下载日志已记录，仅供内部参考

传媒

发布时间：2026-09-29

证券研究报告 / 行业深度报告

中国大模型的钱，最终归谁？

---大模型行业深度之二：国内篇

优于大势

上次评级:优于大势

报告摘要：

国内大模型行业在 2026 年进入商业化元年，并在 8 月完成第一轮集中兑现。字节跳动大模型业务 7 月 ARR 约 40 亿美元；智谱 2026 年上半年收入 9.54 亿元、同比增长 399.7%，8 月底 ARR 达 16 亿美元；MiniMax 2026 年上半年收入 1.166 亿美元、同比增长 283.1%，8 月 ARR 超 8 亿美元；月之暗面 8 月 ARR 突破 10 亿美元；DeepSeek 2026 年前 7 个月营收约 4.75 亿元、约为 2025 年全年的 10 倍。

国内大模型的商业模式与海外同框架、不同配比。沿用本系列第一篇“谁承担算力成本、谁占有客户关系”的两变量框架，国内四条变现路径齐备，但呈现四个结构性差异：开源主导地位显著更强；API 价格较海外低一至两个数量级；政企私有化部署是海外未成规模的变现路径；出海成为独立变现路线。

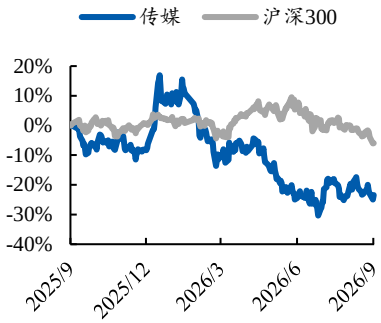
旗舰上移与轻量探底并存，定价体系与海外“存量通缩+档位上移”的双层趋势趋同。据摩根士丹利统计，2026 年二季度中国大模型平均 API 输入/输出价格较 2025 年一季度分别上涨约 48%和 80%。8 月演进进一步加速，DeepSeek 引入峰谷分时定价并预告整体上调，Kimi K3 以输入 20 元/输出 100 元成为当时最贵的国产模型，而智谱 GLM-5.3-Flash 与阿里 Qwen3.8-Flash 把轻量档压至 1 元/百万 Token 以下。旗舰上移与轻量探底并存，国内定价结构与海外“存量通缩+档位上移”的双层趋势趋同。

大厂靠生态闭环，创业公司靠开源与出海。阿里云与算力分部 2027 财年一季度收入同比增长 45%、经调整 EBITA 利润率 12%，是目前唯一披露 AI 分部规模盈利数据的大厂；创业公司中，DeepSeek 前 7 个月整体毛利率 44.6%、API 业务毛利率 82.9%，最接近盈亏平衡；智谱的 API 业务毛利率在 2026 年上半年转正至 24.6%。行业整体仍处于“收入拐点已确认、利润拐点未至”阶段，基准情景下智谱的盈亏平衡指向 2028 年前后。

本报告的核心结论：**中国大模型行业已经形成可持续的商业模式。**API 层面的单位经济已被证明，亏损全部来自训练军备竞赛而非推理亏损；但利润不归“模型”本身，将沿开源标准位（DeepSeek/Qwen）、云生态位（阿里/字节/腾讯）、旗舰定价权位（Kimi/智谱 MaaS）三个节点集中。占不到节点的公司拥有的不是利润来源，而是成本负担。

风险提示：AI 技术与产品进展不及预期的风险；商业化落地与企业需求不及预期的风险；行业竞争与价格战反复的风险；算力供给与出口管制的风险；资本开支与现金流风险；C 端变现进展不及预期的风险；出海与地缘政治风险；数据口径与第三方估算偏差的风险；二级市场估值波动的风险；测算假设不确定性的风险；监管与合规风险。

历史收益率曲线



涨跌幅（%）	1M	3M	12M
绝对收益	0%	1%	-23%
相对收益	6%	13%	-18%

行业数据

成分股数量（只）	171
总市值（亿）	9350
流通市值（亿）	8290
市盈率（倍）	49.23
市净率（倍）	2.52
成分股总营收（亿）	
成分股总净利润（亿）	
成分股资产负债率（%）	253.68

相关报告

《海外大模型如何赚钱》

--20260815

《大模型迭代竞争加剧，AI 应用爆发在即》

--20260814

《AI 视频生成技术成熟，内容产出产业升级》

--20260728

《AI 真人剧：奇点已至，星辰大海》

--20260526

证券分析师：赵明扬
执业证书编号：S0550525070009
zhaomy@nesc.cn

研究助理：刘昱亨
执业证书编号：S0550123100008
liuyh2@nesc.cn

下载日志已记录，仅供内部参考

目 录

1. 引言 4
2. 国内变现路径拆解：同框架、不同配比 5
2.1. 四种模式全景：同一框架下的中国分布 5
2.2. API 定价体系：从价格战到分层重构 6
2.3. C 端订阅：3.45 亿月活的变现困局 8
2.4. 开源范式：从生态杠杆到变现漏斗的顶层入口 9
2.5. 政企私有化部署：国内特有的第五条路径 10
3. 创业公司四样本：DeepSeek、Kimi、智谱、MiniMax 11
3.1. DeepSeek：开源+API 的第三条路 11
3.2. Kimi：API 驱动+出海，增长引擎最像 Anthropic 13
3.3. 智谱：MaaS 转型与成本结构 15
3.4. MiniMax：海外收入六成的生而全球化，B 端切换已兑现 17
3.5. 路径分化与出清 18
4. 大厂生态：模型为杠杆的中国版本 18
4.1. 阿里：AI 的两类业务，云已盈利、模型仍在投入 18
4.2. 字节跳动：2000 亿资本开支换规模第一 19
4.3. 腾讯&百度：生态嵌入与全栈布局 20
4.4. 资金闭环：大厂 CapEx 挤压与创业公司的 IPO 供粮 20
5. 国内外对标：结构性异同 21
5.1. 趋同的三条规律 21
5.2. 五个结构性差异 22
6. 未来发展探讨：盈利路径与格局推演 23
6.1. 盈利路径测算：以智谱为锚 23
6.2. 格局推演：五个情景与各自的证伪信号 25
7. 总结与展望 27
8. 风险提示 28

图表目录

图 1：国内大模型变现模式的两变量定位 5
图 2：国内大模型平均 API 价格由降转升 6
图 3：国内 AI 原生 App 月活规模（2026Q1） 9
图 4：OpenRouter 平台中国模型 Token 份额跃升 10
图 4：DeepSeek 估值跃迁（2026.04-08） 12
图 5：月之暗面 Kimi ARR（2026.03-06） 13
图 6：月之暗面估值阶梯（2023-2026，对数轴） 14
图 7：智谱营业收入与亏损（2022-2026H1） 15
图 8：智谱研发投入与研发占收入比（2022-2026H1） 16
图 9：MiniMax 总收入与海外收入占比（2023-2026H1） 17
图 10：国内主要厂商大模型业务 ARR 对比（2026 年 7-8 月） 19
图 11：国内外代表模型 API 输出价对比（2026 年中，对数轴） 22
图 12：智谱收入三情景测算（2025A-2028E） ..... 错误!未定义书签。

下载日志已记录，仅供内部参考

表 1：国内头部厂商商业模式对比 6
表 2：国内主要厂商 API 分档定价对比（2026 年 8 月底，单位元/百万 Token） 8
表 3：海内外大模型商业模式对照（2026 年中） 23
表 4：国内大模型商业化跟踪指标清单 26

下载日志已记录，仅供内部参考

1. 引言

过去十二个月，中国大模型行业完成了从“百模大战”到“商业化元年”的切换。到 2026 年 8 月底，头部厂商的收入曲线第一次呈现出可追踪、可交叉验证的斜率：据每日经济新闻按 7 月平均消耗折算，字节跳动大模型业务的年化收入（ARR）约 40 亿美元；智谱 2026 年上半年收入 9.54 亿元、同比增长 399.7%，8 月底 ARR 达 16 亿美元、最近一周年化已超 20 亿美元；MiniMax 2026 年上半年收入 1.166 亿美元、同比增长 283.1%，半年即超过 2025 年全年，8 月 ARR 超 8 亿美元；月之暗面 8 月 ARR 突破 10 亿美元、内部目标年底年化 20 亿美元，并已于 9 月 2 日保密递交港交所 A1 申请；DeepSeek 2026 年前 7 个月营收约 4.75 亿元、约为 2025 年全年的 10 倍，已聘请中信证券筹备科创板上市。与收入放量相伴的，是一级与二级市场的剧烈定价：月之暗面 Pre-IPO 轮投前估值锚定 500 亿美元、DeepSeek 第二轮 500 亿元融资对应投前估值约 5000 亿元，而港股已上市的两家大模型公司在 8 月经历深度回撤。收入从哪来、成本花在哪、利润何时出现，同样是理解国内厂商估值锚的前提。

回答这一问题在国内比在海外更困难，原因在于国内商业模式的分化发生在三个叠加的战场上。其一，定价体系经历了海外未曾出现的完整周期：2024 年 5 月豆包以低于行业均价 99.3%的定价打响价格战，两年内行业均价压至“分厘时代”，而 2026 年又出现集体涨价。据摩根士丹利统计，2026 年二季度中国大模型平均 API 输入价格较 2025 年一季度上涨约 48%、输出价格上涨约 80%。其二，开源的主导地位远超海外：DeepSeek 以 MIT 协议开放全部权重，阿里 Qwen 系列全球下载量超 30 亿次，第三方 API 路由平台 OpenRouter 上中国模型的 token 份额从 2024 年不足 2%升至 2026 年中约 45–61%。其三，国内存在海外未成规模的变现路径：央国企与地方政府的私有化部署采购曾支撑智谱约七成的收入，尽管 2026 年上半年智谱的收入结构已反转为 API 主导，但政企市场本身仍在扩容。

本报告作为大模型行业深度系列的第二篇，聚焦国内厂商的商业模式。全篇围绕一个核心问题展开：**中国大模型行业能否形成可持续的商业模式？如果是，收入与利润最终归谁？**这一核心问题拆解为三个子问题：其一，**钱从哪来**：国内头部公司沿哪些路径变现，定价体系如何演化；其二，**结构有何不同**：国内市场结构（开源主导、政企市场、价格战周期、出海、算力受限）与海外的异同如何改变收入池的分配；其三，**利润何时出现、归谁所有**：盈利路径与终局格局指向何方。

报告余下部分的结构如下：第二章沿用第一篇的两变量框架拆解国内变现路径，依次讨论 API 定价的“价格战—涨价”周期、C 端订阅的变现困局、开源范式的中国版本，以及政企私有化部署这一特有路径；第三章以 DeepSeek、Kimi、智谱、MiniMax 四个样本拆解创业公司的商业模式分化；第四章分析阿里、字节、腾讯、百度的生态变现模式及其资金闭环；第五章系统对照海内外商业模式的结构性异同；第六章探讨未来发展，包括以智谱为锚的三情景盈利测算与格局推演。

下载日志已记录，仅供内部参考

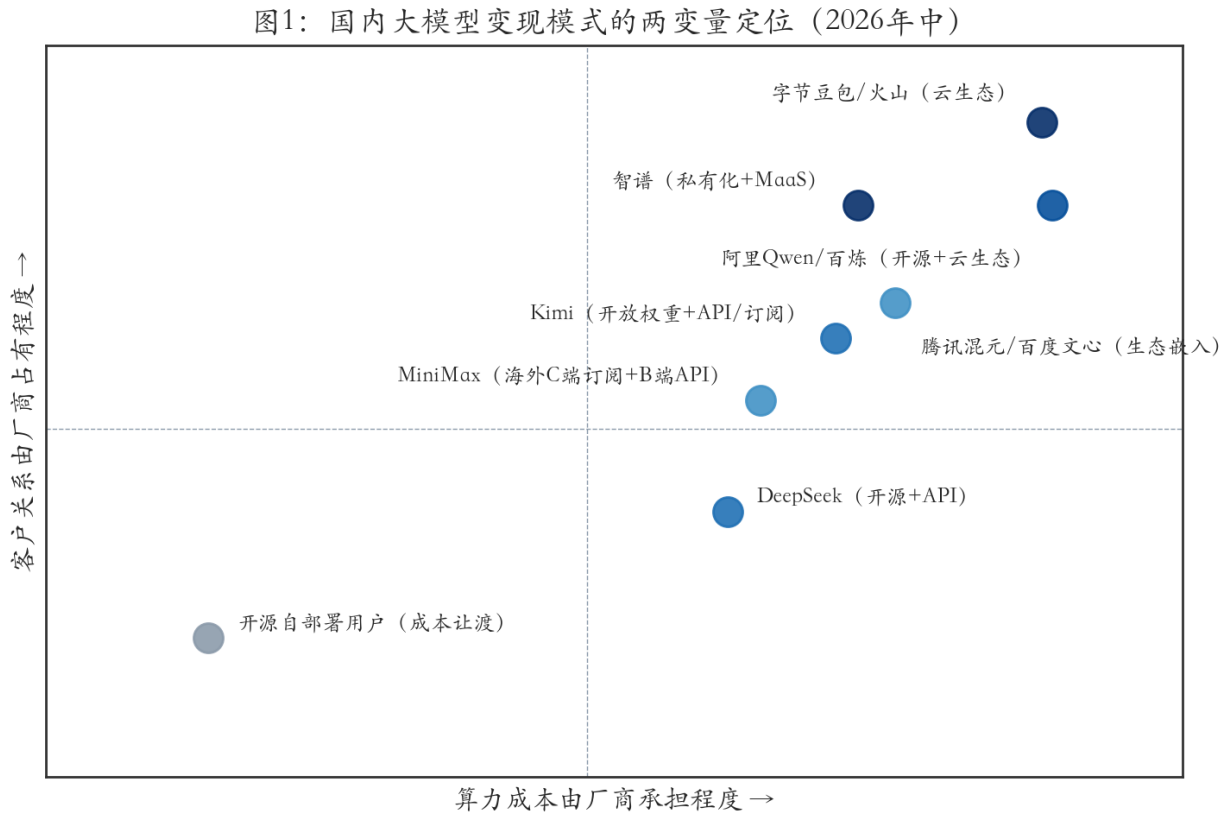
2. 国内变现路径拆解：同框架、不同配比

本章核心结论：**国内大模型厂商的变现路径与海外同为四条：API 按量计费、订阅制、云生态整合、开源间接变现，但配比显著不同。**开源在国内从 Meta 式“与变现脱钩的生态杠杆”变为“变现漏斗的顶层入口”，直接决定收入归谁；API 定价经历了海外没有的完整“价格战—涨价”周期，定价权正在重建；C 端订阅的付费率瓶颈比海外更严重；政企私有化部署构成国内特有的第五条路径。对核心问题的回答：钱从四条主路径加一条国内特有路径来，但路径有厚薄，定价权所在的旗舰档与政企档最厚，走量档最薄。

