2026 年 05 月 24 日

通信行业研究

买入（维持评级）

行业周报

证券研究报告

通信组
分析师：张真桢（执业 S1130524060002）
zhangzhenzhen@gjzq.com.cn

英伟达上调 AI 行业开支，阿里发布 Qwen3.7-Max

通信周观点：

1）英伟达 1Q27 业绩超预期，营收 816 亿美元，同比增长 85%，其中数据中心收入 752 亿美元；公司指引 2Q27 营收 910 亿美元，将 2030 年 AI 行业年度开支预测上调至 3 万亿至 4 万亿美元，预计今年 CPU 收入接近 200 亿美元，Vera CPU 与 Vera Rubin 平台有望打开服务器 CPU 新空间，拉动先进封装及高阶 ABF 载板需求。2）Google I/O 2026 聚焦 Agentic AI，发布 Gemini 3.5、Gemini Omni、第八代 TPU 双芯片架构及 Antigravity 2.0。谷歌月处理 Token 达 3.2 千万亿个，同比增长 7 倍，Gemini 月活达 9 亿。3）Anthropic 预计 2Q26 收入达 109 亿美元并首次实现单季盈利，与 SpaceX 扩大算力合作，SpaecX 预计未来三年支付近 450 亿美元获取算力资源。4）OpenAI 或提交 IPO 申请，最快 9 月上市。5）台积电 CPO 布局升级，COUPE on Substrate 预计 26 年下半年量产，AI 芯片制造正从“先进制程+先进封装”迈向“光电共封装+系统整合”。6）三大运营商均已推出算力 Token 套餐，中国电信最低月租 9.9 元，面向个人、家庭及政企客户提供 Token 服务，通信服务与 AI 算力融合进入规模化商用阶段。7）阿里发布 Qwen3.7-Max，三方盲测中位列国产模型第一。百度 1Q26 智能云收入 88 亿元，同比增长 79%，GPU 云收入同比增长 184%，国产模型与 AI 云持续高景气。8）字节跳动视频生成模型 Seedance 2.1 或将近期发布，生成效果较 2.0 版本提升约 20%，同时低配版价格有望降至约 0.5 元/秒。Seedance 当前按日消耗占比已超八成，视频生成模型和推理算力需求有望继续放量。9）Nebius 受益于 GPU 租赁价格上涨，H100 按需租赁价格由 2.95 美元/小时上调至 3.85 美元/小时，涨幅近 29%，抢占式 GPU 容量价格涨幅达 51%。我们认为老旧 GPU 租赁价格不降反升，说明 AI 算力仍紧缺，利好新型 AI 云及算力租赁厂商。10）智谱发布 GLM-5.1 高速版 API，输出速度达 400 tokens/s，刷新全球大模型厂商 API 速度上限；同时智谱与 MiniMax 将被纳入恒生科技指数，未来有望进入港股通机制。国产大模型在性能迭代、资本市场关注度和流动性层面均持续提升。

细分赛道：

服务器：本周服务器指数+0.86%，本月以来，服务器指数+11.47%。英伟达 1Q27 业绩超预期，数据中心收入 752 亿美元，并将 2030 年 AI 行业年度开支预测上调至 3 万亿至 4 万亿美元。Vera CPU 与 Vera Rubin 平台有望打开服务器 CPU 新空间，拉动先进制程、先进封装及高阶 ABF 载板需求。

光模块：本周光模块指数+0.30%，本月以来光模块指数+15.16%。台积电 CPO 布局升级，COUPE on Substrate 预计 2026 年下半年量产。利好 CPO、光模块及上游光器件环节。

IDC：本周 IDC 指数-2.61%，本月以来，IDC 指数+7.99%。智谱发布 GLM-5.1 高速版 API，刷新全球大模型厂商 API 速度上限;同时智谱与 MiniMax 将被纳入恒生科技指数，看好国产算力链加速向上。

核心数据更新：

电信业务量收增速逐步提升。2026 年 1-3 月电信业务收入累计完成 4394 亿元，同比下降 1.8%。按照上年不变价计算的电信业务总量同比增长 8.3%。2026 年 3 月我国光模块出口金额当月同比-16.6%；累计同比+2.8%。

投资建议与估值

建议关注国内 AI 发展带动的服务器、IDC 等板块，以及海外 AI 发展带动的服务器、光模块等板块。

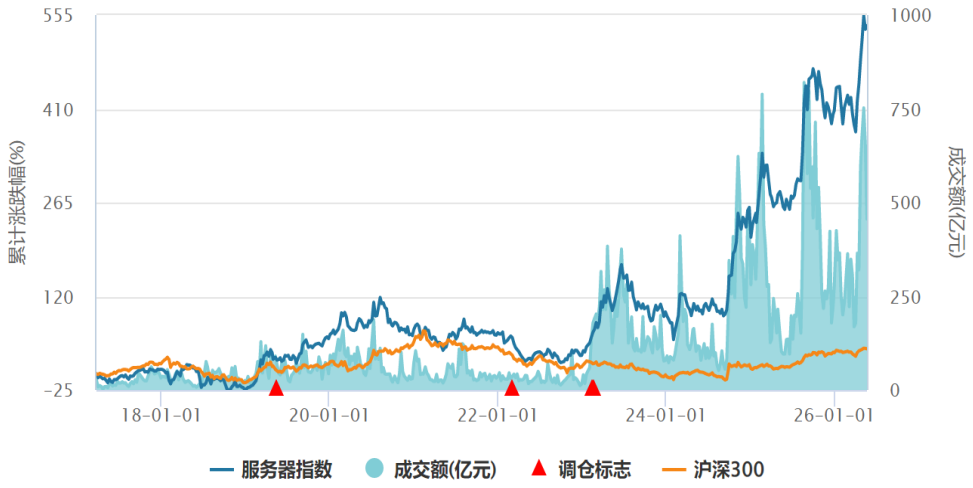
风险提示

AI 建设不及预期、中美关税波动加剧、原材料供应不足。

一、细分行业观点

服务器：本周服务器指数+0.86%，本月以来，服务器指数+11.47%。英伟达 1Q27 业绩超预期，数据中心收入 752 亿美元，并将 2030 年 AI 行业年度开支预测上调至 3 万亿至 4 万亿美元。Vera CPU 与 Vera Rubin 平台有望打开服务器 CPU 新空间，拉动先进制程、先进封装及高阶 ABF 载板需求。

