东吴证券 SOOCHOW SECURITIES

电子行业点评报告

端侧 AI 周跟踪：Google 发布 Gemma 4，模型能力跃迁催化终端硬件升级周期

2026 年 04 月 06 日

证券分析师 陈海进
执业证书：S0600525020001
chenhj@dwzq.com.cn

增持（维持）

投资要点

- **Google 发布 Gemma4 开源模型，Agent 与多模态能力全面增强：**4 月 3 日，Google 发布新一代开源语言模型 Gemma 4，包括 E2B、E4B、26B（MoE）及 31B（Dense）四个版本。Gemma4 全系模型支持以下能力：
  - **Agent 和复杂推理：**支持多步骤推理与复杂逻辑规划，具备面向 Agent 场景的自主工作流执行能力，可调用多种工具与 API。
  - **多模态：**所有模型原生支持图像与视频处理，并在 OCR 与图表理解等任务中表现突出，其中 E2B/E4B 版本额外支持原生音频输入。
  - **离线代码生成：**支持本地环境下代码生成。
  - **长上下文：**小模型支持 128K 上下文窗口，大模型最高支持 256K 上下文，显著提升长文档与复杂任务处理能力。
  - **多语言能力：**已在超过 140 种语言上进行原生训练。
- **技术迭代聚焦内存效率与多模态能力下沉，提升端侧任务承载能力并扩大设备覆盖范围：**从技术演进路径来看，Gemma 4 的迭代围绕内存、交互能力等端侧部署的核心瓶颈进行优化。具体来看，**1）在模型架构层面**延续 Per-Layer Embeddings（PLE）机制，以 E2B 为例，总参数约 5B，但实际推理仅需加载约 2B 核心权重，其余通过 CPU 按需调用，这一变化降低终端硬件的使用门槛，使模型可在当前存量中端设备上运行，扩展了端侧 AI 的可触达设备基数。**2）在长上下文能力上，**通过“交替式滑动窗口+全局注意力”以及 Shared KV Cache 设计，大幅优化内存使用效率：多数层仅处理局部 token，少数层负责全局建模，同时复用缓存避免重复计算，使 KV 缓存需求较传统全注意力机制下降 74%。在端侧内存受限的背景下，这一优化直接决定模型是否具备处理长文档、多轮对话等真实工作负载的能力，是端侧 AI 走向生产力工具的关键。**3）在能力边界上，**Gemma 4 将视觉+音频的原生多模态能力首次下沉至 2B 级模型，为手机端实现理解屏幕、语音交流、跨应用操作等常用功能提供技术基础。**整体来看，我们认为 Gemma 4 通过架构创新一方面显著提升端侧模型对日常多模态任务的处理能力，另一方面有效降低硬件门槛、扩大可触达设备范围，对端侧 AI 产业节奏具有加速意义。**
- **开源协议全面放开叠加 Android 体系落地，驱动端侧硬件升级与新一轮换机周期开启：**从生态角度看，Gemma 系列前几代版本使用 Google 自定义许可证，对商用场景存在一定限制。本次 Gemma 4 切换至 Apache 2.0 协议，在无强制使用政策约束的前提下提供完全商业自由，显著降低企业采用门槛，有望吸引更多开发者与商业客户回流。另一方面，Gemma 4 将作为 Gemini Nano 4 的基础模型，并计划于年内落地新一代旗舰 Android 设备，承担下一代端侧模型基座角色。据官方披露，自首代发布以来 Gemma 累计下载量已超过 4 亿次，拥有超过 10 万个衍生模型，初步形成 Gemmaverse 开发者生态。我们认为在开源协议放宽与 Android 体系导入的双重驱动下，Gemma 4 所代表的端侧模型能力升级有望显著拓展端侧 AI 能力边界，并进一步催化终端硬件性能升级与新形态产品创新，带动新一轮换机周期与品类突破。
- **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

相关研究

《国产算力周跟踪：坚定看好国产超节点产业趋势》

2026-03-29

《AI ASIC：从台系 ASIC 厂商发展历程看国产产业链机遇》

2026-03-25

东吴证券 SOOCHOW SECURITIES

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。

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