2.1. 四种模式全景：同一框架下的中国分布

沿用第一篇的分析框架：**四种模式卖的都是同一批模型的推理能力，区别在于算力成本由谁承担、客户关系由谁占有。**国内玩家在这一框架下的分布如图 1 所示，与海外版图的差异集中在两处：一是开源象限的玩家密度远高于海外，DeepSeek、Qwen、混元、GLM 均开放权重；二是政企私有化部署在国内是独立的收入支柱，而在海外，企业需求几乎全部经由云 API 承接。表 1 对比了国内头部厂商的模式归属。

图 1：国内大模型变现模式的两变量定位



数据来源：各公司公告/招股书、官网定价页、东北证券

下载日志已记录，仅供内部参考

表 1：国内头部厂商商业模式对比

厂商	主导模式	算力成本承担	客户关系	收入锚点
DeepSeek	开源引流+API 按量计费	官方 API 自营；自部署部分让渡	API 客户直接占有	API 调用量
月之暗面 Kimi	开放权重+API/订阅	厂商承担	API 与订阅直接占有	API 收入占比超七成
智谱	MaaS 主导（2026H1 API 占 86.5%）+私有化	私有化部分转嫁客户算力	政企客户+开发者	MaaS ARR 放量+本地部署存量
MiniMax	海外 C 端订阅+B 端 API	厂商承担	海外用户与开发者	海外收入占比 73%
字节跳动	云生态（火山引擎）+C 端流量	厂商承担（2000 亿+资本开支）	云客户+3.45 亿月活	大模型 ARR 约 40 亿美元（媒体折算）
阿里巴巴	开源+云生态（百炼）	厂商承担	云客户	AI 产品收入连续 12 季度三位数增长
腾讯/百度	生态嵌入	厂商承担	存量产品用户	AI 对既有业务的间接提升

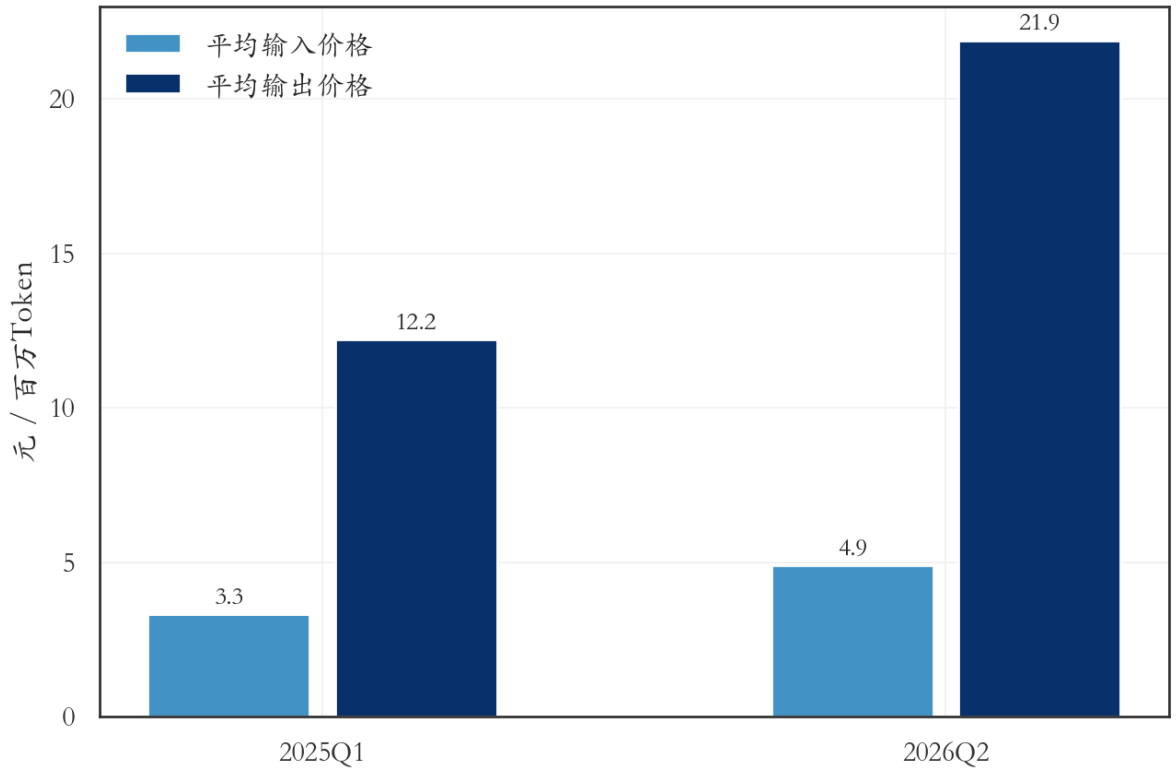
数据来源：各公司公告、招股书、官网及公开报道，东北证券整理

定价权的国内分布与海外同向，开源最低、云生态最高，但绝对水平低一至两个数量级，这决定了国内厂商的毛利率天花板更低、对用量的依赖更强。

2.2. API 定价体系：从价格战到分层重构

国内 API 定价走出了海外未曾出现的完整周期。第一阶段是单边降价：2024 年 5 月，字节豆包率先将主力模型定价降至 0.0008 元/千 Token、较行业均价低 99.3%，随后阿里通义千问降价 97%，百度两款主力模型直接免费，腾讯混元最高降幅 87.5%；到 2026 年 4 月，DeepSeek 两天内两次降价，将 V4 Pro 缓存命中输入价锁定在 0.025 元/百万 Token，行业进入“分厘时代”。

图 2：国内大模型平均 API 价格由降转升



数据来源：字节、阿里、百度、腾讯、MiniMax、智谱、月之暗面、DeepSeek 等官网、东北证券

下载日志已记录，仅供内部参考

第二阶段是 2026 年的结构性涨价。根据对字节、阿里、百度、腾讯、MiniMax、智谱、月之暗面、DeepSeek 官方报价的统计，2026 年二季度中国大模型平均 API 输入价格升至 4.9 元、输出价格升至 21.9 元/百万 Token，较 2025 年一季度的 3.3 元和 12.2 元分别上涨约 48%和 80%（图 2）。标志性事件包括：腾讯 2026 年 3-4 月两轮上调模型 API 价格，部分涨幅高达 463%；智谱 2026 年 2 月 GLM-5 发布后对部分服务提价 30-60%、部分 API 涨幅 67-100%，提价后调用量仍增长超过 400%；连价格战的发动者也在转向，DeepSeek 于 5 月 22 日将 V4-Pro 的 75%折扣转为长期价格的同时，行业整体价格中枢已明显抬升。

第三阶段自 2026 年 8 月开始，定价体系进入“分层重构”：旗舰档继续上移与轻量档再度探底同时发生。上移的一端，Kimi K3 以输入 20 元/输出 100 元每百万 Token 成为最贵国产模型，较上一代 K2.6 提价约 2.1 倍/2.7 倍，直接进入海外主流旗舰的收费区间；DeepSeek 于 8 月 6 日预告“较大幅度”整体调价，8 月 17 日起引入峰谷分时定价，高峰时段（北京时间 9:00-12:00、14:00-18:00）价格为空闲时段的两倍，V4 Pro 高峰输出价最高达 27 元/百万 Token。探底的一端，智谱 GLM-5.3-Flash 以输入 0.8 元/输出 2.8 元每百万 Token 上线（仅为旗舰 GLM-5.3 的 1/10、Claude Opus 4.8 的约 1/40），阿里 Qwen3.8-Flash 上线次日即将价格下调至输入 0.8 元/输出 2.7 元每百万 Token，MiniMax M3 则以“永久五折”把实际结算价压至 2.1 元/8.4 元每百万 Token。9 月分层进一步深化：DeepSeek 9 月 10 日发布 V4.1 Flash（KV 缓存降至前代约 1/4）并再度下调轻量档价格，闲时输入 1 元/输出 4 元、缓存命中低至 0.02 元/百万 Token，同时将 V4 Pro 请求全部路由至 V4.1 Flash 按低价计费，形成“旗舰变相降价、轻量继续探底”的组合；腾讯 8 月 28 日开源 Hy4 preview（770B/49B，输入 6 元/输出 18 元每百万 Token）补齐旗舰档；阶跃星辰 9 月 20 日发布 Step 5 Preview（AA 智能指数 44 分、跻身开源前三），称单任务成本为 Opus 5 的 1/8。这一套结构与第一篇所述海外“存量通缩+档位上移”已经同构：旗舰档按价值定价上移，轻量档在成本曲线上继续探底，混合均价的走向取决于用量结构。值得注意的是 DeepSeek 的定价并非单边：8 月 17 日峰谷定价实为整体上调，高峰缓存命中价涨幅达 400%，9 月 10 日又下调闲时价，价格已成为其调节算力负荷与产品结构的主动工具。

涨价未吓退用户，说明定价权正在重建。智谱一季度 API 定价较 2025 年底提升 83%、调用量反增 400%，表明企业级需求的价格弹性低于预期，客户为确定性付费的意愿已经形成。涨价的驱动因素有三：一是长上下文、深度推理与 Agent 任务显著增加推理负载，前期补贴价难以长期维持；二是高峰算力供给仍然紧张。峰谷定价的本质就是把电力市场的负荷定价引入推理市场；三是更强模型开始按其创造的业务价值而非单纯的计算成本定价。

下载日志已记录，仅供内部参考

表 2：国内主要厂商 API 分档定价对比（2026 年 8 月底，单位元/百万 Token）

厂商	档位	模型	输入价	输出价	缓存命中价	定价机制备注
DeepSeek	旗舰	V4 Pro 0813	9.0（峰）/ 4.5（闲）	27 / 13.5	0.30 / 0.15	8 月 17 日起峰谷分时定价
DeepSeek	轻量	V4.1 Flash	2.0（峰）/ 1.0（闲）	8 / 4	0.04 / 0.02	峰谷分时；V4 Pro 流量已切入该档
月之暗面	旗舰	Kimi K3	20	100	2	最贵国产模型；编程场景缓存率超 90%
月之暗面	中档	K2.7 Code	约 6.7（$0.95）	约 28（$4）	约 1.3（$0.19）	编码专用，美元区定价折算
智谱	旗舰	GLM-5.3	约 9.8（$1.4）	约 31（$4.4）	约 1.8（$0.26）	8 月 19 日 API 上线
智谱	轻量	GLM-5.3-Flash	0.8	2.8	0.23	为旗舰 1/10，限时折扣再减半
阿里	旗舰	Qwen3.8-Max	12	36	1.5	百炼平台国内价
阿里	轻量	Qwen3.8-Flash	0.8	2.7	0.1	8 月 27 日调价后
字节跳动	旗舰	豆包 Seed 2.1 Pro	6	30	1.2	闭源旗舰
字节跳动	中档	豆包 Seed 2.0 Pro	3.2	16	0.8	Batch 调用五折
MiniMax	旗舰	M3（≤512K 上下文）	4.2	16.8	0.84	永久五折后 2.1/8.4
MiniMax	中档	M2.7	2.1	8.4	0.42	Token 计划套餐价更低
腾讯	旗舰（预览）	混元 Hy4 preview	6	18	0.3	8 月 28 日开源，770B/49B
腾讯	轻量	混元 Hy3	1	4	0.25	Apache 2.0 开源
阶跃星辰	旗舰（预览）	Step 5 Preview	未披露	未披露	—	9 月 20 日发布，AA 44 分，权重 10 月开源

数据来源：各厂商官方定价页，东北证券
注：标注$的为美元区定价按 1:7 折算

从表 2 可以看出国内定价的完整三层结构：旗舰档（输出价 16-100 元）由 Kimi K3、Qwen3.8-Max、豆包 Seed 2.1 Pro、GLM-5.3 占据，K3 以 100 元输出价直接对标海外旗舰；中档（8-17 元）是各家的性价比主力；轻量档（2-5 元以下）则是 DeepSeek Flash、GLM-5.3-Flash、Qwen3.8-Flash、混元 Hy3 的走量区间。轻量档的激烈程度不降反升，因为 Agent 工作负载的放量主要发生在这一档。折扣机制方面，缓存命中价普遍为标准输入价的 1/10 至 1/50，峰谷分时与“永久五折”则在 8 月成为新的价格武器。

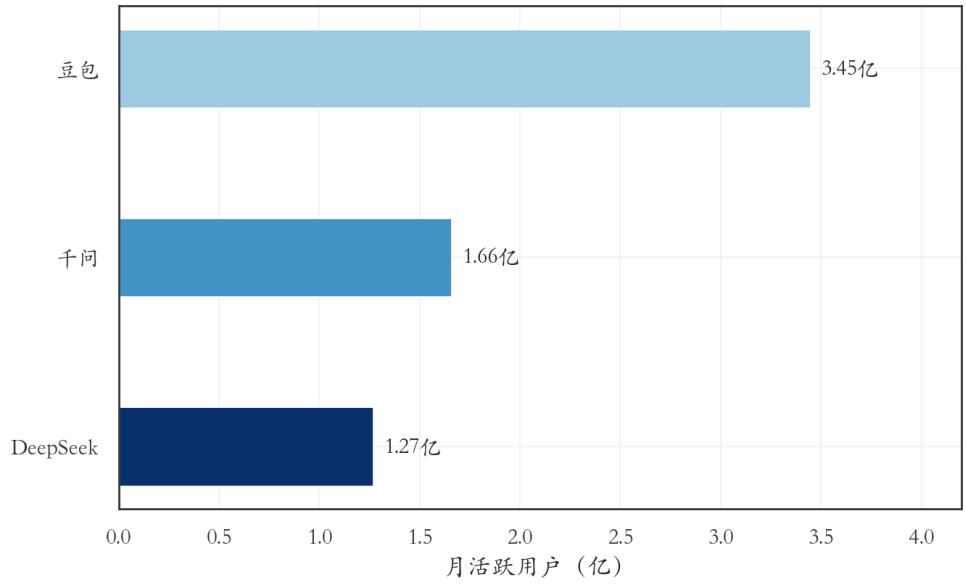
2.3. C 端订阅：3.45 亿月活的变现困局

国内 C 端订阅的瓶颈比海外更严重。据 QuestMobile 数据，2026 年一季度国内 AI 原生 App 月活达 4.4 亿，豆包以 3.45 亿月活居第一，超过第二名千问（1.66 亿）与第三名 DeepSeek（1.27 亿）之和（图 3）。但用户规模没有转化为收入：据晚点 LatePost 报道，截至 2026 年上半年豆包 App 每天收入不足百万元、主要来自电商佣金，换算下来每个日活用户每天贡献不足 0.003 元；而据 CSDN 的成本拆解估算，每百万

