MINIMAX-W（00100.HK）

全模态迭代，工程化打破算力成本边界

领先的多模态能力，造就 MiniMax 丰富的产品矩阵。Minimax 自成立以来便专注全模态模型的研发，其大语言、语音与视频模型在多项全球权威评测榜单中位居前列。基于完备的模型布局，MiniMax 已构建起面向 B/C 端的产品与服务体系。一方面，公司推出星野/Talkie、海螺 AI 等 AI 原生应用，直接面向 C 端用户提供多模态与智能体相关服务；另一方面，通过开放平台向企业客户及开发者提供模型能力，覆盖多种行业与应用场景。2025 年公司收入达 0.79 亿美元，其中 AI 原生产品/开放平台分别实现收入 0.53/0.26 亿美元，同比增长 143%/198%。分地区来看，中国内地/海外的占比分别为 27%/73%。

模型范式持续演进，Agent 级应用开启生产力变革。在模型技术方面，原生多模态、推理、记忆、计算与推理效率等底层技术仍处于研发探索期。随着技术范式的迭代，模型有望解锁更复杂的业务场景，驱动行业天花板持续上行。Agent 能力加速生产力释放，编程与办公场景率先受益。2026 年初，以本地优先为特征的 AI Agent OpenClaw 火爆出圈，实现了模型能力与用户本地工具链的深度融合。

坚持第一性原理，极致优化工程与算力效率。MiniMax 始终秉持“效率与效果对等”的第一性原理，通过全栈自研技术（如闪电注意力机制、MoE 架构、Agent RL、自研基础设施实现软硬协同等）在大模型算力、稳定性与性能的三角约束中找到优化平衡。MiniMax 推理模型算力使用效率 MFU 超过 75%，显著优于行业平均的 40%-50%水平。

编程、办公、多模态创作是未来重点发展的方向。根据 2025 年 MiniMax 业绩会披露，1）编程将出现 L4-L5 级别智能，从工具走向同事级协作。2）办公领域，AI Agent 的交付能力与渗透率持续提升。3）多模态创作走向“直出可交付”的内容，出现更接近流式、实时输出的形态。

投资建议：我们预计 MiniMax 2026-2028 年收入为 2.7、7.0、13.2 亿美元，同比增长 240.8%、161.0%、88.0%；Non-GAAP 归母净亏损为 6.3、8.1、6.4 亿美元。由于公司尚未盈利，我们采用 PS 的方法对公司进行估值。模型性能/价格/速度全球领先、全模态模型布局完善以及细分市场成长空间广阔的因素，我们给予公司 2027 年 75 倍 P/S，目标价 1317 港元。预计公司远期收入达 56.1 亿美元，稳态净利润为 19.6 亿美元，目标市值对应远期 PE 约 27 倍。首次覆盖，给予“买入”评级。

风险提示：模型迭代不及预期的风险、行业竞争激烈的风险、盈利改善不及预期的风险等。

财务指标	2024A	2025A	2026E	2027E	2028E
营业收入（百万美元）	31	79	269	703	1,321
增长率 yoy（%）	782.2	158.9	240.8	161.0	88.0
调整后归母净利润（百万美元）	-244	-251	-629	-812	-644
增长率 yoy（%）	-174.2	-2.7	-150.9	-29.1	20.7
EPS 最新摊薄（美元/股）	-0.8	-0.8	-2.0	-2.6	-2.1
净资产收益率（%）	58.2	70.7	24.1	24.5	17.7
P/E（倍）	-138.9	-135.2	-53.9	-41.8	-52.6
P/S（倍）	1111.1	429.1	125.9	48.3	25.7

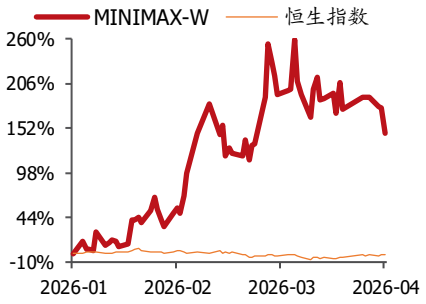
资料来源：Wind，国盛证券研究所　注：股价为 2026 年 04 月 15 日收盘价

买入（首次）

股票信息

行业	海外
04 月 15 日收盘价（港元）	847.50
总市值（百万港元）	265,805.92
总股本（百万股）	313.64
其中自由流通股（%）	100.00
30 日日均成交量（百万股）	1.92

股价走势



作者

分析师　夏君
执业证书编号：S0680519100004
邮箱：xiajun@gszq.com
分析师　刘玲
执业证书编号：S0680524070003
邮箱：liuling3@gszq.com
分析师　孙行臻
执业证书编号：S0680526010001
邮箱：sunxingzhen1@gszq.com

相关研究

财务报表和主要财务比率

资产负债表（百万美元）

会计年度	2024A	2025A	2026E	2027E	2028E
流动资产	806	1007	1617	1461	1658
现金	289	508	872	438	306
应收票据及应收账款	20	27	132	271	461
定期存款	26	14	14	14	14
其他流动资产	470	459	598	738	877
非流动资产	105	81	81	81	81
固定资产	1	2	2	1	1
无形资产	3	2	2	2	2
其他非流动资产	101	77	77	77	77
资产总计	911	1088	1698	1542	1739
流动负债	1708	3734	4413	5140	6113
短期借款	19	35	235	435	635
应付票据及应付账款	103	92	556	1053	1798
租赁负债	2	1	1	1	1
其他流动负债	1584	3605	3620	3650	3678
非流动负债	2	3	3	3	3
租赁负债	1	1	1	1	1
其他非流动负债	1	2	2	2	2
负债合计	1710	3737	4416	5143	6116
股本	0	0	0	0	0
留存收益	-799	-2648	-2718	-3601	-4377
归属母公司股东权益	-799	-2648	-2718	-3601	-4377
负债和股东权益	911	1088	1698	1542	1739

利润表（百万美元）

会计年度	2024A	2025A	2026E	2027E	2028E
营业收入	31	79	269	703	1321
营业成本	27	59	196	488	864
销售费用	87	52	107	206	244
管理费用	14	37	106	171	149
研发费用	189	253	529	726	846
其他收入	36	40	18	18	18
金融资产减值亏损	214	1590	0	0	0
营业利润	-465	-1871	-651	-870	-764
财务费用	1	1	5	12	12
其他收益	0	0	0	0	0
利润总额	-465	-1872	-656	-883	-776
所得税	0	0	0	0	0
净利润	-465	-1872	-656	-883	-776
少数股东损益	0	0	0	0	0
归属母公司净利润	-465	-1872	-656	-883	-776
Non-GAAP 归母净利润	-244	-251	-629	-812	-644
Non-GAAPEPS（美元/股）	-1	-1	-2	-3	-2