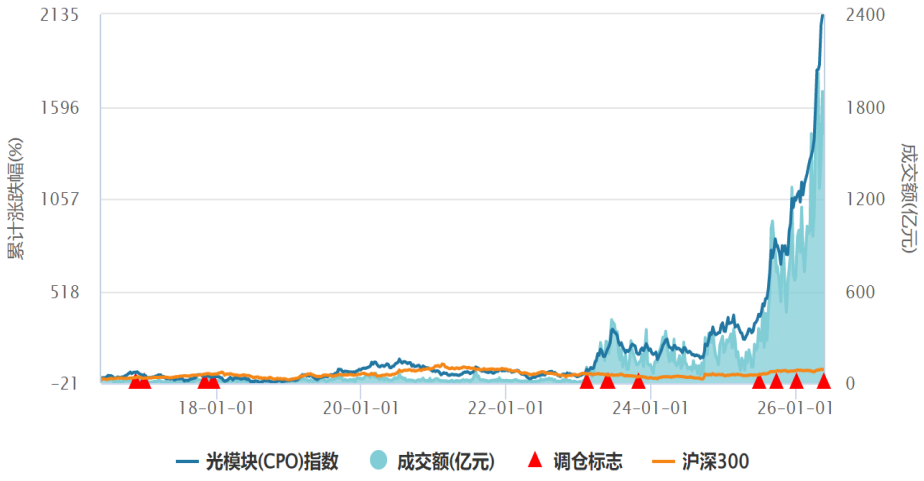
图表1：服务器指数（8841058.WI）走势



来源：wind，国金证券研究所

光模块：本周光模块指数+0.30%，本月以来光模块指数+15.16%。台积电 CPO 布局升级，COUPE on Substrate 预计 2026 年下半年量产。利好 CPO、光模块及上游光器件环节。

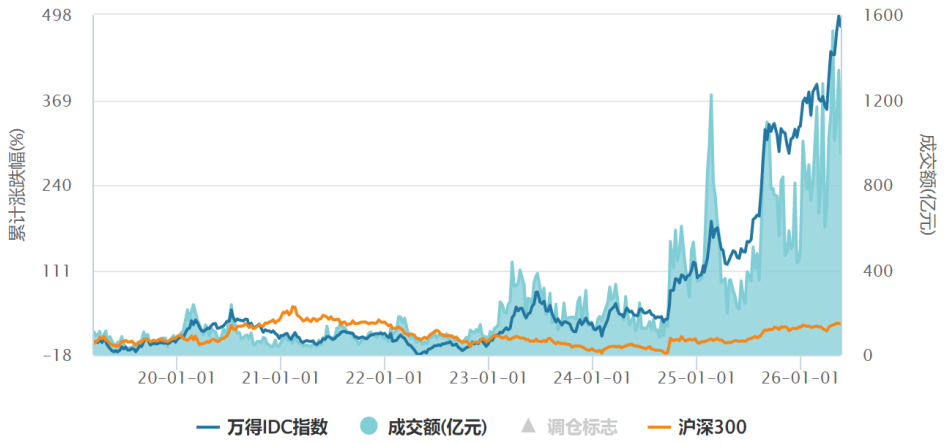
图表2：光模块(CPO)指数（8841258.WI）走势



来源：wind，国金证券研究所

IDC：本周 IDC 指数-2.61%，本月以来，IDC 指数+7.99%。智谱发布 GLM-5.1 高速版 API，刷新全球大模型厂商 API 速度上限；同时智谱与 MiniMax 将被纳入恒生科技指数，看好国产算力链加速向上。

图表3：IDC 指数（866052.WI）走势



来源：wind，国金证券研究所

图表4：本周通信板块景气度

板块	景气度指标	本期景气度说明
运营商	稳健向上	5G 投资周期结束，电信行业端承压，但云与 IDC 业务放量接力成长，整体景气度稳健向上。
光模块	稳健向上	台积电 CPO 布局升级，COUPE on Substrate 预计 26 年下半年量产。
服务器	稳健向上	英伟达 1Q27 业绩超预期，将 2030 年 AI 行业年度开支预测上调至 3 万亿至 4 万亿美元，AI 需求持续加速。
交换机	稳健向上	思科已从超大规模客户获得 53 亿美元订单，并将全年订单预期上调至 90 亿美元。看好交换机作为 AI 重要硬件放量。
连接器	稳健向上	康宁计划将美国光连接产品制造产能提升 10 倍并将美国光纤产能扩大 50%以上，同时新建三座先进制造工厂。推动光纤、连接器等需求增长。
IDC	加速向上	智谱发布 GLM-5.1 高速版 API，刷新全球大模型厂商 API 速度上限；同时智谱与 MiniMax 将被纳入恒生科技指数，看好国产算力链加速向上。
物联网	加速向上	谷歌正式开源 Gemma 4 系列，覆盖从端侧到服务器的多种部署场景，AI 端侧加速向上。
液冷	高景气维持	谷歌正与中国英维克等企业洽谈采购数据中心液冷系统，液冷板块再受催化。

来源：工商时报，新浪财经，NVDIA Newsroom，华尔街见闻，IT 之家，央视新闻，国金证券研究所

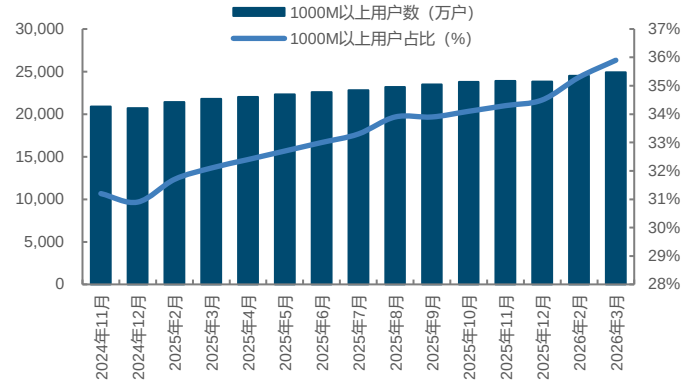
二、核心数据更新

运营商：运营商数据维持稳健增长

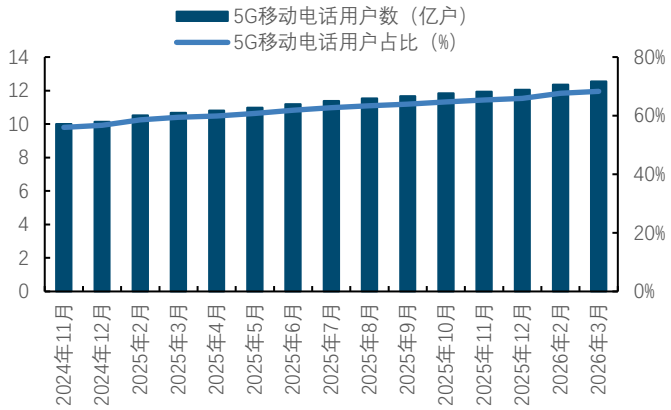
截至 3 月末，三家基础电信企业及中国广电的移动电话用户总数达 18.36 亿户，比上年末净增 920.1 万户。其中，5G 移动电话用户达 12.54 亿户，比上年末净增 4953 万户，占移动电话用户的 68.3%。