下载日志已记录，仅供内部参考

Token 普通对话综合成本约 4 元，豆包日均 180 万亿 Token 的单日成本约 7.2 亿元。收入与成本之间的剪刀差，构成豆包收费的直接动因。

图 3：国内 AI 原生 App 月活规模（2026Q1）



数据来源：QuestMobile、东北证券

2026 年 5 月 4 日，豆包在 App Store 更新付费服务声明：6 月 24 日正式上线三档订阅（标准版 68 元/月、加强版 200 元/月、专业版 500 元/月），付费功能专注数据分析、长文本处理、影视制作等复杂生产力场景，基础对话维持免费。摩根士丹利按保守付费转化率测算，豆包年化订阅收入可达 1.01 亿至 15 亿美元，中性情景约 4.26 亿至 6.84 亿美元。约束条件在用户结构：豆包月活中学生与中老年群体占大头。花旗覆盖 1800 名受访者的调研显示，45%的受访者愿意为 AI 高级功能付费，但可接受的月均价格仅 48.3 元，低于豆包 68 元的起步价。对比海外 ChatGPT 约 5%的付费率，国内 C 端的支付意愿与价格容忍度均更低，“垄断级流量入口+广告/交易变现”的期权价值在国内反而更重要。8 月 11 日，豆包已在酒店订单场景开启 12%的商业化抽佣。

2.4. 开源范式：从生态杠杆到变现漏斗的顶层入口

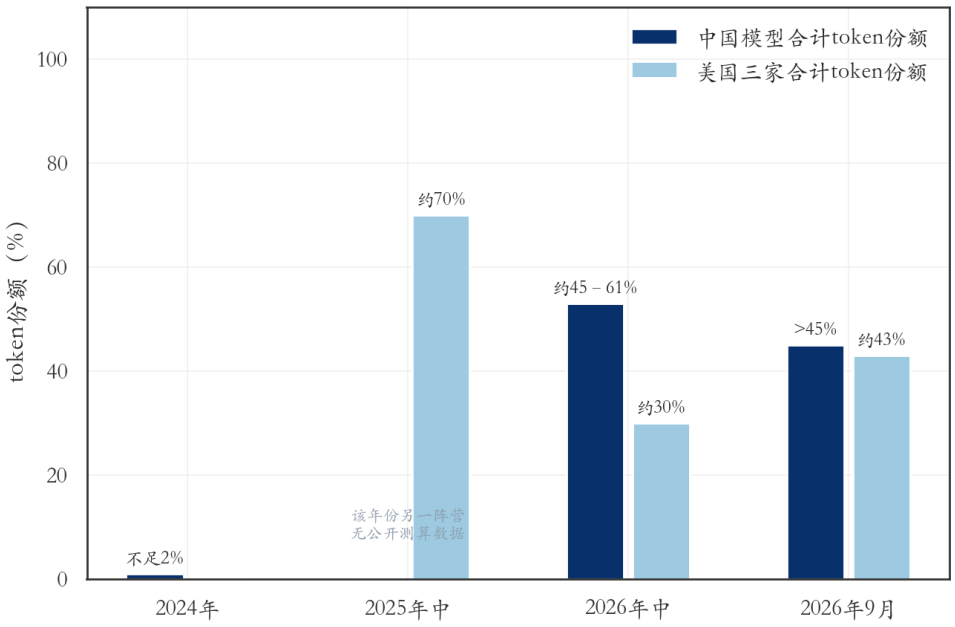
需要先澄清一个概念边界：开源权重本身不收费，开源模型的收入仍然由官方 API、云算力或企业服务承接。因此，开源在国内的角色变化，准确地说不是“从间接变现变为直接变现”，而是开源与变现的耦合度发生了质变。海外的开源者是脱钩的：Meta 从未指望 Llama 产生可归因的收入，其价值经由广告系统间接体现、无法量化剥离；而中国的开源阵营把开源变成了变现漏斗的顶层入口。**开源承担获客与信任建立的功能，收入由漏斗下游的官方 API 与云承接**，且转化路径短、可归因。

这一耦合关系在三个样本上可验证。DeepSeek 的闭环是“开源引流—API 承接—生态反哺”：以 MIT 协议开放全部权重，企业可直接下载验证性能，信任成本与切换成本双降；没有自部署算力的客户回到官方 API，按 token 付费；硅基流动、阿里百炼等算力平台围绕 DeepSeek 建模，又反过来验证其质量、强化品牌。第三方路由平台的数据提供了流量侧证据，且 9 月进一步强化：OpenRouter 数据显示，2026 年 9 月 7-13 日当周中国模型周调用量约 61.2 万亿 Token、接近美国（21.8 万亿）的 3 倍，连续 20 周超过美国；全球前十模型中国占 7 席（一年前仅 2 席）；中国厂商整体份额超 45%对美国约 43%，DeepSeek 以 25.4%的份额登顶全球厂商第一。更早的轨迹是：中国模型合计 token 份额从 2024 年不足 2%升至 2026 年中约 45–61%。需要注意该口径为第三方 API 路由流量，不含厂商一方产品与 C 端流量，不能外推全球；

下载日志已记录，仅供内部参考

且 token 份额不等于收入份额。Anthropic 在该平台仅占约 12-13%的 token 量、却拿走 40-50%以上的收入，开源 API 单价低一至两个数量级，份额证据只能证明漏斗顶层足够大，不能证明收入同比例放大。

图 4：OpenRouter 平台中国模型 Token 份额跃升



数据来源：东北证券

注：该口径为第三方 API 路由流量，不含厂商一方产品与 C 端流量，不能外推全球，且 token 份额不等于收入份额。

阿里 Qwen 的耦合路径是“开源聚生态、生态喂云”：Qwen 系列全球下载量超 30 亿次、衍生模型超 10 万个（截至 2025 年 4 月口径，公司披露），构成全球最大开源模型家族；开发者基于 Qwen 构建应用时，天然调用阿里云百炼的算力，2026 年年中百炼 API 日均调用量达 12 万亿 Token，最终沉淀为财报上的 AI 相关产品收入。腾讯混元 Hy3 于 2026 年 7 月以商业友好度最高的 Apache 2.0 协议开源，逻辑同构；百度则采取“基础版开源+专业版闭源”的双轨策略。开源还承担定价锚的功能：当同档性能可免费自部署，闭源 API 的溢价空间被持续压缩，这是国内 API 均价长期低于海外的结构性原因。小结：开源在国内仍不直接产生收入，但它直接决定收入归谁、以什么价格归谁。这是它与 Meta 式开源的本质区别。

2.5. 政企私有化部署：国内特有的第五条路径

这里需要先澄清一个容易混淆的概念：**海外的 B 端收入不是私有化部署。**Anthropic 约 75-85%的 ARR 来自企业 API，OpenAI 企业收入占比超 40%，但这些企业需求全部经由公有云 API、SaaS 席位与云平台分销（AWS Bedrock、Azure Foundry）承接。模型推理跑在厂商自己的算力上，按 token 或按席位持续计费，厂商承担算力成本。Anthropic 并不向企业交付前沿模型的本地部署。

下载日志已记录，仅供内部参考

而国内政企私有化部署是完全不同的交付形态：模型授权给客户、跑在客户自己的算力上，收入形态是项目制的授权费加实施服务费。两者的差异可以沿六个维度展开：

1）交付形态：海外是企业调用公有云 API，国内政企是模型进内网、本地部署；
2）收入确认：海外为持续性按量收入（ARR 属性强），国内私有化为项目制一次性确认加维保（ARR 属性弱）；
3）算力成本承担：海外 API 模式下厂商背负推理成本，私有化部署下算力由客户自担，模型厂商的成本包袱更轻；
4）毛利结构：智谱本地部署业务毛利率达 59%（招股书口径），高于多数公有云 API 业务，但代价是人力密集、实施成本随项目数线性增长；
5）需求根源：海外企业接受公有云与数据托管，国内政企受数据不出内网、信创合规与采购制度约束，公有云 API 再便宜也无法承接这部分需求；
6）可扩展性：API 模式零边际扩张，私有化部署规模天花板受制于交付人力。

海外并非没有企业级本地部署的软件生意（Palantir、IBM 等），但头部大模型厂商基本不开展这一业务。对它们而言，私有化部署会同时放弃用量收入的复利属性和推理规模的成本摊薄。国内则相反：2025 年以来，政务、金融、能源、制造业的大模型私有化部署需求集中释放，央国企与地方政府密集出台的数字化采购指南直接催生了对国产大模型的规模化采购。智谱曾是这一路径的最具代表性的受益者：按招股书口径，其收入中本地部署约占七成、毛利率 59%，覆盖政务与金融等高价值行业。但 2026 年中期业绩显示结构已经反转：上半年开放平台及 API 收入 8.25 亿元、占比 86.5%，本地部署收入降至 1.29 亿元（同比-20.5%）、占比仅 13.5%。从项目制向用量制的迁移在半年内基本完成，速度快于市场预期。需要强调的是，反转的是智谱的收入结构而非政企市场本身：私有化部署在两国都存在，但在国内能够形成规模化的独立市场。这是数据主权与采购制度的结构差异造成的，智谱本地部署收入的绝对规模（年化约 2.6 亿元）仍在，只是其增长引擎已经切换。因此准确的表述是：**政企私有化是国内特有的变现路径，但头部厂商不会停在这条路径上。**

3. 创业公司四样本：DeepSeek、Kimi、智谱、MiniMax

本章核心结论：四家头部创业公司走出了四条截然不同的路线。DeepSeek 的开源+API、Kimi 的 API 驱动+出海、智谱的政企私有化+MaaS 转型、MiniMax 的生而全球化。四家的共同点在于：B 端/API 收入均已跑赢 C 端订阅，与海外“当前阶段 B 端>C 端”的判断一致；差异在于单位经济的兑现程度。DeepSeek 的 API 毛利率已达 82.9%，智谱的公开财报则显示研发军备竞赛仍在吞噬全部收入。对核心问题的回答：四个样本共同证明**大模型的商业模式在 API 层面已经成立，但成立的位置各不相同。开源标准、旗舰定价、政企、出海，各占一个节点。**

3.1. DeepSeek：开源+API 的第三条路

DeepSeek 在全球 AI 版图中的独特性在于，它同时是开源派与商业派：模型权重全部以 MIT 协议开放，自营 API 按 token 收费且单价层长期为全球最低档；峰谷定价后旗舰档上移、轻量档仍处全球最低档梯队。这条“第三条路”由三段构成：开源吸引信任（企业可直接验证性能，决策成本与切换成本双降）、API 承接流量（无自部署算力的客户回到官方 API）、生态反哺（硅基流动、阿里百炼等算力平台围绕 DeepSeek 建模，验证其质量并强化品牌）。值得注意的是它的商业化克制：在多模态、语音、视频等更容易被普通用户感知的产品形态上，DeepSeek 并未像其他公司那样冲到最前面（首个 API 视觉模型 V4-Flash-Vision-Exp 直至 2026 年 8 月 21 日才上线），而是把资源集中在模型层与推理成本优化上。它要的不是应用层账单，而是基础设施层的议价权。

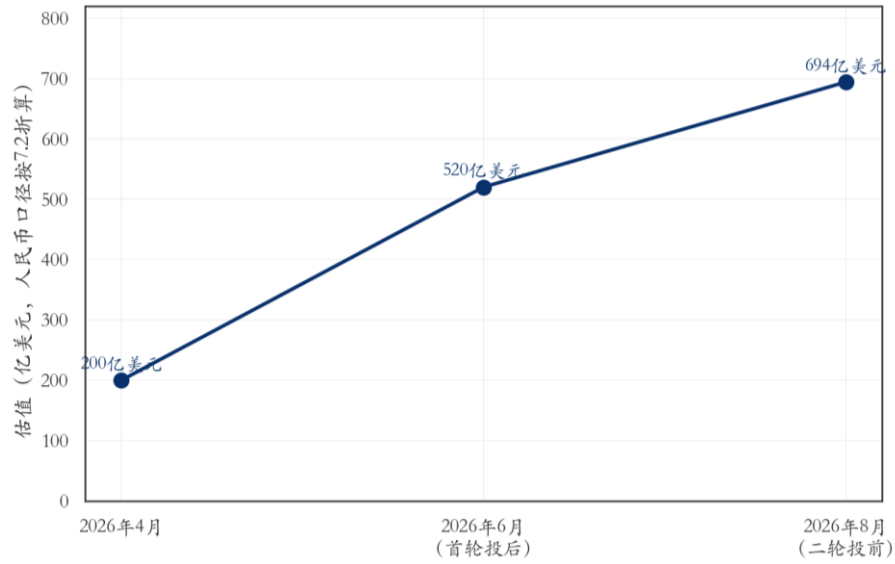
下载日志已记录，仅供内部参考

成本控制是这条路线的根基。据相关报道，DeepSeek 以 OpenAI o1 模型 3-5%的成本达到同等性能，V3 通过 MoE 架构、MLA 注意力机制与低精度训练等手段大幅压缩训练开销。8 月 13 日公司无发布会上线 V4 Pro-0813 版本，Terminal Bench 2.1 从 72.1 分跳升至 87.9 分、DeepSWE 从 12.8 分升至 62.7 分，编程能力大幅跃升而价格不变；8 月 17 日起引入峰谷分时定价，高峰时段（9:00-12:00、14:00-18:00）价格为空闲时段两倍。把电力市场的负荷定价引入推理市场，是 DeepSeek 对高峰算力紧张的直接回应。9 月 10 日，公司发布 V4.1 Flash 并 MIT 开源：通过 FP4 KV Cache 重构与跨层注意力复用，KV 缓存降至每 token 890 字节、约为前代的 1/4，官方称性能、费用、速度全面超越 V4 Pro，并将 V4 Pro 请求全部路由至 V4.1 Flash 按更低单价计费；同步下调闲时价格（输入 1 元/输出 4 元、缓存命中低至 0.02 元/百万 Token）。旗舰流量主动切向轻量档，是“以效率换规模”定价哲学的最激进实践。

财务上，DeepSeek 未披露任何经审计数据，当前可得的画像是三次媒体递进的知情人士报道（公司均未置评，本报告按低置信度数据处理）：The Information 7 月 14 日报道其年化营收接近 5 亿美元；8 月 25 日进一步报道，2026 年前 7 个月营收约 4.75 亿元（约 7070 万美元），约为 2025 年全年营收的十倍（2025 年全年约 4750 万元系“十倍”反推口径，与研究机构 Sacra 估算约 700 万美元，量级互洽）；前 7 个月净亏损约 7.15 亿元，已达 2025 年全年亏损（9.35 亿元）的约 76%；整体毛利率 44.6%，其中 API 对外模型调用服务毛利率 82.9%；同期 AI 基础设施支出约 110 亿元（2025 年全年约 12 亿元）。9 月 8 日，FT 进一步报道其 8 月 ARR 约达 5 亿美元、AI 基础设施累计投入约 16 亿美元。

