仅供内部参考，请勿外传！

华泰证券 HUATAI SECURITIES

从海外模型看国内基础模型发展趋势

华泰研究

2026 年 8 月 10 日 | 中国内地　　动态点评

基础模型竞争正由产品探索逐步走向高价值场景、企业市场与使用量驱动的商业化竞争。OpenAI 早期依托相对自下而上的产品创新机制，通过 ChatGPT 建立大规模 C 端入口；Anthropic 则持续聚焦模型能力、Coding 和企业 API，并随着 Coding 场景成熟逐步体现资源集中优势。据 SemiAnalysis 于 2026 年 7 月 8 日发布的《Anthropic 3Q26 Profit Over $1B》估算，Anthropic 年化运行收入于 2026 年 4 月超过 OpenAI，并于 6 月 28 日达到约 690 亿美元。与此同时，智谱、MiniMax 亦逐步向 API、Coding 和 Agent 收敛。我们认为，基础模型行业的竞争维度正在由模型能力和用户规模进一步扩展至高价值 Token 消耗、企业渗透、单位经济性及产品分发能力。国内基础模型相关公司包括：智谱、MiniMax。

模型厂商组织模式优势或随产业阶段切换

据张小珺与真格基金管理合伙人戴雨森于 2026 年 5 月 26 日发布的播客，OpenAI 内部研究与产品团队自主性较强，有利于扩大产品探索范围并通过用户反馈寻找 PMF，ChatGPT 是这一机制的重要代表；Anthropic 战略方向相对集中，更强调安全、模型能力和企业市场，并将资源集中于 Coding 等少数核心领域。OpenAI 由此较早积累 C 端流量与品牌优势，而 Anthropic 较轻的免费用户负担和企业 API 收入结构，则有利于提高资源配置效率。据 SemiAnalysis 估算，截至 26 年 7 月，Anthropic 约 75%-85%的收入来自按量计费 API。我们认为，随着 Coding 等高价值场景逐步明确，模型厂商资源协同和执行效率的重要性正在提升。

Coding 正在成为关键的商业化场景

Coding 具有高频调用、长上下文、多步骤工具调用和任务价值可量化等特征，模型能力提升更容易转化为开发者付费与企业采购。Anthropic 自 Claude 3.5 Sonnet 起逐步形成 Coding 能力优势，并以 Claude Code 直接承接开发者与企业需求；OpenAI 则依托 ChatGPT 入口、开发者生态和企业渠道推动 Codex 上量。据 OpenAI 于 2026 年 6 月 2 日发布的《Codex is becoming a productivity tool for everyone》，截至 2026 年 6 月 2 日 Codex 周活跃用户已超过 500 万。我们认为，Coding 正在由代码辅助工具延伸至长时程任务执行平台，并成为两家公司商业化竞争的重要方向。

国内厂商商业化路径逐步收敛

MiniMax 早期侧重 C 端 AI 原生应用和海外市场，据 MiniMax 2025 年度业绩公告，AI 原生产品收入占比由 23 年的约 22%升至 25 年的约 67%，海外收入占比提升至约 73%；智谱则较早布局企业客户和本地化部署，据智谱于 2026 年 3 月 31 日发布的 2025 年度业绩公告，，2025 年本地化部署收入占比仍为 74%，但云端部署占比已提升至 26%。随着模型能力和 Coding 场景发展，两家公司均加快 API 及 Agent 布局。我们认为，两家公司均在形成模型能力提升带动 Token 调用、企业采购和 ARR 增长的商业化链条。

投资结论

国内基础模型相关公司包括：智谱、MiniMax。两家公司或已进入模型能力提升带动 API 及 Coding/Agent 需求增长的阶段。据智谱 2025 年报业绩会、36 氪及 MiniMax 投资者交流纪要：智谱 2026 年 3 月 MaaS API ARR 约 2.5 亿美元，7 月已增至约 10 亿美元；MiniMax 4 月 ARR 约 4 亿美元，并维持年底 10 亿美元 ARR 目标。我们认为，国内基础模型商业化正在由项目制、C 端流量进一步向标准化 API 和高价值生产力场景延伸。

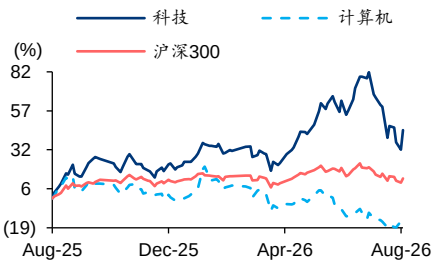
风险提示：模型迭代不及预期、商业化进展不及预期、行业竞争加剧。

科技	增持（维持）
计算机	增持（维持）

郭雅丽　研究员
SAC No. S0570515060003　guoyali@htsc.com
SFC No. BQB164　+(86) 21 3847 6016

袁泽世*，PhD　研究员
SAC No. S0570524090001　yuanzeshi@htsc.com
+(86) 21 2897 2228

行业走势图



资料来源：Wind，华泰研究

仅供内部参考，请勿外传！

正文目录

OpenAI 与 Anthropic：两种组织方式，两条早期路径 3

企业文化或影响公司早期选择 3

OpenAI 为什么率先赢得前半场 4

Anthropic 为什么厚积薄发 4

当前市场或更加认可 API 与 Token 商业模式 5

Coding 能力如何成为商业化拐点 7

Coding 较普通聊天具备更清晰的商业化路径 7

Claude 通过持续迭代逐步建立 Coding 优势 7

Cursor 带动 Claude API，Anthropic 躬身入局 8

OpenAI Codex 快速上量，Coding 进入双强竞争 9

基础模型或会多厂商共存 11

Ramp AI Index：企业模型份额持续变化 11

企业正在同时使用多款模型 12

智谱与 MiniMax 的路径如何演变 13

MiniMax 早期商业化更偏 C 端产品和海外市场 13

智谱早期商业化更偏企业、本地化部署和项目制 14

两家公司商业化重心逐步向 API、Coding 和 Agent 收敛 15

未来基础模型公司最应该看什么 17

风险提示 17

仅供内部参考，请勿外传！

OpenAI 与 Anthropic：两种组织方式，两条早期路径

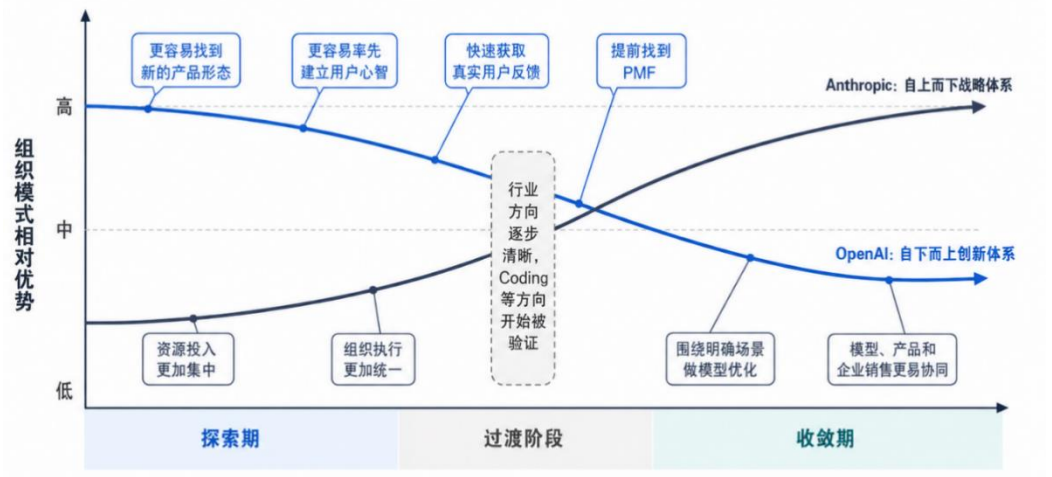
企业文化或影响公司早期选择

OpenAI 相对自下而上的创新体系，或更适合行业方向尚未清晰的探索阶段。据张小珺与真格基金管理合伙人戴雨森于2026年5月26日发布的播客《雨森的创投观察第2集：Harness、下一个字节、2026大机会和Stanley Druckenmiller》对OpenAI组织机制的概括，公司内部研究人员与产品团队拥有较强自主性，产品方向可以由团队内部自然产生，组织也允许较大范围的试错与非预设性探索，ChatGPT可视为该模式下产生的代表性产品。其优势主要体现在：1）扩大产品探索范围，有助于发现新的产品形态并较早建立用户认知。2）通过产品快速获取真实用户反馈，缩短技术研发与市场需求之间的验证链条。3）在技术成熟度仍然有限的情况下，提前识别产品市场匹配（PMF）的可能方向。我们认为，该模式能够提高产品探索的广度，但其有效性更依赖行业仍处于需求定义和产品形态尚未收敛的阶段。

Anthropic 相对自上而下的战略体系，更强调方向聚焦与资源协同。上述播客访谈指出，Anthropic战略重点相对集中，管理层对安全、模型能力及企业市场方向具有较强主导作用，并将资源优先配置于少数核心领域。在行业方向尚未清晰时，该体系覆盖的产品探索范围可能相对有限；随着Coding等应用方向逐步得到验证，公司可以更快集中模型、算力、产品和销售资源，围绕明确场景持续推进能力优化。其优势主要体现在：1）资源投入更加集中，有助于降低多方向探索造成的分散消耗。2）组织目标与执行口径更加统一，有利于形成连续的产品迭代路径。3）模型研发、产品设计与企业销售更容易形成协同。我们认为，该模式更适合商业化方向逐步明确、竞争重心转向执行效率与规模化落地的阶段。

组织模式的有效性或取决于行业发展阶段。在产品探索期，行业需求、技术边界和商业模式尚未稳定，企业更需要通过分散探索、快速试错和用户反馈发现潜在方向，自下而上的创新机制通常具备更强适应性。随着技术路线和商业场景逐步收敛，资源配置效率、战略一致性和组织执行能力的重要性随之提升，自上而下的体系更有利于围绕明确场景持续投入。因此，比较OpenAI与Anthropic的组织模式，需要重点判断AI行业仍处于产品形态持续扩散的阶段，还是已经进入商业化路径逐渐清晰、资源投入趋于集中的阶段。我们认为，两种模式各有适用条件，其相对优势可能随产业周期变化而动态切换。