图表5：千兆用户占比超三成

图表6：截至 3 月末 5G 用户占比超六成



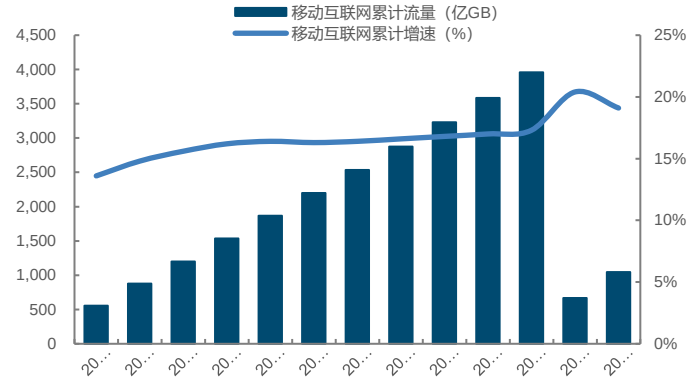
来源：工信部，国金证券研究所



来源：工信部，国金证券研究所

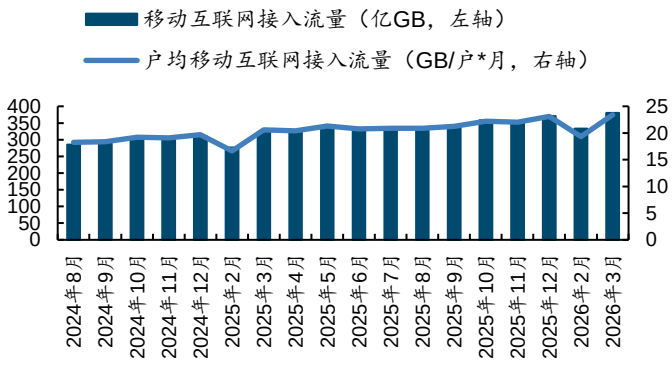
一季度，移动互联网累计流量达 1044 亿 GB，同比增长 19.1%。截至 3 月末，移动互联网用户数达 16.23 亿户，比上年末净增 1348 万户。3 月当月户均移动互联网接入流量(DOU)达到 23.4GB/户·月，同比增长 13.8%，比上年底高 0.36GB/户·月。

图表7：一季度移动互联网累计流量同比增长 19.1%



来源：工信部，国金证券研究所

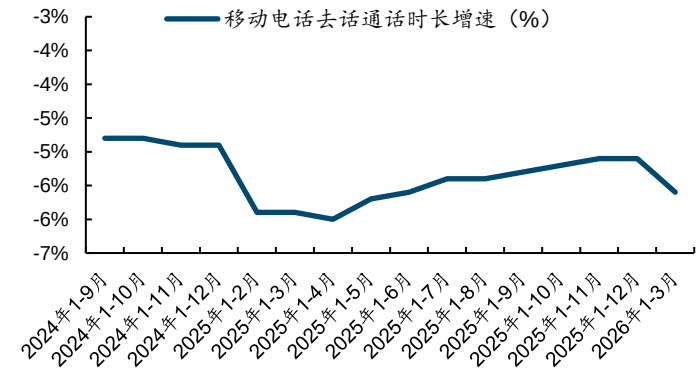
图表8：3 月当月 DOU 达 23.4GB/户·月



来源：工信部，国金证券研究所

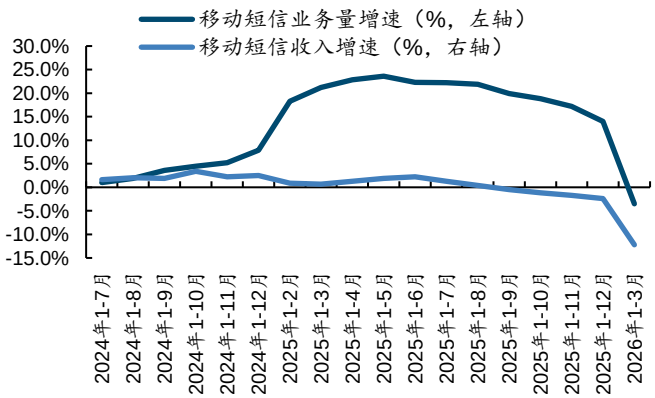
一季度，移动电话去话通话时长完成 4683 亿分钟，同比下降 5.6%；固定电话主叫通话时长完成 132.6 亿分钟，同比下降 21.2%。一季度，全国移动短信业务量同比下降 3.5%；移动短信业务收入同比下降 12.2%。

图表9：电话通话量持续下滑



来源：工信部，国金证券研究所

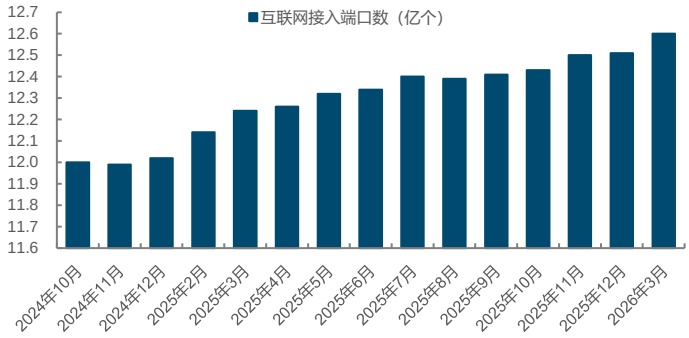
图表10：移动短信业务量大幅下降



来源：工信部，国金证券研究所

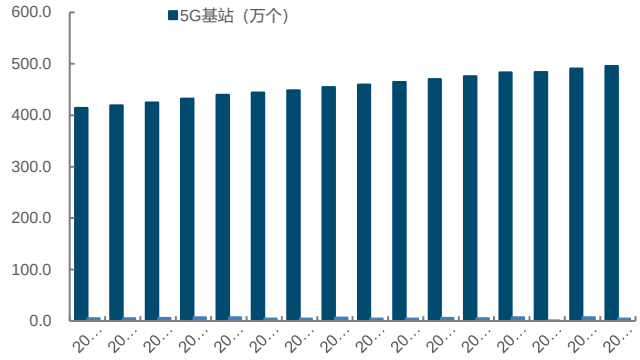
截至 3 月末，全国互联网宽带接入端口数量达 12.6 亿个，比上年末净增 1328 万个。其中，光纤接入（FTTH/O）端口达到 12.2 亿个，比上年末净增 1217 万个，占互联网宽带接入端口的 96.7%。截至 3 月末，具备千兆网络服务能力的 10G PON 端口数达 3201 万个，比上年末净增 38.7 万个。