资本层面，DeepSeek 在 2026 年完成了中国 AI 初创史上罕见的估值跃迁：4 月 The Information 报道估值从 100 亿美元跳至 200 亿美元；6 月完成首轮 500 亿元融资、投后估值超 3500 亿元，首轮意向资金超过 1000 亿元，投资方包括国家人工智能产业投资基金、腾讯（出资 100 亿元）、宁德时代（出资 50 亿元）、网易、京东、IDG 资本等，创始人梁文锋个人出资 200 亿元；8 月重启第二轮 500 亿元融资、投前估值约 5000 亿元，据《财经》报道于 8 月下旬完成签约；9 月上市进程进一步实质化：路透社 9 月 9 日报道公司已聘请中信证券筹备科创板 IPO、计划 2026 年内启动上市流程，21 世纪经济报道确认中信证券已进场尽调（尚未签署正式辅导协议）。

图 4：DeepSeek 估值跃迁（2026.04-08）



数据来源：The Information、新浪财经、《财经》等报道、东北证券

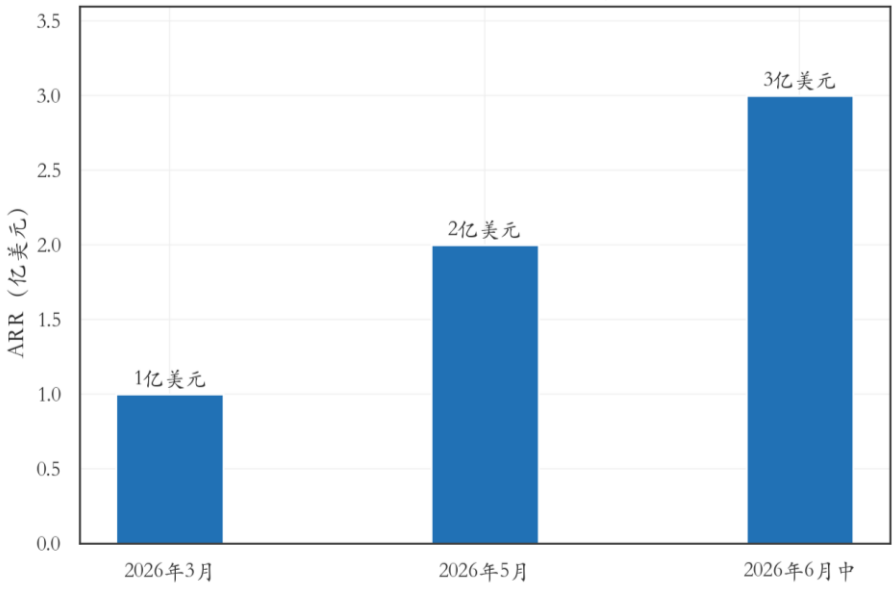
下载日志已记录，仅供内部参考

此外，DeepSeek 计划在内蒙古乌兰察布通过“自建+租赁”增加 1GW 算力；彭博 9 月 4 日报道其向华为订购约 16 万颗昇腾 950DT 芯片（估算约 25.6 亿美元）用于该数据中心。开源者也开始自建基础设施，说明其瓶颈已从需求侧转向供给侧。风险点同样清晰：开源让渡了定价权，峰谷提价能否被企业级客户消化尚待验证；训练投入随代际重置，即便毛利率 82.9%的 API 业务也无法单独供养。

3.2. Kimi：API 驱动+出海，增长引擎最像 Anthropic

月之暗面的商业化轨迹呈现出与 Anthropic 早期相似的特征：开发者调用放量、API 占比提升、海外付费用户增长，并由模型能力迭代带动价格体系上移。公司 ARR 从 2026 年 3 月初突破 1 亿美元、5 月超 2 亿美元、6 月中旬站稳 3 亿美元，三个月翻三倍；K3 模型发布后企业 ARR 进一步数倍增长（公司总裁张予彤 7 月 18 日在社交媒体披露，公司口径）；彭博 9 月 11 日报道其 8 月 ARR 已突破 10 亿美元，内部目标 2026 年底年化收入 20 亿美元。若兑现，Kimi 将用一年时间走完 Anthropic 两年的 ARR 路径。收入结构上，API 收入已占整体收入七成以上并持续提升。产品纪律上，公司明确表示仅布局生产力属性的 AI 应用、重点发力 Coding，不涉足视频生成等娱乐赛道。Kimi 的月活一度腰斩而收入翻倍，说明其增长已从投流获客切换到付费调用驱动。

图 5：月之暗面 Kimi ARR（2026.03-06）



数据来源：彭博社、上海证券报、东北证券

Kimi K3 是其定价权的支点：2.8 万亿参数的开放权重模型（全球首个落地的 3 万亿参数级开源模型，7 月 27 日完整权重开源），在 Artificial Analysis 综合智能指数（v4.1.1 测试版本）中取得 57 分、与 GLM-5.3 并列开源第一；定价上输入 20 元/输出 100 元/缓存命中 2 元每百万 Token，是当前最贵国产模型，较上一代 K2.6 提价约 2.1 倍/2.7 倍。官方口径下编程场景缓存率超 90%、实际输入成本可降至标准价的约四分之一。更早的 K2.7 Code 于 7 月 3 日接入 GitHub Copilot，成为该平台首个纳入的开源权重模型。

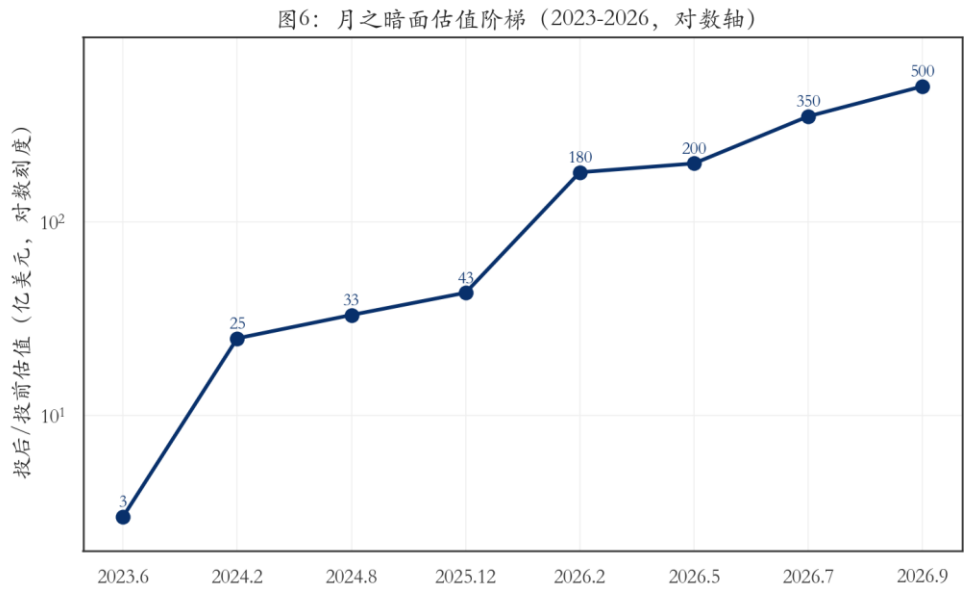
需求旺盛的另一面是供给瓶颈：因 K3 调用激增导致算力告急，公司于 7 月 20 日暂停 C 端新用户订阅，成为算力紧缺时代国内厂商“以限售保体验”的标志性事件。商业化配套也在铺开：7 月与美国运通、农业银行合作发行 Kimi 信用卡；9 月 10 日启动企业合作伙伴“登月计划”（FDE 模式，首批包括中软国际、金山云等五家）；就 K3 进入美国三大云（微软/亚马逊/谷歌）谈判收入分成（据报道初期寻求最高 30%

下载日志已记录，仅供内部参考

分成）；K3 每月在 OpenRouter 上生成约 3000 亿 token，Cognition 于 9 月发布的编程模型 SWE-2 以 K3 为后训练基座，开源模型的生态反哺开始兑现。9 月 11 日全量上线 K2.8 Preview，将 1M 上下文下放至全部会员档位。

资本层面，根据多家媒体报道，公司估值呈阶梯式跳升：2023 年 4 月成立两个月即完成超 2 亿美元天使轮（估值 3 亿美元）；2024 年 2 月 A+轮超 10 亿美元（阿里领投，估值 25 亿美元）、8 月 B 轮后 33 亿美元；2025 年 12 月 C 轮后 43 亿美元；2026 年 1-2 月连续三轮、估值升至 180 亿美元；5 月 D 轮约 20 亿美元后突破 200 亿美元；7 月 29 日 F 轮超 35 亿美元官宣，认购金额超出原定目标 3 倍、融资提前关闭，投后估值 350 亿美元；原定 8 月启动的 Pre-IPO 轮（G 轮）连夜提前、投前估值锚定 500 亿美元，参与方需在 8 月 15 日前完成打款，直接进入股东名册要求机构管理规模 5 亿美元以上（科创板日报报道）；8 月下旬，公司已完成股份制改造（杨植麟任董事长），正寻求监管批准赴港上市，最快 6 个月内登陆港股。

图 6：月之暗面估值阶梯（2023-2026，对数轴）



数据来源：彭博等媒体报道、东北证券

关于“谁最像 Anthropic”需要分层回答。市场上“中国版 Anthropic”的标签多贴在智谱身上，依据是客户结构：两者都是 B 端企业级收入绝对主导、C 端克制。但从增长引擎的形态看，Kimi 与 Anthropic 的同构性更强：收入以公有云 API 按量计费为引擎（占比均超七成）、由开发者 PLG 驱动、以 Coding 旗舰产品（K2.7 Code / Claude Code）打开企业市场。主要差异在于 Anthropic 闭源而 Kimi 开放权重。Kimi 更像是“开源版的 Anthropic 增长模型”。需要提示的是估值透支风险：G 轮投前 500 亿美元对应 ARR 3 亿美元（6 月中口径），PS 倍数超过 160 倍，远高于 Anthropic 同期水平，估值兑现高度依赖 K3 发布后收入斜率的持续验证。下一节将说明，智谱与 Anthropic 的差异恰恰在收入形态上。

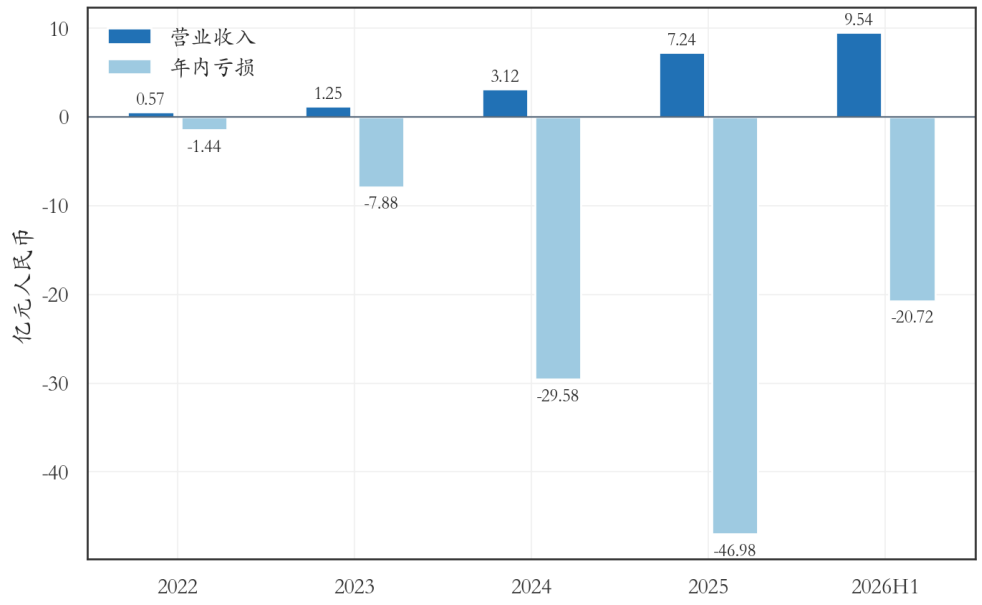
下载日志已记录，仅供内部参考

3.3. 智谱：MaaS 转型与成本结构

智谱（2513.HK）于 2026 年 1 月 8 日在港交所上市，发行价 116.20 港元、募资总额约 43 亿港元，超额配股权行使后净额约 48.96 亿港元，为全球首家以通用大模型为核心业务的上市公司。招股书显示，智谱拥有 270 万+开发者、服务全球 1.2 万家企业，研发人员占比 74%，募集资金净额的 70%（约 29 亿港元）投向通用大模型研发。弗若斯特沙利文数据显示，按 2024 年收入计算，智谱在中国独立通用大模型开发商中排名第一。

收入端，2022-2025 年营收分别为 0.57 亿、1.25 亿、3.12 亿、7.24 亿元，连续三年翻倍（2025 年同比+131.9%）；2026 年上半年收入 9.54 亿元、同比增长 399.7%，半年即超 2025 年全年。其中开放平台及 API 收入 8.25 亿元、同比增长 2735.7%，占比达 86.5%。MaaS 取代本地部署成为绝对主业；MaaS 注册用户超 740 万（较年初+144%），Token 调用量较年初增长 40 倍，API 均价提升 101%，量价齐升。利润端，2022-2025 年年内亏损分别为 1.44 亿、7.88 亿、29.58 亿、46.98 亿元；2026 年上半年期内亏损 20.72 亿元、经调整净亏损 19.64 亿元，研发投入 21.31 亿元（+33.6%，为收入的 2.2 倍）。研发军备竞赛吞噬全部收入的结构未变，但研发占收入比已从 2024 年 702.7%的峰值降至 223%，规模效应开始在费用率上显现。毛利率结构同步切换：综合毛利率 26.4%（同比下降 23.6 个百分点，因低毛利的 API 占比骤升），而 API 业务毛利率由-0.4%转正至 24.6%。Token 单位经济第一次在公司报表上得到验证。

