持200B模型推理。**2025年3月18日英伟达发布桌面级个人AI超算DGX Spark，搭载GB10 Grace Blackwell超级芯片，提供最高1PFLOP FP4稀疏算力、128GB统一内存和4TB SSD，官方售价4699美元。产品支持最高200B参数模型本地推理及70B参数模型微调，可运行Qwen、Llama、FLUX、Cosmos等开源模型，并兼容英伟达软件栈，已经在高校科研、机器人、医疗评估、工程仿真和卫星图像分析中实现落地。

➢ 英伟达已为DGX Spark提供覆盖大模型推理、微调、量化、AI Agent、多模态生成、数据科学、机器人仿真及多机扩展的官方开发者案例，兼容Ollama、vLLM、PyTorch、LLaMAFactory、ComfyUI等主流开源生态。

图：英伟达DGX Spark已在高校科研、医疗AI及机器人等场景落地

场景	官方披露的代表项目
大语言模型	ISTA利用搭载相同GB10平台的HPZGXNano训练和微调最高7B参数模型；Stanford在本地运行120B参数gpt-oss模型并验证生物医学Agent流程
医疗AI	NYUICARE项目使用多Agent评估AI生成的放射学报告，并将医疗影像数据保留在本地
生物医学研究	哈佛KempnerInstitute使用DGX Spark分析约6,000种神经元基因突变，研究癫痫相关机制
机器人	ASU开展机器人感知研究，探索搜救机器狗和视障人士辅助工具
基础科学	威斯康星大学麦迪逊分校在南极IceCube中微子观测站本地运行AI分析
高校综合科研	ASU将多台DGX Spark用于记忆护理、交通安全、可持续能源等研究
沿海与体育研究	特拉华大学将GB10平台用于体育分析和沿海科学研究

资料来源：英伟达、国信证券经济研究所整理

图：英伟达提供量化、AI Agent、多模态生成、机器人仿真等开发者案例

分类	能力
大模型推理	本地运行大模型，支持图形化对话、连续批处理、量化模型及OpenAI兼容API
模型微调	支持PyTorch、NeMo、LLaMAFactory、Unsloth及LoRA等本地微调工作流
模型压缩与加速	支持NVFP4量化、TensorRT-LLM优化及推测解码，降低内存占用并提升潜在推理效率
AI Agent与知识应用	支持编程Agent、个人助理、RAG、多Agent协作及知识图谱构建
Agent安全	支持OpenShell沙箱隔离、系统权限控制及本地私有模型调用
图像及视频生成	支持FLUX、StableDiffusion、Wan和HunyuanVideo等模型工作流
视觉与多模态	支持实时视频理解、视觉问答、视频搜索及内容摘要
数据科学	支持GPU加速的数据处理、聚类、降维和机器学习
行业计算	支持单细胞RNA测序、投资组合优化等行业计算案例
编程开发	支持VSCode远程开发、自然语言编程和本地编程助手
机器人与具身智能	支持IsaacSim、IsaacLab机器人仿真及强化学习
多机扩展	支持DGX Spark互联、NCCL通信及分布式推理和微调
设备管理	支持系统监控、JupyterLab、SSH及跨

资料来源：英伟达、国信证券经济研究所整理

2.3 小米AI Cube：三芯异构集成，本地部署120B/3B双模型

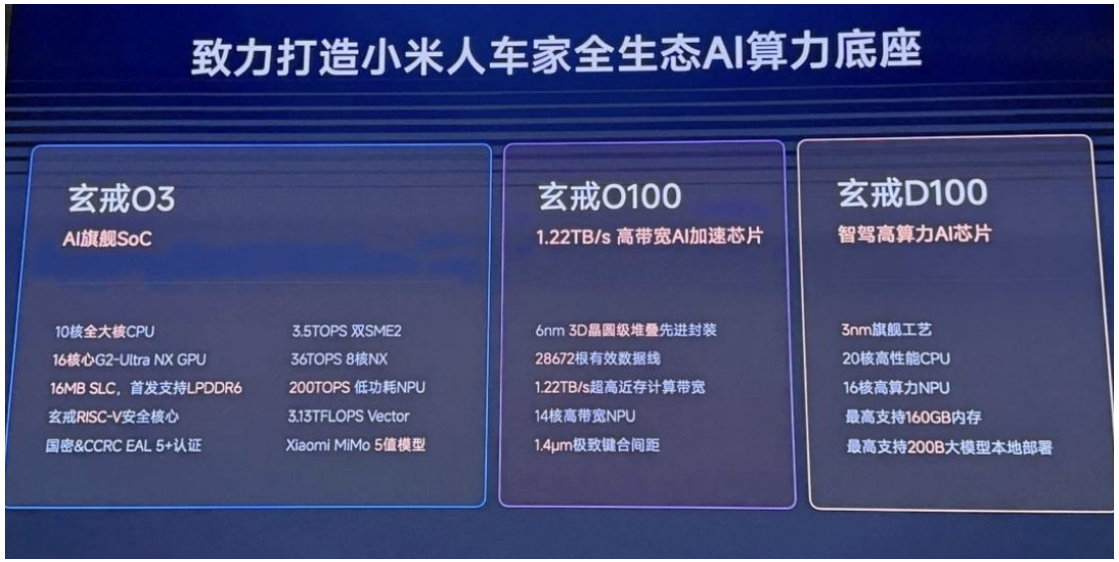
- **小米AI Cube采用三芯异构集成，本地部署120B/3B双模型。**小米AI Cube Prototype于2026年8月玄戒芯片技术沟通会上亮相，采用玄戒O3、O100和D100三芯协同架构，分别承担通用计算、近存AI加速和大容量模型计算，支持120B大模型与3B轻量模型的快慢系统协同。

➢ **三芯立体布局本地大模型及“人车家”智能生态。**玄戒O3集成10核CPU、16核GPU和200TOPS低功耗NPU，支持小米MiMo模型；玄戒O100采用6nm 3D晶圆级堆叠封装，提供1.22TB/s近存计算带宽；玄戒D100采用3nm工艺，集成20核CPU和16核NPU，平台最高支持160GB内存以及200B大模型本地部署。

图：小米AI Cube产品形态

资料来源：IT之家、国信证券经济研究所整理

图：小米AI Cube算力底座芯片性能



资料来源：IT之家、国信证券经济研究所整理

2.4 联想AI主机P7：本地支持122B模型，提供智能体/大模型双模式

- **联想AI主机P7以30W功耗运行122B模型，定位个人/家庭Agent算力节点。**联想AI主机P7是面向个人及家庭用户的AI原生边缘设备，搭载此芯P1 SoC+后摩智能M50（dNPU）算力核心，提供190TOPS本地AI算力、最高80GB内存，并支持122B参数模型和128K长上下文；本地推理速度最高可达50Token/s，满载功耗低于30W、重量约300g，可由普通移动电源驱动。

➢ P7提供智能体和大模型两种模式：智能体模式预装天禧Claw并执行长期任务；大模型模式可通过API向其他AI应用提供本地推理能力。

图：联想AI主机P7以低于30W功耗运行122B模型，提供190TOPS本地AI算力、最高80GB内存，并支持122B参数模型和128K长上下文；本地推理速度最高可达50Token/s

资料来源：联想、国信证券经济研究所整理

目录

01 端侧模型能力加速跃迁，参数效率与Agent能力持续提升

02 端侧AI算力设备加速落地，个人/办公推理任务向端侧迁移

03 **产业链公司加码端侧AI，芯片、终端与场景协同升级**

04 端侧与边缘推理需求释放，中国芯片市场规模高增

05 风险提示

3.1 苹果：软硬件体系完善，2026年集中换代升级

国信证券 GUOSEN SECURITIES

- **苹果端侧AI软硬件体系完善，2026年产品、技术能力显著提升。**苹果通过自研芯片、端侧及云端模型、Apple Intelligence系统编排和Siri AI入口形成完整的AI软硬件体系。功能实现方面，用户侧由Siri AI接收用户多模态上下文和屏幕内容，并将任务传达给Apple Intelligence动态选择模型和工具，完成内容生成和跨应用操作；外部开发者则可通过Foundation Models framework调用苹果自有或第三方模型构建AI应用。2026年苹果芯片、模型能力、操作系统、Siri AI等均换代升级，且开放第三方模型调用，推动产品从“内置AI功能”升级为“个人专属智能AI中枢”。