图表1：“探索期—收敛期”组织优势变化示意图



注：该图机制示意，并非定量测算。

资料来源：张小珺、戴雨森《雨森的创投观察第2集：Harness、下一个字节、2026大机会和Stanley Druckenmiller》（2026年5月26日），华泰研究

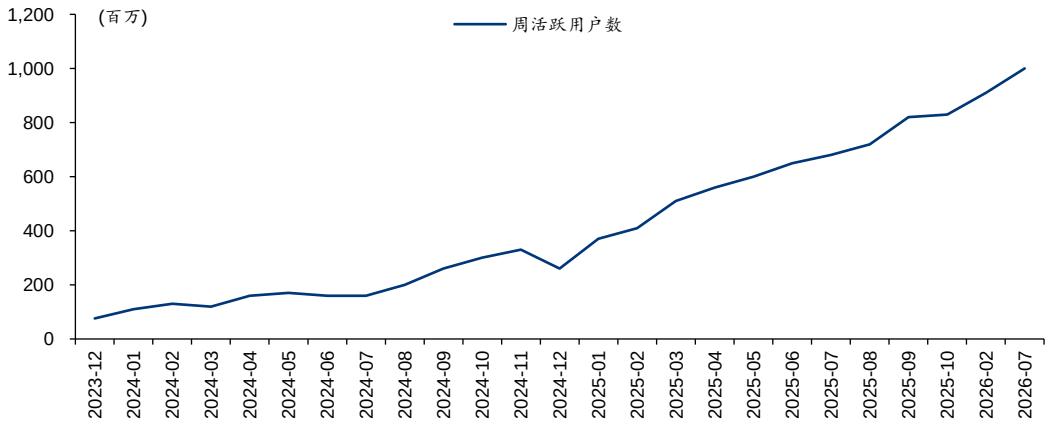
仅供内部参考，请勿外传！

OpenAI 为什么率先赢得前半场

OpenAI 依托 ChatGPT 较早建立 C 端入口，并逐步积累用户与商业化优势。ChatGPT 将底层模型能力转化为可直接触达用户的产品形态，帮助 OpenAI 较早打通技术研发、用户使用与商业变现之间的链条。其价值主要体现在：1）快速扩大用户规模，强化品牌认知。2）通过付费订阅和 API 调用形成收入来源，并带动企业客户拓展。3）持续获取真实用户反馈和交互数据，为模型能力优化及产品迭代提供支持。4）吸引开发者围绕模型能力构建应用，推动开发者生态逐步形成。据 The Information 于 2026 年 4 月 28 日发布的《OpenAI Sees $8 ChatGPT Driving Consumer Subscribers to 122 Million This Year》，目前已有超过 5000 万人付费使用 ChatGPT，OpenAI 预计 2026 年约 50%的收入将来自企业客户。我们认为，ChatGPT 帮助 OpenAI 较早形成了模型能力向用户入口、商业收入和迭代数据转化的路径，并逐步推动商业化结构由个人订阅向企业市场延伸。

OpenAI 持续扩大 ChatGPT 用户规模，并逐步向 Agent 及专业工作流延伸。据 The Information 于 2026 年 7 月 28 日发布的《OpenAI's ChatGPT Nears 1 Billion Weekly Active Users Seven Months After Target》汇总的 OpenAI 历史披露，ChatGPT 周活跃用户由 2023 年 11 月约 1 亿增至 2024 年 8 月超过 2 亿，2024 年 12 月、2025 年 2 月和 3 月分别超过 3 亿、4 亿和达到约 5 亿，2025 年 8 月和 10 月进一步增至约 7 亿和 8 亿，2026 年 2 月超过 9 亿。该报道同时指出，截至 2026 年 7 月 28 日，ChatGPT 周活跃用户目前已接近 10 亿，但较 OpenAI 原定目标晚约 7 个月。与此同时，据 OpenAI 于 2026 年 6 月 2 日发布的《Codex is becoming a productivity tool for everyone》，Codex 周活跃用户由 2026 年 2 月的 160 万增至 6 月超过 500 万，产品使用场景由 Coding 进一步向研究、数据分析和知识工作延伸。我们认为，尽管 ChatGPT 用户增长节奏较原计划有所放缓，其大规模用户入口仍有助于模型能力升级向新增场景传导，并通过 Codex、Agent 等产品逐步向高价值工作流延伸。

图表2：ChatGPT 周活跃用户数（百万）



注：2023 年 11 月至 2026 年 2 月数据按 The Information 2026 年 7 月 28 日图表读取并取近似值；2026 年 7 月数据为该报道正文口径

资料来源：The Information《OpenAI's ChatGPT Nears 1 Billion Weekly Active Users Seven Months After Target》（2026 年 7 月 28 日），华泰研究

Anthropic 为什么厚积薄发

Anthropic 早期 C 端入口相对有限，用户与收入积累慢于 OpenAI。OpenAI 通过 ChatGPT 快速触达消费者，并将模型能力转化为用户规模、订阅收入、品牌流量和产品反馈；Anthropic 早期未形成同等量级的 C 端产品入口，消费者用户规模、订阅收入及自然流量相对有限，因而未能形成与 OpenAI 相似的消费端增长路径，早期 ARR 规模亦相对较低。我们认为，Anthropic 早期在用户和收入规模上的差距，可能更多与缺少高流量产品入口有关，模型能力差距并非唯一解释。

仅供内部参考，请勿外传！

Anthropic 将资源进一步集中于模型能力与企业市场。由于早期消费端入口相对有限，公司重点投入基础模型训练、模型安全与对齐、长上下文、Coding 能力、API 服务、企业客户、云厂商渠道及开发者生态，增长路径更加依赖开发者和企业按量付费。我们认为，Anthropic 早期未优先争夺消费端流量，可能影响短期用户和收入规模，但也有助于将资源集中于企业级模型、产品与渠道能力建设。

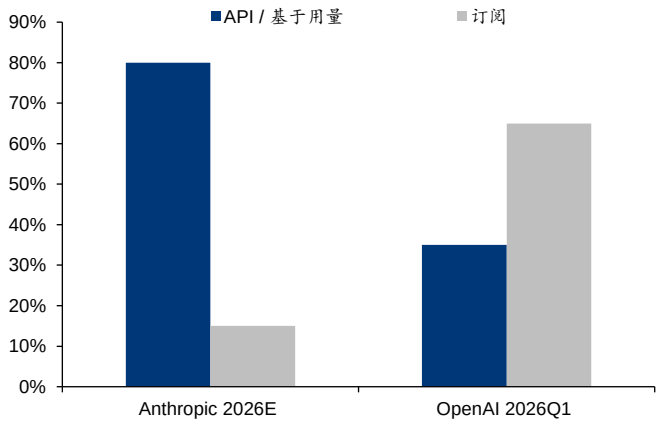
较轻的免费用户负担有助于 Anthropic 控制推理成本。生成式 AI 每次对话和模型调用均会产生芯片折旧、能源消耗及数据中心运营成本，免费用户规模扩大仍会带来相应的边际成本。据 SemiAnalysis《Anthropic 3Q26 Profit Over $1B》估算，OpenAI 需要服务超过 9 亿免费用户，每位免费用户月均服务成本略低于 0.70 美元，相关支出历史上可能对整体毛利率造成约 20%-30%的拖累；若 OpenAI 与 Anthropic 均实现 1000 亿美元 ARR，OpenAI 毛利润可能较 Anthropic 少约 250 亿美元。Anthropic 收入主要来自企业 API 和高回报开发场景，承担的免费消费端推理需求相对较少，因此可以将更多算力配置于付费意愿较为明确的客户。我们认为，较轻的免费用户成本可能是 Anthropic 利润率相对占优的重要原因之一。

当前市场或更加认可 API 与 Token 商业模式

API 计费为 Anthropic 建立了使用量驱动的商业化路径。API 具有标准化、可计量和易分发等特征，企业可以按照实际调用量付费，并将模型嵌入研发、客服、营销和数据分析等既有工作流。随着 Coding、Agent 及多模态应用发展，模型调用逐步由单轮对话延伸至长上下文、多步骤推理和工具调用，收入可以随客户数量、使用频率、任务复杂度及业务覆盖范围提升。据 SemiAnalysis《Anthropic 3Q26 Profit Over $1B》估算，2026 年 Anthropic 约 75%-85%的收入来自按量计费 API，订阅收入约占 ARR 的 15%，其中消费订阅约占 5%；OpenAI 在 2026Q1 超过 65%的收入来自订阅，消费端与企业端分别约占 ARR 的 60%和 40%，至 2026Q2 末可能转为约 40%和 60%。我们认为，Anthropic 以企业 API 为主的收入结构，使其商业化表现与 Token 调用量和企业使用深度形成了较为直接的联系。

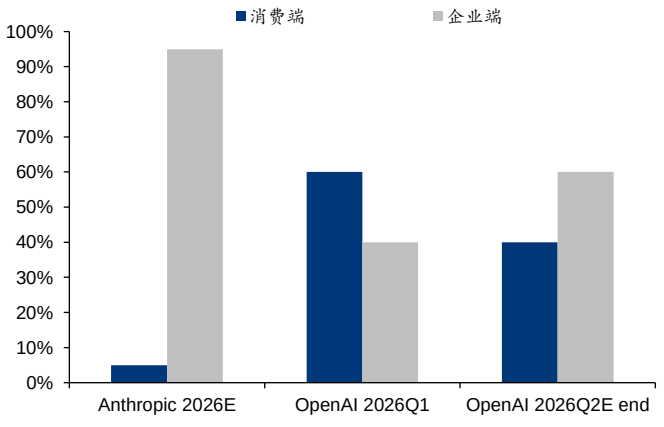
企业导向的收入结构有助于改善 Anthropic 的盈利模型。B2B/API 模式主要服务具有较高投资回报的开发者和企业客户，定价可以更多参考具体任务价值；大规模免费用户和固定价格订阅模式在使用强度提升时，则可能面临更明显的成本压力。据上述 SemiAnalysis 报告估算，Anthropic 预计 2026Q3 实现超过 10 亿美元的 GAAP EBIT，对应 EBIT 利润率约 6%，增长动力主要来自 B2B 和 API 业务，SemiAnalysis 据此将其视为较早验证盈利路径的头部 AI Lab 之一。与此同时，到 2030 年，Anthropic 训练与推理算力占比预计分别约为 48%和 52%，随着收入规模扩大，推理侧资源投入及其利用效率的重要性仍将提升。我们认为，Anthropic 盈利能力的改善可能同时来自企业客户较强的付费意愿、使用量增长以及较轻的免费消费端推理负担。