图表11：千兆光纤宽带网络建设稳步推进



来源：工信部，国金证券研究所

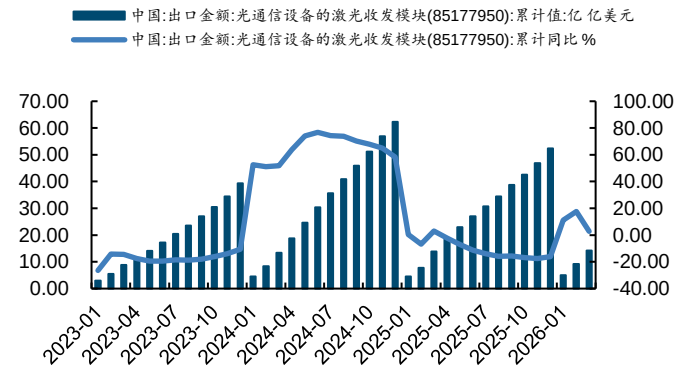
图表12：5G网络建设持续深化



来源：工信部，国金证券研究所

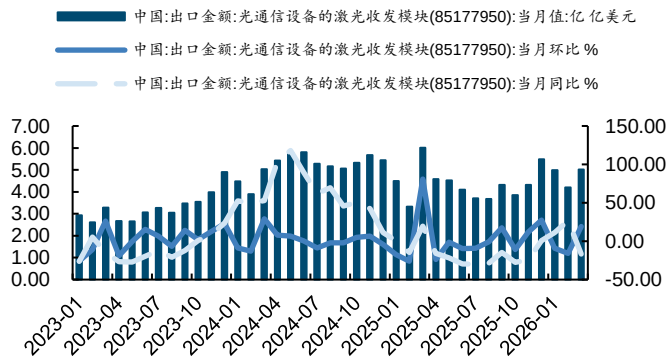
光模块数据：2026年3月我国光模块出口数据同比-16.55%；1-3月累计同比+2.83%。

图表13：3月光模块出口金额累计同比增加2.8%



来源：wind，国金证券研究所

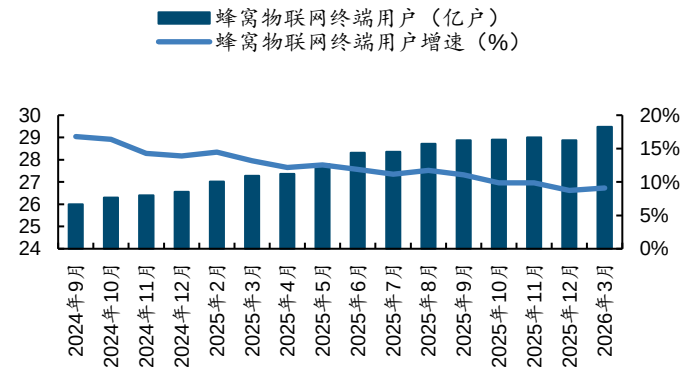
图表14：3月光模块出口金额当月同比-16.55%



来源：wind，国金证券研究所

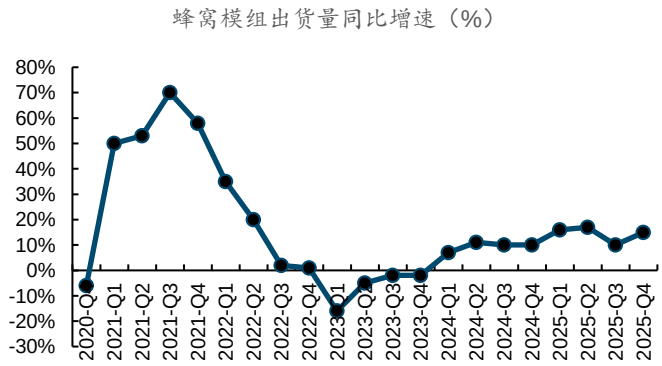
物联网数据：截至3月末，三家基础电信企业发展移动物联网终端用户29.48亿户，比上年末净增5954万户。互联网电视（IPTV、OTT）用户数达4.1亿户，比上年末净增188.7万户。

图表15：截至3月末蜂窝物联网终端用户数同比增长8.11%



来源：工信部，国金证券研究所

图表16：2025年全球蜂窝物联网模组出货量同比增长15%

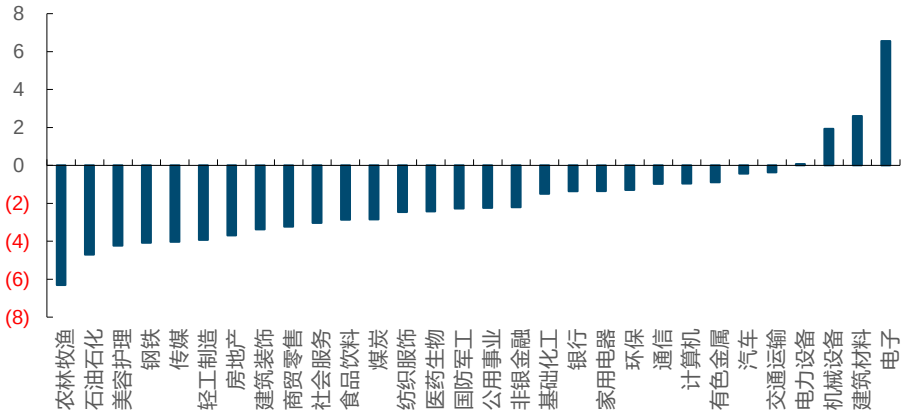


来源：国金证券数字未来实验室，RFID世界网，C114通信网，国金证券研究所

三、本周行情

回顾本周行情（5 月 18 日-5 月 22 日），参考申万一级行业划分，通信板块涨跌幅为-0.97%，排名全行业第 9。

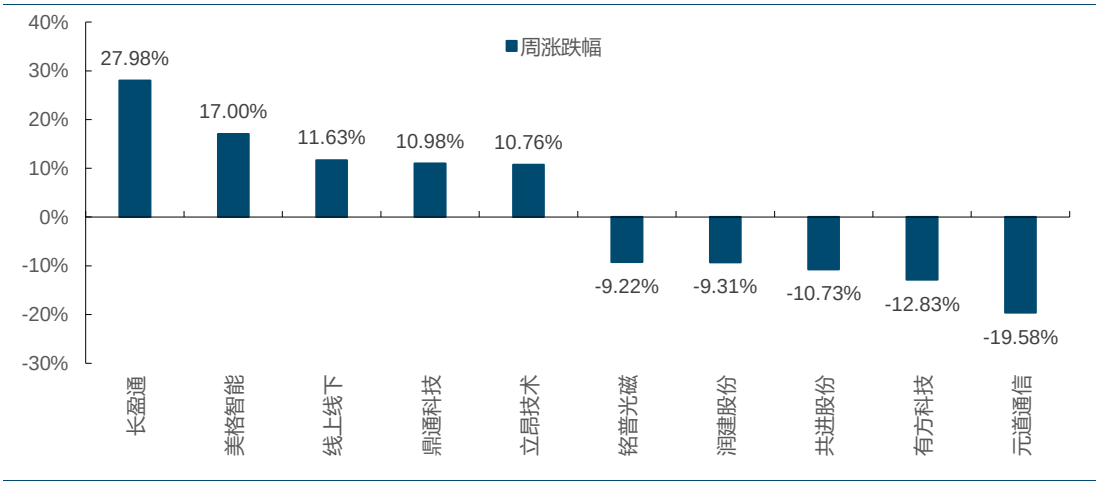
图表17：板块周涨跌幅排序（%）



来源：wind，国金证券研究所

从个股情况来看，本周长盈通、美格智能、线上线下、鼎通科技、立昂技术为通信（申万）涨幅前五大公司，涨跌幅分别为+27.98%、+17.00%、+11.63%、+10.98%、+10.76%。通信（申万）跌幅前五为铭普光磁、润健股份、共进股份、有方科技、元道通信，涨跌幅分别为-9.22%、-9.31%、-10.73%、-12.83%、-19.58%。