图：苹果端侧AI软硬件体系完善，2026年产品、技术能力显著提升

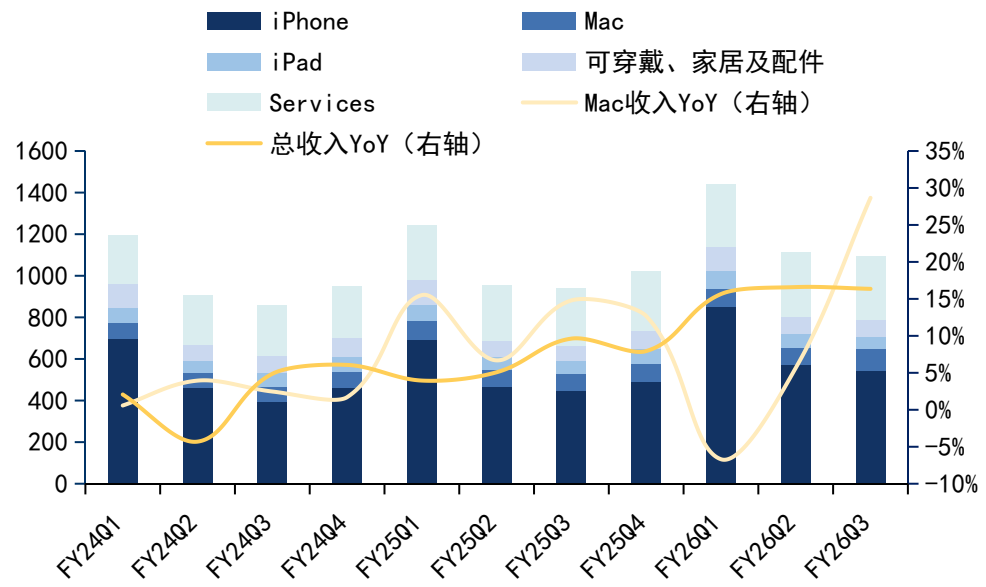
软硬件维度	最新技术/产品	更新时间	较上一代更新内容
Siri、操作系统与设备入口	Siri AI； iOS27、iPadOS27、macOS27、watchOS27、visionOS27； 覆盖iPhone、iPad、Mac、Apple Watch和Apple Vision Pro	2026/6/8	较2024版，Siri AI设备入口拓展至Apple Watch和Apple Vision Pro，新增个人上下文理解、屏幕感知、世界知识、持续对话和更多跨应用操作，并集成至Spotlight、系统上下文菜单及Vision Pro空间界面，同时支持跨设备同步Siri AI对话历史
芯片	A20 Pro	2026/9/9	较A19 Pro，A20 Pro采用2纳米制程，配备6核CPU、7核GPU和双16核Neural Engine；统一内存带宽提升50%，GPU性能最高提升40%，设备端AI处理性能提升至2倍；配合新一代VC均热板，持续性能最高提升40%
	M5、M5 Pro/Max/Ultra、M6	M5：2025/10/15； M5 Pro/Max：2026/3/3； M5Ultra/M6：2026/8/25	较M4，M5的GPU核心新增NeuralAccelerator，GPU AI峰值计算性能超过4倍；M5 Pro/Max/Ultra和M6进一步提高GPU规模、统一内存容量及带宽
端侧基础模型	新一代苹果Foundation Models	2026/6/8	较2024版3B参数模型，新模型优化了指令执行和复杂场景处理能力，提升Siri语音表现力、听写准确度以及图像与文字多模态输入能力
云端模型	PrivateCloudCompute、PrivateCloudComputeLanguageModel接口	2026/6/8	较2024版，新版PrivateCloudCompute首次向开发者开放标准模型接口，可在应用中调用具有更大上下文和更强推理能力的苹果云端模型
端侧/云端开发及调用框架	Foundation Models framework、Core AI、MLX	2026/6/8	较OS26版，新版Foundation Models framework支持图像输入，并统一接入设备端/云端苹果模型、Core AI及MLX本地模型和Claude、Gemini等兼容第三方模型；Core AI新增模型转换、预先编译及苹果silicon优化能力
Agent开发	Xcode 27	2026/6/8	较Xcode26.3，Xcode 27新增计划模式、多会话、UI验证、端到端应用操作、多Agent编排及GoogleAgent支持

资料来源：苹果、国信证券经济研究所整理

3.1 财报：FY26Q3收入高增，Mac增速领跑，研发费用率持续提升

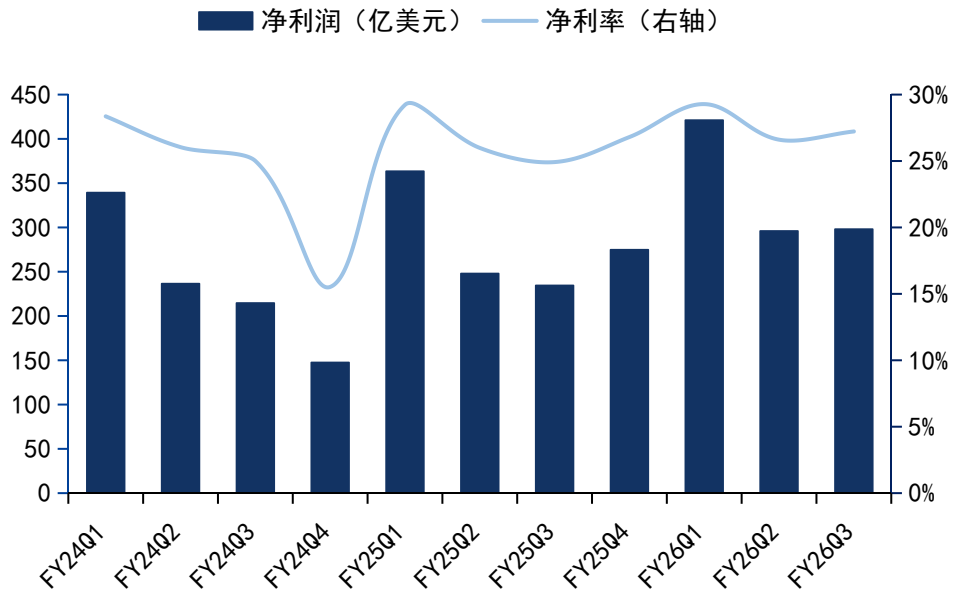
- **FY26Q3收入端，iPhone贡献高增量、Mac增速领跑各品类。**苹果FY26Q3收入1094.2亿美元，同比+16.4%。其中iPhone/Mac/iPad/可穿戴设备、家居及配件/Services分别收入542.5/103.5/61.9/78.8/307.4，分别同比+21.7%/+28.7%/-5.9%/+6.5%/+12.1%；iPhone贡献最高增量，全球份额稳步提升；Mac收入增速领先，主要由笔记本电脑驱动，未来有望受益于端侧大模型部署需求。
- **FY26Q3利润端，关税退税收益带动利润率改善，研发费用率持续提升。**FY26Q3毛利率同比+3.6pct至50.1%，预计得益于一次性的关税退税收益以及产品结构升级；研发/销售及管理费用率分别同比+1.3/-0.4pct至10.7%/6.7%；最终净利率同比+2.3pct至27.2%，净利润同比+27.1%至297.9亿美元。

图：FY26Q3收入端iPhone贡献高增量、Mac增速领先大盘（亿美元）



资料来源：苹果公司公告、国信证券经济研究所整理

图：FY26Q3利润率改善预计得益于一次性的关税退税收益以及产品结构升级



资料来源：苹果公司公告、国信证券经济研究所整理

3.1 人事变动：John Ternus接任CEO，AI团队协同增强

- **自2025年末起苹果陆续调整AI研发团队架构，加强硬件技术、产品及服务的整体协同**：引入AI助手工程人才、按职能重组AI团队，并统一芯片与整机管理，为AI能力转化为终端体验创造条件：
  - **Amar Subramanya任AI副总裁，补充AI产品化经验。**Amar曾任微软AI企业副总裁、Google Gemini Assistant工程负责人；加入后向软件工程高级副总裁Federighi汇报，有望加强基础模型研发与Siri、Apple Intelligence系统效果。
  - **John Giannandrea离任，原AI职责重新分配。**原AI职责分别划归AI副总裁Amar、首席运营官Khan和服务业务高级副总裁Cue，更贴近落地场景，有望降低跨部门协调成本。
  - **Johny Srouji升任首席硬件官，统一领导硬件技术与硬件工程团队。**有利于协调端侧AI的芯片算力、功耗设计与整机散热、续航的需求协同。
  - **John Ternus接任CEO，强化产品工程视角。**拥有25年苹果产品与硬件经验，有望提升AI硬件升级在公司产品战略中的重要性；Cook转任执行董事长，为管理交接提供连续性。