图表3：Anthropic 和 OpenAI 的收入结构（按结构）



资料来源：SemiAnalysis《Anthropic 3Q26 Profit Over $1B: The Anthropic IPO Financials Sneak Peak》（2026 年 7 月 8 日），华泰研究

图表4：Anthropic 和 OpenAI 的收入结构（按客户）

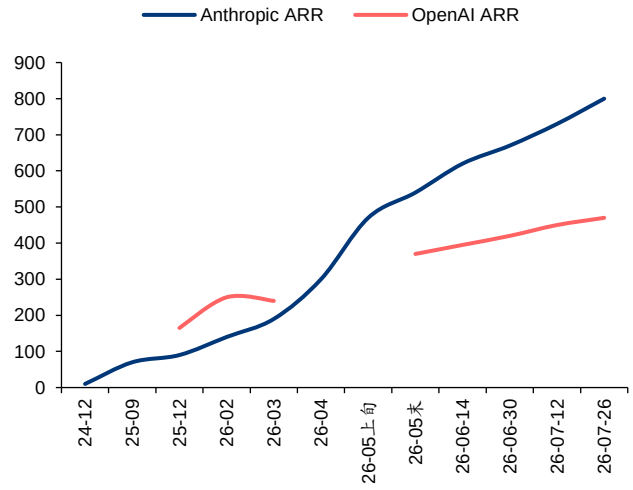


资料来源：SemiAnalysis《Anthropic 3Q26 Profit Over $1B: The Anthropic IPO Financials Sneak Peak》（2026 年 7 月 8 日），华泰研究

仅供内部参考，请勿外传！

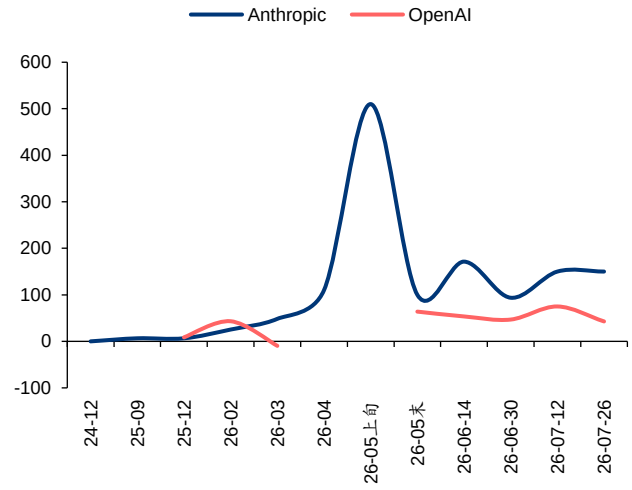
Claude Code 推动 Anthropic 年化运行收入快速增长。据 SemiAnalysis《Anthropic 3Q26 Profit Over $1B》估算，Claude Code 已占 GitHub 代码提交量的 7%以上。2026 年 4 月，Anthropic 年化运行收入约为 300 亿美元，超过同期 OpenAI 约 250 亿美元；5 月初进一步超过 470 亿美元，至 6 月 28 日达到约 690 亿美元，同期 OpenAI 约为 420 亿美元，两家公司合计约 1110 亿美元。Claude Code 通过嵌入企业开发流程，推动开发者持续产生长上下文、工具调用和输出 Token，使收入能够随使用频率与任务深度较快扩张。我们认为，Anthropic 在 2026 年 4 月超过 OpenAI，反映 Coding 与企业 API 已成为基础模型商业化的重要增长来源。

图表5： Anthropic 和 OpenAI 的 ARR



资料来源：SemiAnalysis《Anthropic 3Q26 Profit Over $1B: The Anthropic IPO Financials Sneak Peak》（2026 年 7 月 8 日）、YipitData（数据截至 2026 年 6 月 28 日），华泰研究

图表6： Anthropic 和 OpenAI 的 30 天净新增 ARR



注：30 天净新增 ARR 的计算方法是，将两个相邻时间点公布的 ARR 平均到每一天的 ARR，然后再乘以 30。

资料来源：SemiAnalysis《Anthropic 3Q26 Profit Over $1B: The Anthropic IPO Financials Sneak Peak》（2026 年 7 月 8 日）、YipitData（数据截至 2026 年 6 月 28 日），华泰研究

仅供内部参考，请勿外传！

Coding 能力如何成为商业化拐点

Coding 较普通聊天具备更清晰的商业化路径

Coding 的使用特征使其更容易形成可持续收入。开发者在代码补全、调试、测试、重构和项目级开发中具有较高使用频率，单用户通常会持续产生输入、输出、长上下文及工具调用 Token，Agent 任务还可能连续运行数十分钟甚至数小时。与此同时，Coding 任务的商业价值相对容易量化，企业可以通过开发周期、代码产出、缺陷率和研发人效评估投入回报，模型能力差异也会直接影响任务完成质量和工程效率，因此用户可能愿意为能力更强、稳定性更高的模型支付一定溢价。产品形态还可由个人代码补全逐步扩展至代码库理解、测试、审核和修复，并进一步进入企业采购。我们认为，Coding 能够将模型能力与可量化的生产力提升直接连接，可能成为 API 及 Agent 收入增长的重要场景。

普通聊天具备用户入口价值，但商业化约束相对更多。通用聊天产品通常包含较高比例的免费用户，单次问答的任务价值和付费回报较难量化，尤其在国内市场，消费者直接订阅意愿相对有限，因此用户规模增长未必同步转化为收入增长。与此同时，生成式 AI 每次对话均会产生实际推理成本，大规模免费调用可能持续占用算力资源，而消费者可以在多个免费产品之间切换，用户留存和价格提升均面临一定压力。Coding 产品在代码库上下文、工具配置和团队协作流程中的嵌入程度相对更深，模型切换通常还涉及工作环境及协作流程调整。我们认为，普通聊天更有利于建立用户入口和品牌认知，Coding 则更容易形成高频调用、企业采购和较高价值的 Token 收入。

Claude 通过持续迭代逐步建立 Coding 优势

Claude 的代码能力在 Claude 2 至 3 Opus 阶段持续积累。据 Anthropic 于 2023 年 7 月 11 日发布的 Claude 2 模型，Claude 2 在 HumanEval 上的得分由 Claude 1.3 的 56.0%提升至 71.2%，反映其单函数代码生成能力有所增强。Anthropic 于 2024 年 3 月 4 日发布的《Introducing the next generation of Claude》中的 Claude 3 Opus 进一步提升代码生成、复杂推理和长上下文能力，但整体仍定位为通用智能模型。Anthropic 于 2024 年 10 月 22 日发布的《Introducing computer use, a new Claude 3.5 Sonnet, and Claude 3.5 Haiku》所列统一测试结果显示，Claude 3 Opus 在 SWE-bench Verified 上的任务解决率约为 22%，表明其已经具备一定的代码库理解和问题修复能力，但在真实软件工程任务上的稳定性仍有提升空间。

Claude 3.5 Sonnet 构成 Coding 能力的重要拐点。据 Anthropic 于 2024 年 6 月 20 日发布的 Claude 3.5 Sonnet 模型，初代 Claude 3.5 Sonnet 在内部 Agent Coding 评测中的任务解决率达到 64%，高于 Claude 3 Opus 的 38%，并能够在工具支持下自主读取、编辑、执行和调试代码；同时，其运行速度约为 Claude 3 Opus 的 2 倍，价格维持 Sonnet 档位，有助于代码能力以相对较低的成本进入开发者产品。Anthropic 于 2024 年 10 月 22 日发布的升级版本进一步将 SWE-bench Verified 得分由 33.4%提升至 49.0%，同期 Claude 3 Opus 约为 22%，工具调用和多步骤编程能力亦有所增强。随后 GitHub Copilot 引入 Claude 3.5 Sonnet，表明其代码能力开始获得主流开发者平台采用，Coding 口碑和第三方集成由此逐步形成。

Agent 化产品推动 Coding 优势向商业化转化。Anthropic 于 2024 年 10 月 22 日推出的 Computer Use 使 Claude 能够通过屏幕、鼠标和键盘执行外部操作，将代码生成能力延伸至工具调用和多步骤任务。2025 年 2 月 24 日，Claude 3.7 Sonnet 加入扩展思考能力，Anthropic 同期推出 Claude Code 研究预览，使模型能够搜索代码库、修改文件、运行测试并使用命令行工具，Coding 由单次问答逐步进入完整软件工程流程。此后，Claude Code 在 2025 年 2 月 24 日至 5 月 22 日逐步产品化，Sonnet 4 在 SWE-bench 上的得分进一步提升至 72.7%，后续迭代重点转向长任务执行、代码库理解和 Agent 可靠性。我们认为，2024 年 6 月的 Claude 3.5 Sonnet 可视为 Coding 能力拐点，10 月的版本升级是优势得到进一步验证的节点，2025 年 2 月至 5 月的 Claude Code 则构成产品化和商业化拐点。

仅供内部参考，请勿外传！

图表7：Claude 模型的主要更新节奏

时间	模型/产品节点	Coding 能力与排名信号	第三方采用
2023 年 7 月	Claude 2 发布	HumanEval 约 71.2%，较早期版本提升，说明已具备较强代码生成能力	以开发者试用为主，尚未形成广泛平台级采用
2024 年 3 月	Claude 3 Opus 发布	复杂推理、长上下文和代码生成能力继续提升，SWE-bench Verified 约 22%	开始进入部分开发者工作流，但尚未形成“Coding 首选”口碑
2024 年 6 月	Claude 3.5 Sonnet 发布	内部 Agent Coding 任务完成率约 64%，显著高于 Claude 3 Opus 约 38%；速度更快、成本更低	开发者口碑明显提升，Claude 开始在 Coding 场景形成代表性优势
2024 年 7-9 月	3.5 Sonnet 进入主流开发工具	在真实编程任务中的可用性明显增强	Cursor 等产品将 Claude 3.5 Sonnet 作为核心模型之一，第三方应用带动调用量上升
2024 年 10 月	Claude 3.5 Sonnet 升级	SWE-bench Verified 提升至约 49.0%，较初期版本和 Opus 均明显提高	GitHub Copilot 等主流平台开始引入 Claude 3.5 Sonnet
2024 年 10 月	Computer Use 发布	能力由代码生成进一步延伸至工具调用、界面操作和多步骤任务	开发者开始将 Claude 用于更复杂的 Agent 式任务
2025 年 2 月	Claude 3.7 Sonnet+Claude Code 研究预览	Claude 从模型能力转向完整软件工程流程，支持读代码、改文件、运行测试和调用命令行	开发者开始直接使用 Anthropic 自有 Coding 产品，而非仅通过第三方接入
2025 年 5 月	Claude Code 正式开放	Coding Agent 能力进一步成熟	Claude Code 接入更多开发者和团队场景
2025 年 9 月	Claude Code 规模化扩张	产品从个人开发工具向团队协作扩展	企业订阅和团队使用开始增加
2025 年 11 月	Claude Code 继续放量	Coding Agent 工作流进一步稳定	企业采用加速
2026 年 Q1	Claude Code 成为商业化核心引擎	Claude 在 Coding 场景的优势开始直接反映为企业工作流基础设施价值	开发者和企业双侧采用扩大
2026 年 2 月以后	Claude Code 企业化深化	多 Agent、复杂工程任务和长时程任务能力继续增强	企业客户贡献超过产品收入的一半