现金流量表（百万美元）

会计年度	2024A	2025E	2026E	2027E	2028E
经营活动现金流	-258	-286	-282	-494	-192
税前利润	-465	-1872	-656	-883	-776
折旧摊销	0	1	1	1	1
营运资金变动	41	-12	374	388	583
其他经营现金流	165	1597	0	0	0
投资活动现金流	-431	-140	-140	-140	-140
资本支出	-1	-1	-1	-1	-1
长期投资	72	0	0	0	0
其他投资现金流	-503	-139	-139	-139	-139
筹资活动现金流	771	645	786	200	200
短期借款	19	200	200	200	200
其他筹资现金流	752	445	586	0	0
现金净增加额	81	219	364	-434	-132

会计年度	2024A	2025A	2026E	2027E	2028E
成长能力					
营业收入(%)	782	159	241	161	88
营业利润(%)	-73	-303	65	-34	12
Non-GAAP 归母净利润(%)	-174	-3	-151	-29	21
获利能力					
毛利率(%)	12	25	27	31	35
净利率(%)	-1524	-2368	-244	-126	-59
ROE(%)	58	71	24	25	18
ROIC(%)	60	72	26	27	20
偿债能力					
资产负债率(%)	188	343	260	333	352
净负债比率(%)	186	340	246	305	315
流动比率	0	0	0	0	0
营运能力					
总资产周转率	0	0	0	0	1
应收账款周转率	7	9	9	9	9
应付账款周转率	1	1	1	1	1
每股指标（美元）					
每股收益(最新摊薄)	-1	-1	-2	-3	-2
每股经营现金流(最新摊薄)	-1	-1	-1	-2	-1
每股净资产(最新摊薄)	-3	-8	-9	-11	-14
估值比率					
P/E	-139	-135	-54	-42	-53
P/S	1111	429	126	48	26
P/B	-42	-13	-12	-9	-8

资料来源：Wind，国盛证券研究所　注：股价为 2026 年 04 月 15 日收盘价

内容目录

图表目录

1、MINIMAX-W：多模态技术储备深厚，产品矩阵完备

1.1 历程：深耕全模态模型研发，产品矩阵驱动全球化扩张

注重全模态布局，技术实力位居全球前列。MiniMax 自 2022 年初成立以来，持续围绕用户价值、全球化布局与技术驱动三大核心原则推进业务发展，并以“与所有人共创智能”为使命，聚焦人工智能前沿技术演进，长期目标指向通用人工智能（AGI）的实现。公司自成立以来便专注全模态模型的研发，其大语言（M 系列）、语音（Speech 系列）与视频模型（Hailuo 系列）在多项全球权威评测榜单中位居前列。截至 2025 年 9 月 30 日，MiniMax 的 AI 原生产品累计为全球范围内 2 亿个人用户以及超 10 万家企业以及开发者提供服务，中国内地以外的区域收入占比达 73.1%。

图表1：MiniMax 模型汇总盘点

类别	模型	发布时间	参数	架构特点	简介
语言模型	abab 1-abab 4	2022	-	-	完成模型训练，对话、语义理解、逻辑推理速度与成本持续迭代
	abab 5.5	2023.05	-	-	-
	abab 6.0	2024.01	超 100B	MoE	国内首个参数达千亿以上的基于 MoE 架构的大语言模型，提升了处理复杂任务的能力以及单位时间内能够训练的数据量。
	abab 6.5	2024.04	1000B	MoE	改进模型架构、重构数据 pipeline、训练算法及并行训练策略优化。
	Text-01	2025.01	456B（单次激活 45.9B）	MoE+线性注意力	业内首次把线性注意力机制扩展到商用级别的模型。综合性能对标海外顶尖模型，支持最长 400 万 token 的上下文。
	M1 系列	2025.06	456B（单次激活 45.9B）	MoE+混合注意力	基于 Text-01 训练，以自研的闪电注意力机制为主的混合架构，配合强化学习算法 CISPO，显著提升上下文（100 万 token）和成本效率。
	M2 系列	2025.10	230B（单次激活 10B）	MoE+全注意力+交错思维链	专为 Agent 和 Coding 设计，12 月更新的 MiniMax M2.1，重点聚焦于更多编程语言和办公场景的可用性，兼顾速度和性价比。2026 年 2 月推出的 M2.5 在编程、工具调用和搜索、办公等生产力场景都达到或者刷新行业 SOTA。3 月发布 M2.7，模型能够自行构建复杂 Agent Harness，并基于 Agent Teams、复杂 Skills、Tool Search tool 等能力，完成高度复杂的生产力任务。
视频模型	Hailuo-01	2024.08	-	-	支持图生视频、文生视频功能，后续更新的 S2V-01、I2V-01-Director 分别新增了单图主体参考架构和镜头控制功能。
	Hailuo-02	2025.06	3 倍于 Hailuo-01	NCR	核心架构 Noise-aware Compute Redistribution，提升了同等的参数量级下的训练和推理效率。后续更新的 Hailuo 2.3 在物理表现与指令遵循进一步增强，Media Agent 支持全模态全能创作。在生成价格方面，优于国内外模型。
语音模型	Speech-01	2023.11	-	-	支持语音合成，2025 年 1 月更新的 T2A-01 支持 17 国语言以及超 300 种预置音色，可实现自由配置输出语音的情绪、语速、音高、音色，满足复杂场景的精细化需求。
	Speech-02	2025.04	-	AR Transformer	核心创新是内在的 Zero-Shot 与 AR Transformer 共同训练，提供任意语言×任意口音×任意音色的无限组合。2025 年 10 月更新的 Speech 2.6 适用 Voice Agent 场景、低延时、高自然度，支持超 40 个语种。
音乐模型	Music-01	2024.08	-	-	端到端音乐生成，支持纯音乐、清唱作品等形式。2025 年 9 月升级的 Music 1.5，生成时长升至 4 分钟，具备强控制力、人声自然饱满、编曲层次丰富、歌曲结构清晰四大特点。
	Music-02	2025.10	-	-	模型对音乐的理解与表达实现跃升，精准捕捉与还原人声的细腻情绪于器乐的动态张力。2026 年 1 月更新的 Music 2.5，在段落级强控制与物理级高保真两大技术难题上实现突破。

资料来源：MiniMax 招股书、MiniMax 稀宇科技公众号、MiniMax 官网、Github、Hugging Face，国盛证券研究所

MiniMax 的发展历程可以分为三个阶段：

第一阶段（2022-2023 年）：多模态能力积累期，产品由探索走向市场。

2022 年，MiniMax 完成了 abab 1 至 abab 4 多代语言模型训练，并在国内推出其首个 AI 聊天产品 Glow。Glow 上线仅四个月，注册用户即突破五百万。2023 年，模型研发持续

