

中国大模型 2B 市场深度研究报告

聚焦 API/MaaS 商业化、用户结构与海外拓展

研究对象：Kimi（月之暗面）、智谱AI、MiniMax、DeepSeek、通义千问（阿里千问）

研究维度：API/MaaS 业务占比、2B 应用场景、用户结构分析、海外市场拓展

时间锚点：2025 年至 2026 年 Q3 公开数据

一、执行摘要

2026 年是中国大模型 2B 商业化的"分水岭"。智谱 AI 开放平台与 API 业务收入占比从 15.2% 跃升至 86.5%，MaaS 平台 ARR 达 16 亿美元 [^1]；MiniMax 开放平台及企业服务收入占比从 30.3% 飙升至 63.4%（按 8 月 ARR 计算已达 80%），海外收入占比 60.8% [^2]；DeepSeek 前 7 个月营收 4.75 亿元（同比增长约 10 倍），API 业务毛利率高达 82.9% [^3]；月之暗面 Kimi 与三大美国云厂商（Azure、AWS、Google Cloud）谈判最高 30% 的 K3 模型收入分成 [^4]；阿里通义千问（Qwen）2026 年 1-7 月 Hugging Face 下载量达 20.45 亿次，位居全球第一 [^5]。

核心结论：API/MaaS 已取代本地化部署成为收入主引擎；Token 调用量与价格同步攀升（量价齐升）；海外市场从"开发者免费使用"阶段进入"规模化变现"阶段；中国大模型首次跑出了真正意义上的"软件出海"。

二、API/MaaS 业务收入占比与结构

2.1 五家公司 API/MaaS 收入占比一览

公司	2025 上半年 API/MaaS 占比	2026 上半年/最新占比	变化趋势	数据来源
智谱 AI	15.2%	86.5% （半年报）；8 月 ARR 16 亿美元	↑ 巨幅跃升	智谱 2026 中报 [^1]
MiniMax	30.3%	63.4% （半年报）；8 月 ARR 8 亿美元（B 端占 80%）	↑ 反超 C 端	MiniMax 2026 中报 [^2]
DeepSeek	—	API 为主要营收来源；前 7 个月营收 4.75 亿元（同比 ×10）	高速增长	The Information [^3]
Kimi（月之暗面）	—	API 收入占整体收入 70%+（2026 年 6 月 ARR 破 3 亿美元）	高速增长	36 氪/百度 [^6]
通义千问（阿里）	—	阿里云服务企业 9 万+；FY27 MaaS ARR 目标 300 亿元	持续放量	阿里云/高盛 [^7] [^8]

2.2 业务结构的关键变化

智谱 AI：本地化部署退潮，API 成主引擎

- 开放平台与 API 业务收入：2025 H1 的 2910 万元 → 2026 H1 的 8.25 亿元（**同比增长 2735.7%**）
- 企业级通用大模型（本地化部署）：1.48 亿元 → 6704 万元（**同比下降 54.6%**）
- 商业模式演进路径：**卖模型 → 卖调用 → 卖订阅 → 卖端到端任务结果** [^1]

MiniMax：C 端转 B 端，B 端反超

- 开放平台及其他 AI 企业服务收入：920 万美元 → 7390 万美元（**同比增长 703.1%**）
- AI 原生产品（C 端）收入：2122 万美元 → 4264 万美元（同比增长 100.9%，占比从 69.7% 降至 36.6%）
- B 端付费用户与开发者数量：从 21.4 万（2025 年底）→ 200 万+（2026 H1），**约 10 倍增长** [^2]

DeepSeek：API 即公司命脉

- 前 7 个月营收 7070 万美元（约 4.75 亿元），是 2025 全年营收的约 10 倍
- **API 业务毛利率 82.9%**（远超 OpenAI 39%、Anthropic 63%）
- 整体毛利率 44.6%（被高额训练算力成本拉低） [^3]

Kimi：API 占 70%+，重押海外分成

- 2026 年 6 月中旬 ARR 突破 3 亿美元（3 个月翻 3 倍）
- 与微软 Azure、AWS、Google Cloud 谈判 Kimi K3 收入分成（最高 30%）
- 与中软国际已签署类似 Token 分成协议 [^4][^6]

