下载日志已记录，仅供内部参考

国盛证券 GUOSHENG SECURITIES

证券研究报告|行业专题研究

2026年09月23日



AI投资进入回报兑现期：宏大叙事延续，CAPEX/ARR剪刀差收敛

分析师 杨然

执业证书编号：S0680518050002

邮箱：yangran@gszq.com

分析师 蔺亚婴

执业证书编号：S0680526060002

邮箱：linyaying@gszq.com

打造极致专业与效率

调研日志已记录，仅供内部参考

01

AI的宏大叙事或可持续很久？

Token指数增长 / 渗透率早期 / Agent需求跃迁 / 定价长协化

AI投资分析：1）叙事仍长：Token放量、渗透处于早期、定价长协化

广发日志已记录，仅供内部参考

国盛证券 GUOSHENG SECURITIES

- **分析1：下游需求仍处高增阶段。**Token调用量一年增长约23倍（OpenRouter周度4.9→113万亿），国内日均Token两年增长超千倍（140万亿，截至2026年3月）；Agent将Token需求从"人类交互驱动"推向"机器流程驱动"，单任务消耗为传统对话的3-50倍。

- **分析2：渗透率仍处早期。**ChatGPT月活9.61亿仅约为全球互联网用户15.7%（截至2026年3月）；2025年，活跃AI Agent约2860万个，IDC预测2030年达22.16亿（CAGR 139%）；商业化空间远未打开。

- **分析3：定价模式进入变革期，提升收入确定性。**AI收费从Tokens按量计费转向长协订阅，云厂CAPEX从"脉冲式"变为"平滑可预测"，订单长协化（26Q2谷歌云Backlog达5139亿美元）提升收入确定性。

