

电子行业深度报告

端云协同驱动 AI 入口重塑与硬件范式重构

增持（维持）

2026 年 02 月 27 日

证券分析师 陈海进

执业证书：S0600525020001

chenhj@dwzq.com.cn

投资要点

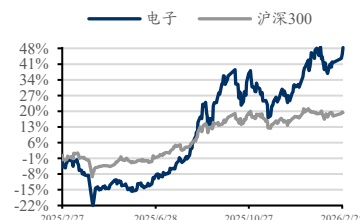
■ **云端模型：能力边界外扩与成本重构并行。**云端大模型作为端侧 AI 能力演进的源头变量，其评价体系正在从单纯能力指标转向能否真正把任务完成。基于这一目标，2026 年以来海外头部厂商正围绕代码能力与多 Agent 体系展开密集布局。**代码模型方面**，智能体时代的推理需求正沿着长链复杂推理与实时交互两大优化方向同步演进，以 **OpenAI 的 Codex-Spark 为代表的低延迟优先型 Agent** 追求交互式 AI 智能体的低延迟体验，让开发者能在模型生成途中随时打断、纠偏并快速迭代；**Claude 4.6 为代表的长链复杂推理型 Agent** 通过提高上下文长度，推动 AI 在高价值复杂任务中的成功率改善，并有望带动推理侧算力消耗中枢持续上移。我们判断未来一段时间内，“快交互+长推理”双能力栈将成为通用型 Agent 的重要演进方向。多智能体框架亦加速走向主流架构选择，有望成为下一阶段 Agent 化落地的重要产业趋势。与此同时，春节期间国内模型厂商同步密集更新，呈现出“性能逼近海外头部、价格快速下探”的特征，同时应用侧需求弹性开始释放，云端模型能力的验证为端侧模型提供可参考模板。

■ **端侧模型：端云协同主线下的效率优化与能力压缩。**端侧模型的终局并非替代云端大模型，而是与云端形成分工明确的协同架构：**高频、轻量、强隐私任务优先在端侧完成本地闭环处理；重推理、长生成和高算力任务经端侧打包与调度后上云执行。**当前端侧模型的演进方向可以归纳为两个核心维度：1）多模态能力为端侧模型关键竞争要点，端侧为多模态零延迟交互方面的理想技术实现路径，当前全双工流式架构逐渐成为主流交互范式；伴随多模态 token 压缩技术环节带宽和算力约束，提高端侧交互的实时性和效率。2）算法侧压缩主要用于对抗功耗和内存等硬件约束，目前主要通过模型架构优化（Edge MoE 和其它替代架构）、低比特量化和推理优化（包括 Attention 效率优化、KV Cache 优化、并行解码和 Diffusion 模型等）等算法手段将推理时计算和存储的开销压缩至最低。

■ **端侧模型牵引硬件重构：算力、存力与散热协同升级。**从整机 AI 功能看，2024 年行业整体仍以高频刚需场景为切入点，重点围绕图像消除、文本摘要等低门槛功能；进入 2025 年，厂商明显加速向多模态创作能力延展，覆盖语音、生成式图像等更复杂交互形态，并进一步向操作系统底层渗透。整机 AI 竞争正从功能数量比拼，转向多模态体验与系统级整合深度的综合较量。**在整机级 AI 能力向多模态等方向升级的背景下，端侧核心部件也正围绕内存与功耗等制约端侧体验的关键变量上进行新一轮升级。**在存储侧，三星 LPDDR6 产品在支持更高数据传输速率和内存带宽的情况下，还从电路架构到电源管理进行了系统性重构，使 LPDDR6 在保持高速性能的同时，实现较上一代约 21% 的能效提升。在散热侧，三星于 2025 年 12 月 19 日发布 Exynos 2600 芯片，首次在移动 SoC 中引入 High-k EMC 材料优化热传输路径，使热阻较 Exynos 2500 降低约 16%。在重载场景（如游戏与端侧 AI 推理）下，持续性能表现显著提升，有效缓解以往因发热导致的降频节流问题。展望未来，高通 Snapdragon 8 Elite Gen 6 等下一代旗舰 SoC 平台或将实现算力、存储与功耗散热同步升级，为端侧 AI 功能进一步复杂化、多模态化及持续运行提供更充足的硬件支撑空间。

■ **风险提示：**模型能力提升不及预期；端侧 AI 商业化落地节奏低于预期；终端硬件升级与需求释放不及预期。

行业走势



相关研究

《AI 基建，光板铜电—GTC 前瞻 Serdes, Rubin Ultra&CPO 交换机详解》

2026-02-25

《2026 年端侧 AI 产业深度：应用迭代驱动终端重构，见证端侧 SoC 芯片的价值重估与位阶提升》

2026-02-23

内容目录

1. 云端模型：能力边界外扩与成本重构并行	4
1.1. 海外：大模型加速迭代，Agent 能力边界持续外扩	4
1.2. 国内：性能快速追赶+性价比优势扩大，带动需求加速释放	6
2. 端侧模型：端云协同主线下的效率优化与能力压缩	10
2.1. 范式收敛：端云协同成端侧模型主流	10
2.2. 多模态：端侧实时交互与执行闭环的关键能力	11
2.3. 模型算法优化：效率优化与能力压缩	11
2.3.1. 模型架构：MoE 在端侧受限于内存瓶颈，EdgeMoE 与新架构并行探索	11
2.3.2. 低比特量化：4-bit 为行业标准配置，2-bit 等更低精度量化技术探索中	12
2.3.3. 推理优化：Attention 效率、KV Cache 管理与并行解码重塑端侧体验上限	13
3. 端侧模型牵引硬件重构：算力、存力与散热协同升级	15
3.1. 整机 AI 功能：从单点功能走向多模态与系统级整合	15
3.2. 端侧算力方案升级：存储、算力与散热协同演进	15
4. 风险提示	17

图表目录

图 1: 2026 年以来海外大模型重要发布事件汇总.....4

图 2: 头部厂商推理模型在低延迟响应与长链推理两大方向上同步演进.....5

图 3: Grok 4.20 四大 Agent 角色分工.....6

图 4: 2026 年以来国内大模型重要发布事件汇总.....7

图 5: 国产大模型对标性能相近的海外模型时，价格优势更加突出.....8

图 6: 智谱上调 Coding Plan 定价约 30%9

图 7: MiniMax Agent 发布后关注度快速升温.....9

图 8: Google Gemma 模型家族拓展垂直专精小模型矩阵10

图 9: 面壁智能 MiniCPM 系列模型发布时间线11

图 10: Liquid AI 端侧模型以小参数实现更高性能表现.....12

图 11: 英伟达 Nemotron-3 模型在 MoE 上的创新突破.....12

图 12: 模型量化方案的性能分析与核心应用场景对比.....13

图 13: Diffusion LLM 原理示意图.....14

图 14: 主要智能手机厂商 AI 功能推出时间表15

图 15: LPDDR6 通过根据使用环境微调工作电压来优化能源效率16

图 16: 三星 Exynos 2600 芯片引入 HPB 技术.....16

1. 云端模型：能力边界外扩与成本重构并行

1.1. 海外：大模型加速迭代，Agent 能力边界持续外扩

云端大模型作为端侧 AI 能力与架构演进的源头变量，2026 年以来正围绕智能体、多模态与成本优化进入新一轮加速迭代期。从产业演进路径看，端侧模型并非孤立发展，其能力边界、架构形态与成本曲线，本质上由云端大模型的技术前沿所锚定。