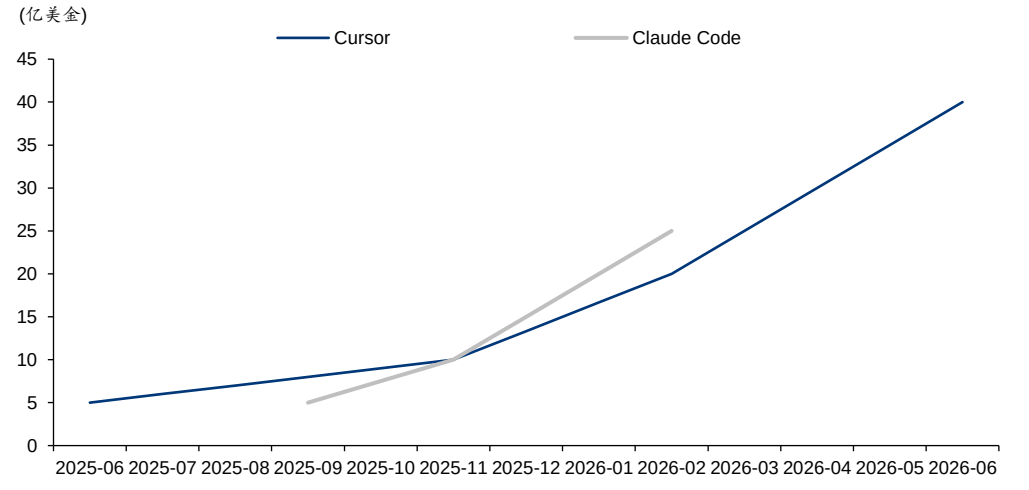
资料来源：Anthropic 官网、华泰研究

Cursor 带动 Claude API，Anthropic 躬身入局

Cursor 在产品起量阶段较多依托 Claude 模型，并验证了 Coding 能力的高频商业化需求。据 Cursor 于 2024 年 3 月 12 日发布的更新日志《Claude Support》，平台接入 Claude 3 Opus，随后陆续接入 Claude 3.5 Sonnet，平台于 2024 年 3 月接入 Claude 3 Opus，2024 年 7 月将 Claude 3.5 Sonnet 设为默认模型，并通过代码库索引、上下文筛选和 Composer 等能力，实现跨文件读取、代码修改及多步骤任务执行；此后又陆续接入 Claude 3.7 Sonnet 和 Claude 4 系列。Cursor 承担开发者交互、上下文管理和任务执行，Claude 提供代码理解、生成及推理能力，使单次用户请求转化为多轮模型调用、长上下文输入和持续 Token 消耗。据 Cursor《Series C and Scale》（2025 年 6 月 6 日）和《Past, Present, and Future》（2025 年 11 月 13 日），公司 ARR 分别超过 5 亿美元和 10 亿美元；据 Forbes 2026 年 6 月 8 日报道，其年化收入由 2026 年 2 月约 20 亿美元增至 4 月末约 30 亿美元，并于 6 月初超过 40 亿美元，截至 2026 年 8 月 3 日尚未出现更晚的可信公开口径。我们认为，Cursor 早期增长为 Claude Coding 能力嵌入开发者工作流后的商业价值提供了较强验证。

Claude Code 正式发布后，Anthropic 开始直接承接 Coding Agent 需求。据 Anthropic 于 2025 年 2 月 24 日发布的《Claude 3.7 Sonnet and Claude Code》及 2025 年 5 月 22 日发布的 Claude Code 正式版公告，Claude Code 先以研究预览形式推出，随后正式开放。根据 Anthropic 2025 年 9 月 2 日、12 月 3 日和 2026 年 2 月 12 日发布的公司公告，Claude Code 的 run-rate revenue 依次超过 5 亿美元、10 亿美元和 25 亿美元；同期周活跃用户数量较年初增长 1 倍，企业订阅数量增至年初的 4 倍，企业客户贡献超过产品收入的一半。我们认为，Cursor 早期验证了 Claude Coding 能力的市场需求，Claude Code 则进一步将部分第三方 API 需求转化为 Anthropic 直接掌握的产品收入、用户入口和企业客户关系。

图表8：Claude Code 和 Cursor 的 ARR 比较



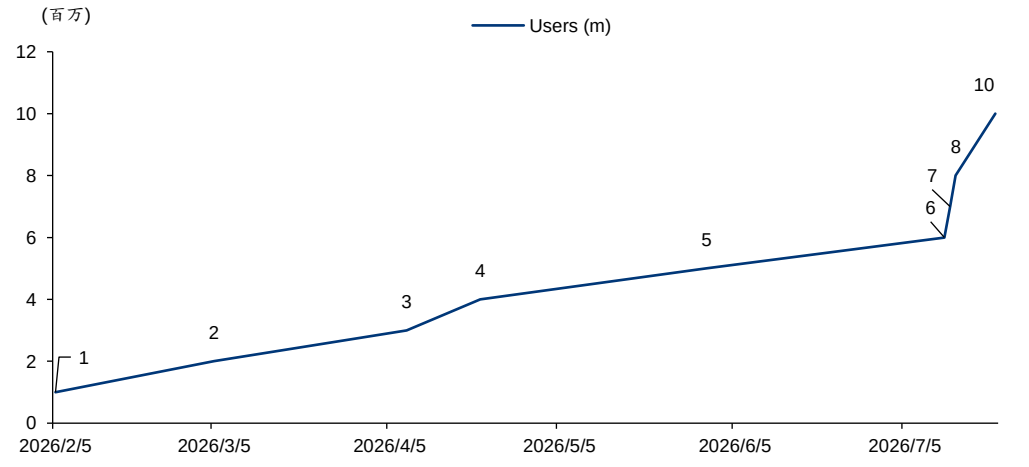
资料来源：Anthropic 官网、Forbes、华泰研究

Claude Code 正由个人命令行工具向企业级 Agent 平台延伸。产品初期主要支持在终端内读取代码库、修改文件和运行命令，正式发布后进一步接入 VS Code、JetBrains、GitHub Actions 及 Claude Code SDK，并持续增加子 Agent、后台任务、Hooks、MCP、插件、网页版并行任务、Agent Teams 和 Claude Code Security 等能力，产品边界由辅助编程逐步延伸至自主执行软件工程任务。Cursor 主要通过 IDE、Composer 及代码库上下文管理承接开发者交互，并同时支持 Claude 和其他模型；Claude Code 则逐步向 IDE、企业工作流和多 Agent 协作扩展。我们认为，两者仍具有渠道合作和能力互补关系，但随着 Anthropic 加强对开发者入口及企业工作流的直接覆盖，产品边界可能进一步重叠，其关系正逐步转向渠道合作与产品竞争并存。

OpenAI Codex 快速上量，Coding 进入双强竞争

Codex 用户规模在桌面端发布后快速增长。据 OpenAI 于 2026 年 6 月 2 日发布的《Codex is becoming a productivity tool for everyone》，Codex 周活跃用户已超过 500 万，较 2 月初桌面端发布时增长超过 6 倍。Peter Gostev 于 2026 年 7 月 14 日在 X 平台发布的修正图表显示，Codex 与 ChatGPT Work 合计活跃用户的增长轨迹进一步上移；据 OpenAI Codex 产品负责人 Tibo Sottiaux 于 2026 年 7 月 14 日及 21 日在 X 平台发布的信息，二者合计活跃用户于 7 月 14 日达到 800 万，并于 7 月 21 日增至 1000 万。该口径包含 ChatGPT Work，仍能够反映 OpenAI Agent 产品的分发和上量速度。我们认为，Codex 已成为 Coding Agent 市场的重要参与者。

图表9：Codex 活跃用户增长情况



资料来源：OpenAI 官网、X 平台、华泰研究

仅供内部参考，请勿外传！

产品入口与专用模型的连续升级支撑 Codex 上量。Codex CLI 于 2025 年 4 月进入本地终端，5 月进一步推出由 codex-1 驱动的云端 Agent，支持并行完成功能开发、代码库问答、Bug 修复、测试和 Pull Request 等任务；同年 9 月至 10 月，OpenAI 陆续上线 IDE 插件、代码审核、Slack 集成、Codex SDK 及企业管理工具，逐步打通 CLI、IDE、云端和企业协作入口。2026 年 2 月推出的桌面端进一步支持多 Agent 并行和长时程任务，随后扩展至 Windows、Computer Use、浏览器、插件、记忆、远程开发环境及移动端远程控制，7 月又并入新版 ChatGPT 桌面应用。我们认为，Codex 通过连续完善产品入口，逐步形成覆盖个人开发者和企业工作流的分发体系。

Codex 正由编程助手向长时程任务执行平台延伸。模型体系由 codex-1 逐步升级至 GPT-5-Codex、GPT-5.1-Codex-Max、GPT-5.2-Codex 和 GPT-5.3-Codex，能力重点由代码生成和审核扩展至大型代码库理解、上下文压缩、多小时调试、项目级重构、工具调用、部署监控及计算机操作。桌面端通过独立线程和 Git worktree 管理多个并行 Agent，并结合 Skills、Automations 及 Symphony 等编排方式，将团队规范、重复任务和外部任务系统纳入统一流程。据 OpenAI 于 2026 年 6 月 2 日发布的《Codex is becoming a productivity tool for everyone》，2026 年 5 月超过 70%的用户曾向 Codex 提交预计需要人工工作 1 小时以上的任务，其使用场景亦开始延伸至报告、表格、演示文稿、研究和数据分析。我们认为，Codex 的产品边界正在由 Coding Agent 向通用任务执行平台扩展。