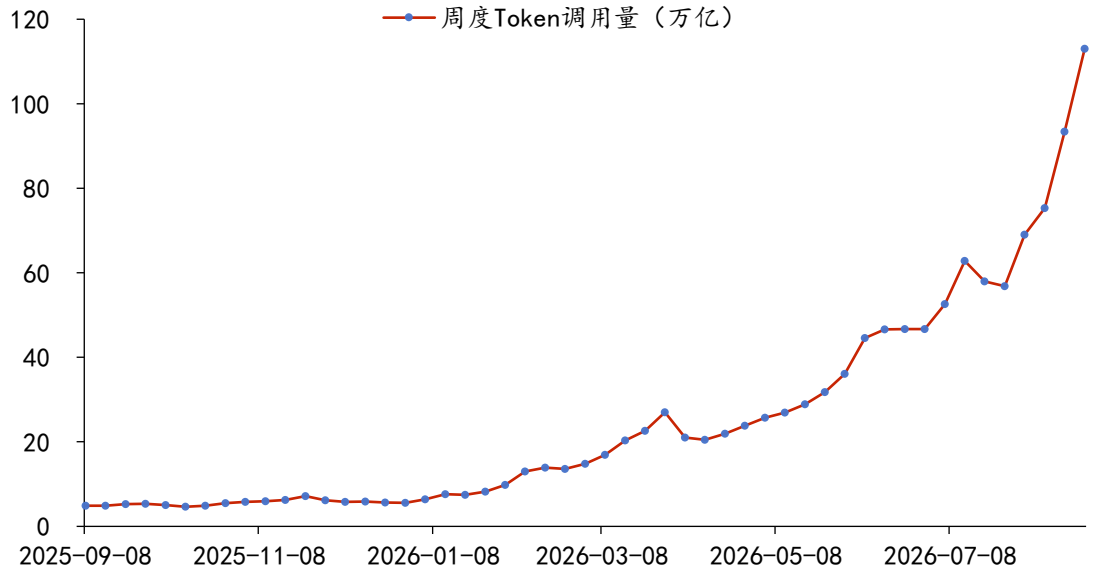
调研日志已记录，仅供内部参考

Token消耗呈指数级增长

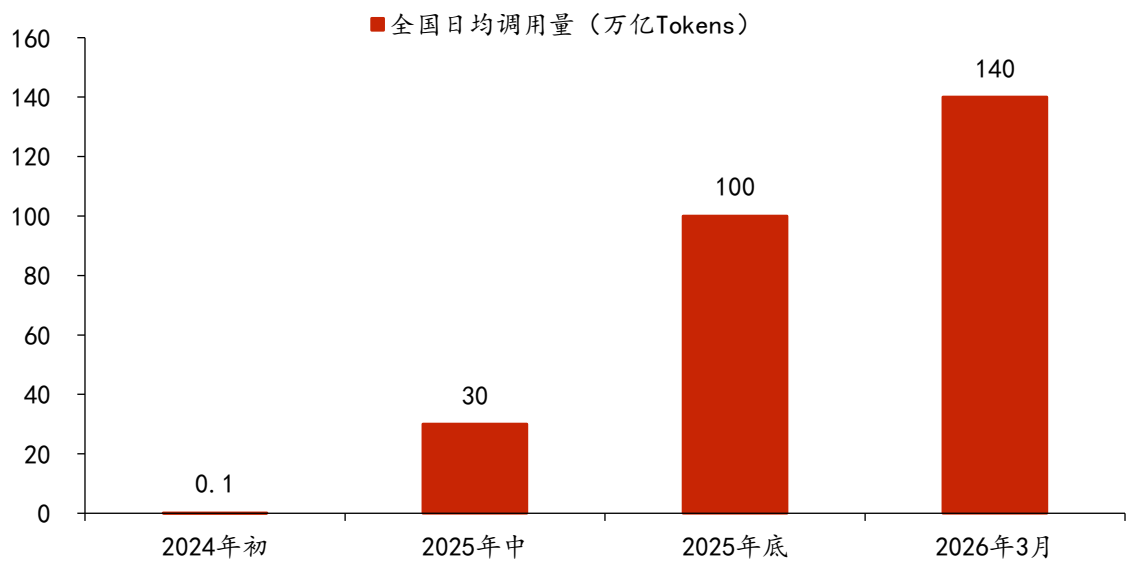
➢ Token是AI时代的"用电量"——总量增长23倍，需求持续扩张

- OpenRouter平台周度Token调用量从2025年9月8日当周的4.9万亿增至2026年8月24日当周的113万亿，近一年增至约23倍；8月以来连续4周上涨，较7月27日当周增长约99%，调用需求加速释放。
- 根据国家数据局，国内日均Token调用量从2024年初约0.1万亿增至2026年3月的140万亿，两年增幅超千倍。
- IDC预测全球活跃Agent 2025→2030年CAGR达139%，年任务执行量CAGR达524%。

OpenRouter周度Token调用量：
近一年增至约23倍（4.9→113万亿）



中国日均Token调用量：两年增长超千倍



资料来源：OpenRouter、国家数据局，国盛证券研究所

广发日志已记录，仅供内部参考

驱动力切换，Token需求从线性走向数量级放大

➢ **Token 总量在涨只是表象，更重要的是"谁在消耗 Token"——从"人类聊天驱动"切换为"机器/Agent 流程驱动"。**

- **需求引擎从"人"换成"机器"**

第一阶段：人驱动。 Token 增长主要靠 ChatGPT 类对话产品的用户渗透——用户数 × 人均对话次数决定用量。这种增长有天花板：人类只有几十亿，每天只有 24 小时，注意力有限，交互频率是"分钟级"的。

第二阶段：机器驱动。 Agent 围绕一个目标自主进行任务拆解、多次调用模型、调用外部工具、自我校验，全程无需人在场，且可以后台 7×24 小时持续执行、多个 Agent 之间高频联动。

这个切换的投资含义是：Token 需求增长与用户数解耦了。第一阶段看渗透率（用户还能涨多少），第二阶段看任务复杂度和 Agent 数量（机器可以无限并发）——后者的增长天花板远高于前者。这正是"叙事还会持续很久"的最强支撑。

- **切换的量级：单任务消耗放大近30倍**

传统 Chatbot：单轮问答约 1000 tokens，LLM 调用 1 次，无工具调用；

Agentic 任务：单任务调用模型 10-20次，消耗Token是标准聊天机器人的5-30倍。

也就是说，即使用户数不增加，只要交互形态从 Chat 迁移到 Agent，Token 总盘就有 1-2 个数量级的放大空间。

资料来源：AXE、Openreview、Naoma，国盛证券研究所

调研日志已记录，仅供内部参考

驱动力切换，Token需求从线性走向数量级放大

➢ **Token 总量在涨只是表象，更重要的是"谁在消耗 Token"——从"人类聊天驱动"切换为"机器/Agent 流程驱动"。**

- **驱动力切换在数据中的体现**

推理模型占比：OpenRouter 平台上推理类模型（会产生大量中间思考 token）的调用占比从 2025 年初几乎为 0 升至年底约 50%——单次请求的 token 密度在系统性上升。

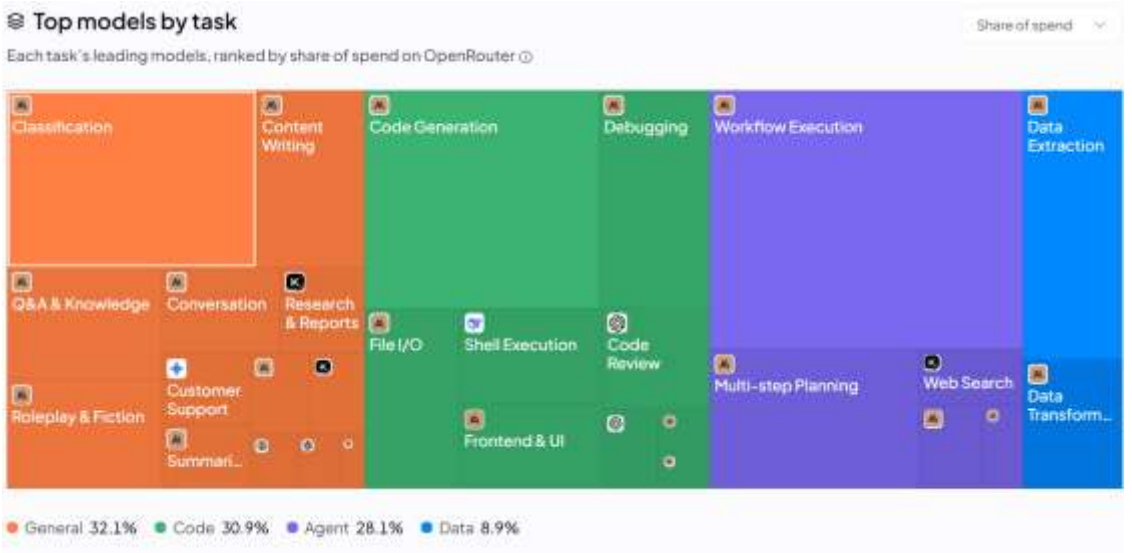
编程/Agent 场景领跑：编程与Agent已成为OpenRouter最核心的使用场景，Code与Agent支出占比分别达30.9%和28.1%，合计接近六成。