我们认为，2026 年大模型竞争范式从算力和参数竞赛加速转向以 ROI 为核心的任务能力比拼，代码模型因而成为海外厂商兑现模型生产力与 Agent 落地能力的核心突破口。在这一框架下，一方面，代码作为 Agent 工具调用与系统操作的通用语言，是连接模型智能与数字世界执行力的理想接口，推动模型从对话式助手升级为具备执行闭环能力的操作型 Agent；另一方面，多 Agent 架构亦加速向产品化与 C 端场景渗透，通过自我校验与任务拆解机制，显著强化复杂任务的闭环完成能力。在二者协同演进下，大模型正由对话式助手升级为操作型智能体。

图1：2026 年以来海外大模型重要发布事件汇总

时间	实体	事件
2026/1/9	Midjourney	Niji 7 动漫专项模型上线，强化亚洲/二次元风格生成能力，进一步细分文生图模型产品矩阵
2026/2/4	Mistral	Voxtral Transcribe 2 发布，完善语音转写模型家族，并开放 Voxtral Realtime 权重以推动实时语音生态
2026/2/5	OpenAI	GPT-5.3-Codex 发布，定位 coding agentic 模型，强化代码生成与自主执行能力
2026/2/5	Anthropic	Claude Opus 4.6 发布并提供 1M token 上下文 beta，继续拉升长上下文能力上限
2026/2/12	Google	Gemini 3 Deep Think 迎来重大升级，推理模式能力增强，并与 Gemini App/Ultra 订阅体系联动
2026/2/17	Anthropic	Claude Sonnet 4.6 发布并引入 1M token context beta，推动中高端模型长上下文能力下沉
2026/2/17	xAI	Grok 4.2 (4.20) 进入 public beta 阶段，持续迭代多智能体与推理能力
2026/2/18	Google	Lyria 3 音乐生成模型接入 Gemini App，推动多模态生成能力向消费级入口渗透

数据来源：各公司官网，TechCommunity，Huggingface，东吴证券研究所

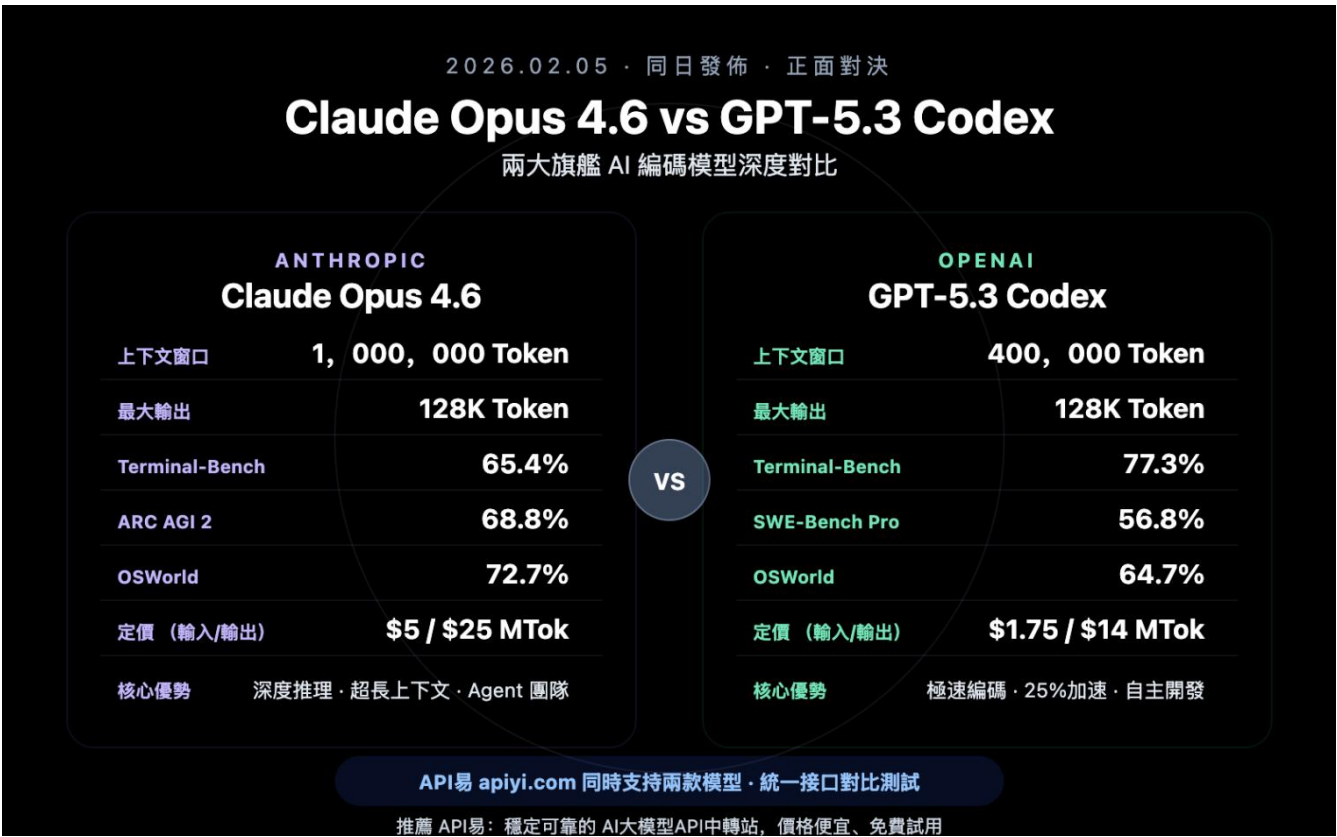
代码模型方面，智能体时代的推理需求正沿着长链复杂推理与实时交互两大优化方向同步演进。

- **低延迟路线（交互型 Agent）。**以 OpenAI 的 Codex-Spark 为代表，追求交互式 AI 智能体的低延迟体验，展现出的“近乎即时”（每秒超 1000 tokens）响应速度，让开发者能在模型生成途中随时打断、纠偏并快速迭代。我们认为这种高度实时的交互形态体现了“一个人即一个开发团队”的产品叙事上，显著强化了用户的掌控感。我们判断该类低延迟路线在需求侧或契合独立开发者、小型工作室及个人高频生产场景，有望形成高黏性的使用闭环。
- **长链复杂推理路线（任务型 Agent）。**Claude 4.6 在长链复杂推理上取得进展：提出了一百万 Token 长上下文的工程设计，使多个 Agent 能够在统一上下文中处理大规模代码库、长周期财务数据及历史交互记录。我们认为这一设计有