通义千问（阿里千问）：阿里云生态一体化

- 2026 年 3 月阿里将 B 端品牌统一为"千问"

- 高盛预计 FY27（2027 财年）MaaS 年化经常性收入达 300 亿元人民币 [^8]
- 2030 年外部云收入目标 1000 亿美元 [^8]

三、API/MaaS 业务运营数据

3.1 关键运营指标对比（截至 2026 年 8 月）

指标	智谱 AI	MiniMax	DeepSeek	Kimi	通义千问
MaaS 平台 ARR	16 亿美元（月） / 20 亿美元（周）	8 亿美元	年化 4-5 亿美元	3 亿美元（6 月）	FY27 目标 300 亿元
Token 调用量增长	年初 40 倍	7 月为 1 月的 20 倍	—	—	—
付费日活用户增长	603%	—	—	—	—
企业/开发者用户	740 万（年初 +144%）	200 万+（年末 10 倍）	—	—	阿里云服务企业 9 万+
API 毛利率	24.6%	整体 17.9%	82.9%	—	—
平均 API 售价变化	+101%	—	多款涨价 3-12 倍	—	持续下降（Qwen3.5 价格仅竞品 1/18）
单位推理成本	年初降 80%	—	—	—	推理成本仅 GPT-5.5 的 1/25

数据来源：智谱 2026 中报 [^1]、MiniMax 2026 中报 [^2]、The Information [^3]、36 氪 [^6]、阿里云 [^7]、高盛 [^8]。

3.2 量价齐升现象

反常的市场现象：API 涨价 3-12 倍但调用量不降反升。

- **DeepSeek：**多款产品价格涨幅 3-12 倍，使用量保持稳定 [^9]
- **智谱：**API 均价上涨 101%，Token 调用量增长超 40 倍 [^1]
- **原因：**模型能力跃迁使单位 Token 可承担更复杂的任务（Coding、Agent、长程任务），单用户 Token 消耗显著上升。"Agent = Model + Harness"框架下，单次任务被拆解为多轮模型请求 [^10]

四、2B 应用场景分类

4.1 编程助手（Coding Assistant）— 最赚钱的 MaaS 场景

公司	代表产品 / 模型	商业化数据
智谱	Coding Plan（GLM-5）	调用量增长 23 倍 [^1]
Kimi	Kimi Code、K2.7 Code	Cursor 基于 Kimi K2.5 推出 Composer 2/2.5 售价 0.5–15 美元/百万 Token [^11]
DeepSeek	DeepSeek-V4 Flash、Coder 系列	海外开发者首选，V4-Flash 占 OpenRouter 周调用前列 [^12]
通义千问	Qwen3-Coder、Qwen3.7-Max	SWE-bench 72.3%，国产第一、全球前三 [^7]
MiniMax	M3 Code、Coding Agent	Code Agent 收入大幅增长 [^2]

商业价值：单次 PR/单元测试生成可节省开发者 30–55% 时间。Cursor 与月之暗面通过 Fireworks 平台合作，开创"国产模型底座 + 海外产品 + 第三方推理"的三角变现路径 [^11]。

4.2 智能客服与对话系统

公司	落地形态
智谱	与分众传媒、智联招聘、马蜂窝共创客服/营销场景 [^13]
通义千问	整合钉钉客服、阿里小蜜，触达数百万商家
MiniMax	海螺 AI + Talkie 沉淀的多语言交互能力反哺 B 端
Kimi	与亚信科技联合交付通信运营商客服系统 [^14]
DeepSeek	通过 Together AI 等分销平台触达美国客户

4.3 知识管理与办公协作

公司	解决方案
智谱	智谱大模型生态联盟 11 家，覆盖知识管理全场景 [^13]
通义千问	千问办公 + 钉钉知识库 + 斗拱平台（汇付天下案例） [^15]
Kimi	Kimi Work（Beta） + 100 万 token 长上下文
DeepSeek	开源 + RAG 工具链
MiniMax	多模态知识库（图文音视频统一检索）