OpenAI 与 Anthropic 分别验证了两条 Coding 商业化路径。Anthropic 依托模型 Coding 能力、长上下文和工具调用优势，先通过 Cursor 等第三方产品扩大 API 调用和 Token 消耗，再以 Claude Code 建立自有产品入口，推动企业采购和年度化收入增长，形成“模型能力提升—第三方生态采用—API 需求增长—自有入口建设—企业市场放量”的发展路径。OpenAI 则依托 ChatGPT 用户基础、开发者平台和企业渠道，将既有用户进一步导向 Codex，并通过专用模型、CLI、IDE、云端和桌面端形成全栈分发体系。我们认为，Anthropic 较早验证了高价值场景驱动 API 商业化的可行性，OpenAI 则进一步体现了 C 端入口与企业级 Agent 产品之间的协同价值。

Coding Agent 市场正在进入双强竞争阶段。Anthropic 率先在 Coding 场景形成开发者口碑和商业化突破，并通过 Claude Code 增强对产品体验、客户关系和收入的控制；OpenAI 则凭借 ChatGPT 入口、开发者基础、企业渠道和完整产品矩阵推动 Codex 快速上量，双方的竞争范围已由模型代码能力扩展至长时程任务执行、企业工作流渗透、高价值 Token 消耗和开发者入口。我们认为，当前尚难判断双方的长期竞争格局，但 OpenAI 的分发优势与 Anthropic 的场景积累均已形成较强基础，Coding 或将成为两家公司后续商业化竞争的核心方向之一。

仅供内部参考，请勿外传！

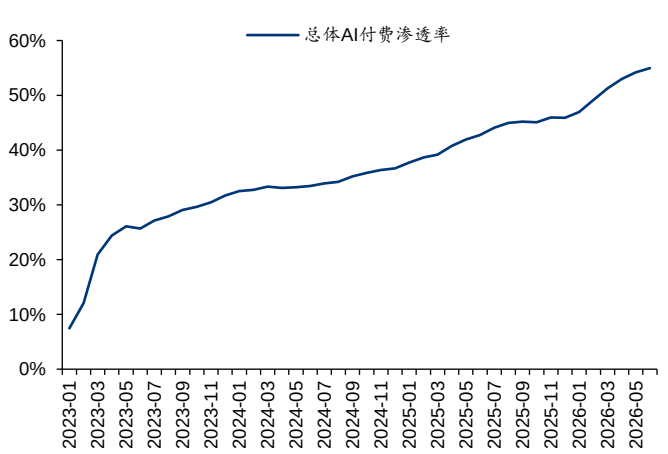
基础模型或会多厂商共存

Ramp AI Index：企业模型份额持续变化

据 Ramp Economics Lab 于 2026 年 6 月 26 日发布的《How much does it cost to be AI-pilled?》及 Ramp AI Index 历史月度数据，企业模型采用格局仍在快速变化。Ramp 基于企业卡和账单支付记录统计付费采用率，反映使用某厂商产品的企业比例；同一家企业可以同时采用多家模型，因此各厂商采用率无需加总至 100%，也不等同于收入或 Token 份额。2025 年 7 月，OpenAI 和 Anthropic 采用率分别为 35.1%和 11.1%，12 月分别升至 36.8%和 16.7%；2026 年 1—3 月，Anthropic 进一步由 19.5%升至 30.6%，并于 4 月以 34.4%超过 OpenAI 的 32.3%。截至 2026 年 6 月，Anthropic 和 OpenAI 采用率分别达到 42.4%和 39.5%。Ramp 于 6 月调整统计方法，以更充分捕捉企业支出，因此跨期数据更适合观察趋势和相对排名。

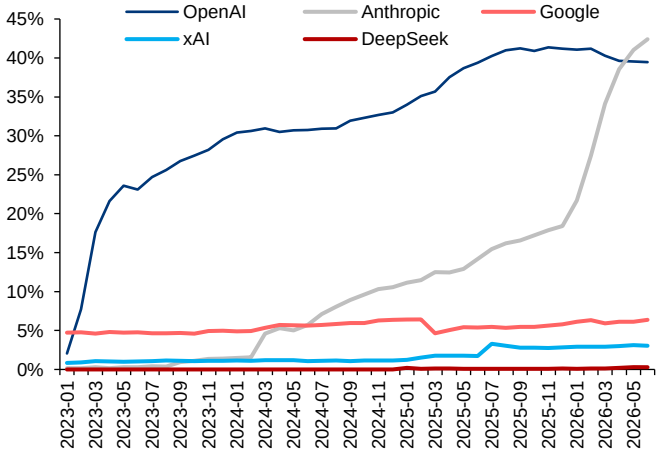
企业模型采用差异与行业结构、客户成熟度及产品能力密切相关。据 Ramp Economics Lab 于 2026 年 4 月 11 日发布的《What drives AI adoption?》（数据截至 2026 年 3 月），Anthropic 在信息技术、金融和专业服务行业的采用率分别为 63%、52%和 47%，同期 OpenAI 分别为 54%、46%和 44%；OpenAI 在其他行业仍保持领先。在 VC 支持企业中，Anthropic 和 OpenAI 采用率分别达到 66%和 59%，反映高技术密度企业具有更明显的多模型采用特征。我们认为，企业模型采用格局仍未固化，模型能力、使用成本和分发渠道变化均可能推动客户采用率持续迁移。

图表10：总体 AI 付费渗透率（%）



资料来源：Ramp AI Index、华泰研究

图表11：主要模型厂商付费采用率（%）



资料来源：Ramp AI Index、华泰研究

仅供内部参考，请勿外传！

企业正在同时使用多款模型

企业多模型采购正在成为常态化技术架构。据 Ramp Economics Lab 于 2026 年 2 月 11 日发布的《Anthropic's breakout month》(基于 2026 年 1 月 Ramp 数据)，约 79%的 Anthropic 客户同时向 OpenAI 付费，16%的 Ramp 企业同时采用 OpenAI 和 Anthropic，较一年前的 8%提高一倍，表明 Anthropic 的增长较大程度来自企业增加第二家模型供应商。据 a16z 于 2025 年 6 月 10 日发布的《How 100 Enterprise CIOs Are Building and Buying Gen AI in 2025》，使用 5 款及以上模型的企业占比由 29%升至 37%，据 a16z 于 2026 年 1 月 30 日发布的《Leaders, gainers and unexpected winners in the Enterprise AI arms race》，81%的企业已在测试或生产环境使用 3 个及以上模型家族。F5 于 2026 年 5 月 5 日发布的《The State of AI Delivery and Security in 2026》显示，78%的组织采用多款模型，管理分布式推理服务的企业平均涉及约 7 款模型；Docker 于 2026 年 3 月 12 日发布的《Flexibility Over Lock-In: The Enterprise Shift in Agent Strategy》显示，46%的受访组织在 Agent 中使用 4-6 款模型，仅 2%依赖单一模型。尽管不同调查的样本和口径存在差异，其结果均表明多模型采购正在由试验行为逐步转向企业常态。

企业通常在探索阶段扩大测试范围，并在生产环境形成相对集中的模型组合。据 Melissa Z. Pan 等于 2025 年 12 月 2 日在 arXiv 发布的《Measuring Agents in Production》，尚未部署的试验和研究型 Agent 更倾向组合 4 款及以上模型，进入生产环境后，企业通常会减少模型数量，以控制成本、延迟、治理复杂度及故障风险。大型企业由于业务部门、应用场景、云平台环境和数据合规要求更加多样，通常具有更强的多供应商需求；中小企业则可能更多依赖统一套件或少数模型，降低集成和运维成本。因此，企业级多模型主要体现为整体保留多家供应商，并为不同工作流配置相对稳定的模型组合，并不意味着每项任务均需同时调用大量模型。

任务差异是企业采用多模型的核心原因。具体来看：1）复杂 Coding、长链路 Agent、深度推理及失败代价较高的任务，更倾向采用能力较强的前沿模型，其模型成本在整体任务价值中的占比相对有限。2）成熟、可重复及高并发任务，更适合采用开源模型、小模型或低价 Flash 模型，以降低成本和延迟。3）涉及敏感数据、地区合规和供应安全的任务，可能采用主权模型、私有部署模型或特定云平台模型。4）语音、客服、金融、Coding 和数据分析等场景，可能采用垂直模型或专用 Agent，实际效果更多取决于工作流数据、验证机制和系统集成能力。企业最终会综合比较模型能力、价格、延迟、稳定性、数据安全、服务可用性及既有技术栈。

部分企业可能遵循"先闭后开"的模型迁移路径，并通过动态路由持续调整供应商结构。在业务探索阶段，企业可先使用前沿模型验证产品形态和工作流；当任务流程逐渐成熟并积累真实数据后，再通过后训练、蒸馏、开源模型或自研模型降低推理成本。前沿模型既可能长期承担复杂任务，也可以为垂直模型训练和工作流优化提供能力基础。企业还可通过路由系统将复杂任务分配给前沿模型，将标准化任务转向低价模型，并保留第二供应商以应对配额、服务中断、地区合规和模型版本风险。我们认为，基础模型市场可能形成多家头部厂商长期共存的竞争格局，各厂商在不同任务中建立差异化优势，市场份额则随模型迭代、价格变化和企业工作流迁移持续调整。

仅供内部参考，请勿外传！

智谱与 MiniMax 的路径如何演变

MiniMax 早期商业化更偏 C 端产品和海外市场

MiniMax 早期商业化更偏向 AI 原生应用和产品探索。公司围绕 Talkie/星野、海螺 AI、MiniMax 及语音产品形成 C 端矩阵，覆盖 AI 陪伴、社交、内容创作和多模态生成等场景，同时以开放平台承接开发者和企业需求。据 MiniMax 于 2026 年 3 月 2 日发布的 2025 年度业绩公告，公司总收入由 2023 年的约 346 万美元增至 2025 年的约 7,904 万美元，AI 原生产品收入占比由约 22%提升至约 67%。我们认为，MiniMax 早期主要通过多款产品验证交互形态、用户需求和 PMF，与 OpenAI 依托 ChatGPT 获取真实用户反馈的产品探索思路具有一定相似性。