助于在金融、法律等对长文本理解与跨文档推理要求较高的 B 端复杂业务场景中显著提升任务成功率。

上述技术路线分化更多体现为场景侧的权重差异而非技术路径的二选一。在实际 Agent 系统中，前端人机交互通常要求低延迟响应，而后台复杂任务执行则依赖长链推理能力，当前头部厂商亦在持续补齐两条能力曲线。我们判断未来一段时间内，这两种能力将共同推动通用模型加速向 Agent 化员工与生产力工具形态对齐。

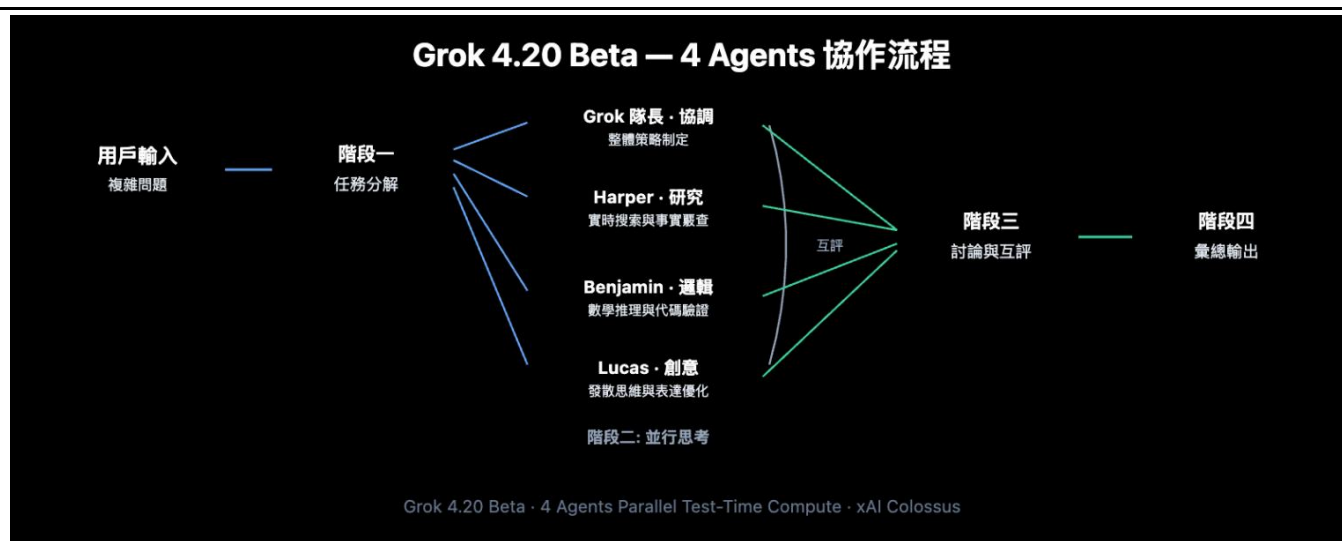
图2：头部厂商推理模型在低延迟响应与长链推理两大方向上同步演进



数据来源：APIYI，东吴证券研究所

多智能体框架加速迈向通用型 Agent 的核心能力底座。多智能体协作并非由 Grok 4.20 首创，行业此前已出现多种探索路径。例如，OpenAI 于 2024 年 10 月开源的 Swarm 多 Agent 编排框架；xAI 亦早在 2025 年 7 月推出的 Grok 4 Heavy 版本中即引入多 Agent 机制。但我们认为 Grok 4.20 以 C 端免费形态大规模推广多智能体能力，具备明显破圈效应。官方披露，其内部由四个具备鲜明认知分工的专家体在同一模型权重与共享上下文下协同运行（并非 4 个独立模型，推理成本仅约 1.5-2.5 倍），通过经强化学习优化的多轮内部辩论机制实现内置自我批判与观点碰撞，使复杂推理准确率显著提升、幻觉率下降约 65%（MMLU-Pro 达 95%）。OpenAI 创始人 Sam Altman 亦指出多代理之间的交互与协作将成为重要演进方向，并有望较快进入 OpenAI 产品体系。我们认为这一表态与头部厂商的产品路径形成相互印证，多 Agent 正加速走向主流架构选择，有望成为下一阶段 Agent 化落地的重要产业趋势。

图3: Grok 4.20 四大 Agent 角色分工



数据来源: APIYI, 东吴证券研究所

模型迭代周期明显进入加速区间。此前的行业认知中,传统的基座模型更新周期通常是6-12个月,但是从目前的模型更新节奏看,模型的迭代周期明显缩短。从具体案例来看: **Google** 在推出 Gemini 3 Pro 后仅约三个月,即进一步发布 Gemini 3.1 Pro,并官方宣称实现推理能力翻倍; **xAI** 创始人 Elon Musk 在介绍 Grok 4.20 时明确提出,该模型能够基于 X (原 Twitter) 平台的实时数据与用户反馈进行高频持续学习,并给出“每周版本更新”的节奏指引; **OpenAI** 研发团队则披露,其已使用 GPT-5.3-Codex 的早期版本参与解决自身训练流程中的工程问题。我们判断,这种“AI 辅助 AI 研发”的闭环一旦成熟,有望系统性压缩模型开发与优化周期。

1.2. 国内: 性能快速追赶+性价比优势扩大, 带动需求加速释放

本轮国产大模型在性能快速追赶的同时性价比优势持续扩大,正从供给端拉低行业推理成本,并开始实质性带动下游需求释放。如果说以 OpenAI、Anthropic 等为代表的海外厂商决定了 Agentic AI 的技术演进方向,那么春节期间阿里通义千问 (Qwen)、字节豆包、智谱 GLM、MiniMax 等国内厂商的密集更新,则凸显出“性能逼近海外头部、价格快速下探”的特征。在成本曲线下移与能力边界外扩的双重驱动下,应用侧需求弹性已开始释放,我们判断模型调用与 AI 应用渗透率有望进入加速上行通道。