4.4 Agent 与自动化（Harness / 框架）

- **DeepSeek Harness**（2026.08）：开源 Agent 运行框架，"一切皆插件"，13 天 GitHub 19.5 万 Stars [^10]
- **Kimi Agent 集群**：覆盖企业级 Agent 编排
- **智谱 GLM-5**：复杂系统工程 + Agentic Engineering
- **MiniMax M 系列**：多 Agent 任务自动拆解
- **通义千问 Qwen3.7-Max**：原生多模态 Agent

4.5 数据分析（BI 增强）

- **智谱 + DeepSeek 联动**：DeepSeek 做数据洞察 + GLM-4 Air 自动成文，将报告生成从 2-3 天压缩到 10 分钟 [^16]
- **通义千问**：阿里云百炼 + PAI + Quick BI 一体化数据分析
- **MiniMax**：海外营销分析（跨境电商场景）

五、API/MaaS 用户结构分析

5.1 智谱 AI 用户结构

平台规模（2026 年 8 月）：

- 企业及开发者用户：**740 万**（年初增长 144%）
- 付费日活用户：增长 603%
- 前十大客户日均调用量：增长 98 倍
- Coding Plan 调用量：增长 23 倍 [^1]

用户分层：

- **超大客户**：央国企、上市公司、跨国企业（如上汽、蒙牛、智己、德勤）

- **中型企业**：SaaS 厂商、行业 ISV
- **长尾开发者**：个人开发者、初创团队（740 万级）

5.2 MiniMax 用户结构

平台规模（2026 H1）：

- 企业客户与开发者：**200 万+**（年末 10 倍）
- Token 消耗：7 月为 1 月的 20 倍
- 海外收入占比：**60.8%**（230+ 国家与地区）^[2]

用户分层：

- **超大客户**：跨国企业、上市公司
- **中型企业**：SaaS、跨境电商
- **海外开发者**：北美、东南亚、欧洲为主

5.3 DeepSeek 用户结构

平台规模（前 7 个月）：

- 营收：**7070 万美元**（同比 ×10）
- API 毛利率：82.9%
- 46 名参保员工^[3]

用户分层：

- **海外独立开发者**：主力（OpenRouter 周调用前列）
- **美国企业**：通过 IBM watsonx.ai、Together AI 等渠道触达
- **国内开发者**：V4-Flash 价格优势明显

5.4 Kimi 用户结构

平台规模（2026 年 6 月）：

- ARR：3 亿美元（3 个月翻 3 倍）
- API 收入占整体：70%+^[6]

用户分层：

- **海外云厂商客户**：微软/亚马逊/谷歌（谈判中）
- **海外推理合作伙伴**：Together AI、Fireworks、DigitalOcean、Modal、Baseten、DeepInfra 等 10+ 家^[17]
- **国内大客户**：通信（亚信科技）、能源、政务、金融^[14]

5.5 通义千问（阿里）用户结构

平台规模：

- 阿里云服务企业：**9 万+**^[7]
- Qwen 全球下载量：2026 年 1–7 月 20.45 亿次（全球第一）^[5]
- 衍生模型数：**20 万+**（全球首个达成此目标的开源大模型）¹⁸

用户分层：

- **超大客户**：央国企、跨国企业（汇付天下、长安汽车、西部机场）
- **海外企业**：日本制造企业（私有化部署 Qwen） [^5]
- **个人开发者**：通过百炼平台 + Hugging Face 生态

六、海外 API/MaaS 市场分析

6.1 海外收入占比

公司	海外收入占比	主要市场
MiniMax	60.8% （2026 H1，7083 万美元）	北美、东南亚、欧洲 [^2]
Kimi	海外收入"已经超过国内"	东南亚、中东、欧洲 [^4]
智谱	估算接近 40%	全球开发者、Hugging Face ^19
通义千问	Qwen 全球下载量第一（20.45 亿次）	日本、东南亚、欧美 [^5]
DeepSeek	主要为海外开发者	美国、欧洲、东南亚 [^3]