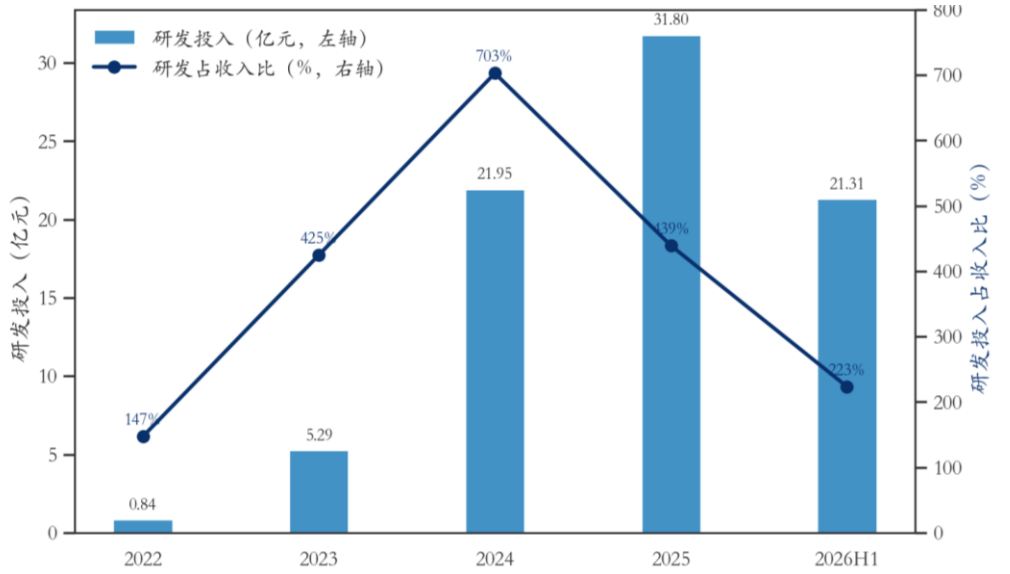
图 7：智谱营业收入与亏损（2022-2026H1）



数据来源：Wind、东北证券

下载日志已记录，仅供内部参考

图 8：智谱研发投入与研发占收入比（2022-2026H1）



数据来源：公司招股书、定期报告、东北证券

智谱 2026 年最重要的边际变化是定价权的验证：2 月 GLM-5 发布后公司对 MaaS 平台结构性提价 67-100%，提价后调用量仍增长超 400%，“涨价—放量—收入预期抬升—估值抬升—支撑下一代研发”的正向闭环初步形成。此外，企业级智能体收入从 4749 万元增至 1.66 亿元（+248.8%）、毛利率 52.3%，Agent 产品 AutoClaw 上线 20 天订阅突破 40 万，公司年报分部数据显示，企业级智能体为增速最快、利润最稳的业务线；公司年报还新增了按业务形态的收入分类（开放平台及 API、企业级智能体、企业级通用大模型、技术服务），官方解释为“原有分类已无法涵盖日益丰富的产品形态与商业场景”。即从项目制公司向平台型公司迁移的信号。

作为上市公司，智谱是观察“二级市场如何给大模型定价”的样本：上市首日收涨 11.7%，此后股价最高触及 2980 港元，市值一度突破万亿港元，成为“中国首个万亿 AI 公司”；但 8 月 18 日智谱与 MiniMax 双双重挫逾 13%，至 9 月中旬市值约 5200 亿港元，距峰值回撤仍超四成。一个值得记录的倒挂是：智谱二级市值（约 670 亿美元）已低于月之暗面一级市场 G 轮投前估值（500 亿美元）附近的可比水平。公开市场与一级市场对同赛道公司的定价出现明显分歧。回撤的导火索是 GLM-5.3 被大和证券评价为“仅为 GLM-5.2 的迭代优化、并非更大参数旗舰”，叠加板块获利回吐。高波动说明二级市场对“收入兑现 vs 研发投入”的定价尚无锚。公司资本运作同步推进：7 月 13 日完成 1978 万股新 H 股配售、募资约 314.11 亿港元，管理层明确表示两年内不追求变现；A 股科创板 IPO 已进入辅导验收阶段（据媒体报道拟募资 150 亿元）。

三季度以来的三组数据进一步强化了智谱的样本价值。其一，ARR 加速：7 月 ARR 达 10 亿美元、提前完成全年目标；8 月底 ARR 达 16 亿美元。从 1 亿到 16 亿美元仅用约七个月。其二，成本曲线：8 月 26 日开源的 GLM-5.3-Flash（即匿名登顶 OpenRouter 的“Ox Alpha/牛来”）全部线上流量由约 10 万张国产芯片承载，公司称单位 Token 推理成本较年初下降 80%、单位算力对应收入提升 14 倍，定价仅为旗舰的 1/10。其三，资本运作：7 月 13 日配售募资约 314.11 亿港元后，9 月 11 日再以 714 港元/股配售约 20 亿美元并发行 30 亿美元零息可转债（转股价 892.5 港元）。两个月合计募资约 393 亿港元，但第二次配售价较 7 月的 1588 港元接近“腰斩”；科创板 IPO 辅导状态已为“辅导工作完成”（财新报道）。

回应上一节的问题：智谱与 Anthropic 相似的是客户结构，B 端政企主导、不依赖 C 端订阅；不同的是收入形态。Anthropic 的企业收入是公有云 API 的持续按量收入

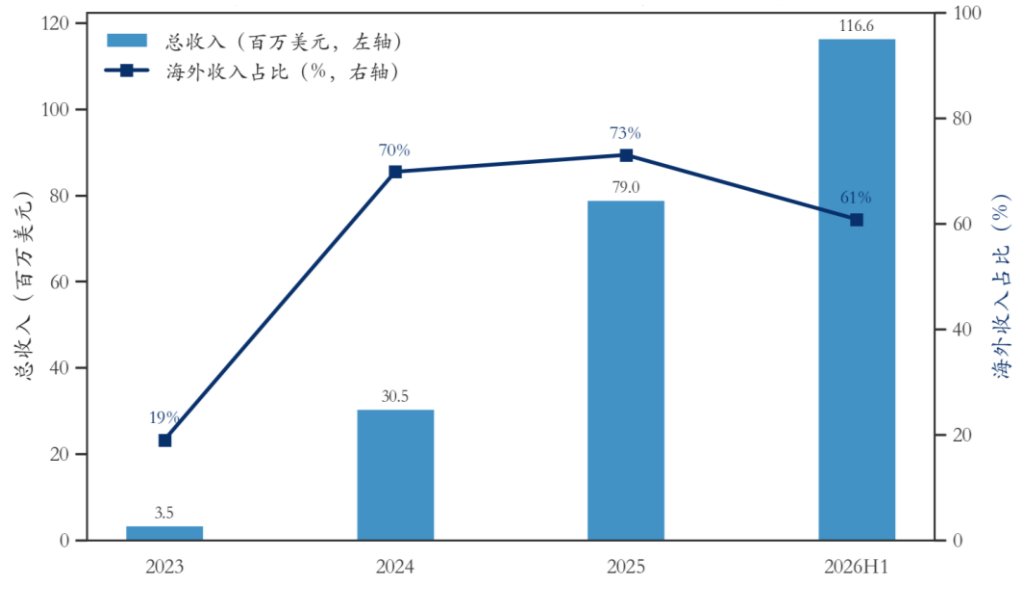
下载日志已记录，仅供内部参考

（ARR 复利属性强、零边际扩张），智谱的七成收入是私有化部署的项目制收入（毛利率高但 ARR 属性弱、人力密集）。因此“中国版 Anthropic”的标签对智谱只贴对了一半：它的客户像 Anthropic，它的商业模式更像“AI 时代的传统软件集成商+MaaS 平台”的混合体。7 月 ARR 10 亿美元的构成中 MaaS 占比的提升速度，是智谱从后者走向前者的关键变量。

3.4. MiniMax：海外收入六成的生而全球化，B 端切换已兑现

MiniMax 证明了国内大模型公司可以绕过国内内卷、直接在全球市场变现。2026 年中期业绩显示，公司上半年总收入 1.166 亿美元、同比增长 283.1%，半年即超过 2025 年全年（7903.8 万美元）；其中 B 端开放平台及企业服务收入 7390 万美元、同比增 703.1%，占比从 30.3%升至 63.4%成为第一大收入来源；C 端 AI 原生产品收入 4260 万美元、同比增 100.9%。海外收入 7083 万美元、占比 60.8%。海外占比从 2025 年的 73%回落，原因是国内收入以 434%的更高增速放量，而非海外收缩。截至 6 月末现金结余 13.23 亿美元。

图 9：MiniMax 总收入与海外收入占比（2023-2026H1）



数据来源：公司招股书、定期报告、东北证券

业务结构的切换在 ARR 口径上更为彻底，根据公司公开业绩会披露，8 月 ARR 已超过 8 亿美元，其中 To B 业务占比约 80%、To C 约 20%，与上年同期 B 端仅约 30%形成反转；7 月平台 Token 消耗量已达 1 月的 20 倍，企业客户与开发者数量突破 200 万、为 2025 年底的 10 倍。模型侧，M3 保持价格不变而智能水平提升；业绩会预告近期推出 M3.1（验证重建后基础设施的稳定性与 Agent 泛化）、为约 3T 参数的 M3 Pro 铺路，视频模型 H3.1 将接续 H3，大规模国产算力集群“很快上线”；H3 视频模型开源三周下载量超 2400 万次、产生超 300 个公开衍生模型，为 2026 年全球下载量最高的模型。开源生态开始外溢：沙特 HUMAIN 于 9 月基于 M3 发布阿拉伯语模型 humain-m3，Genspark 基于 M3 开放权重后训练发布 PPT 专用模型 Gen-1 Slides。这组数据修正了市场此前“MiniMax 的 C 端基因过重”的担忧，B 端切换已经完成。

公司 2026 年上半年经调整净亏损从 1.387 亿美元扩大至 2.93 亿美元，主因研发开支同比增 138.8%至 2.969 亿美元（为同期收入的 2.55 倍）。MiniMax 正处于“以利润换模型代际”的投入期。积极信号是费用效率：销售及分销开支同比下降 17.9%至 2700 万美元，而收入接近四倍增长，增长已不依赖买量。

下载日志已记录，仅供内部参考

风险点在于投入产出的时间差。上半年期内亏损 3.58 亿美元（同比收窄 11%），收窄主要来自金融负债公允价值变动的基数效应，经调整口径亏损仍在扩大；M3 Pro 与 H3 的下一代训练意味着研发开支的绝对额仍将上行。公司创始人闫俊杰在业绩会上给出的解题思路是“Minimize the Cost, Maximize the Intelligence”。只有单位智能成本持续下降，更高水平的智能才能进入更广泛的生产场景。该逻辑能否兑现，取决于 M3 系列的性价比能否在 Kimi K3 与 GLM-5.3-Flash 的两头挤压下守住开发者基本盘。

3.5. 路径分化与出清

并非所有人都能留在通用大模型的牌桌上。曾经的“AI 大模型六小虎”中，留在通用模型牌桌上的还剩四家：月之暗面、智谱、MiniMax，以及 8 月完成近 25 亿美元 Pre-IPO 轮融资、已拆除红筹并完成股改的阶跃星辰。9 月 20 日其发布 Step 5 Preview（600B/27B），AA 智能指数 44 分跻身开源前三，官方称单任务成本为 Opus 5 的 1/8，权重将于 10 月 15 日开源。百川智能已放弃通用大模型赛道，裁撤金融、教育等 B 端团队，全力押注 AI 医疗，其医疗大模型 Baichuan-M3 在权威评测中以 65.1 分位列全球第一，但代价是退出一线竞争；零一万物收缩业务，转向轻量化行业定制模型。在资本开支与训练成本持续上行的行业里，没有清晰变现路径的公司最先出清。

当前四家头部大模型厂商各自走出了自己的商业化路径：DeepSeek 占据“开源标准位”，成本与生态是壁垒，API 毛利率 82.9%证明了单位经济，亏损/收入比显著改善但未扭亏；Kimi 占据“API+出海位”，增长引擎最接近 Anthropic，估值透支风险最高（G 轮对应 PS 超 160 倍）；智谱占据“政企位”，收入可审计、ARR 斜率最陡（8 月 16 亿美元），项目制向用量制的迁移已基本完成（API 占 86.5%），新看点是 API 毛利率转正后的爬坡斜率；MiniMax 占据“海外 C+B 位”，人效与现金储备最好，B 端切换已经兑现（ARR 的 To B 占比 80%），新变量是 M3 系列的性价比防守。四条路径没有对错之分，部分证伪信号已列入表 4 的跟踪框架。此外值得跟踪的是一二级市场的定价倒挂，智谱二级市值与月之暗面一级估值的相对位置，将影响后续所有 IPO 的定价锚。

4. 大厂生态：模型为杠杆的中国版本

本章核心结论：**国内大厂与 Google、Meta 同属“模型为杠杆、靠生态赚钱”的阵营，但兑现路径更偏云基础设施与交易佣金**：模型能力经由云（IaaS/MaaS）和 AI 助手促成的电商、本地生活交易抽佣兑现，与海外以广告为主的间接兑现路径不同。阿里是目前唯一披露 AI 分部规模盈利数据的大厂，云与算力分部季度收入同比+45%、经调整 EBITA 利润率 12%；字节以 2000 亿元以上的年度资本开支换取大模型 ARR 约 40 亿美元的规模第一；腾讯、百度的模型价值仍主要经由存量产品间接体现。资金闭环方面，国内大厂同样进入 CapEx 挤压 FCF 的阶段，而创业公司的供粮渠道是一级市场+港股 IPO，与海外的私募巨头+云厂商绑定结构不同。对核心问题的回答：**在大厂环节，利润经由云与生态兑现。模型是杠杆而非商品，云生态位是三个利润节点中最厚的一个。**

4.1. 阿里：AI 的两类业务，云已盈利、模型仍在投入

阿里巴巴 2027 财年第一季度（截至 2026 年 6 月 30 日）财报首次完整呈现了国内大厂 AI 业务的两层结构。第一层是“卖算力”：AI 云与算力分部季度收入 484.37 亿元、同比增长 45%，经调整 EBITA 同比增长 133%至 56.28 亿元，利润率从 9.1%升至 12%。重资产业务已经盈利且利润率在抬升。第二层是“卖模型”：AI 相关产品季度收入 123.76 亿元、连续第十二个季度三位数同比增长，占云外部商业化收入比

下载日志已记录，仅供内部参考

重首次突破 30%，管理层预计未来一年将突破 50%。CEO 吴泳铭 8 月披露，模型与应用服务 ARR 已突破 160 亿元，其中大部分来自自研模型，且自研模型与托管开源第三方模型的毛利水平接近。但包含千问 App 的 AI 实验室与应用分部仍处投入期：分部收入 33.38 亿元(+16%)，经调整 EBITA 亏损 138.61 亿元、同比扩大 330%，公司称系 AI 能力投入与千问 App 相关推理成本增加所致。