推进，相继发布了 abab 5.5 及语音模型 Speech-01。在产品侧，B 端企业开放平台于 2023 年 3 月正式推出，对外提供模型 API 服务。C 端应用 Glow 因内容审核要求在国内下架，公司随即将重心转向海外市场，推出 AI 社交应用 Talkie，随后面向国内用户上线星野。

第二阶段（2024 年）：全模态雏形确立，海螺 AI 平台成立。
2024 年，MiniMax 先后发布超大规模参数的 MoE 语言模型 abab 6 与 abab 6.5、视频模型 Hailuo-01 以及音乐模型 Music-01，并与此前推出的 Speech-01 语音模型共同构建起较为完整的全模态模型矩阵。在产品侧，海螺 AI 平台正式成立，用户可在统一平台上调用 MiniMax 的视频生成能力，标志着公司全模态产品体系的初步成型。

第三阶段（2025 年至今）：多项模型性能达到 SOTA，驱动产品形态升级。
自 2025 年起，MiniMax 在模型能力上实现明显跃升。语言模型历经多次迭代，由 abab 6.5—MiniMax-M1—M2—M2.1—M2.5—M2.7，视频模型升级至 Hailuo 2.3，音乐/语音模型升级至 Music-2.5/Speech-2.6。模型的持续迭代，使得 MiniMax 跻身全球少数在多模态方向实现突破的大模型公司之列。模型能力的持续增强直接推动产品形态升级，如 MiniMax Agent 应用已能够通过自然语言完成研究、PPT 制作等复杂任务。2026 年 1 月发布的 AI-native Workspace 进一步将能力延展至桌面端，用户只需输入想法，即可由系统协助完成多种工作场景下的任务。根据公司官方披露的信息，模型端在 2026 年仍将保持高频迭代节奏，计划推出新一代语言模型 MiniMax-M3 以及视频模型 Hailuo-03 等，进一步巩固其在多模态模型领域的技术领先地位。

图表2：公司发展历程

资料来源：MiniMax 招股书、财联社、晚点 latepost、稀土掘金 MiniMax 官方账号、腾讯科技、MiniMax 稀宇科技公众号，国盛证券研究所

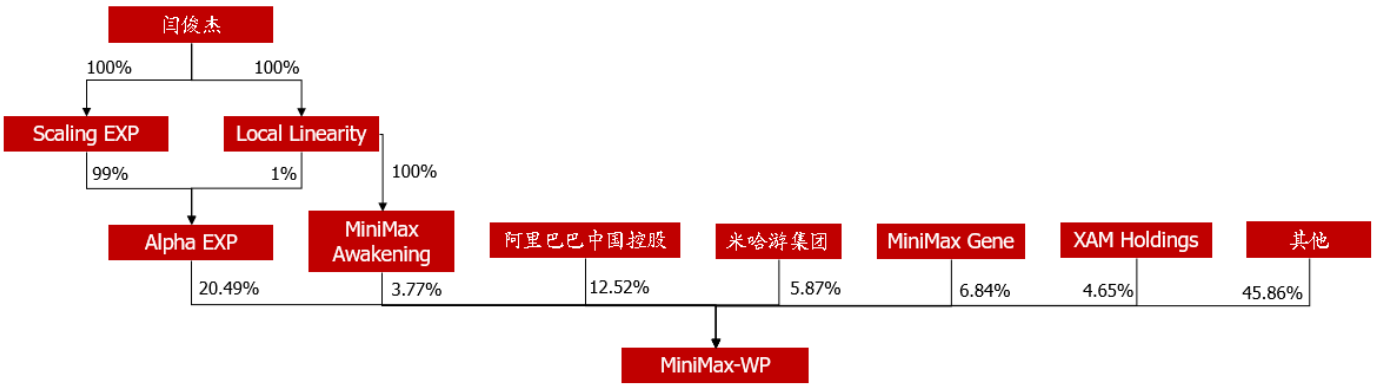
1.2 团队：创始人拥有最高决策权，精英团队与高效组织共振

CEO 闫俊杰是公司最大持股股东，具有超七成的投票权。MiniMax 全球发售后，前五大股东分别为 Alpha EXP、阿里巴巴中国控股、员工持股平台 MiniMax Gene、米哈游集团和 XAM Holding。其中，Alpha EXP 为最大股东，持股比例为 20.49%，该公司由 CEO 闫俊杰通过其全资公司间接持有。此外，公司采用 A/B 股结构，虽然闫俊杰个人仅持有 25.36% 的股份，但其投票权占比高达 72.05%。

扁平化的组织架构、敢于放权，促进创新与执行力的高效融合。MiniMax 采纳扁平化、灵活的组织结构，CEO 之下最多设三层级，以确保决策的高效性，进而推动智能大模型的快速迭代。同时，公司通过项目制模式组建跨职能团队，打破技术、产品与业务之间

的传统隔阂，各团队围绕统一目标协作，致力于提升模型智能水平，并使其更加普及易用。公司鼓励员工超越既定角色，承担更大责任。部分团队负责人在担任领导职务时，年龄均在 30 岁以下，给予年轻员工充分的信任与支持。

图表3：公司股权结构（截至全球发售后）



资料来源：MiniMax 招股书，国盛证券研究所

图表4：MiniMax 核心人员持股权及投票比例

股东	持股比例	投票权比例
闫俊杰（CEO）	25.36%	72.05%
贠烨祎（COO）	2.83%	6.76%

资料来源：MiniMax 招股书，国盛证券研究所

创始人及部分高管具有深度的产业经验。MiniMax 创始人、首席执行官兼首席技术官闫俊杰曾在商汤集团任职超六年，担任副总裁及研究院副院长等职位。在 MiniMax 负责公司的监督整体管理和业务运营、董事会事务，制定战略和运营计划（AI 研发方面、产品及商业化方面）等。执行董事及首席运营官贠烨祎女士也曾在商汤集团股份有限公司担任首席执行官行政助理及战略部总监、创新业务部总监等多个职位。此外，公司在视觉模型和大语言模型研发方面均有专人负责（周彧聪先生负责视觉模型，赵鹏宇先生负责大语言模型），展现出在多模态 AI 技术领域的全面布局和强大实力，这支兼具学术深度和产业经验的顶尖团队，正推动公司在人工智能浪潮中快速崛起。