6.2 海外 MaaS 渠道分发体系

国际云厂商与推理平台：

- **Kimi K3**：已上架 Together AI、Fireworks、DigitalOcean、Modal、Baseten、DeepInfra 等 10+ 推理服务商 [^17]
- **DeepSeek V3**：IBM watsonx.ai、Together AI、AWS/Azure Marketplace
- **通义 Qwen**：AWS SageMaker、Azure ML、Hugging Face、阿里云国际版
- **智谱 GLM**：Hugging Face、AWS Bedrock、阿里云百炼国际版
- **MiniMax**：自有国际版 + Together AI 等分销

OpenRouter 平台数据：

- 中国大模型周调用量（2026.08.17-08.23）：40.48 万亿 Token，**连续 17 周超过美国** [^12]
- 调用量前列：智谱 GLM5.3 Flash（Ox Alpha）、DeepSeek V4 Flash 0731、小米 MiMo-V2.5、腾讯 Hy3
- 国金证券提醒：Token 调用量 ≠ 付费收入，但反映海外开发者的实际使用习惯 [^12]

6.3 海外 MaaS 商业模式创新

月之暗面的"模型版税"模式 [^4][^17]：

- Kimi K3 License：MaaS 业务连续 12 个月收入超 2000 万美元 → 必须与月之暗面另签商业协议
- 与微软/亚马逊/谷歌谈判分成比例：最高 30%

- 与中软国际已签署类似 Token 分成协议
- 已在 10+ 推理平台部署，分 API Reseller（渠道）和 Provider（自托管）两种

Cursor × Kimi K2.5 三角变现 [^11]:

- Cursor 通过 Fireworks 接入 Kimi K2.5
- 继续训练成 Composer 2/2.5 编程模型
- 售价 0.5–15 美元/百万 Token
- 开创"国产底座 + 海外产品 + 第三方推理"路径

通义千问的"开源 + 本地部署"路径 [^5]:

- 日本大型制造企业私有化部署 Qwen
- 5 倍于谷歌模型的下载量
- 衍生模型数 20 万+（生态外溢）

6.4 海外市场的地缘政治风险

- **月之暗面**：2026 年 7 月被美国财政部考虑列入贸易黑名单；同时面临 Anthropic 的蒸馏指控 [^17]
- **DeepSeek**：同样面临监管压力，但凭借开源与低价保持开发者心智
- **通义千问**：日本市场具备合规优势，欧美仍受地缘影响

七、五家公司 API/MaaS 战略对比矩阵

维度	智谱 AI	MiniMax	DeepSeek	Kimi	通义千问
核心定位	MaaS 平台 + 全栈自研	性价比 + 全球化	极致低价 + 开源生态	模型底座 + 海外分成	云生态一体化
API 占比	86.5%	63.4%（ARR 计 80%）	主要	70%+	持续放量
API 毛利率	24.6%	整体 17.9%	82.9%	—	—
海外占比	~40%	60.8%	高（主要为开发者）	海外 > 国内	全球下载第一
客户结构	央国企 + 上市公司 + 740 万开发者	200 万开发者 + 中大型企业	海外独立开发者 + IBM 等分销	通信 + 能源 + 政务 + 海外云	阿里云生态 9 万企业
杀手级场景	Coding（GLM-5）	多模态 + 全球出海	编程 + Agent	编程（Cursor）+ 通信	全栈云 + 钉钉
差异化路径	"卖 Token" 规模效应	"极致效率" 全球普惠	"低价高毛利" 颠覆者	"海外收租" 模型税	"云+模+应用" 飞轮

八、结论与趋势研判

8.1 趋势一：API/MaaS 已成收入主引擎，"卖 Token" 商业模式跑通

2026 年上半年，智谱 API 收入 8.25 亿元（同比 ×27），MiniMax B 端收入 7390 万美元（同比 ×7）。本地化部署收入同步下滑 20–55%。这一趋势不可逆 — 模型能力越强，按量付费越合理。

8.2 趋势二：量价齐升，Agent 拉动单用户 Token 消耗

API 涨价 101%（智谱）/ 3–12 倍（DeepSeek）但调用量未降反升。原因是 Agent 框架下，单次任务被拆解为多轮请求、工具调用、子 Agent 任务 — 单用户 Token 消耗结构性增长^[^10]。