图4：2026 年以来国内大模型重要发布事件汇总

时间	实体	事件
2026/1/14	智谱AI	发布GLM-Image图像生成模型，补齐多模态图像生成能力版图
2026/1/16	MiniMax	推出Music-2.5音乐生成模型，持续完善AIGC多模态产品矩阵
2026/1/19	智谱AI	上线GLM-4.7-Flash免费模型，进一步下探推理成本并扩大开发者渗透
2026/1/22	百度	正式发布文心大模型5.0，强化多模态与推理能力并推进商业化落地
2026/1/22	百川智能	发布Baichuan-M3 Plus医疗模型版本升级，并同步推进价格策略优化
2026/1/25	阿里巴巴	推出Qwen3-Max-Thinking旗舰推理模型，强化复杂推理与Agent场景能力
2026/1/26	腾讯	发布HunyuanImage 3.0-Instruct，增强图像编辑与生成一体化能力
2026/1/27	DeepSeek	开源DeepSeek-OCR 2视觉文本解析模型，推进文档理解开源生态
2026/1/27	月之暗面	发布Kimi K2.5模型版本，持续提升长文本与Agent相关能力
2026/2/2	阶跃星辰	推出Step 3.5 Flash开源Agent基座模型，强化Agent开发生态布局
2026/2/3	智谱AI	上线GLM-OCR图文解析模型，完善多模态文档理解能力体系
2026/2/10	阿里巴巴	发布Qwen-Image-2.0图像生成模型，持续推进文生图能力升级
2026/2/11	科大讯飞	推出星火X2大模型，强调国产算力训练体系与自主可控能力
2026/2/12	MiniMax	发布MiniMax M2.5文本模型，主打高性价比推理与商业化适配
2026/2/12	智谱AI	推出GLM-5旗舰模型版本，进一步提升综合推理与多模态能力
2026/2/12	字节跳动	发布Seedance 2.0视频生成模型，加速视频AIGC能力演进
2026/2/13	字节跳动	推出Seedream 5.0 Lite图像模型，强化轻量化图像生成能力
2026/2/14	字节跳动	正式发布豆包大模型2.0系列，全面升级Agent与多模态能力体系
2026/2/16	阿里巴巴	发布Qwen 3.5模型版本，强化Agent化与视觉理解方向能力布局

数据来源：各公司官网，Huggingface，东吴证券研究所

从供给侧看，春节期间国内模型厂商在能力与成本两端同步推进，整体表现为性能差距缩小、性价比提升的趋势。具体来看：

- **MiniMax M2.5** 定价显著低于行业主流水平。在约 100 Tokens/s 吞吐条件下连续运行一小时成本约 1 美元（50 TPS 约 0.3 美元）。Minimax 在模型宣传页中表示，1 万美元预算理论上可支撑约 4 个 Agent 全年 7×24 小时运行，多 Agent 长期部署的经济可行性明显提升。
- **智谱 GLM-5** 发布后，在多项使用体验维度上已逼近 Claude Opus 4.5 所代表的海外第一梯队水平，显示国产通用模型能力差距持续收敛。
- **字节豆包 2.0 系列** 在维持接近前沿模型（GPT-5 级别能力区间）推理表现的同时，大幅下探 Token 定价。例如 豆包 2.0 Lite 输入价格约 0.6 元 / 百万 tokens，相较行业均值呈数量级下降。
- **阿里通义千问 Qwen 3.5** 引入原生 GUI 理解能力，可精确识别屏幕图标、坐标及空间关系，其计算机控制能力已对齐国际顶尖闭源模型水平。同时官方披露，综合成本较前代下降约 60%，大型工作负载处理能力提升约 8 倍。

图5：国产大模型对标性能相近的海外模型时，价格优势更加突出

模型名称	上下文窗口 (Token)	输入价格(\$/MTok)	输出价格(\$/MTok)
国产模型			
GLM-5	200K	1	3.2
MiniMax M2.5	1M	0.3	1.2
海外模型			
GPT-5	400K	1.25	10
Gemini 3 Pro	1M	2.00 (token数≤200K) / 4.00 (token数>200K)	12.00 (token数≤200K) / 18.00 (token数>200K)
Grok 4	256K	3	15
Claude Opus 4.5		5	25
Claude Sonnet 4.5	1M	3.00 (token数≤200K) / 6.00 (token数>200K)	15.00 (token数≤200K) / 22.50 (token数>200K)

数据来源：各公司官网，东吴证券研究所

在供给侧价格快速下行的背景下，应用与开发者侧已出现若干边际积极变化，显示需求弹性正在被逐步激活。

- **MiniMax M2.5** 发布后，多 Agent 部署开始出现真实落地案例。社交媒体上多位独立开发者将其评价为“首个无需显著考虑调用成本的前沿模型”。据 MiniMax 官方数据，M2.5 在 MiniMax Agent 平台上线不足 24 小时，即有全球用户构建超过 1 万个“专家 Agent”。我们认为，成本下探正在推动多 Agent 协同由 PoC 阶段向可规模部署过渡。
- **智谱 GLM-5** 发布后需求表现强劲。公司一方面将 GLM Coding Plan 价格上调超过 30%，另一方面紧急面向全网招募“算力合伙人”，反映供给侧阶段性承压。同时，在正式发布前，海外聚合平台 OpenRouter 上代号“Pony Alpha”的模型一度登顶热度榜，后被确认即 GLM-5，显示其在海外开发者社区已具备一定关注度。
- **字节 Seedance 2.0** 明确面向专业影视、电商与广告生产场景，产品定位直指商业化内容生成。市场反馈显示，其可生成角色一致、多镜头连贯的视频序列，且对后期制作依赖较低。接入该模型的豆包 App 与即梦 在交互流畅度上明显提升，用户无需复杂提示词，仅通过自然语言或单张图片即可完成高质量内容生成。我们认为，该产品显著降低了 AI 视频创作门槛，有望激活短视频与用户二创生态。

图6：智谱上调 Coding Plan 定价约 30%

基于实际使用情况与资源投入变化，我们决定对 GLM Coding Plan 套餐价格体系进行结构性调整。

调整内容如下：

- 取消首购优惠，保留按季按年订阅优惠
- 套餐价格进行结构性调整，整体涨幅自 30% 起
- 已订阅用户价格保持不变

生效时间：2026 年 2 月 12 日

数据来源：Zai，东吴证券研究所

图7：MiniMax Agent 发布后关注度快速升温

MiniMax-AI/Mini-Agent



A minimal yet professional single agent demo project that showcases the core execution pipeline and production-grade features of agents.

13 Contributors 1 Issue 2k Stars 249 Forks

数据来源：Github，东吴证券研究所

整体来看，我们判断春节以来中国大模型市场已阶段性进入由性价比提升所驱动的需求释放新阶段。在模型能力持续逼近海外头部水平的同时，价格体系快速下移，正在实质改善高调用量与多 Agent 场景的商业可行性。若当前趋势延续，Agent 及 AI 原生应用的渗透率有望进入加速上行通道，并进一步向端侧与行业应用外溢。

2. 端侧模型：端云协同主线下的效率优化与能力压缩

2.1. 范式收敛：端云协同成端侧模型主流

我们认为端侧模型的终局并非替代云端大模型，而是与云端形成分工明确的协同架构。结合上一章对云端模型的复盘可以看到，随着多 Agent 协作逐步成为主流，行业竞争重点已经从单一模型能力的不断做大，转向通过系统工程和任务编排来整体提升完成复杂任务的能力。在这一背景下，端侧本身受限于内存和功耗约束，更不现实也没有必要要求单一模型覆盖全部复杂任务，因此端侧能力演进更合理的路径是做好分工、提高整体效率。我们认为端侧模型负责在用户本地隐私边界内，对物理世界、设备状态及系统环境进行实时感知、理解与初步决策；云端则承担重推理、长上下文与高算力密集型任务。