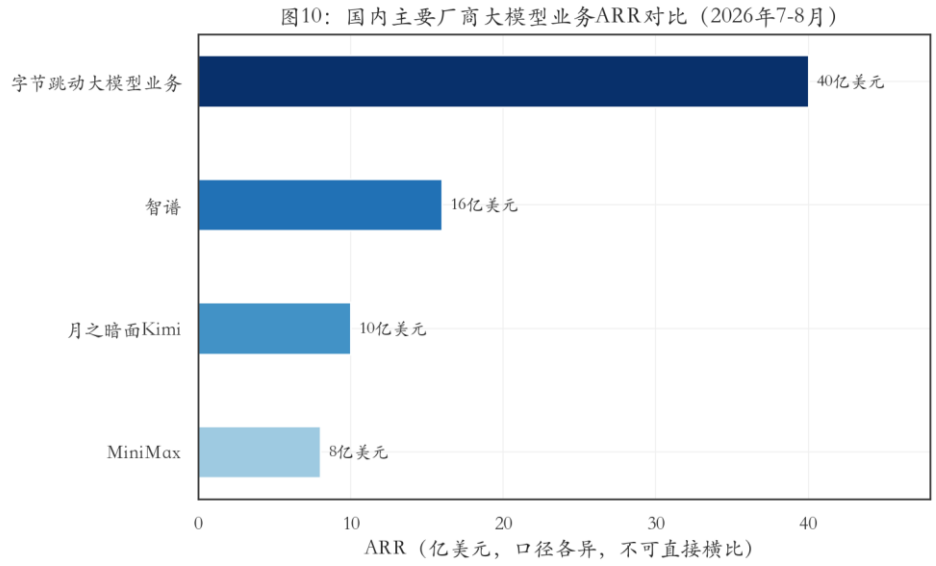
产品侧，阿里 8 月完成了旗舰与轻量两档的同时布局：Qwen3.8-Max（2.4 万亿参数 MoE）国内定价输入 12 元/输出 36 元，为阿里系首次进入旗舰价格带；Qwen3.8-Flash（125B 激活 6B）8 月 26 日发布、次日即降价至输入 0.8 元/输出 2.7 元，训练成本仅为 Qwen3.7-Plus 的 1/9，首发上线“千问办公”生产力平台；9 月 9 日再发全模态的 Qwen3.8-Omni-Flash，并在云峰会发布集成 150 余款模型 API 的 MaaS 平台“千问云”。阿里的结构验证了吴泳铭 2025 年 2 月的判断：“基座大模型商业模式尚不清晰，卖云价值明确”。模型层的价值经由开源生态（Qwen 全球下载超 30 亿次）与百炼 MaaS（API 日均调用 12 万亿 Token）转化为云的收入与利润。

投入端同步加码：本季度资本开支 676.78 亿元（+75%），2025 年 2 月宣布的三年 3800 亿 AI 基建计划已累计投入约 1900 亿元；8 月 23 日公司宣布配售 7.1 亿新股、总额 800 亿港元（港股史上最大后续发行，1 小时内近 3 倍超额认购），净额 100% 投入全栈 AI 能力与 AI 基础设施。管理层基于当前 AI 产品平均毛利水平给出“AI 算力资本开支三年回本”的测算。

4.2. 字节跳动：2000 亿资本开支换规模第一

字节是国内投入强度最高、也是 ARR 规模最大的厂商。官方披露，截至 2026 年 6 月豆包大模型日均 Token 调用量达 180 万亿、过去一年增长超 10 倍；中国公有云 MaaS 市场份额达 49.5%，位居第一；火山方舟企业及个人用户超 110 万。据每日经济新闻按 7 月平均消耗折算，字节大模型业务 ARR 约 40 亿美元，超过当时国内其他模型公司 ARR 总和。代价是投入：据财新报道，字节 2026 年 AI 预算近 1600 亿元，过半用于 AI 芯片采购，另有约 500 亿元算力基础设施投入，并已获得 296 亿美元（约 1987 亿元）三年期贷款。国内最大规模之一的 AI 投入开始引入债权融资工具。

图 10：国内主要厂商大模型业务 ARR 对比（2026 年 7-8 月）



数据来源：每日经济新闻（字节）、公司业绩会（智谱、MiniMax）、彭博（Kimi）、东北证券
注：各公司口径及时间窗口不同，仅可作为量级参照

下载日志已记录，仅供内部参考

字节的商业化难题在 C 端，其 2026 年的解法是组织重组与交易变现两个动作。组织上，豆包、飞书、火山引擎的“三合一”重组于 9 月 15 日完成：飞书产品团队并入豆包、GTM 团队并入火山引擎；飞书发布 8.0（Agent 重构，开放接口从 247 个扩至 767 个），2026 年以来新增客户超九成同步采购飞书 AI 产品，上半年 ARR 增速为去年同期 2.5 倍，显示 AI 能力已成为驱动企业采购、提升单客价值的核心变量。模型层，6 月 23 日发布的豆包 Seed 2.1 Pro 定价输入 6 元/输出 30 元/缓存命中 1.2 元每百万 Token，进入旗舰价格带，9 月 16 日全量上线 0915 版本，强化 Agent 任务交付与多模态 Coding。值得注意的是豆包大模型 2.2 的推迟：原计划 8 月推出，因加强预训练/后训练（重点补 Coding 与 Agent）延期，内部目标年底前在 Coding 上产生类似 GLM-5.2、Kimi K3 的行业冲击力。这与公司 CEO 梁汝波“接受短期落后、坚持自研”的表态互为注脚。火山引擎 MaaS 的 2026 年营收目标则从 100 亿元上调至 150 亿元，视频模型 Seedance 单月收入已超 10 亿元。交易上，8 月 10 日豆包 App 全量上线原生音视频全双工模型 SeedRealtime，8 月 11 日在酒店订单等场景按成交额抽取 12%佣金，从单一 AI 助手向交易撮合平台转型。战略方向上，梁汝波在全员会上将豆包定调为集团未来的主干业务，同时坦承大语言模型被海外领先模型拉大了差距。投入强度与技术差距并存，是字节模式当前的真实状态。

4.3. 腾讯&百度：生态嵌入与全栈布局

腾讯的大模型策略是“生态嵌入”：2026 年 7 月以 Apache 2.0 协议开源混元 Hy3（295B 总参数、21B 激活，输入 1 元/输出 4 元每百万 Token 走量），8 月 28 日再开源 Hy4 preview（770B 总参/49B 激活，1M 上下文，输入 6 元/输出 18 元每百万 Token 进入旗舰档），内部 203 项工程任务盲测略胜 GLM-5.3 与 Kimi K3，正式版计划 2026 年内发布；模型能力接入元宝、WorkBuddy、CodeBuddy、ima 等产品矩阵。

2026 年二季度财报首次量化了 AI 投入的利润代价：剔除新 AI 产品（混元、元宝、CodeBuddy、WorkBuddy、小微）后经营利润同比+19%，即新 AI 产品当季拖累利润约 105 亿元；资本开支 527.84 亿元（+176%），自由现金流-138 亿元（剔除 514 亿元算力采购预付款后为+376 亿元）。腾讯也已进入 CapEx 挤压 FCF 阶段。管理层将 2026 年 AI 投入定调为“翻倍以上”增长，并表示若混元达到 SOTA 水平，可构建规模可观的 Token 业务。腾讯的独特资产是微信，微信 Agent 已进入测试阶段，Agent 框架（OpenClaw 类）的爆发让“Token 调用、工具使用、任务执行、交易分成”每一环节都可能形成收入来源，这是腾讯 AI 叙事的期权所在。

百度坚持“芯（昆仑芯）—云（智能云）—模（文心）—体（智能体）”全栈布局。2026 年 5 月发布的文心 5.1 采用“多维弹性预训练”技术，公司称预训练成本仅为业界同规模模型的 6%（公司口径）；在 LMArena 竞技场中，文心 5.1 登上搜索榜国内第一、全球第四。商业化方面，文心已深度整合进搜索、网盘、文库等存量产品；2026 年二季度核心 AI 驱动业务收入 125 亿元（+25%），其中 GPU 云收入同比增长 283%，但自由现金流为-79.5 亿元。百度同样进入 CapEx 挤压 FCF 阶段。

两家公司的共同点在于：模型本身不直接追求收入最大化，而是作为存量业务的提价与留存杠杆。与 Google 的 Workspace 捆绑提价逻辑同构。

4.4. 资金闭环：大厂 CapEx 挤压与创业公司的 IPO 供粮

本系列深度报告第一篇中指出，海外生态巨头的“内生资金闭环”在 2026 年二季度已被 CapEx 压至极限，国内大厂进入同一阶段：字节 2026 年 AI 资本开支 2000 亿元以上。阿里 2026 财年第四季度（2026Q1）自由现金流转为净流出 173 亿元（上年同期为净流入 37.43 亿元），公司归因于云基础设施支出增加与千问 App 获客成

下载日志已记录，仅供内部参考

本。阿里管理层给出的对冲叙事是“三年回本”测算与云收入的高增速，但该闭环成立的条件（OCF 增速不低于 CapEx 增速）在国内同样紧张。

创业公司的资金结构与海外显著不同。海外独立实验室依赖私募巨头与云厂商的战略投资（Anthropic 与 AWS/Google 的算力换股权结构）；国内创业公司的供粮渠道是“一级市场+港股 IPO”双轨：智谱 IPO 募资净额约 48.96 亿港元并在两个月内再完成两轮合计约 393 亿港元配售及可转债、MiniMax IPO 募资约 48.2 亿港元并在 7 月再融资约 160 亿港元、月之暗面累计融资超 376 亿元并已于 9 月 2 日保密递表、DeepSeek 两轮合计千亿级融资并聘请中信证券筹备科创板。港股在 2026 年事实上承担了国内大模型行业的“公开市场定价与输血”功能，融资工具也从纯股权扩展出债权层。智谱发行 30 亿美元零息可转债、字节获得 296 亿美元贷款、阿里配售 800 亿港元，股权、可转债、贷款三层工具同时启用，说明行业资金需求已超过纯股权融资的容量。这也意味着创业公司的资金链首次暴露于二级市场情绪之下。8 月港股 AI 板块深度回撤、智谱两次配售价接近“腰斩”，即为这种暴露的压力测试。

资金约束之外，更硬的约束是算力供给。据中国信通院数据，2026 年一季度国内 AI 算力需求同比暴涨 417%、而供给增速仅 128%；4 月和 7 月，智谱、阿里、月之暗面接连因算力紧缺暂停或限量开放 C 端新用户订阅。应对方式集中指向自建基础设施：智谱已启动 1GW 级数据中心建设并部分投运、全部采用国产芯片；DeepSeek 计划在内蒙古乌兰察布以“自建+租赁”增加 1GW 算力。算力紧缺一方面压制了 C 端放量节奏，另一方面解释了峰谷定价与轻量档军备竞赛的同步出现。在供给受限的行业里，定价与效率就是产能分配工具。

5. 国内外对标：结构性异同

本章核心结论：海内外大模型商业模式的“形”正在收敛。B 端 API 跑赢 C 端订阅、Agent 驱动用量、计费向价值代理迁移，两地同向；但五个结构性差异决定了国内厂商的财务形态将长期不同于海外：API 价格差一至两个数量级、开源主导权在中国、政企私有化国内特有、出海反向变现、资本结构以港股 IPO 为枢纽。对核心问题的回答：与海外相比，国内的收入池更薄、更依赖用量与生态间接兑现，但当前“B 端>C 端、生态>独立”的方向一致。

5.1. 趋同的三条规律

第一， B 端/API 收入跑赢 C 端订阅在两个市场同时成立。
海外是 Anthropic 对 OpenAI 的 ARR 反超；国内的证据链同样充分：智谱提价 67-100%后调用量增 400%、Kimi 的 API 收入占比超七成、MiniMax 的 B 端增速（+197.8%）反超 C 端，而国内最大的 C 端流量入口豆包日收入曾不足百万元。

第二， Agent 驱动的用量放量同向。
豆包大模型日均 Token 一年增长超 10 倍、阿里百炼日均调用 12 万亿 Token、智谱 Token 调用量 6 个月 15 倍，斜率与 OpenRouter 全球周调用量一年 20 余倍的曲线同构。

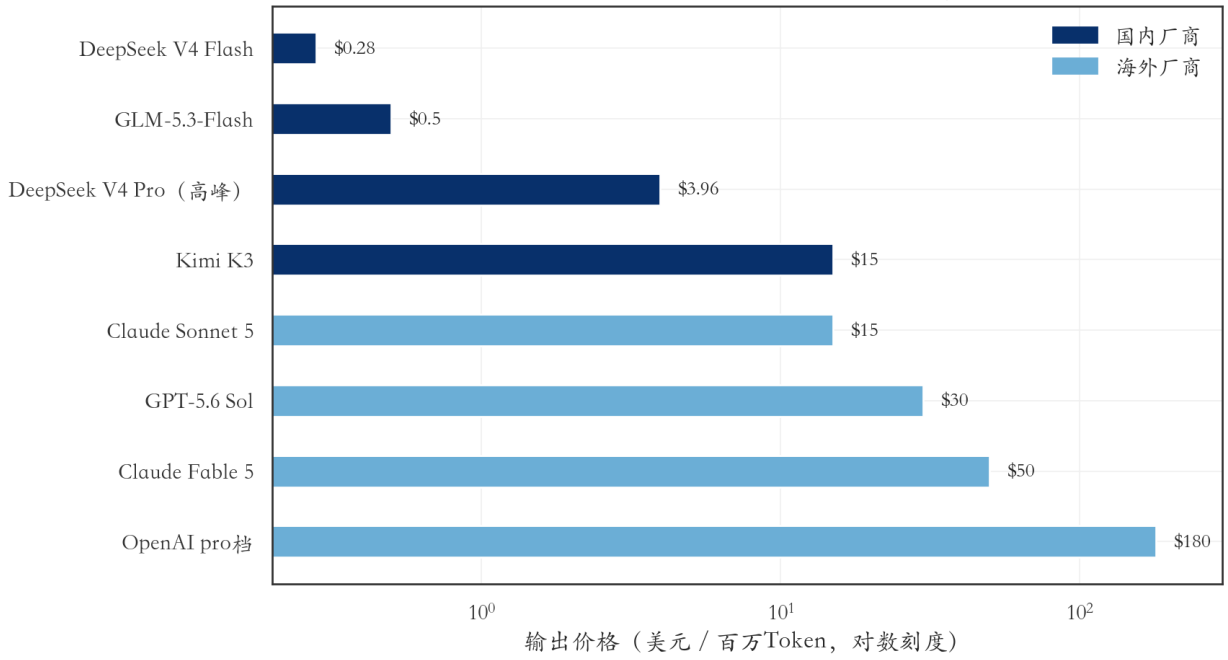
第三， 计费模式向价值代理迁移同向。
智谱推出 Coding Plan 订阅套餐、MiniMax 的 coding plan 消耗增长超 10 倍、Salesforce 式按结果计费的产品也开始在国内出现，token 作为成本代理的局限在国内同样存在。