图：2026年9月1日JohnTernus正式接任CEO

资料来源：苹果公司公告、国信证券经济研究所整理

图：自2025年末起苹果陆续调整AI研发团队架构

生效时间	人事变动	职责变化
2025/12/1	Amar Subramanya加入苹果，任AI副总裁	负责苹果Foundation Models、机器学习研究、AI安全与评估，向CraigFederighi汇报
2025/12/1	John Giannandrea卸任机器学习与AI战略高级副总裁	转任顾问，公告计划于2026年春季退休原组织职责分别划归Amar Subramanya、Sabih Khan及Eddy Cue负责
2026/4/20	John Ternus被宣布为下一任CEO	_
2026/4/20	Johny Srouji升任首席硬件官	接管原由Ternus领导的硬件工程组织，同时继续领导硬件技术组织；两类组织统一由Srouji负责
2026/9/1	John Ternus正式接任CEO；TimCook转任执行董事长	Ternus出任CEO并加入董事会；Cook以执行董事长身份协助部分公司事务，包括与全球政策制定者沟通

资料来源：苹果公司公告、国信证券经济研究所整理

3.2 小米：AI相关产品线丰富，AI Cube完善端侧布局

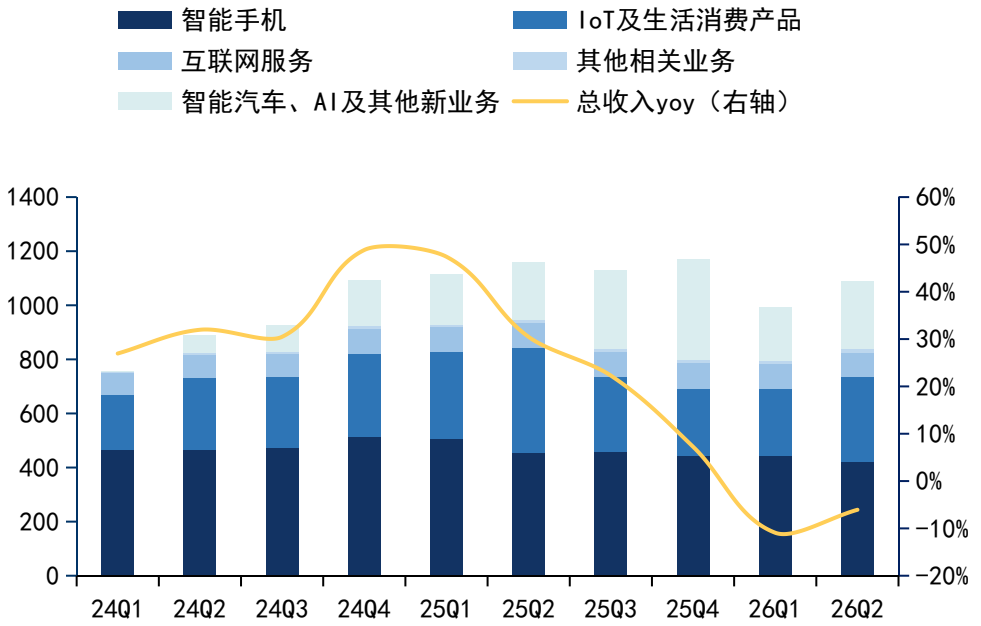
- **小米：AI相关产品线丰富，AI Cube完善端侧布局。**小米AI相关产品布局覆盖智能手机、平板/PC、智能汽车、家庭AIoT、智能穿戴及具身智能，整体采用端云协同架构，部分手机、汽车功能通过端侧算力实现。2026年8月小米推出AI Cube Prototype以及玄戒03、O100和D100芯片，填补端侧大算力产品空白。
- **存储涨价叠加国补退坡，26Q2收入承压。**26Q2公司收入1089.2亿元、同比-6.1%。受存储涨价、全球手机需求疲软及国内国补退坡影响，公司智能手机、IoT与生活消费品收入分别同比-7.5%/-19.2%至421.2/312.8亿元。其中IoT国内收入承压，部分被海外平板、智能电视及可穿戴产品增长抵消。

图：小米AI相关产品布局覆盖智能手机、平板/PC、智能汽车、家庭AIoT、智能穿戴及具身智能

AI相关业务	代表产品线	主要AI功能
手机	小米、REDMI、POCO	AI扩图、图片搜索、计算摄影、画质增强、反光消除、语音唤醒、部分识别与系统调度
平板/PC	小米、REDMI平板、笔记本电脑	AI写作、会议转写、摘要、翻译、搜索、图像生成、跨设备协作
汽车	小米智能汽车	语音交互、车辆控制、导航、座舱服务、环境感知、目标识别、轨迹预测、驾驶决策和泊车
家庭AIoT	智能家电、电视、音箱、摄像头、门锁、清洁机器人	语音控制、人物与环境识别、设备联动、能耗控制、家庭记忆、场景编排和主动执行
穿戴	智能手表、手环、耳机及智能眼镜	运动识别、健康分析、语音入口、音频降噪、通知与跨设备交互
具身智能	工业机器人、CyberDog、CyberOne机器人平台	多模态环境理解、视觉感知、任务规划、移动操作、物体分拣、装配和物流
边缘AI计算	小米AI Cube	大模型本地部署与推理、端侧多模态处理、模型开发验证及端云任务分流

资料来源：小米、IT之家、国信证券经济研究所整理

图：26Q2公司收入1089.2亿元、同比-6.1%，主因手机和国内IoT业务承压（亿元）



资料来源：小米公司公告、国信证券经济研究所整理

3.2 玄戒三芯覆盖手机、端侧AI及智驾，构建“人车家”算力底座

国信证券 GUOSEN SECURITIES

- **小米一次性发布玄戒O3、O100和D100，分别补齐高能效通用计算、高带宽端侧AI加速及智能驾驶高算力能力。**三款芯片于2026年8月24日玄戒芯片技术沟通会上发布，均已完成回片验证；其中O3已进入规模量产，计划首发搭载小米18 Fold，O100与D100预计于2027年正式商用。
- **三颗芯片采用异构分工，支持从3B端侧模型到200B级本地模型。**O3负责手机通用计算和轻量AI，O100通过1.22TB/s近存带宽加速模型推理，D100面向智能驾驶及大容量本地模型。同时小米发布三芯集成的小米AI Cube Prototype，配备80GB统一内存，可本地部署120B/3B快慢双模型，持续性能释放达150W。

图：玄戒O3、O100和D100分别定位通用计算、端侧AI及智驾应用场景

芯片	定位/进度	计算规格	内存/带宽
玄戒O3（AI旗舰SoC）	3nm；已规模量产，首发小米18Fold	200TOPS（A8W4）张量算力；3.13TFLOPS向量算力；4核低功耗NPU；10核CPU；16核G2-UltraNXGPU	LPDDR6；113.8GB/s
玄戒O100（高带宽AI加速芯片）	6nm3D晶圆级堆叠，已完成研发，2027年商用	14核大模型专用NPU	1.22TB/s近存带宽；
玄戒D100（智驾高算力AI芯片）	3nm，已完成研发，2027年商用	20核高性能CPU；16核NPU	最高160GB内存

资料来源：安兔兔、IT之家、国信证券经济研究所整理

图：玄戒三芯支持从3B端侧模型到200B级本地模型

芯片	支持模型	推理表现	落地场景与功能
玄戒O3	小米MiMo3B（5值量化）	MiMo3B：Prefill性能提升40%；Decode性能提升45%；运行功耗降低26%	端侧翻译；音视频理解；硬件AI超分及插帧；4K60AINR夜景视频；首发小米18Fold
玄戒O100	小米MiMo3B	3B模型峰值约330token/s；O3+O100原型机实测约295token/s	弱网或无网高速推理；移动端AI原型机；带宽敏感型端侧推理
玄戒D100	单芯最高支持200B级参数模型	_	高阶智能驾驶；车端大模型推理；AI Cube个人本地计算