总量斜率变陡：OpenRouter周度Token调用量：近一年增至约23倍（4.9→113万亿）——如果只是用户渗透驱动，增速应该是逐渐放缓的 S 曲线，而不是斜率反而变陡。斜率变陡本身就是"新引擎接棒"的证据。

OpenRouter推理模型Token占比过半



OpenRouter编程与Agent场景支出占比接近六成



资料来源：OpenRouter，国盛证券研究所

研究日志已记录，仅供内部参考

渗透率仍处早期，B端与Agent空间远未打开

- **对标移动互联网渗透曲线，AI当前约处于"2011年的智能手机"阶段。**
- **根据过去的经验，渗透率15%-20%是应用高增与商业化的起点，而非终点。**

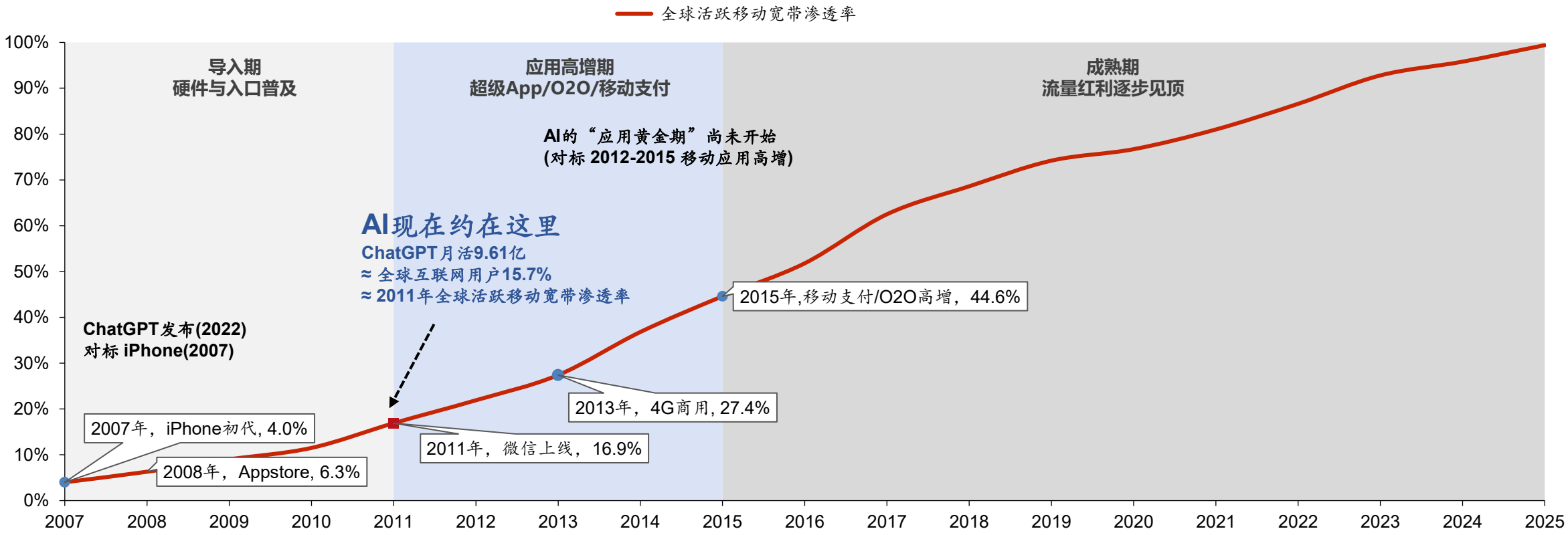
对比各维度数据，我们判断目前的AI渗透率仍处于早期阶段		
C端入口	ChatGPT月活9.61亿（2026.3），约为全球61.2亿互联网用户的15.7%；Gemini App月活9亿+（2026.5），同比翻倍以上。	**渗透率<20%，远低于成熟互联网应用**
B端调用	OpenAI API每分钟处理超150亿tokens（2026.4）、Google约190亿（2026.5）；国内企业级大模型日均调用量49万亿tokens（2025.12）。	**企业AI仍处于试点向生产系统迁移阶段**
Agent数量	全球活跃AI Agent约2860万（2025年），IDC预测2030年达22.16亿（CAGR 139%）；年任务量440亿次→415万亿次。	**agent渗透率较低，数字劳动力刚起步**
供需关系	Anthropic按每年10倍增长规划算力预算，ARR增速（26Q1环比年化）高达约80倍/年，被迫紧急外购算力Colossus 集群、22 万张以上 GPU同等计算能力	**产业界持续低估需求，供给追赶需求**

资料来源：新浪财经、Datareportal、OpenAI、 Google、沙利文分析、头豹研究院、IDC、Anthropic 、Anthropic联合创始人兼CEO Dario Amodei公开发言，国盛证券研究所