图表5：公司管理层背景

董事姓名	职位	简介
闫俊杰	创始人、董事长、执行董事、首席执行官兼首席技术官	曾在商汤集团有限公司担任副总裁及研究院副院长等职位。闫博士拥有东南大学学士学位、中国科学院自动化研究所人工智能领域的博士学位，并在清华大学从事博士后研究。
员烨祎	执行董事兼首席运营官	自 2017 年 9 月至 2018 年 8 月担任商汤集团股份有限公司融资与战略投资部经理，随后自 2018 年 8 月至 2021 年 1 月晋升为首席执行官行政助理及战略部总监。随后自 2021 年 1 月至 2022 年 1 月担任创新业务部总监。 员女士拥有美国约翰霍普金斯大学电子工程理学学士学位，并辅修经济学和数学专业。
赵鹏宇	执行董事兼大语言模型研究与工程负责人	于 2020 年 8 月至 2023 年 7 月在北京葫芦科技有限公司担任研究级软件开发工程师，主要负责推荐算法。 赵先生拥有北京大学计算机科学与技术学士学位和硕士学位。
周彧聪	执行董事兼视觉模型研究与工程负责人	于 2018 年 4 月至 2019 年 7 月任职于商汤集团股份有限公司，并于 2019 年 8 月至 2022 年 3 月任职于华为技术有限公司，专注于算法领域工作。自 2023 年 1 月起亦一直担任上海稀宇的法定代表人及董事。 周先生拥有北京航空航天大学数学与应用数学学士学位和计算机科学硕士学位。

资料来源：MiniMax 招股书，国盛证券研究所

1.3 产品：依托多模态模型能力，建立完善的产品与服务体系

MiniMax 已构建起覆盖全模态的模型矩阵，在大语言、视频及语音三大核心领域均处于全球领先水平：

- **大语言模型：**MiniMax M2 系列专为 Agent 与代码场景深度优化，核心优势体现在较强的代码能力、稳定的 Agentic 表现以及极致的性价比与响应速度。在此基础上，M2.1 版本进一步强化对多编程语言及办公场景的适配能力，模型回复与思维链更加精炼，在实际编程与交互过程中响应速度提升的同时显著降低 Token 消耗。在软件工程、Coding Agent、VIBE 等基准测试中，M2.1 的整体表现已持平或超越 Gemini 3 Pro 与 Claude Sonnet 4.5。2026 年 2 月发布的 M2.5 在编程、工具调用和搜索、办公等生产力场景都达到或者刷新了行业的 SOTA，并且优化了模型对复杂任务的拆解能力和思考过程中 token 的消耗，使其能更快地完成复杂的 Agentic 任务。

- **视频模型：**Hailuo 02 在全球最早能够生成体操等高度复杂场景的模型。其采用 NCR 核心架构，在同等参数规模下显著提升训练与推理效率。模型数据量和参数量的扩大也显著提升了其在复杂指令遵循和物理表现方面的能力。2025 年 10 月发布的 Hailuo 2.3 在肢体动作、风格化表现及人物微表情方面进一步优化，并提升对运动类指令的响应能力，在效果与成本维度上持续刷新全球视频模型性价比纪录。

- **语音模型：**Speech 02 是基于 AR Transformer 构建的高质量 TTS 系统，支持 32 语种及多口音、多情绪的人声生成。最新版本 Speech 2.6 面向 Voice Agent 场景进行优化，实现端到端低于 250 毫秒的超低延时表现，并在专业度与自然度方面进一步提升，整体技术水平处于行业领先区间。根据 Artificial Analysis 数据，截至 2026 年 3 月 6 日，MiniMax 在文生音频领域位列第三，仅次于 Inworld 与 OpenAI。

基于模型能力，MiniMax 已构建起面向 B/C 端的产品与服务体系。一方面，公司推出星野/Talkie、海螺 AI 等 AI 原生应用，直接面向 C 端用户提供多模态与智能体相关服务；另一方面，通过开放平台向企业客户及开发者提供模型能力，覆盖多种行业与应用场景。具体来看：

- **AI 原生产品（2025 年收入占比 67%）：**AI 原生产品体系涵盖 Agent 应用 MiniMax、视频生成平台海螺 AI、音频生成工具 MiniMax 语音以及全模态交互平台 Talkie/星野。在用户规模方面，随着产品矩阵持续完善与应用场景不断扩展，AI 原生产品的累计用户数由 2023 年的 0.1 亿快速增长至 2025 年第三季度的 2.1 亿，用户基础呈现显著的加速扩张趋势。

- **开放平台及其他基于 AI 的企业服务（2025 年收入占比 33%）：**公司通过开放 API 与相关服务向企业客户及开发者输出文本、视频与音频模型能力，支持其在智能终端、医疗健康、文旅、金融及互联网服务等重点行业中实现快速业务部署。以日均 Token 调用量计，平台每日处理规模达数十亿 Token，已发展为全球领先的企业级与开发者开放平台之一。截至 2025 年前三季度，开放平台付费客户数量达到 2,500 家。

图表6：MiniMax 的 AI 原生产品介绍

产品	模型支撑	定位	盈利模式
MiniMax	文本、视频、音频和音乐模型	智能 Agent 应用	免费增值、订阅、基于 token 的应用内购买
海螺 AI	视频模型	视频生成平台	免费增值、订阅、基于 token 的应用内购买
MiniMax 语音	语音和音乐模型	音频生成工具	免费增值、订阅、基于 token 的应用内购买
Talkie/星野	文本、视频、音频和音乐模型	全模态交互平台	免费增值、订阅、在线营销服务、应用内购买

资料来源：MiniMax 招股书，国盛证券研究所

图表7：MiniMax 产品的用户数量

单位：千	2023	2024	2025Q1-Q3
AI 原生产品：			
用户数	11,131	115,378	212,247
MAU	3,144	19,106	27,622
付费用户	119.7	650.3	1,771.6
开放平台：			
用户数	13	42	132
付费用户	0.1	0.7	2.5

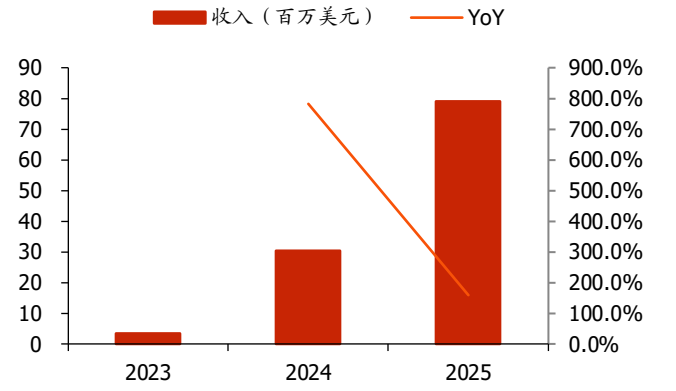
资料来源：MiniMax 招股书，国盛证券研究所

1.4 财务：收入高速增长，现金储备充足

公司收入迈入快速增长阶段。2023-2025 年期间 MiniMax 总收入分别为 3.5、30.5、79.0 百万美元，2024/2025 年同比大幅增长 782.2%/158.9%。细分 2025 年收入结构，AI 原生产品/开放平台及其他基于 AI 的企业服务分别实现收入 53.1/26.0 百万美元，分别同比增长 143.4%/197.8%。