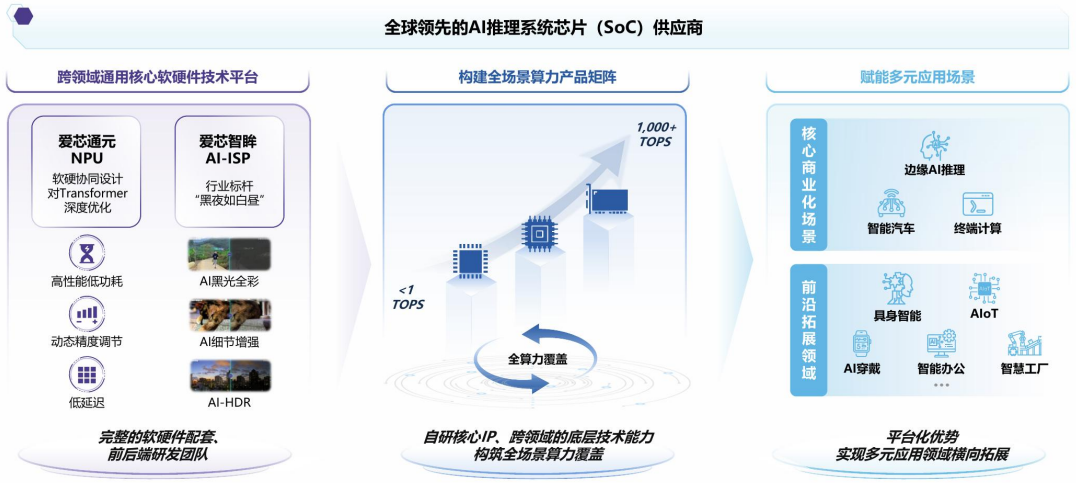
资料来源：安兔兔、IT之家、国信证券经济研究所整理

3.3 爱芯元智：AI推理SoC核心供应商，算力产品矩阵完备

- **爱芯元智AI推理算力产品矩阵完备，26H1终端计算业务高增。**爱芯元智成立于2019年，旗下拥有AI-ISP和通用NPU两大核心IP，并配套AI预处理、工具链及SDK。公司AI推理算力产品覆盖0-1000TOPS算力规格，赋能终端计算、边缘AI、智能汽车三大核心场景，可根据场景功耗和计算需求进行灵活适配。

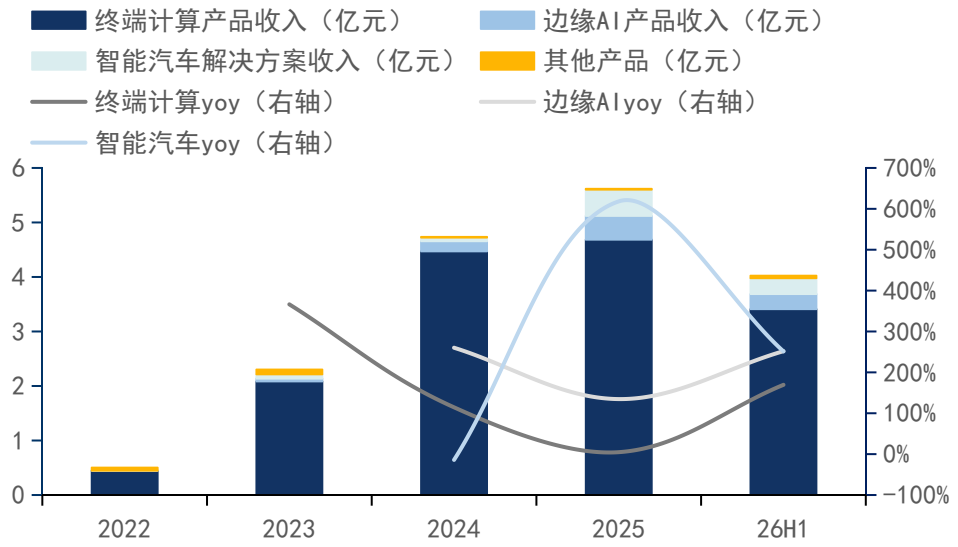
➢ 26H1公司收入4.0亿元、同比+181.8%；分业务看，终端计算产品收入3.4亿元、同比+169.5%，收入占比85%，为核心增长引擎；边缘AI产品/智能汽车解决方案分别收入0.3/0.3亿元、同比+251.9%/+251.3%，低基数下实现高增长。

图：爱芯元智AI推理SoC产品矩阵覆盖0-1000+TOPS，赋能多元应用场景



资料来源：爱芯元智26H1业绩交流、国信证券经济研究所整理

图：26H1爱芯元智收入4.0亿元、同比+181.8%，终端计算业务高增核心驱动



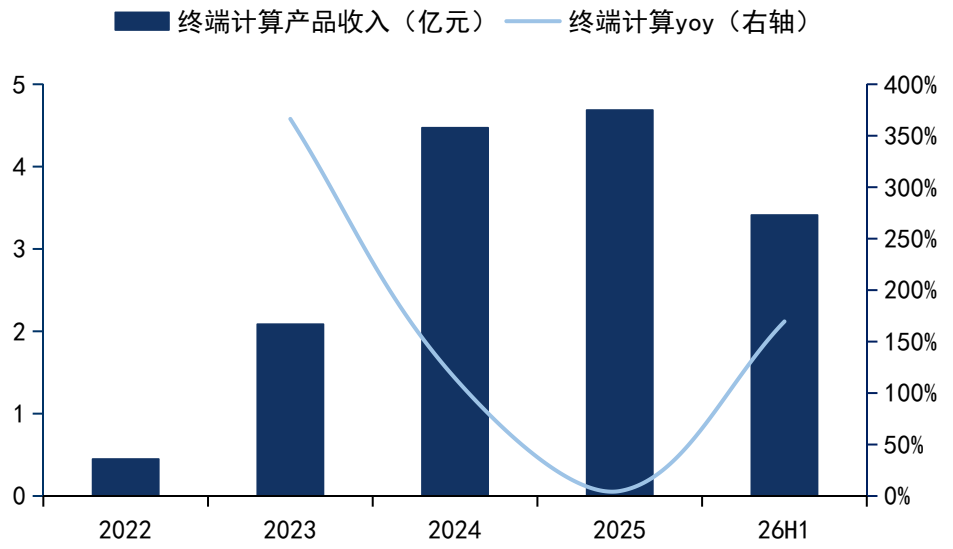
资料来源：爱芯元智公司公告、国信证券经济研究所整理

3.3 终端计算业务：AI-ISP赋能摄像头场景，中高端价格带领军者

- **AI-ISP赋能端侧视觉AI推理场景，中高端市场份额排名第1。**公司终端计算产品主要覆盖摄像头相关场景，在端侧通过AI-ISP对图像噪点和像素进行处理，重点改善低照度场景的成像质量。根据灼识咨询，2024年公司在视觉端侧AI推理芯片市场份额为6.8%（第5）、在中高端市场份额为24.1%（第1），处于核心领导地位。公司从中高端价格带切入市场，并逐步完善低端产品布局，当前产品矩阵覆盖0-100TOPS，下游包括安防、智能穿戴、宠物喂食机和可视电话等场景。

➢ 26H1公司终端计算收入同比+169.5%至3.4亿元，其中出货量同比超+100%、中高端黑光系列销量同比超+200%，市场份额显著提升，主要得益于下游需求增长以及存储大幅涨价驱动产品结构向高端升级。

图：26H1公司终端计算收入同比+169.5%至3.4亿元



资料来源：爱芯元智公司公告、国信证券经济研究所整理

图：公司终端计算产品主要覆盖摄像头相关场景，26H1中高端黑光系列高增

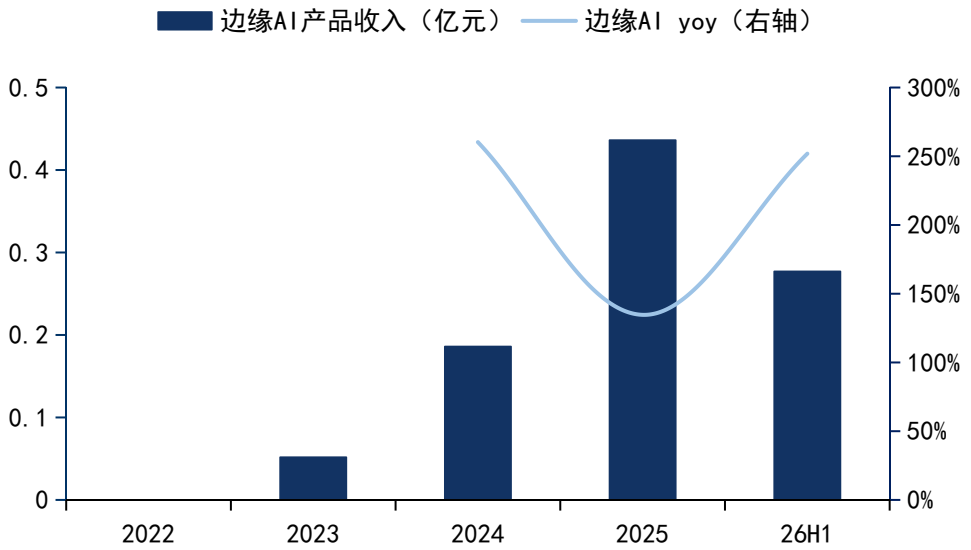
资料来源：爱芯元智26H1业绩交流、国信证券经济研究所整理

3.3 边缘AI业务：面向大模型本地化场景，受益AI Agent需求高增

- **边缘AI业务面向AI服务器、AI NAS、智能体盒子、具身机器人等本地算力场景，受益AI Agent需求高增。**26H1公司推出AXClawBox龙虾盒子以承接AI Agent应用的需求爆发；同时推出AX8850算力卡，面向NAS、PC等终端设备AI升级的场景，提供24TOPS算力、32GB最大内存容量，支持3B以下参数规模模型。同时公司下一代大算力AI芯片已流片，产品算力、带宽较AX8850有显著提升，并可级联成1000+TOPS加速卡，面向边缘服务器等大算力场景，有望打开业务成长空间。