下载日志已记录，仅供内部参考

5.2. 五个结构性差异

差异一：价格水平整体仍差一至两个数量级，但结构正在收敛。
DeepSeek V4 Pro 高峰输出价约 3.96 美元/百万 Token，海外旗舰档为 15-50 美元、OpenAI 超旗舰档达 180 美元；Kimi K3 以 15 美元输出价进入旗舰价格带是国产模型的首次，Qwen3.8-Max（6 美元）、GLM-5.3（4.4 美元）、豆包 Seed 2.1 Pro（约 4.2 美元）紧随其后。国产旗舰正在填补与海外之间的价格空档，而轻量档（GLM-5.3-Flash 0.5 美元、Qwen3.8-Flash 约 0.39 美元）继续把全球推理价格的下沿往下打。低价的后果是双向的：它压低了国内 API 的毛利天花板，也换来了全球开发者流量与开源标准的话语权。

图 21：国内外代表模型 API 输出价对比（2026 年中，对数轴）



数据来源：各厂商官方定
价（2026 年 8 月）、东北证券
注：DeepSeek V4 Pro 为其峰谷定价的高峰输出价，GLM-5.3-Flash 为常规价（限时折扣再减半）

差异二：开源主导权在中国。
海外开源（Llama、Mistral）是生态防御武器，中国开源（DeepSeek、Qwen、GLM、混元）已成为变现漏斗的顶层入口与全球标准，开源本身不收费，但它决定收入经谁的 API 与云承接。OpenRouter 上中国模型 token 份额约 45-61%、Qwen 衍生模型超 10 万个。这意味着全球大模型的“开源底座”事实上由中国厂商供给，其商业反哺经由云算力与官方 API 实现。

差异三：政企私有化部署是国内特有的变现路径。
海外头部模型厂商的企业收入经由公有云 API 承接、基本不做本地部署；国内的数据主权与合规要求创造了本地化部署市场，曾支撑智谱约七成收入。但 2026 年上半年智谱 API 收入占比已达 86.5%、本地部署降至 13.5%，说明该路径是头部厂商的“起步支架”而非“终局形态”；市场本身毛利率高（59%）但复制性弱，且采购节奏受政策周期影响。

差异四：出海是反向变现路线。
MiniMax 收入的 73%来自海外、Kimi 的海外付费与 API 收入持续增长、阿里在新加坡发布面向海外的 Qwen Cloud。中国模型公司把低价高质的能力直接卖给发达市

下载日志已记录，仅供内部参考

场，绕开国内支付意愿的约束，这与 TikTok/Shein 的消费互联网出海逻辑不同：模型出海卖的是开发者生产资料而非用户注意力。

差异五：资本结构不同。
海外是私募巨头+云厂商战略绑定（Anthropic-AWS 的算力换股权），国内是一级市场+港股 IPO 双轨，并在 2026 年三季度扩展出债权层（智谱零息可转债、字节 296 亿美元贷款）。后果是：国内创业公司的估值与资金链更早暴露于公开市场，股价波动会直接反噬融资能力；同时也更早接受披露约束。智谱与 MiniMax 的定期财报已成为行业稀缺的公开财务数据源。

表 3：海内外大模型商业模式对照（2026 年中）

维度	海外	国内
API 价格带（输出，美元/百万 Token）	旗舰 15-50，超旗舰 180	走量档约 0.3-2，Kimi K3 首次进入 15 美元旗舰档
开源地位	生态防御武器（Meta），与变现脱钩	变现漏斗的顶层入口+全球开源底座（DeepSeek/Qwen），收入仍由 API/云承接
C 端变现	订阅为主，付费率约 5%，广告第二曲线启动	订阅刚起步，支付意愿更低；交易抽佣探索中
政企私有化	基本不存在（经云 API 承接）	独立收入支柱（智谱约七成收入）
出海	美国厂商天然全球	模型公司主动出海（MiniMax 海外 73%）
创业公司资金来源	私募巨头+云厂商战略绑定	一级市场+港股 IPO 双轨

数据来源：各公司公告与公开报道，东北证券

6. 未来发展探讨：盈利路径与格局推演

本章核心结论：以智谱的公开财报为锚做三情景测算，基准情景下其盈亏平衡指向 2029 年前后；格局推演上，本报告给出五个情景而非单一基线。双集中为骨架，算力约束与 Agent 应用层决定利润分配，融资窗口决定尾部出清节奏，地缘变量决定出海天花板，各情景均附证伪信号。对核心问题的回答：**利润拐点在 2029 年前后（基准情景），利润归属取决于五个情景中哪些兑现。但三个利润节点的集中趋势在各情景下均成立。**

6.1. 盈利路径测算：以智谱为锚

选择智谱的原因：它是国内最早披露完整经审计财务报表的独立大模型公司，其余厂商（DeepSeek、Kimi、字节模型业务）的收入数字均为媒体折算、知情人士透露或公司选择性披露，口径不足以支撑测算。

测算基数：2025 年收入 7.24 亿元、年内亏损 46.98 亿元；2026 年上半年收入 9.54 亿元（+399.7%，中报口径）；2026 年 8 月底 ARR 达 16 亿美元、最近一周年化超 20 亿美元（业绩会口径，含 MaaS 年化与在手订单的混合口径，确认收入将低于 ARR）；综合毛利率 26.4%、API 毛利率 24.6%转正（中报口径）；上半年研发投入 21.31 亿元；7 月与 9 月完成两轮合计约 393 亿港元配售及可转债。

乐观情景：收入 2026-2028 年约 75/150/240 亿元。2026 年的逐月推导：H1 实际 9.54 亿元；按月度 ARR 口径，7 月收入约 6 亿元、8 月约 9.6 亿元；依据业绩会“最近一周年化超 20 亿美元”对应月收入约 12 亿元，乐观情景假设 9-12 月延续爬坡，在月收入 12 亿元基础上略上浮至月均 13 亿元。全年=9.54+6.0+9.6+4×13≈77 亿元，向下取 75 亿元，留出执行误差与季节性余量。参照 Anthropic 的 ARR 轨迹（2024

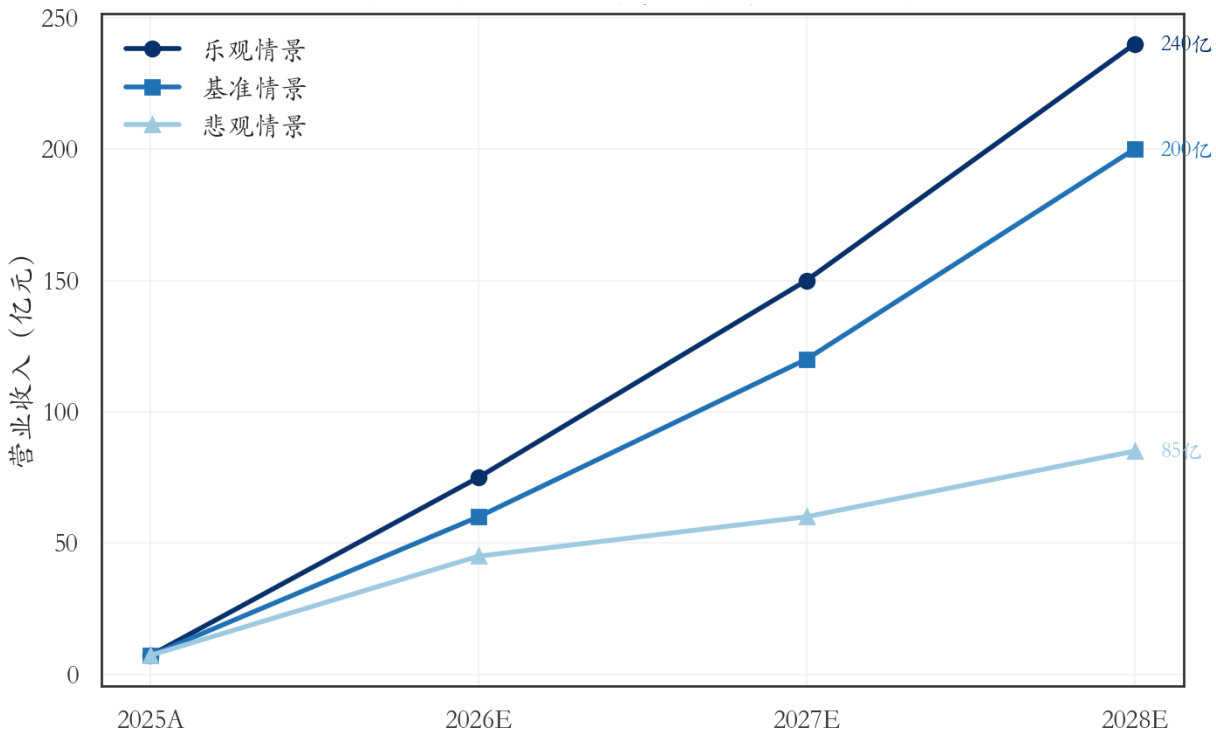
下载日志已记录，仅供内部参考

→2025 年约 10 倍、2025→2026 年约 4 倍），乐观情景下预测智谱 2027 年收入+100%，2028 年规模效应下增速自然递减至+60%；参照 Anthropic 的 blended 毛利率从 40%向 65%的爬坡路径，预测智谱毛利率从 2026H1 的 26.4%起步，提升斜率更都但天花板相对较低，2026-2028 年预测毛利率 35%/50%/60%。关键假设：GLM 保持全球开源第一梯队、Agent 产品线放量、政企采购政策延续、国产算力成本优势兑现。失效条件：ARR 环比转负，或提价后调用量连续下滑。

基准情景：收入 2026-2028 年约 60/120/200 亿元。2026 年的逐月推导：H1 实际 9.54 亿元+7 月 6.0 亿元+8 月 9.6 亿元+9-12 月企稳（月均约 8.4 亿元，介于 7 月与 8 月之间）。企稳而非继续爬坡的理由：牛来/GLM-5.3-Flash 的发布脉冲在 9 月后自然回落，但定价权已重建（提价不掉量），不会回落至 7 月水平以下；2027 年收入+100%，一定程度保持 2026H2 的爬坡增速；2028 年增速自然回落之+67%。预测 2026-2028 年毛利率 30%/45%/55%，低于乐观情景，反映更温和的提价节奏；盈亏平衡指向 2028 年前后。关键假设：本地部署基本盘稳定、MaaS 占比按公司战略持续提升、配售与回 A 募资到位支撑训练投入。

悲观情景：收入 2026-2028 年约 45/60/85 亿元。2026 年的逐月推导：8 月收入含一次性项目确认或峰值外推成分，若牛来/GLM-5.3-Flash 在匿名测试期（免费+新奇效应）产生的冲量不可持续，9 月起回落至月均 4.8 亿元（8 月的一半）。全年=9.54+6.0+9.6+4×4.8≈44 亿元，考虑 9 月暂无回落迹象，向上取 45 亿元。悲观情景下轻量档价格战向旗舰档蔓延，增速显著降档，参照阿里云在 2023-2024 年价格战期增速降至个位数的历史，智谱因基数更小，2027 年收入增速保留+33%。2028 年悲观情景下行业出清后份额向存活者集中，增速回升至+42%。毛利率停滞于 30-40%区间（价格战压制提价空间），盈亏平衡推迟至 2030 年以后。

图 32：智谱收入三情景测算（2025A-2028E）



数据来源：Wind、公司业绩报告、东北证券

下载日志已记录，仅供内部参考

三个情景的共同约束是训练成本曲线：前沿模型单次训练成本以约 2.4 倍/年上行（Epoch AI/斯坦福测算，见系列第一篇），保持在前沿是竞争前提。对国内创业公司而言，该约束更紧。它们的 ARR 基数比海外双寡头低两个数量级，而训练成本曲线是全球同一条。这决定了国内创业公司的盈利路径必然以“窄场景定价权”（Coding、政企、出海）为支点，而非正面军备竞赛。

6.2. 格局推演：五个情景与各自的证伪信号

本报告以两个主轴变量构造情景空间：其一是算力供给斜率。中国信通院数据显示 2026 年一季度国内 AI 算力需求同比+417%而供给仅+128%，供给缺口何时收敛决定价格与产品结构；其二是融资窗口宽窄。港股 8 月深度回撤（与一级市场 Pre-IPO 密集启动并存，窗口的持续性决定独立公司的生存时间。以下五个情景非概率断言，且互不互斥，现实的终局大概率是若干情景的叠加。

情景一（基准）：双集中。开源底座向 DeepSeek/Qwen 集中（全球开发者标准+云算力回流），企业级云服务向阿里/字节/腾讯集中（算力+模型+交付闭环），独立公司保留四条活路：DeepSeek 的开源标准位、Kimi 的 API+出海、智谱的政企+MaaS、MiniMax 的海外 C+B。推导链条：开源份额决定标准归属，云收入（阿里云+45%、字节 MaaS 份额 49.5%）决定企业侧归属，ARR 斜率决定独立公司存亡。四条活路的共同前提是 ARR 持续跑赢训练成本曲线（约 2.4 倍/年）。证伪信号：DeepSeek/Qwen 的开源份额被反超（GLM-5.3-Flash、Hy4 preview 已先后在单周登顶 OpenRouter，榜首开始轮动）、大厂云停止接入第三方开源模型。

情景二：算力约束常态化，轻量档通吃。若国产算力产能爬坡持续慢于需求斜率，旗舰模型的推理成本无法下探，主战场将移向轻量档：8 月智谱 GLM-5.3-Flash 与阿里 Qwen3.8-Flash 把轻量档压至旗舰价格的 1/10-1/20、DeepSeek 引入峰谷定价、多家厂商暂停 C 端新订阅，均是供给受限下的价格与产能分配行为。该情景下的赢家是架构效率领先者（DeepSeek 的 MLA/MoE、智谱的稀疏注意力+mHC、阿里的 GDN+QSA 混合架构）与自有国产算力者（智谱 1GW 数据中心、DeepSeek 乌兰察布项目）；输家是没有算力供给确定性、只能靠 API 转售的中间层。证伪信号：昇腾等国产芯片产能放量后旗舰档价格重启单边下行、厂商恢复 C 端不限量订阅。

情景三：Agent 应用层吞噬模型层利润池。若计费继续从 token 向“席位+用量”乃至结果付费迁移，模型厂自营应用将直接捕获场景价值，纯 API 收入占比下降，竞争从“模型公司之间”升级为“工作流平台之间”。证据已经积累：智谱企业级智能体收入+248.8%、AutoClaw 上线 20 天订阅破 40 万；Kimi 推出 Kimi Work 并将上线 Hosted Agent 平台；MiniMax 的 Token 计划快速渗透；海外参照系是 Claude Code。单一产品线占 Anthropic ARR 约 18%（本系列第一篇）。该情景下，得 Agent 场景者得利润池，模型能力沦为入场券；受益顺序大致为：有 Coding 旗舰产品的公司（Kimi、智谱）>有分发入口的大厂（字节、腾讯）>纯 API 公司。证伪信号：Agent 产品续费率/留存率显著低于 API、企业继续以裸 token 调用为主。