海外市场是 MiniMax 早期产品扩张的重要支撑。据 MiniMax 于 2026 年 3 月 2 日发布的 2025 年度业绩公告，公司海外收入占比由 2023 年的不足 20%提升至 2025 年的约 73%，收入地域结构已由国内为主转向全球市场；截至 2025 年 9 月底，AI 原生产品累计用户超过 2 亿，Talkie/星野和海螺 AI 构成主要用户入口。海外市场扩大了 MiniMax 的可触达用户范围，并为 AI 陪伴、内容生成和多模态产品提供了更多付费环境与文化场景。我们认为，全球化分发与多产品试错共同构成 MiniMax 早期商业化路径的重要特征。

图表12：MiniMax 收入结构拆分

	2023 年	2024 年	2025 年
AI 原生产品—收入	758	21,805	53,075
同比增速		2776.6%	143.4%
收入占比	21.9%	71.4%	67.2%
毛利	-2,882	-1,776	
毛利率	-380.2%	-8.1%	
AI 企业服务—收入	2,702	8,718	25,963
同比增速		222.6%	197.8%
收入占比	78.1%	28.6%	32.8%
毛利	2,028	5,514	
毛利率	75.1%	63.2%	
公司整体—总收入	3,460	30,523	79,038
总毛利	-854	3,738	20,079
同比增速		782.0%	158.9%
总毛利率	-24.7%	12.2%	25.4%
中国内地—收入	2,797	9,217	21,375
收入占比	**80.8%**	**30.2%**	**27.0%**
新加坡—收入	1	11,455	
收入占比	0.0%	37.5%	
美国—收入	575	4,999	
收入占比	16.6%	16.4%	
其他—收入	87	4,852	
收入占比	2.5%	15.9%	
世界其他地区—收入			57,663
收入占比			73.0%

资料来源：MiniMax 招股书、Wind、华泰研究

C 端用户规模与商业价值并不一定同步增长。据《晚点 LatePost》于 2026 年 6 月 16 日发布的《字节跳动的 AI 账本：豆包每天不足百万收入、Seedance 毛利 70%》，截至 2026 年上半年，豆包日活跃用户已超过 2 亿，但日收入不足百万元；据其测算，2026 年 5 月豆包日均算力成本达到数千万元；同期 Seedance 年化收入约 20 亿美元，收入主要来自企业客户，毛利率约 70%。通用 C 端应用有助于积累用户、品牌和使用反馈，但免费或低付费模式可能带来较高推理成本；视频生成、Coding、API 和企业服务等场景的任务价值及付费回报通常更加清晰。我们认为，MiniMax 的 C 端产品矩阵仍具有用户入口和场景探索价值，商业化重心可能逐步向高价值任务和企业需求延伸。

仅供内部参考，请勿外传！

MiniMax 与 OpenAI 的相似性主要体现在早期产品探索机制。两家公司均重视产品入口、用户规模和真实反馈，并通过消费端产品探索新的 AI 交互方式；在具体路径上，ChatGPT 以通用助手为核心入口，MiniMax 则更多通过 AI 陪伴和视频生成等垂直产品触达用户，其收入结构中充值和营销服务的权重也相对更高。此外，两家公司在资金、算力、市场环境和企业客户基础方面存在差异。我们认为，MiniMax 更适合被理解为一条“早期偏重 C 端产品探索，后续逐步提高多模态、API 及企业业务权重”的发展路径。

智谱早期商业化更偏企业、本地化部署和项目制

智谱早期商业化以企业客户和本地化部署为主。据智谱 2025 年 12 月 30 日发布的招股书及 2026 年 3 月 31 日发布的 2025 年度业绩公告，公司 2022-2025 年收入分别约为 0.57 亿元、1.25 亿元、3.12 亿元和 7.24 亿元，本地化部署收入占比分别为 95.5%、90.4%、84.5%和 73.7%，云端部署占比则由 4.5%逐步提高至 26.3%。本地化部署主要面向企业及政企客户，将 GLM 模型部署于客户本地环境，并结合客户数据、算力基础设施和业务需求进行适配。我们认为，该模式帮助智谱较早完成企业市场的商业化验证，但收入扩张相对依赖项目获取、实施及交付能力。

图表13：智谱收入结构拆分

	2022 年	2023 年	2024 年	2025 年
公司整体—总收入	57,409	124,538	312,414	724,334
同比增速		116.9%	150.9%	131.9%
收入成本	-26,049	-44,056	-136,525	-427,678
毛利	31,360	80,482	175,889	296,656
毛利率	54.6%	64.6%	56.3%	41.0%
本地化部署—收入	54,815	112,614	263,930	533,955
同比增速		105.4%	134.4%	102.3%
收入占比	95.5%	90.4%	84.5%	73.7%
收入成本	25,429	35,833	89,674	273,299
毛利	29,386	76,781	174,256	260,656
毛利率	53.6%	68.2%	66.0%	48.8%
云端部署—收入	2,594	11,924	48,484	190,379
同比增速		359.7%	306.6%	292.7%
收入占比	4.5%	9.6%	15.5%	26.3%
收入成本	620	8,223	46,851	154,379
毛利	1,974	3,701	1,633	36,000
毛利率	76.1%	31.0%	3.4%	18.9%
本地化部署—中国内地—收入	54,815	112,614	262,479	
收入占比	100.0%	100.0%	99.5%	
本地化部署—东南亚—收入			1,080	
收入占比			0.4%	
本地化部署—香港—收入			371	
收入占比			0.1%	
本地化部署—总计—收入	54,815	112,614	263,930	

资料来源：智谱招股书、Wind、华泰研究

2025 年智谱标准化产品收入占比进一步提高。据智谱于 2026 年 3 月 31 日披露的 2025 年度业绩公告，公司企业级通用大模型收入约 3.65 亿元，占比 50.4%，主要采用私有化部署模式；企业级智能体收入约 1.65 亿元，占比 22.9%；开放平台及 API 收入约 1.90 亿元，占比 26.3%；技术服务及其他收入占比较低。其中，API 及相关收入由 2024 年的约 0.48 亿元增至 2025 年的 1.90 亿元，同比增长约 293%。我们认为，智谱正在形成“本地化部署基本盘+API 及 Agent 标准化产品”的并行商业化结构，收入的可复制性和使用量驱动特征逐步增强。

仅供内部参考，请勿外传！

智谱与 Anthropic 均偏向企业市场，但商业化方式仍有差异。据 SemiAnalysis《Anthropic 3Q26 Profit Over $1B》估算，Anthropic 约 75%-85%的 ARR 来自按使用量计费的 API，消费订阅约占 5%，并主要通过云平台渠道、Claude Code、企业订阅及开发者生态扩大 Token 调用。智谱早期企业收入较多来自私有化部署、客户本地算力适配和行业解决方案，单个项目对实施及交付能力的要求相对较高。随着 API 和 Agent 业务占比提升，智谱的商业化路径正逐步向标准化模型服务和使用量驱动模式延伸。我们认为，Anthropic 的标准化程度相对更高，智谱则处于由项目交付向标准化模型服务快速迁移的阶段。

两家公司商业化重心逐步向 API、Coding 和 Agent 收敛

智谱与 MiniMax 正由不同起点向相近商业化方向延伸。智谱早期更侧重政企客户、私有化部署和企业模型服务，近年来进一步拓展自有 API、Coding Plan 和 Agent 能力；MiniMax 则从消费级应用和多模态产品起步，随后通过开放平台、Token Plan、MiniMax Code 及 MiniMax Agent 扩大开发者和企业市场。智谱与 MiniMax 均已形成 API 按量计费和固定额度套餐，并通过 Coding、Agent 及第三方开发工具提高模型调用频率和单项任务 Token 消耗。我们认为，两家公司的商业化重心均在向标准化 API、高价值生产力场景和使用量驱动收入迁移。

图表14： 智谱和 MiniMax 旗舰模型价格

公司	当前旗舰模型	模型定位	上下文	API 按量价格：输入/输出/缓存读取
智谱	GLM-5.2	Coding、Agent 及长时程工程任务	1M；最大输出 128K	8/28/2 元/百万 Token
MiniMax	MiniMax-M3	Coding、Agent 及原生多模态任务	最高 1M；保证至少 512K	≤512K：2.1/8.4/0.42 元；>512K：4.2/16.8/0.84 元/百万 Token

资料来源：智谱官网、MiniMax 官网、华泰研究

图表15： 智谱和 MiniMax Coding Plan 套餐对比

公司	套餐	当前月费	5 小时额度	每周额度/峰值时段参考并发数	主要可用模型
智谱	Lite	118 元	2,000 积分	10,000 积分	GLM-5.2、GLM-5-Turbo、GLM-4.7
智谱	Pro	538 元	12,000 积分	60,000 积分	同上
智谱	Max	1,078 元	28,000 积分	140,000 积分	同上
MiniMax	Plus	49 元	5 小时滚动窗口	约 3—4 个并发 Agent	M3、M2.7 及部分多模态模型
MiniMax	Max	119 元	5 小时滚动窗口	约 4—5 个并发 Agent	同上
MiniMax	Ultra	469 元	5 小时滚动窗口	约 6—7 个并发 Agent	同上

资料来源：智谱官网、MiniMax 官网、华泰研究

智谱已形成 Coding 能力提升带动 API 调用增长的初步链条。据智谱 2025 年 7 月 28 日发布的《GLM-4.5: Reasoning, Coding, and Agentic Abilities》，GLM-4.5 开始统一训练推理、Coding 和 Agent 能力，GLM-4.6 及 GLM-4.7 进一步强化代码生成、工具调用和复杂任务执行；此后 GLM-5 将产品定位由 Vibe Coding 推进至 Agentic Engineering，GLM-5.1 和 GLM-5.2 继续提升代码库理解及长时程任务能力。据智谱于 2026 年 6 月 16 日发布的《GLM-5.2: Built for Long-Horizon Tasks》，GLM-5.2 在 Terminal-Bench 2.1 上的得分由上一代的 63.5 提高至 81.0，SWE-bench Pro 由 58.4 提高至 62.1。据智谱 2025 年度业绩说明会（2026 年 3 月 31 日），智谱 2026 年一季度 API 价格累计提高约 83%，同期调用量增长约 400%，反映模型能力和 Coding 需求提升后，价格与使用量仍可同步增长。