研报日志已记录，仅供内部参考

渗透率仍处早期，B端与Agent空间远未打开

- 锚点：ChatGPT月活9.61亿（2026.3），约为全球61.2亿互联网用户的15.7%——与2011年全球活跃移动宽带订阅率（约16.9%）相当；ChatGPT发布（2022）对标iPhone发布（2007），当前恰是“微信上线”前后的位置。
- 含义：移动互联网最大的应用与商业化红利（超级App、O2O、移动支付、4G流量）发生在渗透率由15%增长到50%的2012-2015年。
- AI的"应用黄金期"与收入兑现的主升段大概率还在前面，宏大叙事远未到终局。



资料来源：ITU、Apple、腾讯、中央网络安全和信息化委员会、人民邮电报，国盛证券研究所

研究日志已记录，仅供内部参考

收费模式Tokens→长协化，锁定长期需求

➢ 类比移动互联网"按KB计费→包月无限流量"：定价确定性催生用量繁荣

- **Tokens按量计费的局限：**成本不可预测、账单失控，抑制企业深度应用——类比2G时代0.01元/KB，用户"小心翼翼关图片"；
- **长协订阅：**预购吞吐量单位——成本可预测、资源有保障、客户粘性高；
- **对云厂的意义：**CAPEX从"基于不确定需求的脉冲式投资"变为"基于锁定合同的平滑规划"，26Q2谷歌云Backlog达5139亿美元、AWS长期合同未履约义务达4960亿美元，投资回收可见性显著提升；
- **单价下行×用量高增（Jevons效应）：**DeepSeek V4 Pro空闲时段降至 0.66/1.98 美元/百万 Tokens，推动调用放量；而26Q1智谱API涨价83%后付费调用量反增4倍——行业从"低价换量"进入"量价分层"，总盘持续扩张。

移动互联网 2G/3G：按KB/MB计费	移动互联网 4G：包月不限量	移动互联网 5G：按速率分级
AI映射 → Tokens按量	**AI映射 →长协订阅**	**AI映射 → 分层定价+MaaS**

资料来源：人民网、国资委，国盛证券研究所

调研日志已记录，仅供内部参考

02

AI投资是否进入回报兑现期

Capex/ARR剪刀差 —— 投入端增速下台阶 / 变现端加速兑现 / 剪刀差开始收敛

AI投资分析：2）capex与arr的收敛性

广发日志已记录，仅供内部参考

国盛证券 GUOSHENG SECURITIES

- **关键指标1-- CAPEX：**2026年四大云厂指引合计超7300亿美元（同比约+79%）；CapEx/OCF投入强度接近极限，后续跟踪重点从"投多少"转向"回报与现金流"。

- **关键指标2--ARR：**OpenAI ARR超250亿美元（2026.2），国产模型ARR"三级跳"；ARR是验证CAPEX能否兑现为收入的核心中间变量。