端侧模型进入“自然语言→API 执行”的新范式。以 Gemma 体系为例，其官方定位为“帮助开发者在从工作站、笔记本到手机等任意终端部署 AI 应用”，是面向端侧与边缘场景的模型系列。Gemma 于 2025 年 12 月 18 日发布的 FunctionGemma（约 270M 参数）在产业侧引发较高关注。该模型核心价值在于回应端侧对“既要轻量、又要极致任务完成能力”的诉求，它旨在将自然语言翻译成可执行的 API 操作。FunctionGemma 可以作为完全独立的智能体，处理私密的离线任务，也可以作为大型互联系统的智能分流器。在此模式下，它能够在边缘端即时处理常见指令，同时将更复杂的任务调度至 Gemma 3 27B 等模型进行处理，我们认为有望成为行业的重要参考标杆。

图8：Google Gemma 模型家族拓展垂直专精小模型矩阵

时间	模型	参数	时间	模型	参数
2024/2/21	Gemma	2B/7B	2025/2/19	PaliGemma 2 mix	3B/10B/28B
2024/4/5	Gemma 1.1		2025/3/10	Gemma 3	1B/4B/12B/27B
2024/4/9	CodeGemma		2025/3/10	ShieldGemma 2	
2024/4/9	RecurrentGemma		2025/5/20	MedGemma	4B/27B
2024/5/3	CodeGemma v1.1		2025/6/26	Gemma 3n	E2B/E4B
2024/5/14	PaliGemma		2025/7/9	T5Gemma	
2024/6/11	RecurrentGemma	9B	2025/7/9	MedGemma (multimodal)	27B
2024/6/27	Gemma 2	9B/27B	2025/8/14	Gemma 3	270M
2024/7/31	Gemma 2	2B	2025/9/4	EmbeddingGemma	308M
2024/7/31	ShieldGemma		2025/9/13	VaultGemma	1B
2024/9/12	DataGemma	2B	2025/12/18	FunctionGemma	270M
2024/10/3	Gemma 2 JPN	2B	2025/12/18	T5Gemma v2	270M-270M/1B-1B/4B-4B
2024/10/15	Gemma-APS	2B/7B	2026/1/13	MedGemma 1.5	4B
2024/12/5	PaliGemma 2	3B/10B/28B	2026/1/15	TranslateGemma	4B/12B/27B

数据来源：Google，东吴证券研究所

2.2. 多模态：端侧实时交互与执行闭环的关键能力

多模态能力将成为端侧模型关键竞争要点。在端云协同成为主流架构的背景下，端侧模型承担的核心职责是对物理世界、设备状态及系统环境进行实时感知与初步决策，需要处理大量图像、视频与语音等多模态数据。另外，在手机和 PC 等数字世界自动驾驶场景中，尽管主流观点仍以 API 调用为更高效的终局方案，但大量长尾场景依然不可避免需要读取截图等视觉信息，因此多模态能力是端侧智能体实现能力闭环的重要组成部分。

低延迟多模态正成为端侧竞争的胜负手。云端模型天然受制于网络往返时延，多模态“零延迟”交互正成为端侧的重要差异化优势，这也对端侧模型在多模态交互速度上提出更高要求。从近期模型迭代来看，行业主要围绕以下技术方向展开：

- 全双工流式架构成为主流交互范式。无论是面壁智能还是阿里通义千问，均在主动弱化传统“回合制问答”范式，转向实时多模态交互体系：MiniCPM-o 4.5 强调音视频输入与文本/语音输出的全双工非阻塞处理能力；Qwen2.5-Omni 明确走向端到端多模态+流式输出链路。
- 视觉 Token 压缩成为多模态效率竞争关键。端侧多模态的核心瓶颈在于带宽与算力约束。以面壁智能 MiniCPM 4.5 引入的 3D-Resampler 技术为例，其可将高分辨率视频压缩为极少量视觉 tokens 后再输入主干模型处理。

图9：面壁智能 MiniCPM 系列模型发布时间线

时间	产品	技术特征
2025/1/23	MiniCPM-Embedding	双向注意力改造：Weighted Mean Pooling 向量汇聚
2025/1/23	MiniCPM-Reranker	双向注意力 Cross-Encoder 架构
2025/6/6	MiniCPM4-8B / 0.5B	InfLLM v2 可训练稀疏注意力；LongRoPE 长上下文；FP8+MTP 推理加速；CPM.cu 内核；ArkInfer 端侧推理；Model Tunnel 2.0
2025/6/6	BitCPM4-1B / 0.5B	极限三值（ternary）量化，大幅压缩 bit width
2025/6/6	MiniCPM4-Survey	Plan-Retrieve-Write 多智能体生成框架（Agent 化 RAG）
2025/6/6	MiniCPM4-MCP	MCP 模块化部署与工具调用体系
2025/8/2	MiniCPM-V 4.0	视觉能力强化；端侧效率优化
2025/8/26	MiniCPM-V 4.5	统一 3D-Resampler（高倍视觉 token 压缩）；LLaVA-UHD 文档/OCR；混合 RL 推理优化
2025/9/5	MiniCPM4.1-8B	Trainable sparse attention；Frequency-ranked speculative decoding；EAGLE3 推理加速
2026/2/3	MiniCPM-o 4.5	Full-duplex 实时语音对话；Proactive speaking 主动开口；流式多模态 Omni
2026/2/11	MiniCPM-SALA	稀疏+线性混合注意力（InfLLM v2 + Lightning）；1M 上下文；HyPE 位置编码；冷启动结构迁移

数据来源：面壁智能，东吴证券研究所

2.3. 模型算法优化：效率优化与能力压缩

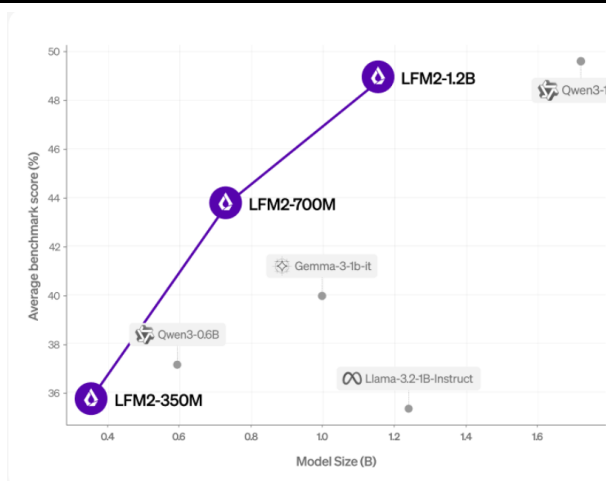
2.3.1. 模型架构：MoE 在端侧受限于内存瓶颈，EdgeMoE 与新架构并行探索

我们认为，MoE 仍是当前主流演进方向，但其在端侧的大规模落地仍受制于存储与带宽瓶颈，行业正处于多路径架构探索阶段。MoE 在云端已成为重要范式。但在端侧场景中，即便采用稀疏激活机制，系统仍需存储全部专家权重。以 Mixtral-8×7B 为例，