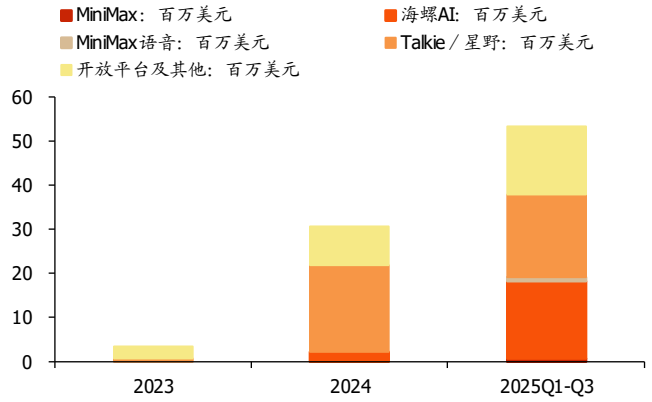
海螺 AI 贡献近三分之一营收，助力收入结构向多元化演进。从产品结构看，2025 年前三季度 MiniMax Agent、海螺 AI、MiniMax 语音、Talkie 与星野、开放平台的企业服务的收入占比分别为 1.4%、32.7%、2.0%、35.1%和 28.9%。海螺 AI 收入占比显著提升，有效降低了公司对单一产品 Talkie 与星野的依赖，收入结构进一步多元化。

图表8：2023-2025 年 MiniMax 收入及同比增速



资料来源：MiniMax 招股书、MiniMax 财报，国盛证券研究所

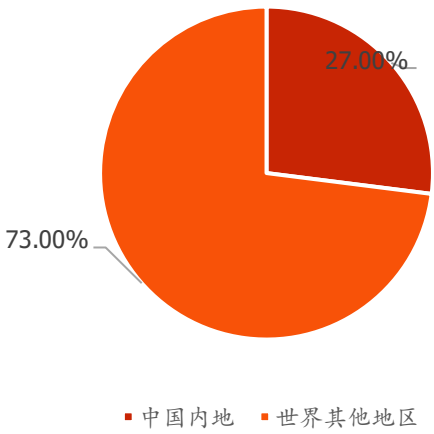
图表9：2023-2025Q1-Q3 MiniMax 分业务收入



资料来源：MiniMax 招股书，国盛证券研究所

公司为全球化企业，2025 年七成以上收入来自海外。根据公司最新公告披露，2025 年公司 73%收入来自海外，27%收入来自中国内地，为公司提供地域均衡多元化的全球收入结构。

图表10：2025 年公司收入结构（按地区）

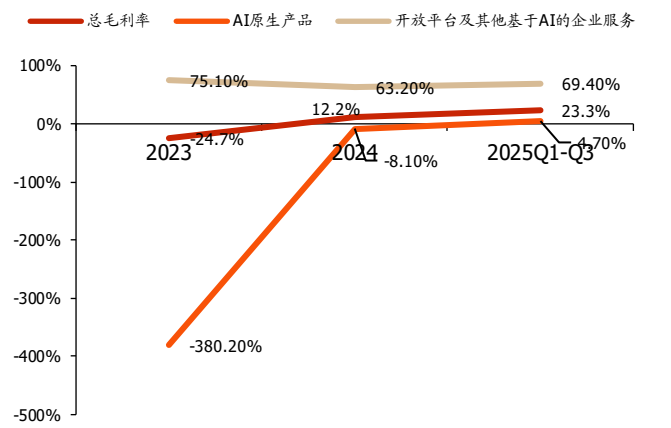


资料来源：MiniMax 财报，国盛证券研究所

规模效应初显，推理成本效率优化驱动毛利空间修复。随着用户规模扩大及推理需求上升，销售成本由 2023 年的 4.3 百万美元增长至 2025 年前三季度的 41.0 百万美元。从销售成本结构上看，与推理活动相关的云计算服务成本占比稳定在 90%以上。但销售成本占收入的比重持续下降，由 2023 年的 124.7%下降至 2025 年前三季度的 76.7%。

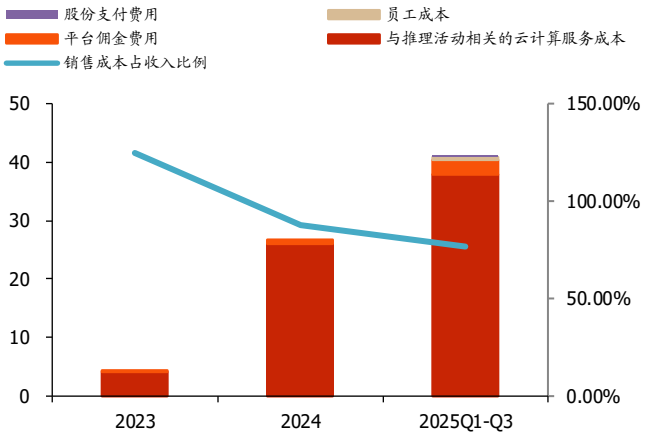
企业服务毛利率稳定，AI 原生产品处于修复通道。MiniMax 的毛利润由 2023 年的-0.9 百万美元提升至 2025 年前三季度的 12.5 百万美元，毛利率由-24.7%提升至 23.3%。毛利率的持续修复，主要来自模型智能水平提升与推理效率改善所带来的单位算力成本下降。分业务来看，AI 原生产品毛利率改善尤为明显，由 2023 年的-380.2%提升至 2025 年前三季度的 4.7%，首次实现转正，商业化路径逐步清晰。开放平台及其他基于 AI 的企业服务业务毛利率长期维持在 60%以上，为整体毛利率提供了稳定支撑。