- **需求真实性验证：**GPU租赁价格全面反转上行，算力由买方市场转为卖方市场，支撑资本开支的合理性。

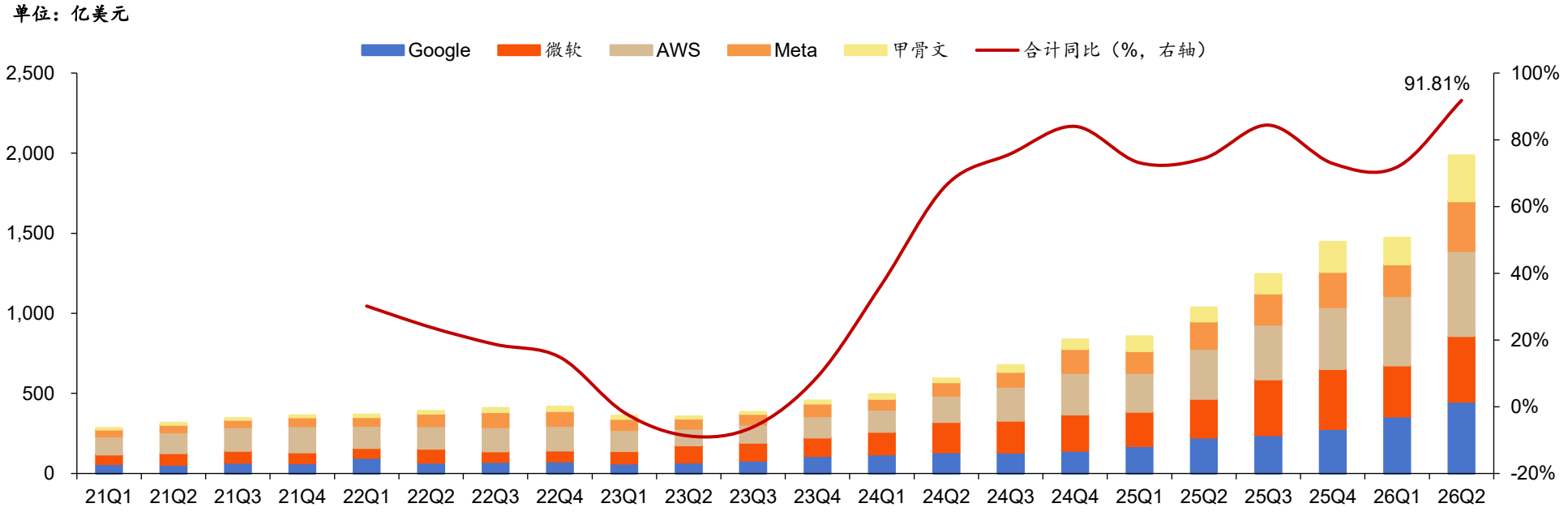
研报日志已记录，仅供内部参考

投入端：CapEx总量仍在高位扩张，但现金流压力显性化、节奏趋于理性

➢ 五大厂26Q2合计约1986亿美元，同比+91.81%，全年指引较2025年再增约79%

- **2026年科技巨头CapEx整体维持高位、以小幅上修为主：**亚马逊全年资本开支预期由约2000亿美元上调至2200亿美元；微软由1900亿美元调整至1750亿美元，主要反映租赁分类变化，剔除该影响后，2026年资本开支投资预期保持不变；谷歌将全年资本开支区间由1800-1900亿美元上调至1950-2050亿美元，并明确预计2027年资本开支将进一步大幅增长；Meta则将区间由1250-1450亿美元小幅调整至1300-1450亿美元。
- **增量驱动因素：**算力短缺（Azure确认需求超供给）+ 组件涨价（存储LTA集中激活）+ Stargate等新兴主体（OpenAI 4年5000亿美元投资计划）。

五大科技厂商季度资本开支：**26Q2合计约1986亿美元，同比+91.81%**



资料来源：各公司公告、业绩会，国盛证券研究所

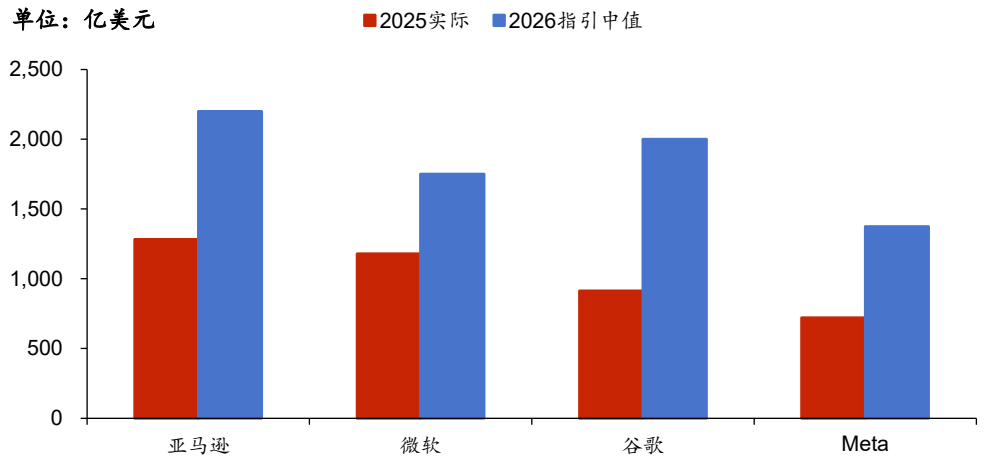
调研日志已记录，仅供内部参考

现金流与回报：从"投多少"到"扛不扛得住"

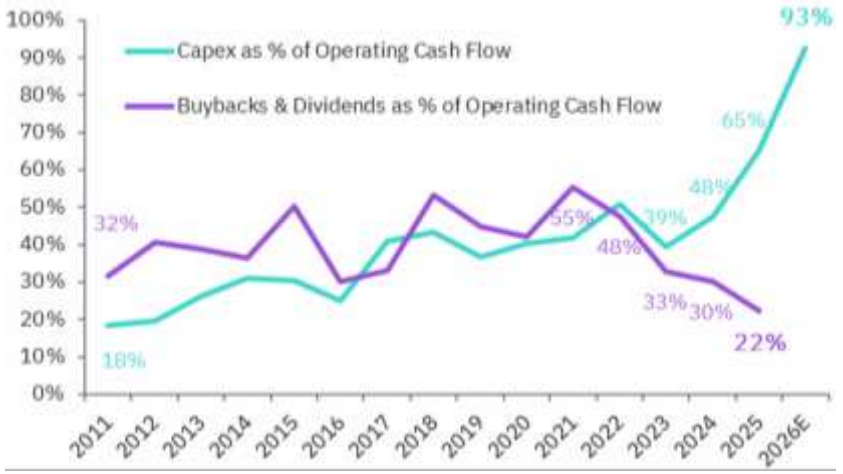
➢ CapEx/OCF不断攀升是本轮周期最重要的压力测试指标

- **全球云厂商资本开支持续加速，AI基础设施投入进入新一轮扩张周期：**2026年亚马逊、微软、谷歌和Meta资本开支指引中值合计超过7,300亿美元，同比增长约79%；其中谷歌、Meta增幅尤为显著，反映AI算力、数据中心及配套基础设施建设仍在提速。大厂已由试探性投入转向规模化部署，算力供给扩张成为推动AI产业发展的核心主线。
- **资本开支对现金流的消耗同步上升，投资逻辑从“投入规模”转向“回报兑现”：**四大云厂商CapEx/经营性现金流由2023年的39%升至2026年预计的93%，意味着经营现金流中用于基础设施建设的比例显著提高。后续市场将更加关注新增算力能否转化为云服务收入、AI产品变现和利润增长；若商业化进度不及资本开支扩张速度，折旧压力与自由现金流约束将逐步显现。