➢ 26H1公司边缘AI收入同比+251.9%至0.3亿元，主要得益于市场需求扩容与公司相应新品发布，后续业务有望持续受益于个人/企业AI Agent需求增长及大模型本地化部署场景扩容。

图：26H1公司边缘AI收入同比+251.9%至0.3亿元



资料来源：爱芯元智公司公告、国信证券经济研究所整理

图：边缘AI业务面向大模型本地化场景，受益AI Agent需求高增

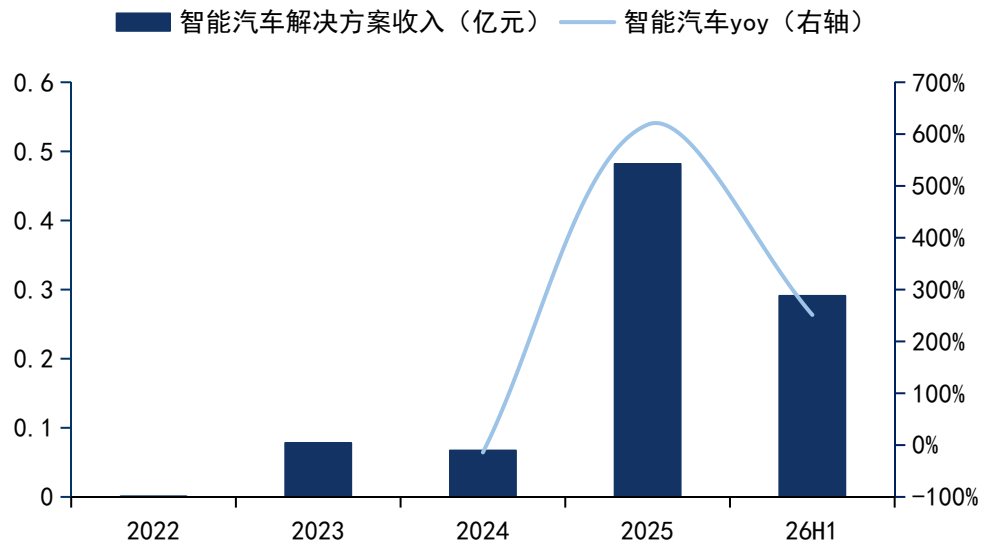
资料来源：爱芯元智26H1业绩交流、国信证券经济研究所整理

3.3 智能汽车业务：构建低阶-高阶智驾产品矩阵，受益AEB强标落地

- **公司智能汽车产品矩阵覆盖低阶-高阶智驾全品类，有望受益于AEB强标落地下的市场需求扩容。**公司当前拥有M55/M57两颗智驾芯片、面向L2场景，同时M97、M95两颗L2+高阶芯片于2026年成功点亮，有望于2027年放量。行业层面，中国计划于2027年要求新车型、2028年要求所有新车强制标配AEB，有望打开智驾芯片市场空间。

➢ 26H1公司智能汽车收入0.3亿元、同比+251.3%，总出货量42万颗、核心驱动业绩增长；同时新增量产车型定点24个，为未来业绩提升提供有力支撑。

图：26H1公司智能汽车收入0.3亿元、同比+251.3%，出货量核心驱动业绩增长



资料来源：爱芯元智公司公告、国信证券经济研究所整理

图：26H1新增量产车型定点24个，高端芯片成功点亮，有望驱动业绩高增

资料来源：爱芯元智26H1业绩交流、国信证券经济研究所整理

3.4 瑞芯微：主芯片与协处理器双轨布局，构建端侧+边缘算力基座

国信证券 GUOSEN SECURITIES

• **端侧+边缘算力基座布局：**公司采用“主芯片+协处理器”双轨架构，主芯片负责通用计算、系统控制、多媒体处理及多传感器接入，协处理器专门负责大模型推理与AI任务处理，并通过RKNN3工具链支持模型跨平台部署。公司已形成1TOPS至16TOPS以上的自研NPU产品梯队，覆盖智能家居、工业控制、机器人、智能座舱、AI Box、NAS及端侧智能体等场景。RK182X具备20TOPS级推理能力，单芯片支持3B/7B模型，RK1828多卡级联后可支持最高27B级稠密模型。

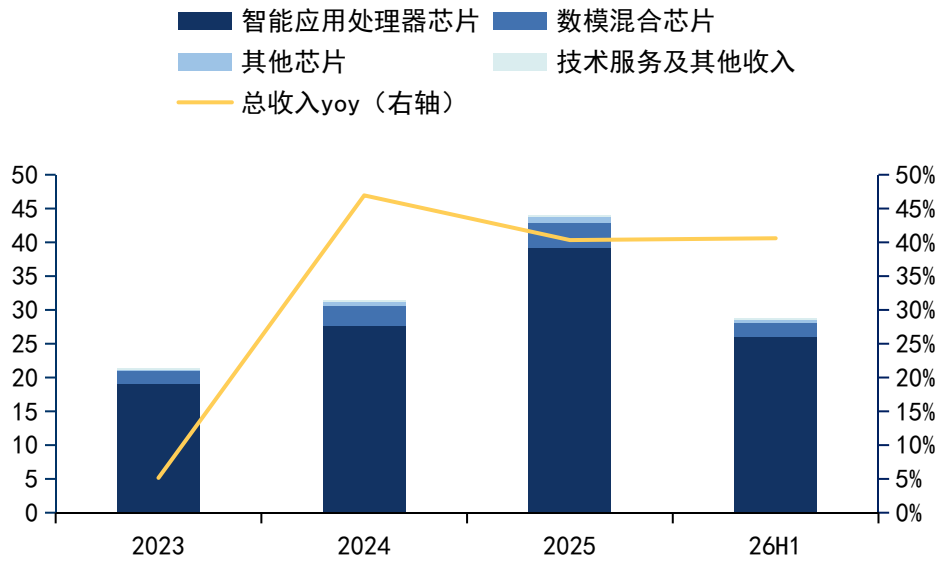
•**产品落地与生态扩张：**2025年7月公司首发RK1820/RK1828，26H1首批客户已进入高端智能电视、商用清洁机器人、陪伴机器人及NAS等产品发布和量产阶段；2026年5月发布4TOPS中阶SoC RK3572。2026年8月Hugging Face旗下Pollen Robotics推出的Microduck开源双足机器人采用RK3566，在端侧运行强化学习运动策略，以50Hz控制15个舵机并处理摄像头、传感器及无线通信，体现公司芯片平台在低成本具身智能开发设备中的落地能力。26H1公司收入28.77亿元，同比+40.6%。

图：瑞芯微边缘、端侧推理产品矩阵

层级/产品	性能与存储	典型应用
协处理器RK1820/RK1828	RK182X 20TOPS；2.5/5GB内置高带宽DRAM；3D堆叠，百GB/s级；支持3B/7B模型，最高适配8B模型	LLM/VLM、多模态推理、视频摘要、座舱AI Box、机器人“大脑”、本地知识库
旗舰主芯片RK3588	6TOPS；8核A76+A55；64-bit LPDDR4/4X/5；8K60解码、8K30编码	机器人、边缘计算、NAS、智能座舱、工业视觉、大屏设备、视频会议系统
中高端主芯片RK3576	6TOPS；8核A72+A53；32-bit LPDDR4/4X/5；8K30解码、4K60编码	智能座舱、机器人、机器视觉、云电脑、工业HMI、平板电脑、交互大屏
中阶主芯片RK3572（2026.05）	4TOPS；8核CPU；LPDDR5X/5/4X/4；8K解码、4K编码、双屏异显	AI摄像头、视频分析、NAS、零售设备、数字标牌、移动终端
视觉/工业RV1126B/RK3568J	RV1126B：3TOPS，支持≤2B模型；RK3568J：1TOPS、4核A55、双千兆网口	IPC、AOV低功耗视觉、工业网关、储能、PLC、AOI检测、数据采集