图表18：通信板块个股周涨跌幅（剔除*ST公司）



来源：wind，国金证券研究所

四、本周重要新闻

4.1 行业新闻

5 月 21 日 Epoch AI Substack 这篇：Frontier labs don’t use most AI compute (yet) - by Josh You 指出顶尖前沿实验室并没有充分利用大量的 AI 计算资源，但几年之内，顶尖实验室可能会占据全球计算资源更大的份额。届时，顶尖实验室的计算增长将与整体计算能力的增长速度更加直接相关，这可能会减缓我们目前看到的模型能力和人工智能部署/收入的快速增长。为了保持规模化发展，整体计算能力的建设速度需要加快。鉴于人工智能资本支出（capex）已接近每年 1 万亿美元，计算能力的加速增长将需要巨大的经济变革。大多数人工智能计算可能不会用于前沿人工智能。

文章附录总结了对 OpenAI、Anthropic、xAI 以及 Google 和 Meta 的前沿实验室在 2025 年底所能获取的 AI 计算资源量的估算。这五家公司可能是全球计算资源最丰富的开发商，尽管它们未必是模型质量的领导者。文中关注的是前沿 AI 公司租用或使用的计算资源量，而不是它们拥有的计算资源量。Anthropic 和 OpenAI 主要从亚马逊、谷歌和微软等云合作伙伴处租用计算资源；谷歌和 Meta 则大多拥有自己的 AI 计算资源，但其中很多被分配给了非前沿的内部用途，或者像谷歌那样，出租给外部客户。

Nebius：

人工智能云基础设施提供商 Nebius Group (NBIS)的股票近来备受华尔街关注，而这背后的主要原因正是芯片巨头英

伟达(NVDA)。在英伟达最近一次的财报电话会议上，首席财务官科莱特·克雷斯透露，其传统 Hopper（H100）GPU 的租赁价格今年迄今已上涨 20%，而更早的 A100 云价格也上涨了近 15%。科技行业出现了一种罕见的趋势：老旧芯片的价格不降反升，而这一趋势对 Nebius 来说却是一个重大利好。作为一家快速发展的新型云运营商，Nebius 的核心业务是购买大量的英伟达 GPU 集群，然后按小时出租给人工智能开发者、初创公司和企业，帮助他们竞相构建人工智能应用。英伟达已经确认 GPU 租赁需求持续超过供应，Nebius 迅速采取行动，利用其新获得的定价权获利。该公司宣布将按需 H100 GPU 租赁价格从每小时 2.95 美元上调至 3.85 美元，涨幅近 29%；而其抢占式 GPU 容量的价格涨幅更为惊人，高达 51%。投资者迅速做出反应，推动 Nebius 股价在 5 月 21 日飙升近 14.7%。但英伟达的定价权仅仅是华尔街对 Nebius 日益高涨热情的部分原因。

三大运营商：

继中国移动、中国联通之后，5 月 17 日世界电信和信息社会日当天，中国电信正式推出系列试商用 Token 套餐。至此，国内三大基础电信运营商均已公布面向个人、家庭及政企市场的算力 Token 套餐价格，标志着通信服务与 AI 算力的融合进入规模化商用阶段。中国电信：最低月租 9.9 元，推出三档 Token Plan 中国电信此次发布的 Token 套餐体系主要覆盖三大人群。面向开发者及中小微企业客户，提供“Token+连接+安全”一体化服务，推出三档 Token Plan 以及宽带上行提速包和安全防护包两种可选包服务。据悉，Token Plan 包括基础版、专业版、旗舰版三种套餐包，每月资费为 39.9 元、159.9 元、299.9 元不等。

4.2 公司新闻

台积电：

5 月 18 日工商时报讯，AI 资料中心高速传输需求爆发，台积电共同封装光学（CPO）布局再升级。台积电近期揭露，结合紧凑型通用光子引擎（COUPE）的「COUPE on Substrate」方案，预计 2026 年下半年进入量产。业界解读，这不仅代表光通讯技术升级，更意味 AI 供应链竞争，正由先进制程、先进封装，进一步延伸至 ABF 载板与光电共封装整合战场。随英伟达新一代 Vera Rubin 平台持续提高 AI GPU、HBM 与高速网路互连整合度，市场预期，高阶载板重要性快速攀升。半导体业者分析，若 CPO 成为未来 AI 伺服器主流架构，英伟达不排除透过长约、预付款甚至策略合作等方式，提前锁定高阶载板产能，避免重演过去 CoWoS 与 HBM 供应吃紧情况。近期英伟达也宣布与康宁（Corning）深化合作，扩大 AI 资料中心光学元件供应布局，确保未来高速光互连所需产能无虞。供应链指出，当光学元件越靠近运算晶片，资料传输距离便越短，可同步降低讯号损耗、延迟与功耗，对 AI 伺服器由单机竞赛，走向整柜、整丛集效能竞赛，具有关键意义。供应链透露，相较传统 CPU 载板，AI GPU 与 ASIC 所需的载板面积与层数大增，ABF 材料消耗量提升 5 至 10 倍；随 AI GPU、ASIC 及高阶网通晶片需求持续攀升，未来高阶 ABF 载板供需结构恐将长期维持紧张。台积电此次将 COUPE 延伸至基板层级，显示 AI 晶片制造由「先进制程＋先进封装」，迈向「光电共封装＋系统整合」新阶段。

英伟达：

5 月 19 日工商时报讯，英伟达执行长黄仁勋即将来台，AI 供应链再度进入备战状态。供应链消息传出，黄仁勋预计于 5 月 27 日抵台，28 日举办「兆元宴」，广邀台积电、鸿海、台达电、联发科等台湾供应链大咖餐叙。然法人更关注的是，辉达是否将透过更积极方式锁定台积电先进制程产能，为 2028 年后新世代 AI GPU 量产预作准备。今年黄仁勋访台时间点备受市场关注，COMPUTEX 2026 将于 6 月 2 日至 5 日登场，而黄仁勋将于 6 月 1 日进行演讲，主轴聚焦 AI 下一阶段发展，涵盖运算、实体 AI 与代理式 AI 系统，阐述从能源、网路、晶片、系统到应用的 AI 基础建设版图。去年黄仁勋 11 月来台，首站造访台南，传出有意出资锁定台积电 Fab 18 厂区旁规划的 P10、P11 预留用地。供应链最新消息指出，南科特定区开发区块 A，将规划作为 Fab 22 P7 厂区，预计第二季开工，未来锁定 2 奈米及更先进制程，同时也保留两个厂的扩充空间。将为英伟达 Feynman 世代晶片生产做准备，此一变化令市场重新解读台积电南科先进制程布局。除既有 Fab 18 承担 5 奈米、3 奈米量产重任外，Fab 22 扩建与新厂区也会坐落于台南，搭配南科三期之 AP 8 先进封装厂，成为 AI 晶片世代交替下的生产重镇。

