

消费电子行业点评报告

端侧 AI 周跟踪: Agent 模型加速落地, 端云协同提升端侧 AI 价值

增持 (维持)

2026 年 08 月 16 日

证券分析师 陈海进

执业证书: S0600525020001

chenhj@dwzq.com.cn

研究助理 郭雄飞

执业证书: S0600126060019

guoxf@dwzq.com.cn

投资要点

■ **Meta 发布 Muse Glimmer, 30B 级 Agent 模型可实现消费级单卡本地部署。**8 月 10 日, Meta 发布约 300 亿参数的开源权重多模态模型 Muse Glimmer, 定位于可长期运行在个人设备上的本地 Agent, 官方表示模型可在单张消费级 GPU 上运行。Glimmer 采用 Dense 结构, 支持文本、图片输入及文本输出, 覆盖 100 余种语言。机器学习研究者 Sebastian Raschka 指出, Glimmer BF16 模式下 KV Cache 仅约 52KiB/token, 低于 Qwen3.6 27B 的 64KiB/token, 并远低于 Gemma 4 31B 约 840KiB/token; 其认为 Glimmer 最突出的特点在于较低内存占用、较快 Prefill/Decode 速度与较完整 Agent 能力的组合。

■ **NVIDIA 发布 Nemotron 3.5 Lightning, 进一步明确“强模型规划、小模型执行”的 Agent 分层架构。**8 月 11 日, NVIDIA 推出 Nemotron 3.5 Lightning 30B-A3B, 底层采用 Mamba-2、MoE 与 Attention 混合架构, 支持最长 1M 上下文, 可运行于 GeForce RTX 5090、DGX Spark、Jetson 等本地设备。NVIDIA 将该模型定位为 Agent 系统中的高频执行层: 由前沿模型负责复杂任务的规划和推理, 再由 Lightning 负责工具调用等高频执行任务, 并配合 NeMo Switchyard 根据任务复杂度动态决定调用大模型或小模型, 从而避免 Agent 每一步均调用高成本 Frontier Model。Artificial Analysis 评测显示, 其智能水平并非同尺寸模型最高, 但输出速度中位数达到约 301 token/s, 显著高于同尺寸开放权重模型约 106 token/s 的中位水平。

■ **Google 发布 Pixel 11 系列及 Pixel Watch 5, 端侧 AI 承担任务进一步增加, AI 应用由被动响应向主动协助与持续感知演进。**8 月 12 日, Google 在 Made by Google 2026 发布 Pixel 11 系列四款手机、Pixel Watch 5 及首款 Pixel Tag。一方面, Google 进一步提高端云协同架构中由终端本地直接承担的 AI 计算比例。新版实时翻译功能采用由 Tensor G6 驱动的端侧生成式 AI 模型, 在本地处理实时语音到语音翻译并对视频、音频进行实时配音; Pixel Watch 5 搭载 Snapdragon W5 Gen 2 Accelerated 及低功耗双芯片架构, 其 Raise to Talk 功能开始直接由端侧 AI 处理, 包括启动锻炼、设定计时器等任务。另一方面, AI 应用形态继续由“用户主动调用助手”向主动 Agent、多模态交互和持续感知延伸。Pixel 11 上的 Gemini Intelligence 已可处理覆盖 40 余个 App 的多步骤任务, Pixel 11 Pro Fold 则利用折叠屏双屏形态引入 Sign-to-Text, 通过前置摄像头识别美国手语并在外屏实时输出文字; Pixel Watch 5 通过 Gemini Intelligence Proactive Suggestions 结合日历等上下文主动在腕上给出下一步操作建议, 另一方面还通过多模态传感器与端侧 AI 持续监测严重、持续的血氧下降, 并基于 Health Foundation Models 提供血压、睡眠呼吸质量及胰岛素抵抗等长期趋势分析。

■ **推理效率诉求提升, 端侧 AI 重要性凸显。**我们认为, 随着 AI 基础设施投入规模持续扩大, CSP 对 ROI 与推理效率的重视度明显提升; 与此同时, Agent 应用使模型调用频次快速增加, 端云协同的重要性进一步凸显。我们看到, NVIDIA 已探索由云端 Frontier Model 承担复杂规划、端侧高效模型承担高频、低延迟执行任务, Google 新品中本地 AI 承担的翻译、语音及感知任务亦持续增加。随着模型效率和端侧算力提升, 未来更多高频任务有望向本地迁移, 端侧 AI 在整体算力体系中的重要性有望持续提升, 建议关注端侧 AI 产业链机会。

■ **风险提示:** 技术创新不及预期风险, 终端需求不足风险, 宏观环境风险。

行业走势



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>