智谱和 MiniMax 的 API 收入扩张已开始反映在 ARR 指标中。据智谱 2025 年度业绩说明会（2026 年 3 月 31 日），智谱 2026 年 3 月 MaaS API ARR 约为 2.5 亿美元；据 36 氪 2026 年 7 月 16 日发布的《独家丨智谱 ARR 达到 10 亿美元，半年增长 15 倍》，截至 2026 年 7 月 ARR 约为 10 亿美元，2026 年 1 月至 7 月增长约 15 倍。MiniMax 同期亦在加快开发者产品布局，据 MiniMax 2025 年度业绩会（2026 年 3 月 2 日）及 2026 年 5 月 28 日公开业务数据，公司 2026 年 2 月 ARR 已超过 1.5 亿美元，此后两个月增长超过 100%，对应规模超过 3 亿美元。另据 2026 年 7 月 4 日发布的 MiniMax 高盛电话会纪要，MiniMax 2026 年 4 月 ARR 已达到约 4 亿美元，管理层对年底实现至少 10 亿美元 ARR 保持信心。我们认为，两家公司均在形成“模型能力提升—开发者采用—Token 调用增长—企业采购与 ARR 提升”的商业化链条。

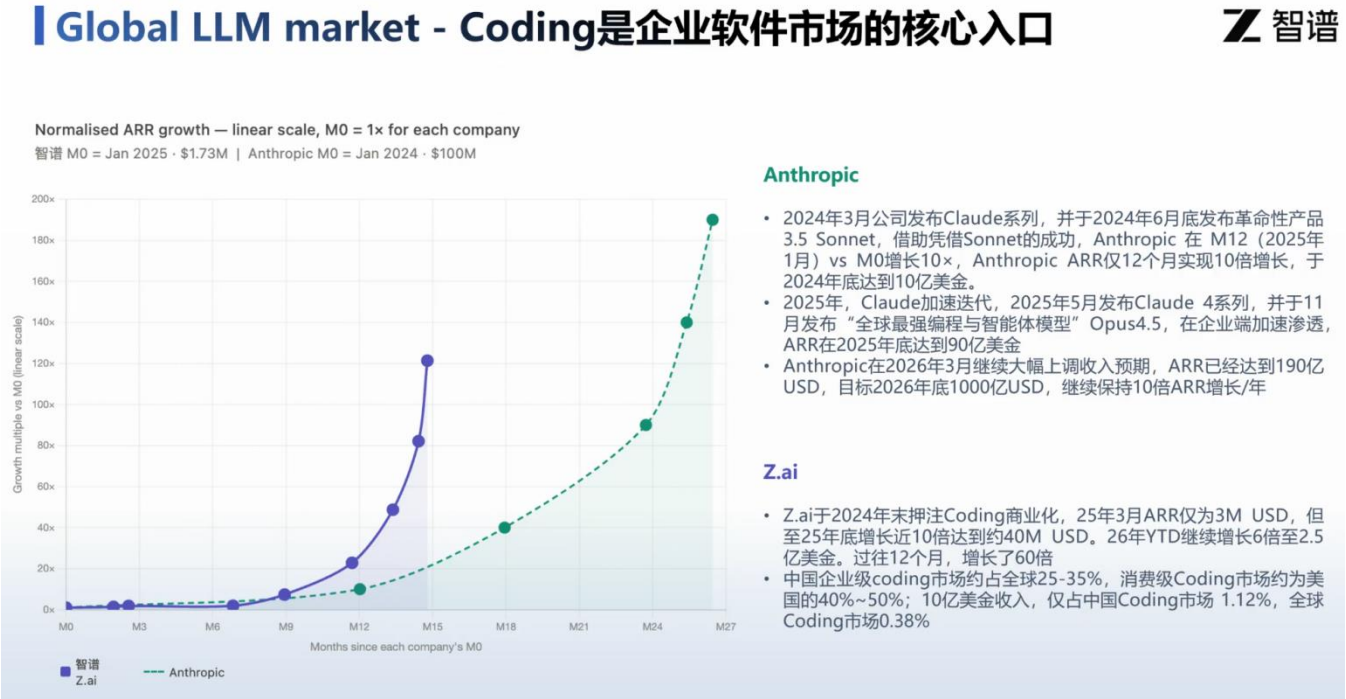
仅供内部参考，请勿外传！

图表16：智谱与 MiniMax ARR 进展

时间	智谱 ARR（亿美元）	MiniMax ARR（亿美元）	来源及说明
2026 年 2 月	—	超过 1.5	MiniMax 2025 年度业绩会（2026 年 3 月 2 日）
2026 年 3 月	约 2.5	—	智谱 2025 年度业绩说明会（2026 年 3 月 31 日）
2026 年 4 月	—	约 4	MiniMax 高盛电话会纪要（2026 年 7 月 4 日）
2026 年 7 月	约 10	—	36 氪《独家丨智谱 ARR 达到 10 亿美元，半年增长 15 倍》（2026 年 7 月 16 日）
2026 年末目标	—	至少 10	MiniMax 高盛电话会纪要（2026 年 7 月 4 日）

资料来源：智谱官网、MiniMax 官网、36 氪、华泰研究

图表17：智谱和 Anthropic ARR 增速对比



资料来源：智谱 25 年报 call back、华泰研究



仅供内部参考，请勿外传！

未来基础模型公司最应该看什么

基础模型公司的竞争力可从模型能力、商业化质量、市场份额及产品分发四个维度持续跟踪。模型能力决定产品上限和单位成本，商业化指标验证客户付费意愿，市场份额反映竞争优势的稳定性，产品与分发则决定模型能力能否转化为规模化用户和收入。

模型能力仍是基础模型公司竞争力和商业化效率的基础。评价体系不宜依赖单一 Benchmark，后续应重点跟踪：1）Coding、Agent、推理、多模态、长上下文及工具调用能力。2）真实代码库修改、复杂任务执行和长时程任务完成率。3）单位任务成本、响应速度、并发吞吐及服务稳定性。4）模型架构、训练方法和推理系统的技术进展。据 DeepSeek-AI 于 2026 年 4 月 26 日发布的《DeepSeek-V4: Towards Highly Efficient Million-Token Context Intelligence》及 Xinyu Cheng 等于 2026 年 7 月 6 日发布的《DSpark: Confidence-Scheduled Speculative Decoding with Semi-Autoregressive Generation》CSA/HCA 混合注意力通过压缩 KV Cache 和稀疏检索降低百万 Token 长上下文成本，DSpark 则通过半自回归生成和动态验证调度提高候选 Token 接受长度，改善实际服务中的吞吐和延迟。相关创新表明，底层技术进展可以同时影响模型能力、推理成本和产品体验。我们认为，后续应更加关注新技术能否在真实生产环境中提高任务完成率、单位算力产出和服务稳定性。

ARR 是观察基础模型商业化进展的重要指标，但仍需结合收入构成和单位经济性判断增长质量。基础模型公司的收入来源包括 API、企业模型服务、消费订阅、Coding、Agent、视频及多模态产品、开放平台、海外业务和本地化项目，不同收入模式在标准化程度、客户留存和毛利率方面存在差异。后续应重点跟踪：1）ARR 规模、增速及季度新增 ARR。2）API 实际消耗、预充值和合同额之间的差异。3）客户留存率、净收入留存率 NRR 及大客户集中度。4）Coding、Agent 等高价值场景的收入贡献。5）毛利率和推理成本。6）项目制收入向标准化、可复制收入转化的进度。我们认为，ARR 能够较快反映业务起量速度，真实 Token 消耗、客户留存和毛利率能否同步改善，将决定商业模式的可持续性。

多模型共存背景下，持续份额及其变化趋势较单一时点排名更具研究价值。海外已有 Ramp AI Index 等指标持续跟踪企业付费采用率，可以观察市场份额如何随模型能力、价格、产品和渠道变化而迁移，并验证 Coding、低成本模型及云平台分发对客户获取的推动作用。国内尚缺少覆盖企业采购、API 调用和开发者使用的统一公开数据，后续需要组合跟踪：1）企业客户和开发者数量。2）开放平台 Token 调用量。3）Coding 工具及第三方应用中的模型采用率。4）云平台和模型聚合平台份额。5）国内外收入及客户结构。6）模型切换率和多模型路由占比。我们认为，基础模型公司无需在全部场景同时领先，在若干高价值任务中建立稳定份额并持续提高客户使用深度，同样可以形成较强商业价值。

风险提示

模型迭代不及预期。若基础模型在 Coding、Agent、长时程任务等能力上的提升低于预期，可能影响开发者采用、Token 调用量及商业化增长。

商业化进展不及预期。若企业 API、Coding 和 Agent 需求增长放缓，或客户付费意愿弱于预期，智谱、MiniMax 等厂商 ARR 增长可能低于预期。

行业竞争加剧。海内外基础模型厂商持续迭代并加大价格竞争，企业同时采用多家模型已逐步常态化，可能对单一厂商的市场份额、定价及盈利能力形成压力。

仅供内部参考，请勿外传！

免责声明

分析师声明

本人，郭雅丽、袁泽世，兹证明本报告所表达的观点准确地反映了分析师对标的证券或发行人的个人意见；彼以往、现在或未来并无就其研究报告所提供的具体建议或所表达的意见直接或间接收取任何报酬。请注意，标*的人员并非香港证券及期货事务监察委员会的注册持牌人，不可在香港从事受监管活动。

一般声明及披露

本报告由华泰证券股份有限公司或其关联机构制作，华泰证券股份有限公司和其关联机构统称为“华泰证券”(华泰证券股份有限公司已具备中国证监会批准的证券投资咨询业务资格) 。本报告所载资料是仅供接收人的严格保密资料。本报告仅供华泰证券及其客户和其关联机构使用。华泰证券不因接收人收到本报告而视其为客户。

本报告基于华泰证券认为可靠的、已公开的信息编制，但华泰证券对该等信息的准确性及完整性不作任何保证。

本报告所载的意见、评估及预测仅反映报告发布当日的观点和判断。在不同时期，华泰证券可能会发出与本报告所载意见、评估及预测不一致的研究报告。同时，本报告所指的证券或投资标的的价格、价值及投资收入可能会波动。以往表现并不能指引未来，未来回报并不能得到保证，并存在损失本金的可能。华泰证券不保证本报告所含信息保持在最新状态。华泰证券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。