情景四：融资窗口关闭，行业进入并购整合期。推导链条：港股估值中枢下移→IPO 与配售定价走弱→未上市公司在 2027 年下一代训练代际到来时资金承压→有算力缺模型的大厂、运营商或国资平台成为整合方。历史路径在海外已完整走过一遍：Inflection、Adept 先后被微软、亚马逊收编，Character.AI 以授权形式并入 Google（公开历史事实）。国内的候选整合方画像清晰：运营商与国资算力平台有算力与政企渠道但缺头部模型，字节之外的互联网大厂有现金但模型梯队不齐。当前的反向证据是 DeepSeek 二轮融资意向资金超 1000 亿元、月之暗面 G 轮认购超目标 3 倍。窗口仍开，但 8 月港股回撤说明窗口的宽度由二级市场定价，而非一级市场的热情。证伪信号：Kimi 与阶跃星辰 IPO 定价强劲、DeepSeek 上市后估值站稳。

下载日志已记录，仅供内部参考

情景五：地缘变量重塑出海路线，方向是双向的。向下路径：若模型出口管制收紧或主要市场限制中国模型，MiniMax（海外收入占 60.8%）与 Kimi 的海外路线天花板提前到来，倒逼回国内红海、加剧情景四的整合压力。这一路径在 9 月已获得第一批实证：美国 NSA/CISA/FBI 于 9 月 8 日联合点名 DeepSeek、月之暗面、阿里巴巴、MiniMax、阶跃星辰、智谱六家公司“恶意蒸馏”美国前沿模型，CIA 随后将中国企业列为“正当目标”；美国会已通过的《远程访问安全法案》若成法，境外调用美国 AI 模型将被视为出口。管制正从芯片实物扩展到“算力与模型的远程调用”。向上路径：若全球南方市场（东南亚、中东、拉美）以成本优先采购模型，中国开源模型有机会成为“AI 一带一路”的默认底座。OpenRouter 上中国模型 45-61%的 token 份额中相当部分来自非美国开发者，是这一路线的早期证据。两条路径的分野不在商业而在政策，可跟踪变量是各国对中国模型的采购限制清单与出海厂商的区域收入结构变化。

收束来看：**最可能的现实是“情景一为骨架、情景二与三决定利润分配、情景四清理尾部、情景五决定天花板”。**双集中指三节点中的前两个（开源标准位+云生态位），旗舰定价权位是独立公司在双集中夹缝中的第三个生存位。对投资者而言，格局推演的使用方式不是押注单一情景，而是盯住各情景的证伪信号。表 4 的跟踪框架已按情景编号标注。另有两个跨情景的持续验证点：其一，字节“2000 亿 CapEx+40 亿美元 ARR”的投入产出比，若 B 端闭环兑现慢于预期，国内最大规模的算力投入将面临与海外巨头相同的“压缩股东回报或转向外部融资”选择；其二，DeepSeek 与月之暗面的 IPO 定价，将成为所有未上市模型公司的估值锚。

表 4：国内大模型商业化跟踪指标清单

跟踪维度	指标	验证的判断
定价权	主要厂商 API 报价变动；行业均价	涨价周期能否延续
创业公司 ARR	Kimi/MiniMax ARR；智谱 MaaS ARR	ARR 斜率 vs 训练成本曲线
公开财报	智谱/MiniMax 季报：收入、毛利率、研发占比	三情景中哪条在兑现
C 端变现	豆包订阅渗透与抽佣 GMV；付费率	C 端第二曲线能否成立
开源格局	OpenRouter 中国模型份额；HuggingFace 下载量	开源主导权能否转化为收入
大厂闭环	阿里云增速与利润率；字节 CapEx/ARR	内生资金闭环是否成立
资金链	港股 AI 板块估值与 IPO/配售节奏	IPO 供粮模式可持续性（情景四）
格局信号	轻量档价格与调用量；Agent 产品续费率；开源份额排名	情景二、三是否在兑现
地缘变量	各国对中国模型的采购限制；出海厂商区域收入结构	情景五的方向分野

数据来源：东北证券

下载日志已记录，仅供内部参考

7. 总结与展望

回到引言的核心问题：中国大模型行业能否形成可持续的商业模式？收入与利润最终归谁？本报告的答案是明确的，中国大模型行业的商业模式已经成立。证据链是完整的：API 层面的单位经济已被证明（智谱提价 67-100%后调用量反增 400%、DeepSeek 的 API 毛利率 82.9%），收入拐点已于 2026 年三季度集体兑现（字节约 40 亿美元、智谱 16 亿、Kimi 与 MiniMax 分别突破 10 亿与 8 亿美元，口径各异但方向一致）；公司层面的亏损全部来自训练军备竞赛与研发投入，而非推理亏损，这与第一篇对海外的判断同构。但利润不归“模型”本身：收入与利润将沿三个节点集中。开源标准位（DeepSeek/Qwen，收割全球开发者流量与云算力回流）、云生态位（阿里/字节/腾讯，以 IaaS 毛利承接 AI 需求）、旗舰定价权位（Kimi、智谱 MaaS，按价值而非成本定价）。占不到这三个节点之一的公司，拥有的不是利润来源，而是成本负担，将沿情景四的路径被出清或收编；利润拐点的时间，基准情景下指向 2028-2029 年。

本报告围绕国内大模型行业的商业模式，得出一组层层递进的结论。在变现端，国内厂商沿 API、订阅、云生态、开源四条路径分化，但配比与海外显著不同：开源从与变现脱钩的生态杠杆变为变现漏斗的顶层入口，政企私有化构成国内特有的第五条路径，出海成为反向变现路线；API 定价经历了完整的“价格战—涨价”周期，2026 年二季度的集体提价与智谱“提价后调用量反增 400%”共同验证了定价权的重建。在成本端，智谱中报与 DeepSeek 的知情人士数据共同给出了成本结构证据：智谱 2026 年上半年 API 收入占比已达 86.5%、API 毛利率转正至 24.6%，但研发开支仍为收入的 2.2 倍；DeepSeek 前 7 个月以 82.9%的 API 毛利率仍录得 7.15 亿元净亏损（The Information 援引知情人士，未经审计）。单位经济已证明，训练与研发投入吞噬全部收入的结构在国内与海外完全一致。在格局端，B 端/API 跑赢 C 端订阅在国内同样成立；大厂中阿里率先把 AI 云做成规模盈利样本，创业公司中四条活路已经清晰，六小虎中未能跑通变现路径的两家已经出清。

整体而言，国内大模型行业正从“补贴换规模”进入“算账时代”。格局上，本报告不以单一基线作结：双集中是骨架，算力约束与 Agent 应用层决定利润分配，融资窗口决定尾部出清节奏，地缘变量决定出海天花板。商业模式的分化将比模型能力的趋同更深刻地决定各家厂商的终局。

展望后续，以下方向值得持续跟踪与深入研究：
涨价周期的可持续性验证。2026 年二季度的均价回升是补贴出清还是定价权重构，取决于提价后调用量是否持续不掉量；可对主要厂商报价与智谱 MaaS 量价数据做季度跟踪。

开源主导权的收入转化。中国模型在 OpenRouter 约 45-61%的 token 份额（第三方路由口径）向官方 API 与云算力收入的转化率，是 DeepSeek/Qwen 模式估值的核心变量。

港股供粮模式的压力测试。智谱回 A、Kimi 赴港 IPO 的定价与认购情况，将决定创业公司融资窗口的宽度；二级估值若系统性下移，行业并购整合可能提前。

出海的合规与地缘变量。MiniMax 与 Kimi 的海外收入占比继续抬升的同时，数据跨境、模型出口管制与目标市场监管将决定这条路线的天花板。

下载日志已记录，仅供内部参考

8. 风险提示

AI 技术与产品进展不及预期的风险。本报告对 Agent 放量、按价值定价迁移等判断，建立在模型能力持续提升的前提上；若模型进步放缓或长任务可靠性未跨过商用阈值，API 用量与提价逻辑均可能低于预期。

商业化落地与企业需求不及预期的风险。国内企业 IT 预算受宏观影响较大，若政企采购放缓或企业 AI 支出从试点转向规模投产的速度不及预期，智谱等依赖政企与 MaaS 的厂商收入增速可能系统性低于本报告假设。

行业竞争与价格战反复的风险。2026 年的集体提价能否延续存在不确定性；若某头部厂商为抢份额重启降价，行业均价与各家毛利率将再度承压。

算力供给与出口管制的风险。国内厂商面临先进芯片获取的外部管制约束；若算力供给收紧，训练成本曲线上行与推理交付能力受限将同时恶化，对创业公司影响尤甚。

资本开支与现金流风险。字节 2000 亿元级 CapEx、阿里自由现金流转负显示国内大厂同样处于高强度投入期；若增量收入确认滞后，相关公司可能压缩股东回报或调整投入节奏，进而影响全产业链需求。

C 端变现进展不及预期的风险。豆包付费订阅刚起步，花旗调研显示用户可接受月均价格低于其起步价；若订阅渗透与交易抽佣均不及预期，国内 C 端流量的估值逻辑将受到挑战。

出海与地缘政治风险。MiniMax、Kimi 的海外收入依赖全球开发者与付费用户；数据跨境监管、模型出口管制或目标市场准入限制，可能直接冲击该变现路线。2026 年 9 月美国 NSA/CISA/FBI 已联合点名六家中国大模型公司“恶意蒸馏”，若《远程访问安全法案》成法或主要市场跟进限制，出海路线的天花板将提前到来。

数据口径与第三方估算偏差的风险。本报告引用的字节 ARR（40 亿美元）、DeepSeek 前 7 个月营收与毛利率、豆包日收入与成本等数据，来自每日经济新闻、晚点 LatePost、The Information、摩根士丹利、CSDN 等媒体报道或第三方测算，存在口径差异（如 ARR 折算方式、是否含内部消耗），未经审计确认，可能与实际情况存在偏差。

二级市场估值波动的风险。2026 年 8 月港股 AI 板块深度回撤，智谱股价距 7 月峰值一度回撤超六成；已上市公司的股价波动直接影响其配售与再融资能力，未上市公司的 Pre-IPO 与 IPO 定价以二级估值为锚。若估值中枢持续下移，情景四（并购整合）的兑现节奏将提前。

测算假设不确定性的风险。本报告以智谱为锚的三情景测算基于收入增速、毛利率改善与研发收敛等假设，实际结果对假设高度敏感；智谱单一样本不能代表全行业，三情景仅供示意，不构成对相关公司未来业绩的保证。

监管与合规风险。生成式 AI 面临内容安全、算法备案、数据隐私等监管要求；国内外监管框架的演进可能影响相关公司的产品形态、上线节奏与商业化路径。

下载日志已记录，仅供内部参考

研究团队简介：

赵明扬：北京大学硕士（网络传播新媒体方向）；曾任行业主管部委负责产业研究，科技领域投资（VC PE）、投行(FA)、实业（上市公司投资总监）十余年，所投企业 A 股、美股、港股皆有上市。2025 年 1 月加入东北证券，现任传媒互联网行业首席分析师。

刘昱亨：北京大学计算机硕士，北京航空航天大学工学学士。2023 年加入东北证券，现任东北证券传媒互联网行业研究助理。

分析师声明

作者具有中国证券业协会授予的证券投资咨询执业资格，并在中国证券业协会注册登记为证券分析师。本报告遵循合规、客观、专业、审慎的制作原则，所采用数据、资料的来源合法合规，文字阐述反映了作者的真实观点，报告结论未受任何第三方的授意或影响，特此声明。

投资评级说明

股票投资评级说明	买入	未来 6 个月内，股价涨幅超越市场基准 15%以上。	投资评级中所涉及的市场基准： A 股市场以沪深 300 指数为市场基准，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）为市场基准；香港市场以摩根士丹利中国指数为市场基准；美国市场以纳斯达克综合指数或标普 500 指数为市场基准。
	增持	未来 6 个月内，股价涨幅超越市场基准 5%至 15%之间。	
	中性	未来 6 个月内，股价涨幅介于市场基准-5%至 5%之间。	
	减持	未来 6 个月内，股价涨幅落后市场基准 5%至 15%之间。	
	卖出	未来 6 个月内，股价涨幅落后市场基准 15%以上。	
行业投资评级说明	优于大势	未来 6 个月内，行业指数的收益超越市场基准。	
	同步大势	未来 6 个月内，行业指数的收益与市场基准持平。	
	落后大势	未来 6 个月内，行业指数的收益落后于市场基准。	



下载日志已记录，仅供内部参考

重要声明

本报告由东北证券股份有限公司（以下称“本公司”）制作并仅向本公司客户发布，本公司不会因任何机构或个人接收到本报告而视其为本公司的当然客户。

本公司具有中国证监会核准的证券投资咨询业务资格。

本报告中的信息均来源于公开资料，本公司对这些信息的准确性和完整性不作任何保证。报告中的内容和意见仅反映本公司于发布本报告当日的判断，不保证所包含的内容和意见不发生变化。

本报告仅供参考，并不构成对所述证券买卖的出价或征价。在任何情况下，本报告中的信息或所表述的意见均不构成对任何人的证券买卖建议。本公司及其雇员不承诺投资者一定获利，不与投资者分享投资收益，在任何情况下，我公司及其雇员对任何人使用本报告及其内容所引发的任何直接或间接损失概不负责。

本公司或其关联机构可能会持有本报告中涉及到的公司所发行的证券头寸并进行交易，并在法律许可的情况下不进行披露；可能为这些公司提供或争取提供投资银行业务、财务顾问等相关服务。

本报告版权归本公司所有。未经本公司书面许可，任何机构和个人不得以任何形式翻版、复制、发表或引用。如征得本公司同意进行引用、刊发的，须在本公司允许的范围内使用，并注明本报告的发布人和发布日期，提示使用本报告的风险。

若本公司客户（以下称“该客户”）向第三方发送本报告，则由该客户独自为此发送行为负责。提醒通过此途径获得本报告的投资者注意，本公司不对通过此种途径获得本报告所引起的任何损失承担任何责任。

东北证券股份有限公司

网址：http://www.nesc.cn 电话：95360,400-600-0686 研究所公众号：dbzqyanjiusuo

地址	邮编
中国吉林省长春市生态大街 6666 号	130119
中国北京市西城区锦什坊街 28 号恒奥中心 D 座	100033
中国上海市浦东新区杨高南路 799 号	200127
中国深圳市福田区福中三路 1006 号诺德中心 34D	518038
中国广东省广州市天河区冼村街道黄埔大道西 122 号之二星辉中心 15 楼	510630