图表11：2023-2025 Q1-Q3 MiniMax 毛利率



资料来源：MiniMax 招股书，国盛证券研究所

图表12：2023-2025Q1-Q3 MiniMax 销售成本（百万美元）

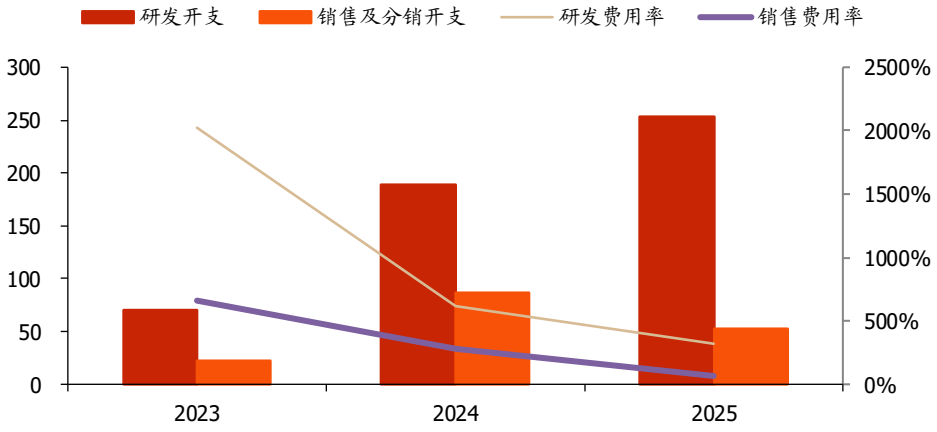


资料来源：MiniMax 招股书，国盛证券研究所

营销重心转向内生驱动，营销开支明显回落。2025 年 MiniMax 的销售及分销开支为 51.9 百万美元，同比下降 40.3%。公司在 2025 年逐步调整营销策略，逐步转向以内生增长为核心的获客模式，销售及分销开支明显回落。

研发杠杆效应逐步释放，技术投入进入质效提升期。研发开支由 2023 年的 70.0 百万美元增长至 2025 年的 252.8 百万美元，增长主要来自两方面因素，一是推进大模型及多模态能力建设带来的与训练相关的云计算服务开支提升，二是内部研发与工程团队扩编后员工成本的增加。2025 年 MiniMax 的研发开支增速为 33.8%，显著低于 158.9%的收入增速。预计未来研发开支的绝对金额将继续增长，但研发效率及收入规模的提升将摊薄研发开支比重。

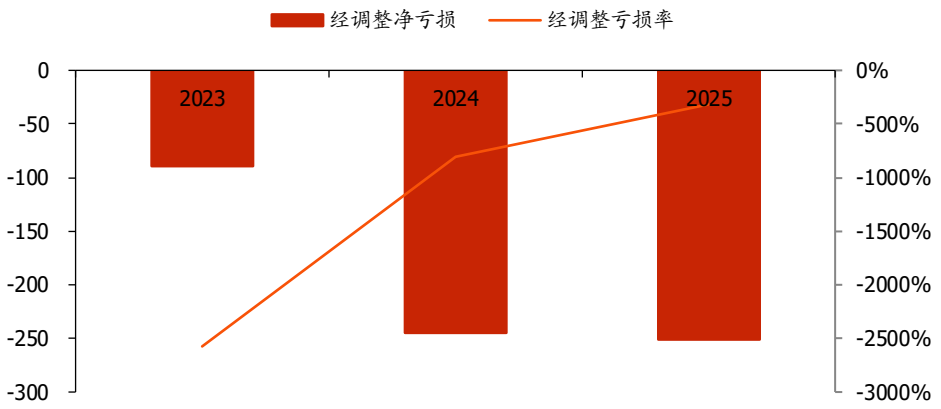
图表13：2023-2025 年 MiniMax 销售与研发开支（百万美元）及费用率



资料来源：MiniMax 招股书、MiniMax 财报，国盛证券研究所

净亏损的绝对金额预计持续扩大，但净亏损率将收窄。2023-2025 年，MiniMax 的 Non-GAAP 归母净亏损分别为 89.1、244.2、250.9 百万美元；Non-GAAP 归母净亏损率为 2574.4%、800.2%、317.4%。考虑到 AI 大模型行业技术迭代速度较快，公司需要持续投入以维持竞争优势，短期内亏损状态预计仍将延续，但净亏损率将持续收窄。

图表14：2023-2025 年 MiniMax 经调整净利润（百万美元）及利润率



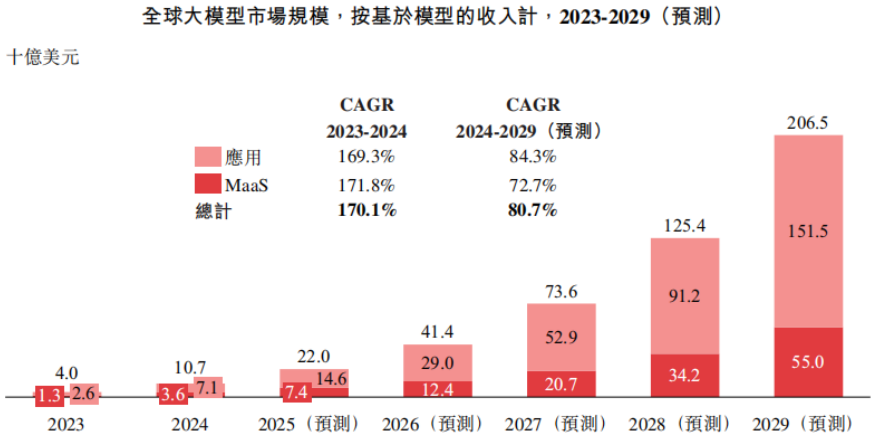
资料来源：MiniMax 招股书、MiniMax 财报，国盛证券研究所

现金储备足以支撑公司 3 年常规运营。从现金储备来看，截至 2025 年前三季度，MiniMax 的现金结余为 10.5 亿美元。按照公司披露的每月现金消耗约 2,810 万美元测算，在不考虑首次公开发售融资的情况下，现有现金储备可支持公司运营约 37 个月。

2. 大模型行业分析：全球市场高速增长，模型能力核心决定竞争优势

根据灼识咨询的数据，全球大模型市场仍处于商业化落地早期阶段。随着技术成熟度提升和用户付费意愿增强，按基于模型的收入口径计算，全球大模型市场规模预计将由 2024 年的 107 亿美元快速增长至 2029 年的 2,065 亿美元，复合年增长率达 80.7%。其中，大模型应用市场规模有望由 2024 年的 71 亿美元增长至 2029 年的 1,515 亿美元，CAGR 达 84.3%，MaaS 市场规模则预计由 36 亿美元增至 550 亿美元，CAGR 达到 72.7%。

图表15：全球大模型市场规模（十亿美元）



资料来源：MiniMax 招股书，国盛证券研究所

目前大模型应用的主要赛道包括生产力、娱乐、视觉生成、音频生成和通用 2B 服务。纵观这些赛道，由于大模型技术的通用性特征，这些市场都实现了用一套高度可扩展的模型、撬动下游规模化与个性化并存的多元需求场景，从大型行业客户到中小创作者全覆盖，实现模型 ROI+转正。这些赛道里能长期保持自然增长的产品，都是底层模型智能持续提升驱动的。长期头部玩家需要有可端到端优化的底层自研模型，且保持模型水平在第一梯队。一个模型智能水平的突破能使其快速突围，当技术突破带来更佳用户体验时，用户就会自然迁移。