5 月 21 日新浪财经报道，美东时间周三盘后，英伟达公布了其第一财季业绩，不仅营收和利润均超出分析师预期，并且对第二财季的展望也优于预期。在财报发布后，英伟达盘后股价一度快速拉升，但涨势并未继续太久。截至发稿，英伟达股价盘后下跌 1.6%。这种情况其实并不少见。回顾过去，英伟达过去 5 次财报公布后的次个交易日，有 4 次出现下跌。财报显示，英伟达 2027 财年 Q1 营收 816 亿美元，同比增长 85%；市场预期为 786.72 亿美元，上年同期为 440.62 亿美元。Q1 净利润 583 亿美元，同比增长 211%；市场预期为 422.44 亿美元，上年同期为 187.75 亿美元。Q1 数据中心收入为 752 亿美元，市场预期 728 亿美元，上年同期为 391.12 亿美元。公司预计 2027 财年 Q2 营收为 910 亿美元；市场预期 867.88 亿美元。公司 2027 财年第一季度，通过股票回购和股息向股东返还了 200 亿美元；宣布新增 800 亿美元的股票回购授权，并将季度股息提高至每股 0.25 美元。

5 月 21 日华尔街见闻讯，英伟达管理层在电话会上表示，当前需求正在加速，并将 2030 年末年度 AI 行业整体开支规模预测，上调至 3 万亿至 4 万亿美元。“智能体 AI”驱动下一波算力基础设施建设狂潮，并首次将 CPU 业务作为未来核心增长引擎推向台前。预计今年 CPU 收入有望达到 200 亿美元。公司管理层在电话会上表示，“智能体 AI”驱动下一波算力基础设施建设狂潮，并首次将 CPU 业务作为未来核心增长引擎推向台前。英伟达预计今年 CPU 收入有望达到 200 亿美元，并表示相信公司能够获得足够供应以支撑持续增长。公司 CFO Colette Kress 表示，公司当前需求正在加速，并将 2030 年末年度 AI 行业整体开支规模预测，上调至 3 万亿至 4 万亿美元。英伟达 CEO 黄仁勋则预计 Vera Rubin 在整个生命周期内都将面临供应受限局面。在利润极其充沛的背景下，英伟达宣布新增高达 800 亿美元的股票

回购授权。同时，季度股息从每股 0.01 美元大幅上调至 0.25 美元。公司计划在今年将 50%的自由现金流返还给股东。

5 月 21 日 inside 发文，NVIDIA 因应「AI 代理」崛起，推出专为其工作负载最佳化的 Vera CPU。目标开拓两千亿美元潜在市场，正式进军并挑战既有的伺服器 CPU 产业，将成长故事从 GPU 延伸至全新战场。财务长 Colette Kress 在法说会上表示，NVIDIA 预计今年 CPU 总营收将接近 200 亿美元，目标是成为全球领先的 CPU 供应商。黄仁勋随后进一步澄清，这 200 亿美元仅指独立贩售的 Vera CPU 营收，不含与 Rubin GPU 搭售的 Vera Rubin 整合平台。换句话说，NVIDIA 在 CPU 领域的实际营收规模将远高于此数字。

美东时间 5 月 22 日，WCCfetch 发文，随着 2026 年台北国际电脑展（Computex 2026）的临近，NVIDIA 将在展会上阐述人工智能计算行业的最新趋势。继英特尔发布第一季度财报后，业界转向使用 CPU 进行智能体 AI 计算的趋势引发了人们的猜测，即 GPU 在 AI 集群中的作用可能会下降。NVIDIA 将在 Computex 上展示其 Vera CPU 的优势，并声称其计算性能比 AMD 和英特尔的 x86 架构竞争对手高出 1.5 倍。关注的焦点是基于 Arm 架构的 CPU 在人工智能计算领域新一轮 CPU 竞争中的作用。在即将于台湾举行的台北国际电脑展（Computex）上，英伟达将重点展示其 Vera Rubin CPU，而英特尔则会重点展示入门级 Wildcat CPU。这两家半导体公司都将借此应对各自面临的竞争挑战。

字节跳动：

5 月 19 日消息，据 AI 普瑞斯独家获悉，字节跳动视频生成模型 Seedance 2.1 将于近期发布，预计生成效果较 2.0 版本提升 20%。此外，字节跳动还将推出 Seedance2.0 低配版，比目前的 fast 版效果更好，但价格更低，预计是 5 毛一秒，可能要对市场上其他厂商的视频模型造成巨大冲击。据行业人士向 AI 普瑞斯透露，按日消耗占比计算，Seedance 已占据市场超八成份额，可灵约占 14%，万相 2.7 约占 4%，HappyHorse 占比不足 1%。今年 2 月，Seedance 2.0 上线后即引爆市场，漫剧、短剧公司蜂拥而至，到 3 月底字节漫剧日消耗突破 7000 万元，首次超过真人短剧。而可灵作为快手旗下旗舰产品，2026 年 1 月年化收入运行率已超 3 亿美元，月活用户突破 1200 万，在视频生成赛道商业化进度仅次于 Seedance。此前，据金融时报发布的中国 AI 实验室最佳视频模型排行榜，在文本转视频、图像转视频、视频编辑三大领域内，字节 Seedance2.0 均排名第一，阿里 HappyHorse1.0 排名第二。

Google：

北京时间 5 月 20 日凌晨 1 点，谷歌 I/O 2026 开发者大会在美国加州山景城开幕。与往年相比，今年这场发布会的重点，不是某一个模型或功能，而是一次系统性转向——谷歌正把 AI 智能体全面“塞进”所有核心入口。从搜索框到 Chrome 浏览器，从 Android 手机到智能眼镜，Gemini 不再只是一个对话助手，而是一个可以持续运行、跨应用执行任务的 AI 代理：它能替用户追踪信息、生成内容、调用工具，甚至直接完成下单和操作流程。在开场演讲中，谷歌 CEO 桑达尔·皮查伊（Sundar Pichai）表示，过去 12 个月是谷歌“飞速发展”的一年。谷歌每月处理的 Token 数量达到 3.2 千万亿个，同比增长 7 倍；Gemini 应用月活跃用户数达 9 亿。

5 月 20 日经济日报综合外电，Alphabet 旗下 Google 和资产管理巨擘黑石集团 18 日宣布，将合资成立一家人工智慧（AI）云端公司，致力于推动 Google 自研晶片的商业化，为各大企业提供辉达以外的算力选择。这家新公司将由黑石出资 50 亿美元作为股本；Google 则提供自研的张量处理器（TPU）等晶片，以及相关软体与服务，Google 资深主管史洛斯（Benjamin Treynor Sloss）将出任执行长。新公司目标是在 2027 年，启用 500MW 的算力基础设施容量，之后也将大幅扩充容量。知情人士表示，两间公司已锁定可能纳入这项合资计画的资料中心，其中部分仍在兴建中。