华泰证券（华泰证券（美国）有限公司除外）不是 FINRA 的注册会员，其研究分析师亦没有注册为 FINRA 的研究分析师/不具有 FINRA 分析师的注册资格。

华泰证券力求报告内容客观、公正，但本报告所载的观点、结论和建议仅供参考，不构成购买或出售所述证券的要约或招揽。该等观点、建议并未考虑到个别投资者的具体投资目的、财务状况以及特定需求，在任何时候均不构成对客户私人投资建议。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。对依据或者使用本报告所造成的一切后果，华泰证券及作者均不承担任何法律责任。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。

除非另行说明，本报告中所引用的关于业绩的数据代表过往表现，过往的业绩表现不应作为日后回报的预示。华泰证券不承诺也不保证任何预示的回报会得以实现，分析中所做的预测可能是基于相应的假设，任何假设的变化可能会显著影响所预测的回报。

华泰证券及作者在自身所知情的范围内，与本报告所指的证券或投资标的不存在法律禁止的利害关系。在法律许可的情况下，华泰证券可能会持有报告中提到的公司所发行的证券头寸并进行交易，为该公司提供投资银行、财务顾问或者金融产品等相关服务或向该公司招揽业务。

华泰证券的销售人员、交易人员或其他专业人士可能会依据不同假设和标准、采用不同的分析方法而口头或书面发表与本报告意见及建议不一致的市场评论和/或交易观点。华泰证券没有将此意见及建议向报告所有接收者进行更新的义务。华泰证券的资产管理部门、自营部门以及其他投资业务部门可能独立做出与本报告中的意见或建议不一致的投资决策。投资者应当考虑到华泰证券及/或其相关人员可能存在影响本报告观点客观性的潜在利益冲突。投资者请勿将本报告视为投资或其他决定的唯一信赖依据。有关该方面的具体披露请参照本报告尾部。

本报告并非意图发送、发布给在当地法律或监管规则下不允许向其发送、发布的机构或人员，也并非意图发送、发布给因可得到、使用本报告的行为而使华泰证券违反或受制于当地法律或监管规则的机构或人员。

本报告版权仅为华泰证券所有。未经华泰证券书面许可，任何机构或个人不得以翻版、复制、发表、引用或再次分发他人(无论整份或部分)等任何形式侵犯华泰证券版权。如征得华泰证券同意进行引用、刊发的，需在允许的范围内使用，并需在使用前获取独立的法律意见，以确定该引用、刊发符合当地适用法规的要求，同时注明出处为“华泰证券研究所”，且不得对本报告进行任何有悖原意的引用、删节和修改。华泰证券保留追究相关责任的权利。所有本报告中使用的商标、服务标记及标记均为华泰证券的商标、服务标记及标记。

中国香港

本报告由华泰证券股份有限公司或其关联机构制作,在香港由华泰金融控股（香港）有限公司向符合《证券及期货条例》及其附属法律规定的机构投资者和专业投资者的客户进行分发。华泰金融控股（香港）有限公司受香港证券及期货事务监察委员会监管，是华泰国际金融控股有限公司的全资子公司，后者为华泰证券股份有限公司的全资子公司。在香港获得本报告的人员若有任何有关本报告的问题,请与华泰金融控股（香港）有限公司联系。

仅供内部参考，请勿外传！

香港-重要监管披露

- 华泰金融控股（香港）有限公司的雇员或其关联人士没有担任本报告中提及的公司或发行人的高级人员。
- 智谱（2513 HK）、MiniMax（100 HK）：华泰金融控股（香港）有限公司、其子公司和/或其关联公司在本报告发布日担任标的公司证券做市商或者证券流动性提供者。
- 有关重要的披露信息，请参华泰金融控股（香港）有限公司的网页 https://www.htsc.com.hk/stock_disclosure

其他信息请参见下方 **“美国-重要监管披露”**。

美国

在美国本报告由华泰证券（美国）有限公司向符合美国监管规定的机构投资者进行发表与分发。华泰证券（美国）有限公司是美国注册经纪商和美国金融业监管局（FINRA）的注册会员。对于其在美国分发的研究报告，华泰证券（美国）有限公司根据《1934 年证券交易法》（修订版）第 15a-6 条规定以及美国证券交易委员会人员解释，对本研究报告内容负责。华泰证券（美国）有限公司联营公司的分析师不具有美国金融监管（FINRA）分析师的注册资格，可能不属于华泰证券（美国）有限公司的关联人员，因此可能不受 FINRA 关于分析师与标的公司沟通、公开露面和所持交易证券的限制。华泰证券（美国）有限公司是华泰国际金融控股有限公司的全资子公司，后者为华泰证券股份有限公司的全资子公司。任何直接从华泰证券（美国）有限公司收到此报告并希望就本报告所述任何证券进行交易的人士，应通过华泰证券（美国）有限公司进行交易。

美国-重要监管披露

- 分析师郭雅丽、袁泽世本人及相关人士并不担任本报告所提及的标的证券或发行人的高级人员、董事或顾问。分析师及相关人士与本报告所提及的标的证券或发行人并无任何相关财务利益。本披露中所提及的“相关人士”包括 FINRA 定义下分析师的家庭成员。分析师根据华泰证券的整体收入和盈利能力获得薪酬，包括源自公司投资银行业务的收入。
- 智谱（2513 HK）：华泰证券股份有限公司、其子公司和/或其联营公司在本报告发布日之前的 12 个月内担任了标的证券公开发行或 144A 条款发行的经办人或联席经办人。
- 智谱（2513 HK）：华泰证券股份有限公司、其子公司和/或其联营公司在本报告发布日之前 12 个月内曾向标的公司提供投资银行服务并收取报酬。
- 智谱（2513 HK）：华泰证券股份有限公司、其子公司和/或其联营公司预计在本报告发布日之后 3 个月内将向标的公司收取或寻求投资银行服务报酬。
- 智谱（2513 HK）、MiniMax（100 HK）：华泰证券股份有限公司、其子公司和/或其联营公司在本报告发布日担任标的公司证券做市商或者证券流动性提供者。
- 华泰证券股份有限公司、其子公司和/或其联营公司，及/或不时会以自身或代理形式向客户出售及购买华泰证券研究所覆盖公司的证券/衍生工具，包括股票及债券（包括衍生品）华泰证券研究所覆盖公司的证券/衍生工具，包括股票及债券（包括衍生品）。
- 华泰证券股份有限公司、其子公司和/或其联营公司，及/或其高级管理层、董事和雇员可能会持有本报告中所提到的任何证券（或任何相关投资）头寸，并可能不时进行增持或减持该证券（或投资）。因此，投资者应该意识到可能存在利益冲突。

新加坡

华泰证券（新加坡）有限公司持有新加坡金融管理局颁发的资本市场服务许可证，可从事资本市场产品交易，包括证券、集体投资计划中的单位、交易所交易的衍生品合约和场外衍生品合约，并且是《财务顾问法》规定的豁免财务顾问，就投资产品向他人提供建议，包括发布或公布研究分析或研究报告。华泰证券（新加坡）有限公司可能会根据《财务顾问条例》第 32C 条的规定分发其在华泰证券内的外国附属公司各自制作的信息/研究。本报告仅供认可投资者、专家投资者或机构投资者使用，华泰证券（新加坡）有限公司不对本报告内容承担法律责任。如果您是非预期接收者，请您立即通知并直接将本报告返回给华泰证券（新加坡）有限公司。本报告的新加坡接收者应联系您的华泰证券（新加坡）有限公司关系经理或客户主管，了解来自或与所分发的信息相关的事宜。

仅供内部参考，请勿外传！

评级说明

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（含此期间的股息回报）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，台湾市场基准为台湾加权指数，日本市场基准为日经 225 指数，新加坡市场基准为海峡时报指数，韩国市场基准为韩国有价证券指数，英国市场基准为富时 100 指数，德国市场基准为 DAX 指数），具体如下：

行业评级

增持：预计行业股票指数超越基准
中性：预计行业股票指数基本与基准持平
减持：预计行业股票指数明显弱于基准

公司评级

买入：预计股价超越基准 15%以上
增持：预计股价超越基准 5%~15%
持有：预计股价相对基准波动在-15%~5%之间
卖出：预计股价弱于基准 15%以上
暂停评级：已暂停评级、目标价及预测，以遵守适用法规及/或公司政策
无评级：股票不在常规研究覆盖范围内。投资者不应期待华泰提供该等证券及/或公司相关的持续或补充信息

法律实体披露

中国：华泰证券股份有限公司具有中国证监会核准的“证券投资咨询”业务资格，经营许可证编号为：91320000704041011J
香港：华泰金融控股（香港）有限公司具有香港证监会核准的“就证券提供意见”业务资格，经营许可证编号为：AOK809
美国：华泰证券（美国）有限公司为美国金融业监管局（FINRA）成员，具有在美国开展经纪交易商业务的资格，经营业务许可编号为：CRD#:298809/SEC#:8-70231
新加坡：华泰证券（新加坡）有限公司具有新加坡金融管理局颁发的资本市场服务许可证，并且是豁免财务顾问，经营许可证编号为：202233398E

华泰证券股份有限公司

南京
南京市建邺区江东中路 228 号华泰证券广场 1 号楼/邮政编码：210019

电话：86 25 83389999/传真：86 25 83387521
电子邮件：ht-rd@htsc.com

北京
北京市西城区太平桥大街丰盛胡同 28 号太平洋保险大厦 A 座 18 层/邮政编码：100032
电话：86 10 63211166/传真：86 10 63211275
电子邮件：ht-rd@htsc.com

深圳
深圳市福田区益田路 5999 号基金大厦 10 楼/邮政编码：518017
电话：86 755 82493932/传真：86 755 82492062
电子邮件：ht-rd@htsc.com

上海
上海市浦东新区东方路 18 号保利广场 E 栋 23 楼/邮政编码：200120
电话：86 21 28972098/传真：86 21 28972068
电子邮件：ht-rd@htsc.com

华泰金融控股（香港）有限公司
香港中环皇后大道中 99 号中环中心 53 楼
电话：+852-3658-6000/传真：+852-2567-6123
电子邮件：research@htsc.com
http://www.htsc.com.hk

华泰证券（美国）有限公司
美国纽约公园大道 280 号 21 楼东（纽约 10017）
电话：+212-763-8160/传真：+917-725-9702
电子邮件: Huatai@htsc-us.com
http://www.htsc-us.com

华泰证券（新加坡）有限公司
滨海湾金融中心 1 号大厦, #08-02, 新加坡 018981
电话：+65 68603600
传真：+65 65091183
https://www.htsc.com.sg

©版权所有2026年华泰证券股份有限公司