资料来源：瑞芯微公司公告、公司官网、国信证券经济研究所整理

图：26H1瑞芯微收入28.77亿元，同比+40.6%，芯片进入量产兑现期（亿元）



资料来源：瑞芯微公司公告、国信证券经济研究所整理

3.5 海康威视：智能视觉安防龙头，建立云边端AI体系

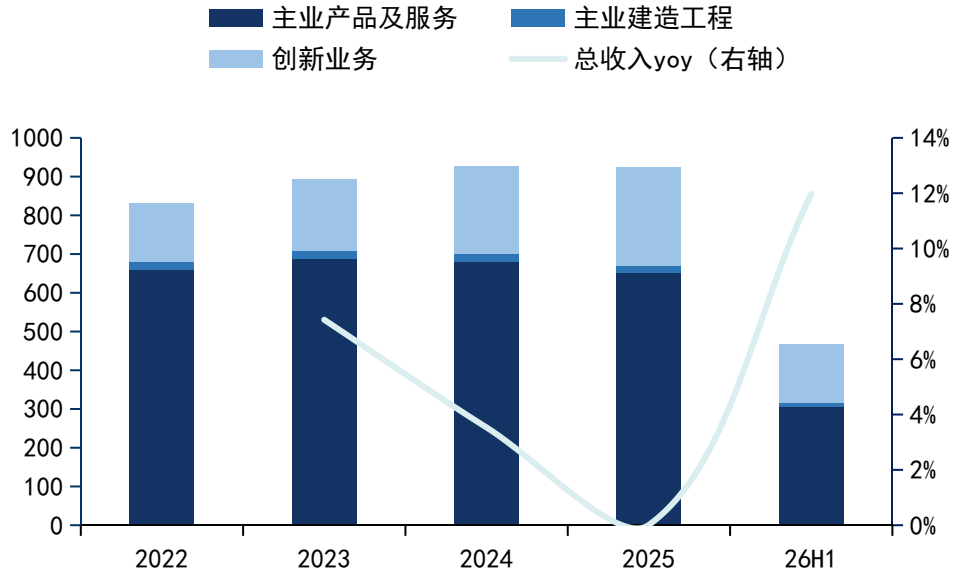
- **海康威视：智能视觉安防龙头，建立云边端AI体系。**公司已形成覆盖“设备-边缘-中心云”的云边端一体化AI架构：1）**设备端**以DeepinView X大模型摄像机、HeatPro热像设备及智能交通、X光安检和车载感知设备为代表，完成视频、热成像等多模态数据采集、低时延分析、AI编码与异常识别；2）**边缘端**通过文搜NVR、文搜超脑、巡检超脑及AI Box，就近实现自然语言文搜、目标检索、周界分析和多算法推理；同时利用智能融合主机、全分析服务器及文搜、编码磁盘阵列，承载区域级多路视频结构化、语义建模、亿级数据检索和智能编码；3）**中心云**则以DS-VG多GPU服务器、训练推理平台和大模型一体化平台，支持模型训练、集中推理及大模型部署。
- **受益“反内卷”及AI创新，26H1收入稳健增长。**26H1公司收入468.2亿元，同比+12.0%，其中主业产品及服务/主业建造工程/创新业务分别同比+4.6%/+32.8%/+28.9%至306.2/10.4/151.7亿元，主要受益于行业“反内卷”及AI产品创新下的增量需求。

图：智能视觉安防龙头，建立云边端AI体系

部署层级	代表产品	AI能力
设备端	DeepinView X大模型摄像机、HeatPro热像设备、智能交通/X光/车载感知设备	多模态感知、低时延AI分析、AI编码、异常检测、安检识别
边缘端-现场级	文搜NVR、文搜主机、文搜超脑、AI开放平台超脑、巡检超脑、AI Box	自然语言文搜、目标检索、周界分析、多算法推理及场景化智能应用
边缘端-区域级	智能融合主机、全分析服务器、文搜型CVR、多模态文搜及观澜编码磁盘阵列	多路结构化分析、语义建模、亿级文搜及AI智能编码
中心云	DS-VGAI服务器、训练平台、推理平台、大模型一体化平台	多GPU训练、集中推理、HPC、模型管理及基础模型—行业模型—任务模型部署能力

资料来源：海康威视公司公告、国信证券经济研究所整理

图：26H1公司收入同比12.0%至468.2亿元，主要受益于行业“反内卷”及AI创新下的增量需求（亿元）



资料来源：海康威视公司公告、国信证券经济研究所整理

目录

01 端侧模型能力加速跃迁，参数效率与Agent能力持续提升

02 端侧AI算力设备加速落地，个人/办公推理任务向端侧迁移

03 产业链公司加码端侧AI，芯片、终端与场景协同升级

04 **端侧与边缘推理需求释放，中国芯片市场规模高增**

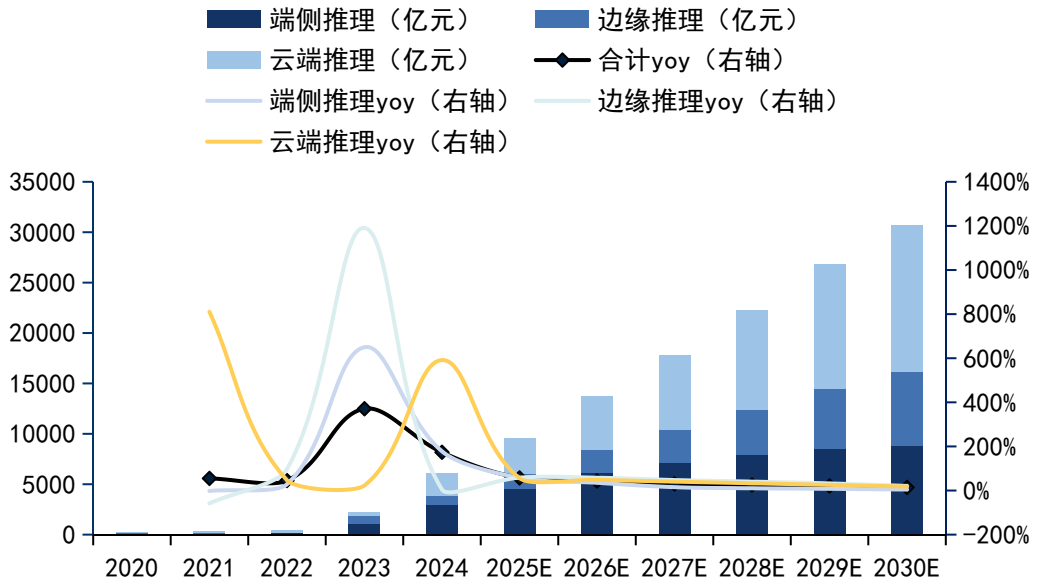
05 风险提示

全球AI推理芯片市场规模高增，边缘、云端场景增速领先

国信证券 GUOSEN SECURITIES

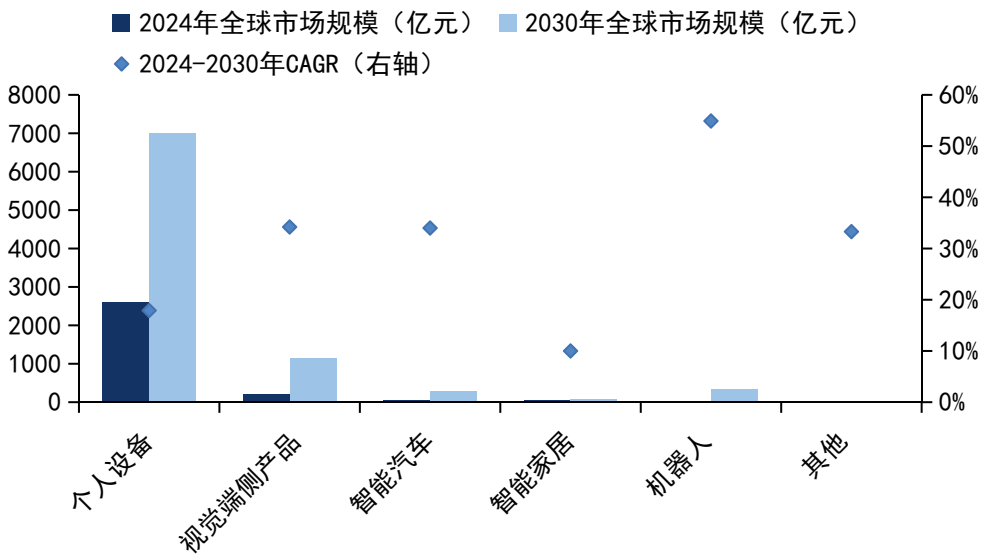
- **轻量化、高性能模型持续推出，推理任务向边缘、端侧扩散。**伴随轻量化、高性能模型持续推出，AI推理从云端向边缘和终端设备扩散，低时延、数据隐私和本地处理需求将加快落地。根据灼识咨询，2024-2030年全球AI推理芯片市场规模预计从6067亿元提升至30696亿元，期间CAGR为31%。细分来看，2030年端侧/边缘/云端推理市场规模分别提升至8861/7262/14573亿元，2024-2030年CAGR分别为20%/42%/36%，边缘和云侧场景芯片规模预计实现更高增速。
- **个人设备/视觉前端为端侧AI两大核心场景，驱动市场持续扩容。**分场景看，个人设备、视觉端侧产品为端侧AI推理的核心场景（2024年推理芯片市场份额占比89%/7%），2030年预计全球市场规模分别为7009/1143亿元，2024-2030年CAGR分别为18%/34%。个人设备增长预计主要来自智能手机、AI PC及XR设备算力升级；视觉端侧产品增长主要得益于工业、城市、家居等场景的数据处理体量和智能化需求提升。