实际推理过程中耗时往往不在算力，而在专家权重的内存读写与加载。针对上述问题，业界已通过 EdgeMoE 等工程化手段进行阶段性优化。EdgeMoE 将专家模型分区存储至外部存储，仅在被激活时按需加载，在部分测试中可带来约 1.2–2.7 倍的推理性能提升，并降低约 5–18% 的内存占用。但我们认为，这些方案本质仍属于工程层面的过渡优化，距离在移动设备（功耗 < 10W、内存 < 8GB）上实现 MoE 的原生高效运行仍有明显距离。

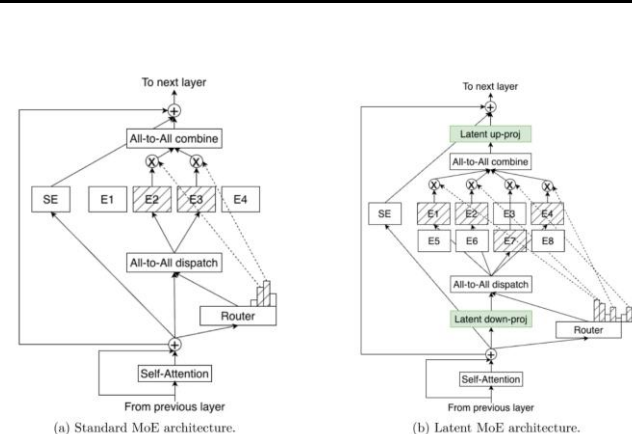
行业也在同步探索 MoE 之外的替代架构。当前 LLM 仍以“注意力机制+前馈网络”为基础，但这种情况正在发生改变，科研文献提出了几种注意力机制的变体：一类方向是对注意力本身进行结构改进以提升效率与表达能力，例如 Qwen 提出的 Gated Delta-Net、DeepSeek 的 Manifold-Constrained Hyper-Connections (mHC) 等；另一类方向则聚焦于在支持长上下文的同时降低延迟，包括 Mamba 与注意力结合的混合架构（如 Qwen3 Next、Nvidia Nemotron3），以及面向低时延场景的替代路径，如 LIV 卷积与线性注意力等。

图10: Liquid AI 端侧模型以小参数实现更高性能表现



数据来源: Liquid AI, 东吴证券研究所

图11: 英伟达 Nemotron-3 模型在 MoE 上的创新突破



数据来源: NVIDIA, 东吴证券研究所

2.3.2. 低比特量化: 4-bit 为行业标准配置, 2-bit 等更低精度量化技术探索中

4-bit 已成为行业默认配置, 但仍存在进一步优化空间。如果模型架构决定了能力上限, 那么量化水平往往决定模型是否真正适配设备侧部署。当前业界的标准路径已基本收敛为“16-bit 训练、4-bit 部署”: GPTQ (2022) 与 AWQ (2023) 验证了 4-bit 后训练量化在实现约 4 倍显存压缩的同时, 仍能较好保持模型质量。但模型推理时部分长尾数值在量化后超出量纲范围, 容易导致精度明显下降甚至崩塌。针对这一痛点, 业界大致形成两类解决思路: 一类是在训练阶段就让模型学会适应低比特, 代表方法包括 ParetoQ (范围学习 QAT); 另一类是在量化阶段重新整理数值分布, 让原模型更容易被压缩, 典型方案包括 SmoothQuant (MIT HAN Lab)、SpinQuant (Meta) 以及 QServe (MIT HAN

Lab, W4A8KV4 体系)。

行业已开始探索 2-bit 等更低比特量化技术。4-bit 已具备良好实用性，继续下探并非简单线性压缩。Microsoft 提出的 BitNet 表明，1.58-bit 量化是可行的，但这一能力无法通过把现有模型直接压缩过去实现，必须从头开始训练模型。ParetoQ 的研究进一步指出，bit 数与精度并非平滑关系：在 3-4 bit 区间，量化更像压缩；而在 2 bit 及以下，模型开始学习不同的表示形式。这意味着，在固定内存空间预算下，将一个更大的模型量化到 2-bit，比将一个参数减半的模型量化到 4-bit 更好。展望未来，若低比特训练能够规模化落地，行业有望打开一条不同于高精度训练的全新效率路径；与此同时，混合精度量化也正逐步成为重要的工程优化方向。

图12：模型量化方案的性能分析与核心应用场景对比

量化位宽	模型压缩比	模型精度表现	核心落地场景
8-bit	缩小至 1/2 (2x smaller)	几乎无损	云端/服务端：适用于无严格内存与算力限制的部署环境。
4-bit	缩小至 1/4 (4x smaller)	精度轻微衰减 1%-3%	云端/移动端/边缘端：常结合量化感知训练 (QAT) 部署，是目前主流的平衡方案。
低于 4-bit	缩小至 1/4 - 1/8	精度衰减约 3%	移动端/边缘端：在极度受限的硬件下为最佳综合折中方案，依赖 QAT 技术。
向量量化	缩小至 1/8 (8x smaller)	精度衰减约 3%	专用硬件加速器：例如苹果神经网络引擎 (Apple Neural Engine) 专属生态优化。

数据来源：《On-Device LLMs: State of the Union, 2026》，东吴证券研究所

2.3.3. 推理优化：Attention 效率、KV Cache 管理与并行解码重塑端侧体验上限

推理正成为决定任务完成效果与智能体验的关键变量。具体到推理优化的工程层面，端侧模型的实际体验差异很大程度上将由 Attention 访存效率、KV Cache 管理以及并行解码共同决定。

在 Attention 效率方面，长序列场景下的主要瓶颈是内存访问而非纯算力。FlashAttention 系列通过 IO-aware 设计与分块机制，显著减少 HBM 与 SRAM 之间的数据搬运，并持续提升 FLOPs 利用率（FlashAttention-2 在 A100 上可达约 72%，FlashAttention-3 在 H100 上约 75%，FlashAttention-4 面向 Blackwell 进一步优化）。对端侧而言，核心原则是尽量减少内存流量、让计算块适配高速缓存，以及在有限并行度下榨干硬件。相应地，local-global attention、grouped query attention 已成为端侧模型常见配置，部分新架构甚至在若干层跳过 attention，以显著降低 KV cache 规模与复杂度。

KV Cache 对内存的占用随 token 序列线性增长，在长上下文场景中甚至可能超过模型权重本身。研究表明，在边缘部署中，KV 压缩的重要性有时高于权重量化，相关工作已验证 KV cache 可压缩至约 3bit 而质量损失有限。MIT HAN Lab 提出“缓存关键