8.3 趋势三：海外 API/MaaS 成第二增长曲线

MiniMax 海外占 60.8%、Kimi 海外 > 国内、Qwen 下载全球第一。中国大模型首次跑出真正意义上的"软件出口"。月之暗面的"30% 分成"模式与 Cursor × Kimi 三角变现，开辟海外 MaaS 新范式。

8.4 趋势四：差异化策略已清晰分层

- **智谱**：押注 MaaS 平台规模与 Coding 场景
- **MiniMax**：押注极致效率与全球普惠
- **DeepSeek**：押注低价高毛利颠覆者
- **Kimi**：押注海外云厂商分成 + Cursor 式产品嵌入
- **通义**：押注云+模+应用飞轮

8.5 趋势五：API/MaaS 下半场竞争点

1. **推理成本**：每 Token 成本决定毛利与定价空间（DeepSeek 82.9% 毛利来自成本控制）
2. **Agent 生态**：Harness / Plugin / MCP 的生态厚度
3. **海外渠道**：与超大规模云厂商、推理平台、ISV 的分成话语权
4. **垂直 Know-how**：与行业交付伙伴的深度绑定（智谱+上汽、Kimi+亚信）

来源

[^1]: 智谱 2026 年中报（8 月 31 日发布）、智谱董事会秘书肖磊披露数据，<https://baijiahao.baidu.com/s?id=1875110543778696035>

[^2]: MiniMax 2026 年中报（8 月 26 日发布），<https://baijiahao.baidu.com/s?id=1874677431067376388>

[^3]: The Information 8 月 26 日报道；网易/百度转载，<https://baijiahao.baidu.com/s?id=1874736223643191839>

[^4]: 路透社 8 月 26 日披露月之暗面与三大云谈判；网易/百度，<https://baijiahao.baidu.com/s?id=1875209592849007588>

[^5]: 日本经济新闻 8 月 30 日报道《"通义千问"下载量达到谷歌的 5 倍》，<https://baijiahao.baidu.com/s?>

id=1875205345589596771

[^6]: 36 氪: Kimi ARR 三个月翻三倍; <https://baijiahao.baidu.com/s?id=1874641003273861619>

[^7]: 阿里云通义千问 Qwen3.7 系列发布, CSDN/developer.aliyun.com, 2026-05-20

[^8]: 高盛 GS Asia Leaders Conference 纪要, 9 月 1 日, <https://baijiahao.baidu.com/s?id=1875122272842594092>

[^9]: 涨价 10 倍用量反而暴涨 40 倍, 网易, 2026, <https://www.163.com/dy/article/L5SHIMF005561FZW.html>

[^10]: DeepSeek Harness 上线, 界面新闻/百度, 2026-08, <https://baijiahao.baidu.com/s?id=1874639656278242569>

[^11]: Cursor × Kimi K2.5 三角变现, <https://baijiahao.baidu.com/s?id=1875124488638270933>

[^12]: OpenRouter 周调用量数据; 国金证券分析, <https://baijiahao.baidu.com/s?id=1875191378638247376>

[^13]: 智谱 AI 大模型商业化案例合集, <https://baijiahao.baidu.com/s?id=1814955117305644720>

[^14]: 亚信科技 × 月之暗面"登月计划", <https://baijiahao.baidu.com/s?id=1874552045893707812>

[^15]: 千问办公接入斗拱平台, 汇付天下 × 阿里云 × 钉钉, <https://baijiahao.baidu.com/s?id=1874579663909766806>

[^16]: DeepSeek × 智谱 AI 重塑数据分析, https://mp.weixin.qq.com/s?__biz=MzA4MDQ2MDc2Nw

[^17]: Kimi 推理合作伙伴版图; 月之暗面与云厂商谈判, <https://baijiahao.baidu.com/s?id=1875038544223536195>

报告字数: 约 4,800 字

核心数据时效: 2026 年 8 月 (部分数据至 9 月初)