2026年CAPEX指引：
四大厂合计超7300亿美元，同比+79%



四大厂CapEx/OCF：2026E升至约93%



资料来源：各公司公告、业绩会、FTFL，国盛证券研究所

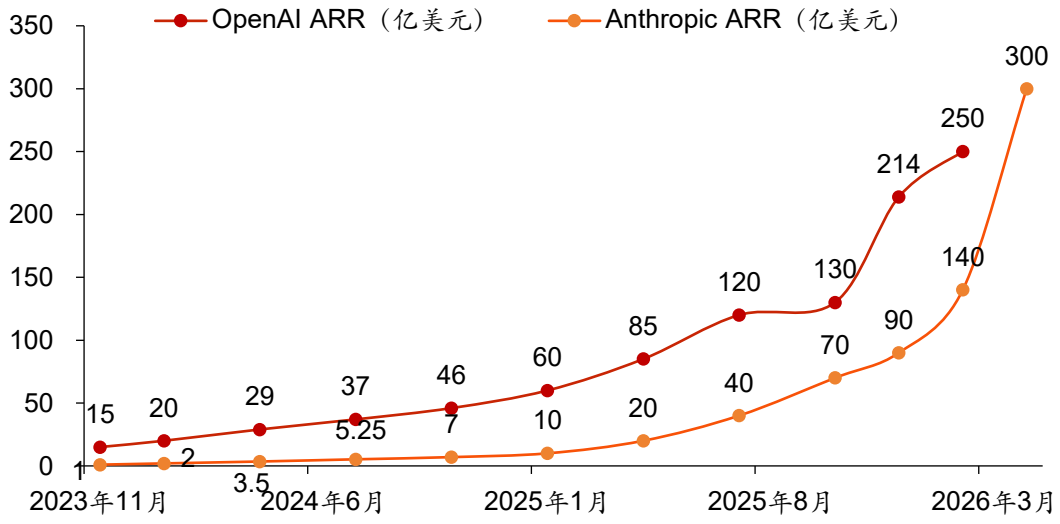
研报日志已记录，仅供内部参考

变现端：ARR加速兑现，"提价不伤量"验证商业化拐点

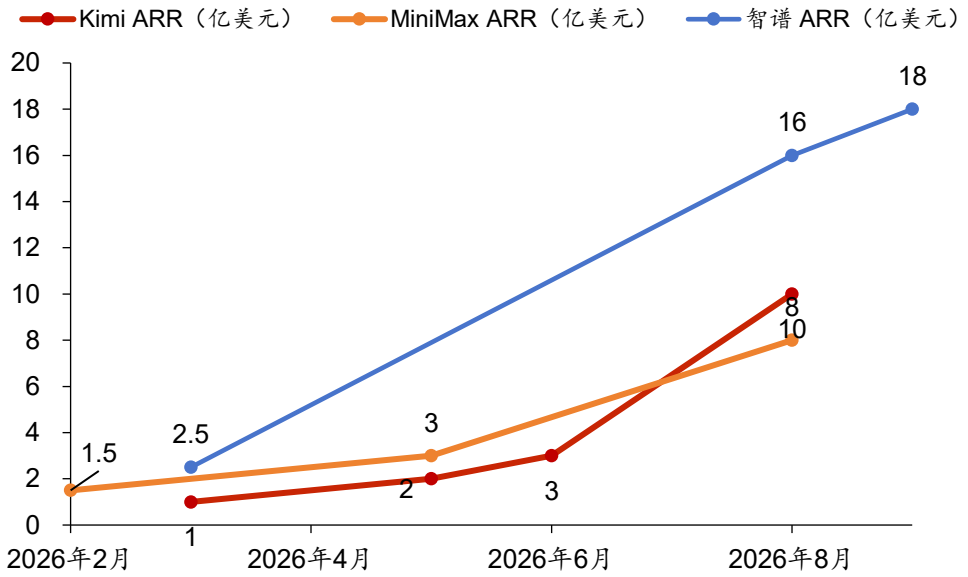
➢ 2026年全球AI Agent软件支出预计约2065亿美元（据Gartner）；Token放量正加速转化为收入。

- **海外：**OpenAI annualized revenue 截至2026年2月底已超过250亿美元，较2025年底的214亿美元继续提升；同时企业业务已贡献超过40%的收入。Anthropic增长更快，年化收入由2025年底约90亿美元升至2026年4月超过300亿美元，海外头部模型厂商商业化持续加速。
- **国产：**Kimi ARR由2026年3月约1亿美元快速升至6月中旬超过3亿美元，且API收入已占整体收入七成以上；MiniMax ARR则由2026年2月约1.5亿美元持续爬坡，2026年8月已超过8亿美元。智谱MaaS API平台2026年3月ARR达约2.5亿美元，同比提升60倍，且Q1 API提价83%后调用量仍增长400%，9月ARR达18亿美元，显示国产模型商业化已进入收入与调用量同步放量阶段。

海外头部模型厂商ARR进入百亿美元级，
OpenAI超250亿美元、Anthropic超300亿美元



国产大模型ARR加速爬坡，
Kimi升至8亿美元、智谱升至18亿美元



资料来源：Open AI官网、智东西、36氪、第一财经、财联社、新浪财经、Wind，国盛证券研究所

研报日志已记录，仅供内部参考

交叉验证：算力租赁价格全面反转上行

➢ 从"担心过剩降价"到"供不应求涨价"——需求真实性的市场化定价证据

- **H100一年期GPU租赁合同的价格已从2025年10月的低点1.70美元/小时/GPU飙升至2026年3月的2.35美元/小时/GPU，涨幅近40%。**
- **涨价已从新卡（B200/B300）蔓延至上一代卡（H100/A100），上一轮"H100租赁价格下滑压制回报"的担忧已被逆转，存量算力资产回报率改善。**