而非缓存全部”的思路：StreamingLLM 发现保留 attention sinks（序列起始 token）即可实现近似无限长度生成；DuoAttention 进一步区分不同注意力头功能（检索型 vs. 流式型）并差异化处理，从而同时降低内存与时延。压缩策略也从简单淘汰演进到结构感知方法，例如 ChunkKV 以语义 chunk 为单位压缩，在保持语言结构的同时带来约 26% 的吞吐提升；EvoKV 通过进化搜索为各层分配最优 cache 预算，在部分任务上以仅约 1.5% 的缓存实现接近甚至优于全 KV 性能。

并行解码的核心思路是用小模型先一次性生成多个 token，再由大模型并行校验，相较于传统的自回归解码大幅提高推理效率。当前两条主流路线包括：Medusa(Princeton, 2024)，通过增加多解码头实现约 2.2–3.6 倍加速且无需改动主模型；以及 EAGLE(SafeAI Lab)，通过隐藏态外推生成草稿 token，在无需微调主模型情况下实现类似加速，EAGLE-3 进一步融合多层语义特征以提升接受率。相关能力已集成进 vLLM 与 TensorRT-LLM 等主流服务框架。ICML 2025 研究还表明，用于打草稿的小模型可跨词表加速任意 LLM，最高可实现约 2.8 倍推理提速。对于端侧而言，该路径尤具吸引力，因为设备侧天然更容易同时运行一个更小的草稿模型。

Diffusion LLM 提供了另一条提升推理效率的路径。以 LLaDA、SBD、TiDAR 为代表的方法将文本生成视为去噪过程，先把所有的词大致生成出来，然后并行地一步步细化修正。研究显示，若与并行解码结合，整体推理速度有望较传统自回归提升约 4–6 倍。尽管该方向仍处于早期阶段，但在对时延敏感的端侧场景中，其潜在价值值得持续关注。

图13: Diffusion LLM 原理示意图

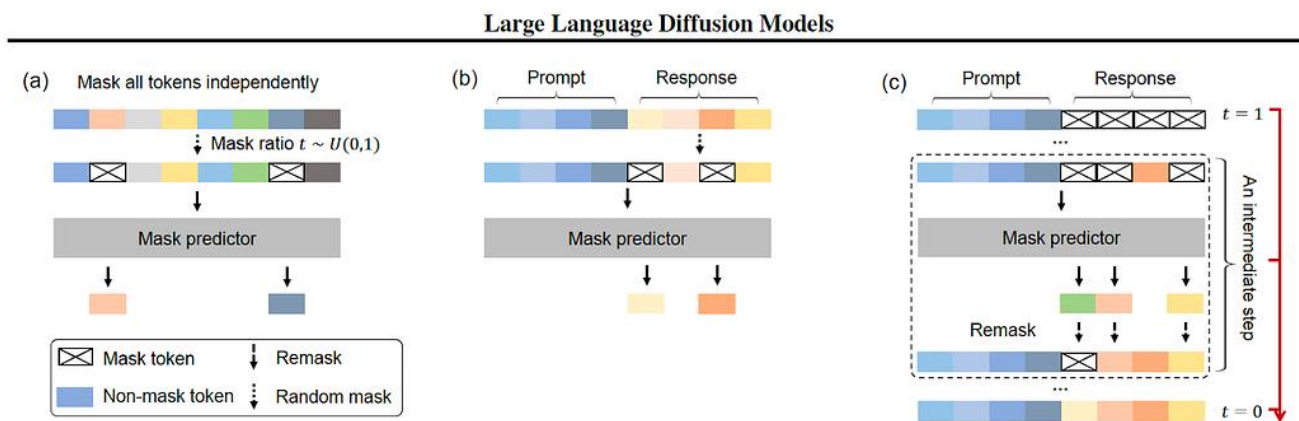


Figure 2. A Conceptual Overview of LLaDA. (a) Pre-training. LLaDA is trained on text with random masks applied independently to all tokens at the same ratio $t \sim U[0, 1]$. (b) SFT. Only response tokens are possibly masked. (c) Sampling. LLaDA simulates a diffusion process from $t = 1$ (fully masked) to $t = 0$ (unmasked), predicting all masks simultaneously at each step with flexible remask strategies.

数据来源：《Large Language Diffusion Models》，东吴证券研究所

3. 端侧模型牵引硬件重构：算力、存力与散热协同升级

3.1. 整机 AI 功能：从单点功能走向多模态与系统级整合

整机 AI 功能正由单点功能迈向多模态能力扩展与系统级深度整合。复盘主要智能手机厂商的 AI 功能推出节奏，我们认为 2024 年行业整体仍以高频刚需场景为切入点，重点围绕图像消除、文本摘要等低门槛功能；进入 2025 年，厂商明显加速向多模态创作能力延展，覆盖语音、生成式图像等更复杂交互形态，并进一步向操作系统底层渗透。在基础功能逐步趋同的背景下，各家厂商也在积极寻找结合自身系统生态或硬件能力的差异化抓手，整机 AI 竞争正从功能数量比拼，转向多模态体验与系统级整合深度的综合较量。

图14：主要智能手机厂商 AI 功能推出时间表

品牌	24H1	24H2	25H1	25H2
Galaxy AI	实时翻译/同传 聊天/笔记助手 圈选即搜 生成式编辑	涂鸦生图 人像工作室	即刻简报 AI 智能选择 音频消除 照片助手 绘画助手	
Apple Intelligence		写作辅助 通知摘要 照片清理/回忆 ChatGPT 集成		实时翻译 生成式表情 (Genmoji) 视觉智能 图像创作空间 (Image Playground) 智能快捷指令
Gemini	圈选即搜 魔术撰写 照片表情 (Photomoji) 防诈骗保护	Gemini Live (实时对话) 合影添加 (Add Me) 通话笔记 通知摘要		魔术提示 (Magic Cue) 语音翻译 相机教练 音乐创作
小米 HyperAI	AI 同传 笔记 照片编辑器 圈选即搜	AI 写作工具 AI 魔法绘画 AI 防诈骗 AI 消除/扩图/胶片	AI 锁屏	
一加 AI	AI 摘要 AIGC 消除 AI 消除	AI 影像套件	AI 翻译 Plus Mind (专属智能体)	AI 语音转写 Gemini + Plus Mind
vivo			AI 消除 实时抠图 转录助手 屏幕翻译 圈选即搜 笔记助手 实况文本	AI 创作 AI 字幕 AI 搜索 文档扫描

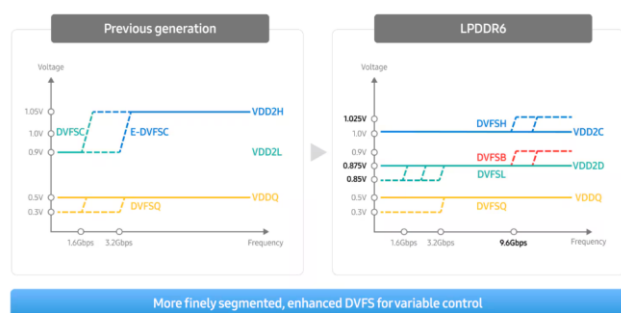
数据来源：Counterpoint，东吴证券研究所

3.2. 端侧算力方案升级：存储、算力与散热协同演进

在整机级 AI 能力向多模态等方向升级的背景下，端侧核心部件也正围绕内存与功耗等制约端侧体验的关键变量上进行新一轮升级。在存储侧，三星推出并荣获 CES 2026 创新奖的 LPDDR6 产品，支持高达 10.7 Gbps 的数据传输速率，并通过扩展 I/O 数量提升整体带宽能力，单颗容量最高支持 16GB。更值得关注的是，LPDDR6 的升级并不仅