图表16：大模型应用市场格局

资料来源：MiniMax 招股书，国盛证券研究所

全球大模型产业格局正由少数科技巨头依托算力、数据和资本优势形成的垄断态势，逐步转向多极竞争与生态共建。随着开源开放降低技术门槛，各国加快算力基础建设、拓展应用领域，竞争主体不断增多，合作网络持续延伸。当前，美国在基础研究和商业化生态方面保持领先，欧洲以开源发展与合规监管构建差异化路径，中国则凭借超大市场规模、丰富应用场景和端侧生态快速迭代，稳居全球发展第一梯队。

图表17：中美欧大模型发展优势比较

区域	发展阶段	战略定位	技术与产业优势	主要特征与趋势
中国	全球第一梯队	凭借超大市场规模与丰富场景	产业链完整，政策体系完善，应用落地能力突出	产业链完善，落地速度快
美国	全球第一梯队	聚焦基础模型、算法创新与 AGI 验证	算法与算力领先，私营科技企业主导，商业化路径成熟	创新驱动强，生态集中度高
欧洲	全球第二梯队	以监管和伦理为导向，强调安全与合规治理	优势在标准制定与学术研究深度，公私合作模式稳健	合规导向，发展节奏平稳

资料来源：36 氪研究院，国盛证券研究所

中国大模型行业的发展进入快速跃升阶段，形成了“技术演进与场景协同并进”的独特路径。中国大模型产业正经历从“技术跟随者”到“创新并行者”的角色转变，部分领域已跻身国际前列。产业层面，千亿级参数已成为国产主流大模型的标配，头部厂商在文本和多模态内容理解与生成、深度推理等核心能力已对标国际顶尖水平。应用层面，庞大的国内市场与丰富的垂直场景为大模型落地提供了天然试验田，金融、政务、教育、医疗等关键行业正加速实现规模化部署，形成了从通用能力到行业应用的双轮驱动。新兴赛道中，端侧大模型与智能体生态快速成长，多智能体协作、多模态融合和复杂任务执行的演进趋势日益显现。

2.1 模型：多模态、推理、记忆、效率

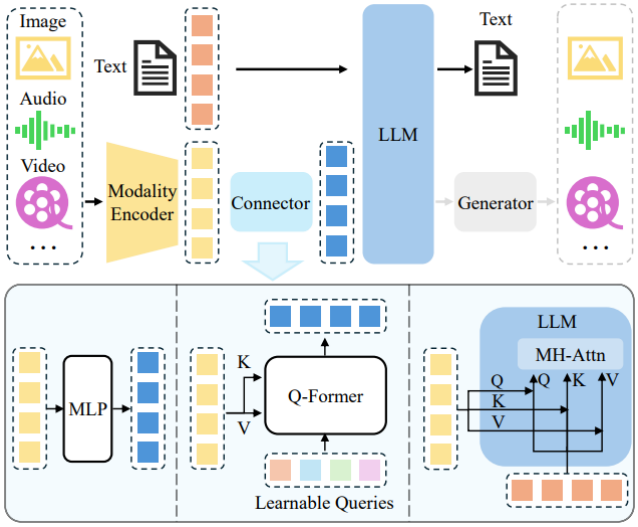
2.1.1 多模态能力：由拼接走向原生

多模态模型可以根据多模态数据训练的方式分为拼接式和原生多模态模型。多数模型采用的是拼接的方式，典型的如 CLIP 通过联合训练“图像编码器”和“文本编码器”，预测训练 Batch 中（图像，文本）训练样本对的正确匹配关系。这种方式容易导致不同模态间理解的断层，且模型在处理非文本输入时效率较低。原生多模态模型指的是，从训练阶段开始便利用不同模态的数据共同进行预训练，在实现输入和输出多模态的同时，还具备多模态推理及跨模态迁移能力。

但原生多模态也面临训练与计算成本方面的挑战：

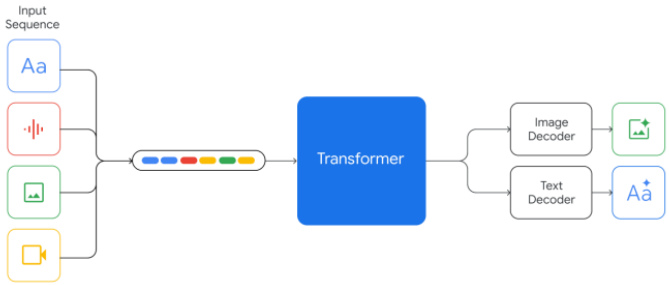
- 训练数据中文本占比远超图像和音频，如何防止模型产生强烈的“文本偏见”是关键。
- 统一的庞大词汇表（包含视觉和音频 Token）极大地增加了内存和计算开销。

图表18：拼接式多模态模型架构



资料来源：《A Survey on Multimodal Large Language Models》-Shukang Yin 等人，国盛证券研究所

图表19：谷歌 Gemini 模型支持的输入输出模态



资料来源：《Gemini: A Family of Highly Capable Multimodal Models》-谷歌 Gemini 团队，国盛证券研究所

谷歌最早采用原生多模态架构，国产模型持续跟进。自 2023 年 12 月推出的 Gemini 1.0 模型，谷歌便开始采用原生多模态的架构，将图像、视频、音频和文本混合训练。随着模型的持续迭代，海外其他大模型厂商如 OpenAI（OpenAI-4o）、Meta（Llama 4）等也逐步采用原生多模态的方式训练模型。国内厂商布局则相对较晚，目前百度的文心 5.0、月之暗面的 Kimi-K2.5 和阿里巴巴的 Qwen3.5 等为原生多模态模型。

2.1.2 推理能力演变：强化学习与长思维链

随着 OpenAI 推出 GPT-o1 等推理模型，大语言模型行业进入了推理阶段。2024 年 9 月发布的 o1 模型通过引入大规模强化学习算法，增强了模型的推理链条，提升了错误识别和修正能力。技术的关键变化在于从“预训练扩展”向“测试时计算”的转变，使得模型具备了更加深入的推理和自我优化能力。在 STEM（科学、技术、工程、数学）领域，o1 模型的推理能力有了显著突破。例如，在国际数学奥林匹克资格考试中，GPT-4o 仅能解决 13%的问题，而 o1 则达到了 83%的正确率。这标志着 AI 已经从简单的对话场景扩展至科学、代码、数学等复杂领域。