图：2024-2030年全球AI推理芯片市场规模预计从6067亿元提升至30696亿元，期间CAGR为31%



资料来源：爱芯元智公司公告、灼识咨询、国信证券经济研究所整理

图：2030年全球个人设备、视觉端侧产品的AI推理芯片市场规模预计达7009/1143亿元

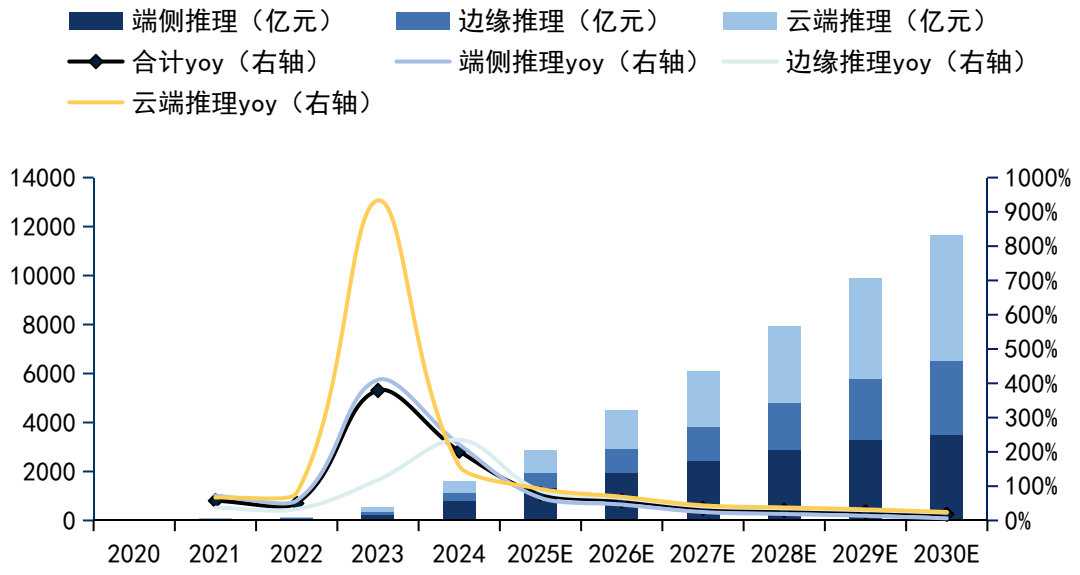


资料来源：爱芯元智公司公告、灼识咨询、国信证券经济研究所整理

中国AI推理芯片规模增速领先，2030年全球份额有望达38%

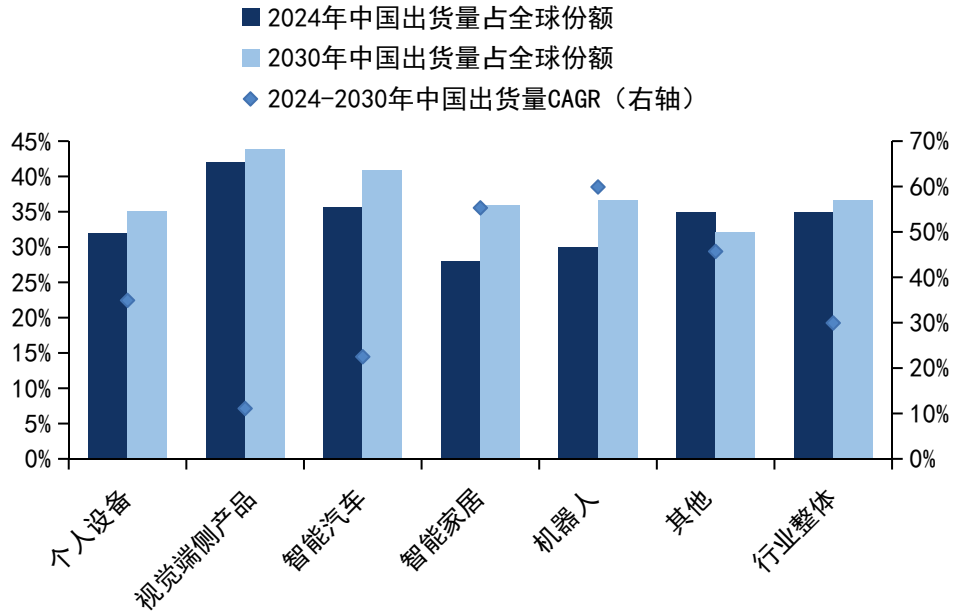
- **中国AI推理芯片市场规模增速领先全球，2030年全球市场份额预计达38%。**根据灼识咨询，2024-2030年中国AI推理芯片市场规模预计从1608亿元提升至11664亿元，期间CAGR为39%，市场规模口径占全球份额从27%提升至38%。细分来看，2030年端侧/边缘/云端推理市场规模分别提升至3499/3033/5132亿元，2024-2030年CAGR分别为28%/45%/48%，均高于全球增速。
- **芯片出货量增价减，视觉端侧产品与智能汽车份额领先。**量价拆分看：1）**出货量维度，**中国AI推理芯片占全球份额有望从2024年的35%提升至2030年的37%。分场景看，个人设备、视觉端侧产品、智能汽车、智能家居、机器人场景份额均有望提升，其中视觉端侧产品与智能汽车场景份额有望超40%，预计部分得益于国内政策扶持以及新能源车出海。2）价格维度，AI推理芯片ASP预计从2024年的493元/颗下降至2030年的445元/颗，期间CAGR为-1.7%。