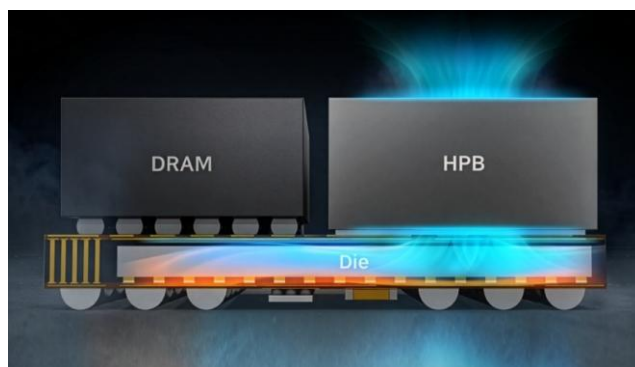
限于制程优化带来的功耗下降，而是从电路架构到电源管理进行了系统性重构。其引入增强型动态电压频率调节（DVFS）与动态效率模式等机制，可根据实时工作负载动态调整功耗与性能；同时，结合三星专有的智能 PMIC，对内外部电源进行精细化管理，使内存在 AI 推理与机器学习等场景下实现更精准的能效匹配。配合全新的低功耗工作范围电路设计，LPDDR6 在保持高速性能的同时，实现较上一代约 21% 的能效提升，更契合 AI 负载波动特征。在散热侧，三星于 2025 年 12 月 19 日发布 Exynos 2600 芯片，首次在移动 SoC 中引入 High-k EMC 材料优化热传输路径，使热阻较 Exynos 2500 降低约 16%。在重载场景（如游戏与端侧 AI 推理）下，持续性能表现显著提升，有效缓解以往因发热导致的降频节流问题。

图15: LPDDR6 通过根据使用环境微调工作电压来优化能源效率



数据来源: Samsung, 东吴证券研究所

图16: 三星 Exynos 2600 芯片引入 HPB 技术



数据来源: Samsung, 东吴证券研究所

新一代旗舰 SoC 或将实现算力、存储与功耗散热同步升级。高通下一代旗舰平台 Snapdragon 8 Elite Gen 6 有望推出标准版与 Pro 版双版本形态。据 Notebookcheck 报道，Pro 版本将配备更强 GPU，并有望率先支持 LPDDR6 内存；其中一款芯片的频率或将达到 5GHz-5.5GHz。更高性能也意味着更高功耗与更严苛的热设计约束，据 Fixed Focus Digital 披露，高通计划在今年晚些时候发布的旗舰芯片中引入三星 HPB（高性能散热方案），覆盖 Snapdragon 8 Elite Gen 6 及其 Pro 版本。

4. 风险提示

模型能力提升不及预期：若大模型在复杂推理、多模态理解或 Agent 执行成功率等关键指标上的进展放缓，可能削弱 AI 功能对用户的实际价值感知，从而影响应用渗透与调用量增长。

端侧 AI 商业化落地节奏低于预期：端侧 AI 仍处于从技术验证走向规模化应用的过渡阶段，若开发者生态、杀手级应用或用户付费意愿培育不及预期，可能导致商业闭环形成速度放缓。同时，隐私、安全与系统权限等因素亦可能在部分区域或场景中制约端侧 AI 的推广进度。

终端硬件升级与需求释放不及预期：端侧 AI 体验高度依赖存储带宽、算力与散热能力的协同升级，若 LPDDR6、先进 SoC 或新型散热方案的导入节奏放缓，可能阶段性限制端侧模型能力释放。同时，在宏观消费需求偏弱或换机周期拉长的情况下，终端侧 AI 功能对出货与 ASP 的拉动效果亦可能低于市场预期。

免责声明

东吴证券股份有限公司经中国证券监督管理委员会批准，已具备证券投资咨询业务资格。

本研究报告仅供东吴证券股份有限公司（以下简称“本公司”）的客户使用。本公司不会因接收人收到本报告而视其为客户。在任何情况下，本报告中的信息或所表述的意见并不构成对任何人的投资建议，本公司及作者不对任何人因使用本报告中的内容所导致的任何后果负任何责任。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。

在法律许可的情况下，东吴证券及其所属关联机构可能会持有报告中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投资银行服务或其他服务。

市场有风险，投资需谨慎。本报告是基于本公司分析师认为可靠且已公开的信息，本公司力求但不保证这些信息的准确性和完整性，也不保证文中观点或陈述不会发生任何变更，在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的报告。

本报告的版权归本公司所有，未经书面许可，任何机构和个人不得以任何形式翻版、复制和发布。经授权刊载、转发本报告或者摘要的，应当注明出处为东吴证券研究所，并注明本报告发布人和发布日期，提示使用本报告的风险，且不得对本报告进行有悖原意的引用、删节和修改。未经授权或未按要求刊载、转发本报告的，应当承担相应的法律责任。本公司将保留向其追究法律责任的权利。

东吴证券投资评级标准

投资评级基于分析师对报告发布日后 6 至 12 个月内行业或公司回报潜力相对基准表现的预期（A 股市场基准为沪深 300 指数，香港市场基准为恒生指数，美国市场基准为标普 500 指数，新三板基准指数为三板成指（针对协议转让标的）或三板做市指数（针对做市转让标的），北交所基准指数为北证 50 指数），具体如下：

公司投资评级：

买入：预期未来 6 个月个股涨跌幅相对基准在 15%以上；

增持：预期未来 6 个月个股涨跌幅相对基准介于 5%与 15%之间；

中性：预期未来 6 个月个股涨跌幅相对基准介于-5%与 5%之间；

减持：预期未来 6 个月个股涨跌幅相对基准介于-15%与-5%之间；

卖出：预期未来 6 个月个股涨跌幅相对基准在-15%以下。

行业投资评级：

增持：预期未来 6 个月内，行业指数相对强于基准 5%以上；

中性：预期未来 6 个月内，行业指数相对基准-5%与 5%；

减持：预期未来 6 个月内，行业指数相对弱于基准 5%以上。

我们在此提醒您，不同证券研究机构采用不同的评级术语及评级标准。我们采用的是相对评级体系，表示投资的相对比重建议。投资者买入或者卖出证券的决定应当充分考虑自身特定状况，如具体投资目的、财务状况以及特定需求等，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。

东吴证券研究所
苏州工业园区星阳街 5 号
邮政编码：215021

传真：（0512）62938527

公司网址：<http://www.dwzq.com.cn>