资料来源：SemiAnalysis，国盛证券研究所

调研日志已记录，仅供内部参考

客观风险评估：两大宏大叙事风险均为"部分事实+前瞻推演"

中长期叙事风险：

1）若模型能力趋同、训练及推理效率持续提升，头部厂商的算力采购结构与扩张节奏可能调整，新增硬件需求增速存在放缓风险。

2）若存量算力扩大对外供给，可能影响算力租赁价格、设备利用率及新增采购需求；其长期影响取决于需求增长能否消化新增可用容量。

事实：

1）部分大型科技企业正在探索算力对外供给。Meta在2026年二季度电话会上表示，潜在企业业务包括直接出售算力，并已收到高于购入成本的报价，但相关披露尚未明确实际成交价格和规模。Anthropic于2026年5月披露与SpaceX达成协议，使用Colossus 1数据中心的全部算力，但这一合作不足以认定xAI整体转型为算力出租商。

2）与此同时，新增算力合作仍在推进。NVIDIA与OpenAI披露至少10GW的部署合作意向；Anthropic披露与Amazon最高5GW、与Google及Broadcom 5GW的算力安排，相关容量计划分期落地。上述披露表明部分企业仍有较大算力需求，但应区分合作意向、已签协议和实际投运容量，不能仅凭个别外售或采购事件判断全行业是否供给过剩。

尚待验证的判断：

1）现有披露尚不支持“大厂普遍停止扩张算力”的判断。按各公司披露口径汇总，2026年第二季度五大厂26Q2合计约1986亿美元，同比+91.81%；按全年指引区间中值汇总，较2025年增长约79%。

2）模型竞争格局尚未形成稳定结论。Kimi K3、DeepSeek V4等模型持续推出，表明多家厂商仍在研发迭代；但模型发布及公司自测结果不足以证明头部厂商缺乏稳定竞争优势，也不足以判断头部与其他厂商之间的能力差距如何变化，仍需结合持续评测、客户留存和商业化表现观察。

需要持续观测的指标：

1）AI商业化收入、利润及现金流回收与投入规模的匹配程度。收入增速与资本开支增速之差可作为观察指标，但不等同于投资回报率，需结合资产使用期、折旧、利用率及后续财报评估。

2）对外提供算力的已签合同、实际成交价格、交付规模、客户用途及合同期限，以判断其对供给结构、租赁价格和新增采购的影响。

3）单位推理成本、服务价格、调用量及实际算力消耗的变化，检验使用量增长能否抵消单位资源消耗下降；效率提升是否带来总需求扩大，仍需数据验证。

研报日志已记录，仅供内部参考

跟踪框架：三项重点观察与验证

在剪刀差数据证伪之前，产业主升阶段尚未结束；跟踪验证优于主观择时

节点一：2026Q3财报

1）跟踪AI相关收入、利润及经营现金流，区分ARR、确认收入与现金回款。

2）关注公司后续披露的2027年投资计划及调整原因，统一会计口径后再作比较。

3）观察CapEx/OCF、自由现金流及融资结构；增速差不等同于投资回报率，单季度数据不足以确认长期趋势。

节点二：算力对外供给的商业化进展

1）跟踪Meta等公司的实际签约、成交价格、交付规模及客户用途，区分报价与已成交业务。

2）结合新增需求、利用率和采购替代情况，评估外售算力对供需结构的影响。

3）观察相关安排是否扩展至更多企业，避免将个别项目直接外推为行业普遍现象。

节点三：租赁合同、价格与交付

1）按GPU型号、地区、服务内容和合同期限，跟踪新签及续约价格、实际交付周期与利用率。

2）区分公开报价、短期现货与长期合同价格，避免由局部样本推断整体租赁价格趋势。

3）结合Token调用量、API价格及实际算力消耗，观察效率提升、需求增长和变现能力的匹配程度。

综合判断：现有公司披露显示，部分企业仍在推进算力合作与商业化，但订单兑现、投入回收及供需状况可能存在差异。仅凭单季度数据、个别外售案例或历史周期类比，尚不足以判断系统性泡沫是否形成或产业上行阶段是否结束；后续需以连续数据动态检验。

研究与估值关注：结合订单约束力、交付与回款、盈利质量、客户集中度及融资能力开展个股分析，并使用一致口径的盈利预测和现金流假设评估估值。PEG、PE等单一指标不足以单独判断投资价值；对依赖远期增长假设的标的，需检验预期变化下的估值敏感性。

广发日志已记录，仅供内部参考

跟踪框架：CAPEX×ARR剪刀差影响行情节奏

同时观察投入扩张与有效需求，将使用量增长与收入、利润及现金回收相互印证。

投入端：CAPEX

1）投入规模：跟踪季度实际支出及年度指引调整，统一财年、自然年和现金支出、融资租赁等口径。

2）资金状况：结合CapEx/OCF、自由现金流及融资结构，观察投入对经营现金流的占用和外部融资需求。

3）回报质量：跟踪折旧、利润及资本回报；ROIC需与匹配口径的资本成本比较，未单独披露的AI业务回报不作直接推算。

中间变量：Token/价格

1）需求样本：参考OpenRouter平台Token使用趋势，区分Token量与请求次数，不将平台样本直接外推为全行业需求。

2）算力价格：按GPU型号、地区和合同期限跟踪价格、交付及利用率，区分挂牌价、实际成交价与续约价。

3）API计价：区分模型、输入、输出、缓存及折扣口径，结合付费使用量观察价格变化，避免由用量增长直接推断收入增长。

变现端：ARR

1）海外厂商：分别跟踪微软、谷歌云业务收入及订单，以及OpenAI、Anthropic公开披露的年化收入指标；区分已确认收入、Backlog/RPO与ARR口径。