图：2024-2030年中国AI推理芯片市场规模预计从1608亿元提升至11664亿元，期间CAGR为39%



资料来源：爱芯元智公司公告、灼识咨询、国信证券经济研究所整理

图：中国AI推理芯片占全球份额有望从2024年的35%提升至2030年的37%

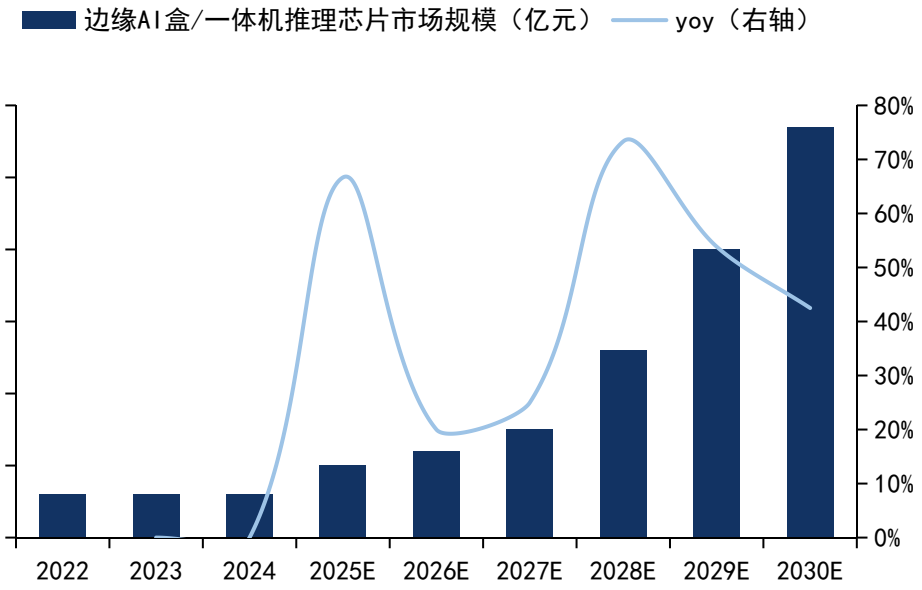


资料来源：爱芯元智公司公告、灼识咨询、国信证券经济研究所整理

边缘AI盒/一体机产品密集推出，有望推动边缘推理芯片放量

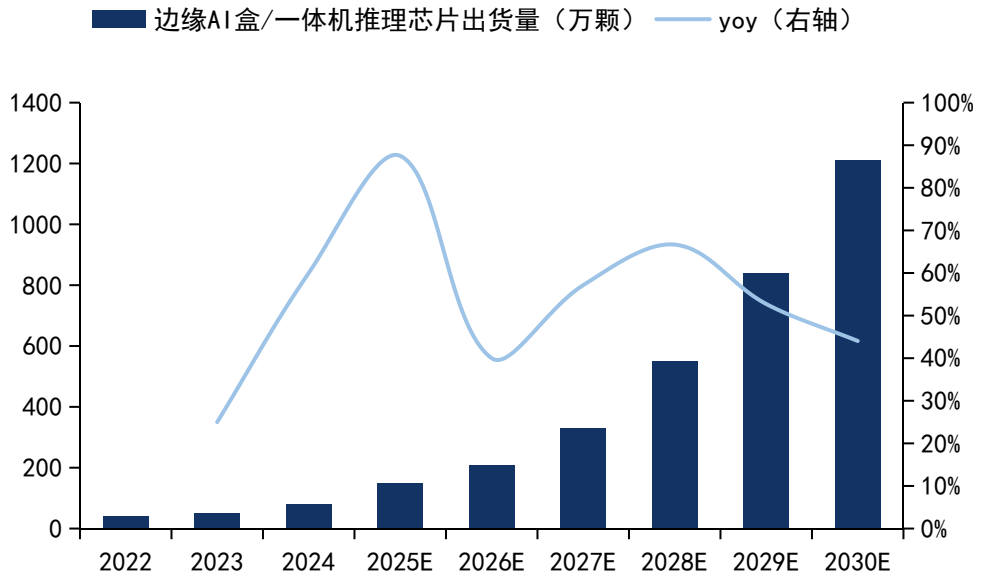
- **边缘AI盒/一体机产品密集推出，有望推动边缘推理芯片放量。**随着30B以下轻量化模型性能跃升，基础个人/办公Agent推理任务加快向AI Box、一体机等边缘设备部署。产业端，2026年英伟达、小米、苹果新品密集加码端侧算力基座、有望加速小参数模型适配及场景普及、推动推理向端侧下沉。根据灼识咨询，2024-2030年中国边缘AI盒/一体机推理芯片出货量有望从80万颗增至1210万颗，期间CAGR达57%。同时产品迭代、规模效应及市场竞争有望引导芯片价格逐步优化、进而加速场景普及，2024-2030年芯片ASP预计从750元/颗降至471元/颗；对应市场规模预计从6亿元增至57亿元，期间CAGR达44%。

图：2024-2030年中国边缘AI盒/一体机推理芯片市场规模预计从6亿元增至57亿元，期间CAGR达44%



资料来源：爱芯元智公司公告、灼识咨询、国信证券经济研究所整理

图：2024-2030年中国边缘AI盒/一体机推理芯片出货量有望从80万颗增至1210万颗，期间CAGR达57%



资料来源：爱芯元智公司公告、灼识咨询、国信证券经济研究所整理

目录

01 端侧模型能力加速跃迁，参数效率与Agent能力持续提升

02 端侧AI算力设备加速落地，个人/办公推理任务向端侧迁移

03 产业链公司加码端侧AI，芯片、终端与场景协同升级

04 端侧与边缘推理需求释放，中国芯片市场规模高增

05 风险提示

风险提示

第一，宏观经济波动风险。若宏观经济出现不利波动，或将影响公司业务发展、产业变革及新技术落地进程。

第二，下游需求不及预期的风险。若下游AI需求不及预期，相关企业的AI研发投入增速或低于预期，以致行业增长不及预期。

第三，核心技术突破不及预期的风险。AIGC相关产业技术壁垒较高，若AI大模型、AI芯片等关键技术研发进展不及预期，或将制约产品性能提升，延缓商业化落地及产业发展进程。

第四，行业竞争加剧。随着AI技术快速迭代与应用普及，若行业竞争加剧，或将导致芯片及AI算力相关业务的利润率下降。

第五，测算误差风险。相关测算基于一定假设，若实际需求、产品价格及技术迭代节奏等与假设不符，或基础数据存在统计口径差异，测算结果或与实际情况存在偏差。

免责声明

国信证券 GUOSEN SECURITIES

国信证券投资评级

投资评级标准	类别	级别	说明
报告中投资建议所涉及的评级（如有）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到12个月内的相对市场表现，也即报告发布日后的6到12个月内公司股价（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。A股市场以沪深300指数（000300.SH）作为基准；新三板市场以三板成指（899001.CSI）为基准；香港市场以恒生指数(HSI.HI)作为基准；美国市场以标普500指数(SPX.GI)或纳斯达克指数（IXIC.GI）为基准。	股票投资评级	优于大市	股价表现优于市场代表性指数10%以上
		中性	股价表现介于市场代表性指数±10%之间
		弱于大市	股价表现弱于市场代表性指数10%以上
		无评级	股价与市场代表性指数相比无明确观点
	行业投资评级	优于大市	行业指数表现优于市场代表性指数10%以上
		中性	行业指数表现介于市场代表性指数±10%之间
		弱于大市	行业指数表现弱于市场代表性指数10%以上

分析师承诺

作者保证报告所采用的数据均来自合规渠道；分析逻辑基于作者的职业理解，通过合理判断并得出结论，力求独立、客观、公正，结论不受任何第三方的授意或影响；作者在过去、现在或未来未就其研究报告所提供的具体建议或所表述的意见直接或间接收取任何报酬，特此声明。

重要声明

本报告由国信证券股份有限公司（已具备中国证监会许可的证券投资咨询业务资格）制作；报告版权归国信证券股份有限公司（以下简称“我公司”）所有。本报告仅供我公司客户使用，本公司不会因接收人收到本报告而视其为客户。未经书面许可，任何机构和个人不得以任何形式使用、复制或传播。任何有关本报告的摘要或节选都不代表本报告正式完整的观点，一切须以我公司向客户发布的本报告完整版本为准。

本报告基于已公开的资料或信息撰写，但我公司不保证该资料及信息的完整性、准确性。本报告所载的信息、资料、建议及推测仅反映我公司于本报告公开发布当日的判断，在不同时期，我公司可能撰写并发布与本报告所载资料、建议及推测不一致的报告。我公司不保证本报告所含信息及资料处于最新状态；我公司可能随时补充、更新和修订有关信息及资料，投资者应当自行关注相关更新和修订内容。我公司或关联机构可能会持有本报告中所提到的公司所发行的证券并进行交易，还可能为这些公司提供或争取提供投资银行、财务顾问或金融产品等相关服务。本公司的资产管理部门、自营部门以及其他投资业务部门可能独立做出与本报告中意见或建议不一致的投资决策。

本报告仅供参考之用，不构成出售或购买证券或其他投资标的要约或邀请。在任何情况下，本报告中的信息和意见均不构成对任何个人的投资建议。任何形式的分享证券投资收益或者分担证券投资损失的书面或口头承诺均为无效。投资者应结合自己的投资目标和财务状况自行判断是否采用本报告所载内容和信息并自行承担风险，我公司及雇员对投资者使用本报告及其内容而造成的一切后果不承担任何法律责任。

证券投资咨询业务的说明

本公司具备中国证监会核准的证券投资咨询业务资格。证券投资咨询，是指从事证券投资咨询业务的机构及其投资咨询人员以下列形式为证券投资人或者客户提供证券投资分析、预测或者建议等直接或者间接有偿咨询服务的活动：接受投资人或者客户委托，提供证券投资咨询服务；举办有关证券投资咨询的讲座、报告会、分析会等；在报刊上发表证券投资咨询的文章、评论、报告，以及通过电台、电视台等公众传播媒体提供证券投资咨询服务；通过电话、传真、电脑网络等电信设备系统，提供证券投资咨询服务；中国证监会认定的其他形式。

发布证券研究报告是证券投资咨询业务的一种基本形式，指证券公司、证券投资咨询机构对证券及证券相关产品的价值、市场走势或者相关影响因素进行分析，形成证券估值、投资评级等投资分析意见，制作证券研究报告，并向客户发布的行为。

国信证券
GUOSEN SECURITIES

国信证券经济研究所

深圳

深圳市福田区福华一路125号国信金融大厦36层

邮编：518046总机：0755-82130833

上海

上海浦东民生路1199弄证大五道口广场1号楼12楼

邮编：200135

北京

北京西城区金融大街兴盛街6号国信证券9层

邮编：100032