21 经济网消息，北京时间 5 月 20 日凌晨，谷歌 CEO Sundar Pichai 在 Google I/O 2026 的舞台上算了一笔账，谷歌头部大客户每天可以处理约 1 万亿个 token，如果把其中 80%的工作负载从其他前沿模型切换到 Google 新发布的 Gemini 3.5 Flash 上，一年能省超过 10 亿美元。2026 年谷歌 I/O 大会的核心主线仍然是 agent，从 agent 平台（Antigravity）到消费者 agent（Spark）到搜索 agent，Google 要把 agent 做成全栈能力。在这场长达两小时的 Keynote 中，Google 发布了新一代 Gemini 3.5 系列模型、全模态世界模型 Gemini Omni、第八代 TPU 双芯片架构，以及从编码工具升级为 agent 管理平台的 Antigravity 2.0。同时，我们也看到了一条新的主线变得更加清晰，Agentic AI 时代已经走向中场，前沿模型的核心战场，也在从拼“最强最聪明”，转向了把 agent 的运行成本压到企业敢大规模部署的门槛以下。

Anthropic：

Anthropic PBC 有望首次实现单季盈利，对其人工智能软件的需求推动收入激增。一位知情人士援引这家初创公司向投资者披露的数据称，这家 Claude 开发商预计第二季度收入将达到 109 亿美元，较前一季度增长逾一倍。知情人士称，Anthropic 预计截至 6 月的季度营业利润将达到 5.59 亿美元。该人士表示，随着公司加大对计算资源和其他成本的投入，Anthropic 并不预期未来几个季度一定会盈利。Anthropic 曾被视为竞争对手 OpenAI 的弱势挑战者，但得益于 AI 智能体工具的进展，该公司收入一直在快速增长，并在企业客户市场不断扩大份额。Anthropic 正洽谈以超过 9000 亿美元的估值进行新一轮融资，这将超过 OpenAI 在私募市场的最新估值。此前报道称，Anthropic 正在考虑最快于 10 月进行首次公开募股。OpenAI 也计划最快于今年秋季上市。

5 月 21 日新浪财经讯，据两名知情业内高管透露，为应对旗下人工智能产品日益增长的市场需求、扩充算力资源，Anthropic 公司目前正洽谈租用搭载微软自研 AI 服务器芯片的算力服务器。若成功拿下 Anthropic 这一客户，对微软而言将是重大利好。此前微软自研芯片项目在去年遭遇研发延期。英伟达 AI 芯片几乎占据了微软面向 AI 开发者出租的全部算力设施，如今微软意图效仿谷歌、亚马逊等云服务竞品，打造可替代英伟达芯片的自研芯片产品。各大云服务商高管均表示，英伟达硬件采购成本居高不下，研发定制化自研芯片，是稳住并提升企业利润空间的必要举措。

微软迈亚芯片能为 Anthropic 运行克劳德大模型提供全新算力选择，Anthropic 还有望推动该芯片新一代产品按自身业务需求定制优化。迈亚芯片主打提速运行现有大模型，性能优于英伟达同类产品，但并不像英伟达芯片那样，适配客户全新大模型的研发搭建工作。

财联社 5 月 21 日电，据一份证券披露文件，Anthropic PBC 已同意在未来三年内向埃隆·马斯克的 SpaceX 支付近 450 亿美元，用于获取计算资源，这是为支持其 Claude 人工智能 AI 软件而扩大协议的一部分。SpaceX 周三在其首次公开招股(IPO)相关文件中披露，Anthropic 预计将每月向马斯克的 SpaceX 支付 12.5 亿美元，直至 2029 年 5 月。文件还指出，任何一方均可提前 90 天通知对方终止协议。Anthropic 本月早些时候表示，签署了一份协议，以从 SpaceX 位于孟菲斯的大型数据中心 Colossus 1 获取超过 300 兆瓦的计算能力，但并未透露具体条款。据 Anthropic 联合创始人暨首席计算官 Tom Brown 的帖子，之后双方扩大合作关系，纳入 SpaceX 第二个数据中心的算力。

OpenAI：

5 月 21 日新浪财经讯，一位知情人士周三透露，OpenAI 正准备在未来几周内秘密提交美国首次公开募股（IPO）申请。这家 ChatGPT 开发商上次估值达 8520 亿美元，其此番计划出炉之际，恰在其成功抵御埃隆·马斯克发起的关乎公司存亡的法律挑战两天之后，且可能抢走马斯克旗下 SpaceX 公司预计于当天晚些时候提交的 IPO 申请的风头。该消息人士称，OpenAI 计划最早于 9 月上市，目前正与高盛(996.73，8.56，0.87%)和摩根士丹利(201.03，0.52，0.26%)合作起草 IPO 招股说明书，并计划很快向监管机构提交。OpenAI 今年早些时候已筹集 122 亿美元，这很可能是硅谷有史以来规模最大的融资轮。但在近几个月，面对谷歌(379.38，-4.09，-1.07%)和 Anthropic 等竞争对手的挑战，该公司已两次调整产品路线图，部分行业观察人士预计，未来数月内 Anthropic 的营收增长将超越 OpenAI。路透去年曾独家报道称，OpenAI 正考虑最早于 2026 年下半年向证券监管机构提交申请。

5 月 22 日 Epoch AI 报告称，OpenAI 在 2023 年开启了 AI 算力建设浪潮。但如今，OpenAI 使用的算力约占全球总量的 10%，而头部实验室合计使用的算力可能还不到全球的一半。在本周的 newsletter 中，Josh You 讨论了这一占比未来可能如何变化，以及什么时候可能触及上限。根据数据中心电力容量和算力支出的披露信息，OpenAI、Anthropic 和 xAI 使用的算力合计可能不到全球总量的 30%。谷歌和 Meta 虽然是大型超大规模云厂商，但它们很大一部分算力用于云服务和推荐系统，而不是用于其前沿模型实验室。2026 年，前沿实验室在 AI 算力中的占比可能会继续提升。例如，Anthropic 受益于其收入的历史性增长，正在积极锁定算力资源。但头部实验室可能在未来几年内吸收大部分新增空间，此后其增长将受到芯片产能的限制。如果少数前沿玩家最终主导整个 AI 行业，那么要么头部实验室的算力增长必须放缓，要么整体 AI 建设速度必须加快。随着 AI 资本开支接近每年 1 万亿美元，后者将带来重大的经济影响。

OpenAI 第一季度营收约 57 亿美元，比主要竞争对手 Anthropic 同期营收高出近 10 亿美元。据两位知情人士透露，编码智能体 Codex、企业销售增长以及 ChatGPT 广告测试推动了第一季度的增长。然而此后，Anthropic 似乎已反超这家 ChatGPT 的创造者，其近期增速可能使两家公司的营收差距在年底前进一步扩大。据一位知情人士透露，Anthropic 的年化收入(过去一个月的营收乘以 12)近期已逼近 450 亿美元。OpenAI 在 2 月份的年化收入为 250 亿美元。Anthropic 还预计，其第二季度营收将增长逾一倍，达到近 110 亿美元，外加近 6 亿美元的意外营业利润。OpenAI 的第二季度展望尚无法获悉，不过该公司核心产品的增长有望拉动营收。据知情人士称，自 4 月以来，OpenAI 发布了 GPT 5.5 模型，在拆解用户复杂请求方面比旧版模型更强；同时，随着最新图像生成模型的推出，其消费者用户增长也进一步加快。此外，OpenAI 的编码模型 Codex 同样出现了显著增长。