2）国内厂商：跟踪智谱、Kimi、MiniMax公开披露的收入、付费客户及API计价；未披露数据保留缺口，不以平台排名替代经营数据。

3）验证重点：结合价格、付费使用量、客户留存及企业客户结构，观察商业化增长能否转化为毛利和现金回收。

- **核心判断：**只要Token用量与ARR保持高增、且CAPEX指引未实质下修，AI硬件与算力主线的宏大叙事仍将延续；当ARR增速系统性跌破CAPEX增速且租赁价格转跌时，需警惕周期拐点。

广发日志已记录，仅供内部参考

风险提示

- **宏观经济不及预期。**若全球经济增长放缓、企业IT预算收缩或融资环境趋紧，可能影响云厂商、运营商及企业客户的资本开支意愿，进而导致AI服务器、网络设备和算力基础设施建设节奏放缓，对产业链相关公司的订单、收入及盈利能力造成不利影响。
- **模型迭代不及预期。**若大模型在推理能力、智能体能力、多模态融合及单位算力效率等方面的升级速度低于预期，或Scaling Law边际效应减弱，可能抑制训练和推理算力需求，并影响国产模型、AI芯片及相关软硬件生态的商业化进程。
- **AI应用落地不及预期。**若AI应用在准确性、可靠性、数据安全、成本回报及行业适配等方面未能达到客户要求，企业部署和付费意愿可能低于预期，导致Token调用量及算力需求增长放缓，相关厂商前期研发和资本投入的回报周期也可能相应延长。

下载日志已记录，仅供内部参考

免责声明

国盛证券股份有限公司（以下简称"本公司"）具有中国证监会许可的证券投资咨询业务资格。本报告仅供本公司的客户使用。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，本公司不对任何人因使用本报告中的任何内容所引致的任何损失负任何责任。

本报告的信息均来源于本公司认为可信的公开资料，但本公司及其研究人员对该等信息的准确性及完整性不作任何保证。本报告中的资料、意见及预测仅反映本公司于发布本报告当日的判断，可能会随时调整。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。本公司不保证本报告所含信息及资料保持在最新状态，对本报告所含信息可在不发出通知的情形下做出修改，投资者应当自行关注相应的更新或修改。

本公司力求报告内容客观、公正，但本报告所载的资料、工具、意见、信息及推测只提供给客户作参考之用，不构成任何投资、法律、会计或税务的最终操作建议，本公司不就报告中的内容对最终操作建议做出任何担保。本报告中所指的投资及服务可能不适合个别客户，不构成客户私人咨询建议。投资者应当充分考虑自身特定状况，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。

投资者应注意，在法律许可的情况下，本公司及其本公司的关联机构可能会持有本报告中涉及的公司所发行的证券并进行交易，也可能为这些公司正在提供或争取提供投资银行、财务顾问和金融产品等各种金融服务。

本报告版权归"国盛证券股份有限公司"所有。未经事先本公司书面授权，任何机构或个人不得对本报告进行任何形式的发布、复制。任何机构或个人如引用、刊发本报告，需注明出处为"国盛证券研究所"，且不得对本报告进行有悖原意的删节或修改。

分析师声明

本报告署名分析师在此声明：我们具有中国证券业协会授予的证券投资咨询执业资格或相当的专业胜任能力，本报告所表述的任何观点均精准地反映了我们对标的证券和发行人的个人看法，结论不受任何第三方的授意或影响。我们所得报酬的任何部分无论是在过去、现在及将来均不会与本报告中的具体投资建议或观点有直接或间接联系。

投资评级说明

投资建议的评级标准		评级	说明
评级标准为报告发布日后的6个月内公司股价（或行业指数）相对同期基准指数的相对市场表现。其中A股市场以沪深300指数为基准；新三板市场以三板成指（针对协议转让标的）或三板做市指数（针对做市转让标的）为基准；香港市场以摩根士丹利中国指数为基准，美股市场以标普500指数或纳斯达克综合指数为基准。	股票评级	买入	相对同期基准指数涨幅在15%以上
		增持	相对同期基准指数涨幅在5%~15%之间
		持有	相对同期基准指数涨幅在-5%~+5%之间
		减持	相对同期基准指数跌幅在5%以上
	行业评级	增持	相对同期基准指数涨幅在10%以上
		中性	相对同期基准指数涨幅在-10%~+10%之间
		减持	相对同期基准指数跌幅在10%以上

国盛证券研究所

北京

地址：北京市丰台区丽泽路22号院新华社国家金融信息大厦西塔B座38层

邮编：100077

邮箱：gsresearch@gszq.com

上海

地址：上海市浦东新区南洋泾路555号陆家嘴金融街区22栋

邮编：200120

电话：021-38124100

邮箱：gsresearch@gszq.com

南昌

地址：南昌市红谷滩新区凤凰中大道1115号北京银行大厦

邮编：330038

传真：0791-86281485

邮箱：gsresearch@gszq.com

深圳

地址：深圳市福田区福华三路100号鼎和大厦24楼

邮编：518033

邮箱：gsresearch@gszq.com