微软：

5 月 22 日新浪财经报道，本周微软取消了内部的 Claude Code 授权，原因是基于 token 的计费方式使得成本过高，即使对于一家拥有近乎无限云资源的公司而言也难以承受。AI 补贴时代正在终结。微软这家公司曾经砸 130 亿美元投资 OpenAI，还为 Anthropic 提供了大部分的 Azure 云计算资源，但现在看到竞争对手 Anthropic 的 Claude Code 工具账单后，却觉得太贵不值得继续付钱了。这不是说 Anthropic 的产品不够好，是因为按 token 计费的方式，让企业必须面对大规模使用 AI 模型的真实成本，而这个成本比之前大家用固定费率试用时想象的高得多。最近半年，Anthropic、OpenAI 和 Google 都悄悄提高了实际价格。很多企业之前乐观地假设 AI 成本会一直下降，于是大力建设各种 AI 工作流程，现在真实账单来了，结果年度预算才几个月就全部烧光。

智谱：

IT 之家 5 月 22 日消息，智谱今日宣布面向部分企业客户提供 GLM-5.1 高速版 API“GLM-5.1-highspeed”。该模型输出速度达到 400 tokens/s，刷新当前全球大模型厂商 API 的速度上限。更重要的是，在过去，“快”往往意味着“小”，高速模型几乎总是轻量级模型。GLM-5.1 高速版打破了这一行业惯例，首次在国产大模型中，将旗舰级能力与低延迟同时带入生产环境。

5 月 23 日商业周刊报道，中国两家领先的 AI 大模型开发商将被纳入香港一项主要科技股指数，这意味着它们有望通过互联互通机制吸引更多内地投资者资金。恒生指数公司 5 月 22 日在季度检讨后宣布，智谱和 MiniMax Group Inc. 将被纳入恒生科技指数以及恒生综合指数。此次纳入指数，可能为这两家公司进入港股通机制铺平道路，市场预计此举将吸引强劲的南向资金流入。自 1 月上市以来，智谱股价已累计上涨约 1000%，MiniMax 上涨约 365%。按照快速纳入规则，被纳入相关指数后，智谱最快可于 6 月 8 日符合港股通交易资格。由于采用同股不同权架构的公司面临更严格要求，MiniMax 则可能要到 8 月才能符合资格。

4.3 海内外大厂重点跟踪

阿里巴巴：

快科技 5 月 20 日消息，阿里巴巴在 2026 阿里云峰会上，正式发布了全新一代千问旗舰模型 Qwen3.7-Max。在三方机构 Arena 全球大模型盲测总榜中，Qwen3.7-Max 超过 Kimi-K2.6、DeepSeek-v4-pro、GLM-5.1，与 GPT、Claude、Gemini 最强模型接近，位列国产模型第一。这是千问旗舰模型近三个月内的第三次重大迭代，从 3.5 到 3.6 再到 3.7，阿里大模型研发节奏明显加速。Qwen3.7-Max 面向智能体（Agent）场景全新设计，在多个核心维度实现突破。阿里云表示，Qwen3.7-Max API 即将上线百炼平台，后续还将推出 Qwen3.7-Plus 等版本，覆盖从编程智能体到视觉智能体的全场景需求。

百度：

5 月 18 日华尔街见闻报道，根据周一公布的最新财报，2026 年一季度，百度总收入为 321 亿元，同比小幅下降约 1%、环比下降 2%，但超过市场预期的 314.9 亿元；归属百度的净利润为 34 亿元，同比下降 55%、环比增长 93%；非公认会计准则下归属百度净利润为 43 亿元，同比下降 33%、环比增长 11%。经调整每 ADS 利润 12.06 元，市场预期仅为 11.84 元。百度核心 AI 新业务收入达到 136 亿元，同比增长 49%、环比增长 21%，占百度一般性业务收入的比例首次超过一半，达到 52%。这意味着，百度收入重心正在从传统搜索广告，进一步转向智能云、AI 应用和 AI 原生营销。其中，智能云基础设施是本季最核心的增长引擎：一季度收入 88 亿元，同比增长 79%、环比增长 52%；GPU 云收入同比增长 184%。相比之下，传统业务收入 102 亿元，同比下滑 29%，在线营销服务收入 126 亿元，同比下滑 22%，反映出传统广告业务仍在承压。

风险提示

1、AI 商业价值不及预期的风险：目前 AI 市场应用仍处于初级阶段，盈利模式仍需探索，市场尚未成熟。若商业模式无法持续发展新客户，需求大幅减弱，或市场接受度偏低，可能对营业收入造成较大负面影响，损害相关公司的盈利能力及产品或服务的商业价值。

2、技术发展速度不及预期的风险：目前 AI 模型的使用仍受限于诸多因素，在特定领域无法达到预期的提高生产力效果。该领域目前仍面临较大的技术挑战，包括模型训练效果不稳定、算法不成熟等问题。若技术落地不及预期，可能影响 AI 的应用领域和运行效率，造成较大的投资损失。

3、供应链集中度过高的风险：AI 行业基础设施建设目前高度依赖某几家核心供应商，极易受到相关供应商供应短缺的影响。此外在训练方面，AI 技术依赖于大量优质数据的输入。不可靠、低质量的数据来源会一定程度上影响 AI 模型训练的性能，同时提高训练过程中的不可控成本。

4、行业监管加剧的风险：目前生成式 AI 工具仍存在法律、伦理、安全风险。AI 生成内容的产权问题仍存在较大争议。各国可能针对 AI 的使用及 AI 生成内容进行更严格的监管及抵制，影响投资预期，并阻碍 AI 技术在产业上进一步落地。公司面临法律诉讼和声誉受损等负面影响风险。

5、市场竞争加剧的风险：在如今巨头科技公司加大 AI 投入，大量创业公司涌入竞争的大环境下，技术的迅速迭代及新算法的涌现可能使得公司技术迅速落后竞争对手，影响相关公司的市场份额和投资回报的稳定性。

行业投资评级的说明：

买入：预期未来 3—6 个月内该行业上涨幅度超过大盘在 15%以上；

增持：预期未来 3—6 个月内该行业上涨幅度超过大盘在 5%—15%；

中性：预期未来 3—6 个月内该行业变动幅度相对大盘在 -5%—5%；

减持：预期未来 3—6 个月内该行业下跌幅度超过大盘在 5%以上。

特别声明：

国金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。

上海	北京	深圳
电话：021-80234211	电话：010-85950438	电话：0755-86695353
邮箱：researchsh@gjzq.com.cn	邮箱：researchbj@gjzq.com.cn	邮箱：researchsz@gjzq.com.cn
邮编：201204	邮编：100005	邮编：518000
地址：上海浦东新区芳甸路 1088 号紫竹国际大厦 5 楼	地址：北京市东城区建内大街 26 号新闻大厦 8 层南侧	地址：深圳市福田区金田路 2028 号皇岗商务中心 18 楼 1806

【小程序】
国金证券研究服务

【公众号】
国金证券研究