DeepSeek R1 的推理能力成为关注的焦点。2025 年初，中国大模型公司深度求索 DeepSeek 发布并开源 R1 模型，模型采用纯深度学习的方法，并发现 AI 自发涌现出推理能力。DeepSeek R1-Zero 模型通过完全依赖强化学习而非监督微调（SFT），实现了“自我反思”和“错误发现”等高级认知行为，标志着大模型真正理解了问题背后的逻辑结构。虽然 DeepSeek R1-Zero 在可读性和语言流畅性上存在一定问题，但通过引入冷启动数据和多阶段训练流程，R1 模型成功保持了推理能力并优化了表达方式，且其训练成本仅为 OpenAI o1 模型的 1/10，两者性能却属于同一级别。

图表20：DeepSeek R1 与其他代表性模型对比

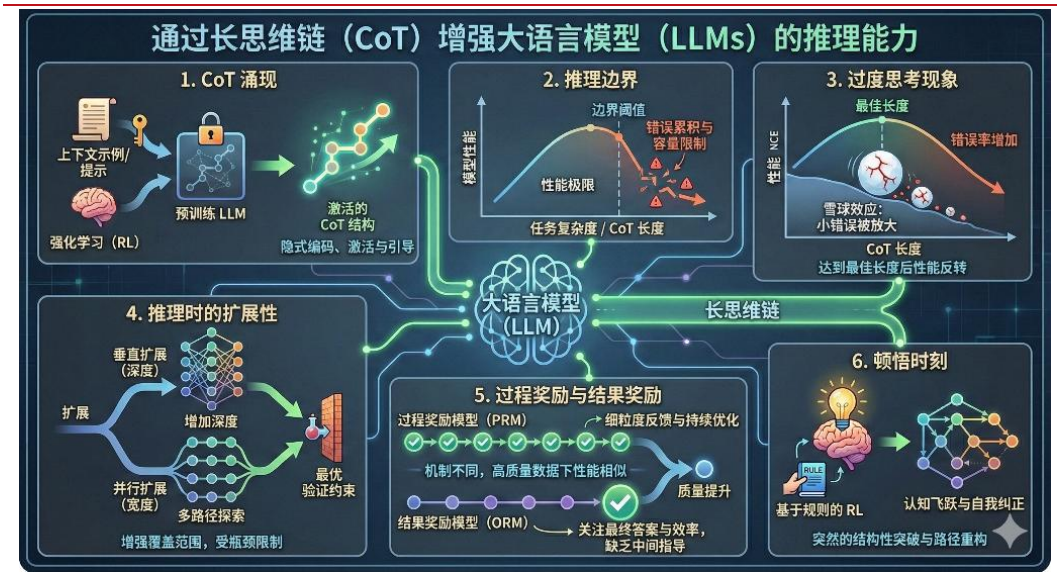
	Benchmark (Metric)	Claude-3.5-Sonnet-1022	GPT-4o 0513	DeepSeek V3	OpenAI o1-mini	OpenAI o1-1217	DeepSeek R1
	Architecture	-	-	MoE	-	-	MoE
	# Activated Params	-	-	37B	-	-	37B
	# Total Params	-	-	671B	-	-	671B
English	MMLU (Pass@1)	88.3	87.2	88.5	85.2	**91.8**	90.8
	MMLU-Redux (EM)	88.9	88.0	89.1	86.7	-	**92.9**
	MMLU-Pro (EM)	78.0	72.6	75.9	80.3	-	**84.0**
	DROP (3-shot F1)	88.3	83.7	91.6	83.9	90.2	**92.2**
	IF-Eval (Prompt Strict)	**86.5**	84.3	86.1	84.8	-	83.3
	GPQA Diamond (Pass@1)	65.0	49.9	59.1	60.0	**75.7**	71.5
	SimpleQA (Correct)	28.4	38.2	24.9	7.0	**47.0**	30.1
	FRAMES (Acc.)	72.5	80.5	73.3	76.9	-	**82.5**
	AlpacaEval2.0 (LC-winrate)	52.0	51.1	70.0	57.8	-	**87.6**
	ArenaHard (GPT-4-1106)	85.2	80.4	85.5	92.0	-	**92.3**
Code	LiveCodeBench (Pass@1-COT)	38.9	32.9	36.2	53.8	63.4	**65.9**
	Codeforces (Percentile)	20.3	23.6	58.7	93.4	**96.6**	96.3
	Codeforces (Rating)	717	759	1134	1820	**2061**	2029
	SWE Verified (Resolved)	**50.8**	38.8	42.0	41.6	48.9	49.2
	Aider-Polyglot (Acc.)	45.3	16.0	49.6	32.9	**61.7**	53.3
Math	AIME 2024 (Pass@1)	16.0	9.3	39.2	63.6	79.2	**79.8**
	MATH-500 (Pass@1)	78.3	74.6	90.2	90.0	96.4	**97.3**
	CNMO 2024 (Pass@1)	13.1	10.8	43.2	67.6	-	**78.8**
Chinese	CLUEWSC (EM)	85.4	87.9	90.9	89.9	-	**92.8**
	C-Eval (EM)	76.7	76.0	86.5	68.9	-	**91.8**
	C-SimpleQA (Correct)	55.4	58.7	**68.0**	40.3	-	63.7

资料来源：腾讯技术工程公众号，国盛证券研究所

思维链为提升模型推理能力的常用手段。思维链可划分为短思维链与长思维链：前者以浅层、线性的推理为特征，侧重于解决逻辑单一、定义明确的基础问题；而长思维链则通过构建“深度推理、探索与反思”的整合框架，确保每个逻辑环节在复杂结构下均能严密执行，并鼓励模型挖掘非显性的潜在路径，同时引入迭代分析机制对结论进行动态重估与修正。这种从线性到演进式逻辑的转变，使得长思维链在处理高难度问题时具备更强的精确度、稳健性以及逻辑挖掘深度。

推理大语言模型在长思维链演进中呈现出六大关键行为特征：1）涌现现象，长思维能力并非凭空产生，而是通过回溯、自检等行为激活；2）推理边界，任务复杂度超越模型逻辑容量阈值会导致准确率下降；3）过度思考，揭示了推理长度与精度并非线性正相关，需防范错误累积；4)推理侧扩展，证明通过增加推理时的计算投入能显著提升逻辑上限；5）在训练范式上，模型面临过程监督与结果监督的路径选择，需平衡监督细粒度与奖励作弊风险；6）模型在强化学习中会产生顿悟时刻，需通过维持“策略熵”来防止思维收敛导致的推理崩塌。

图表21：长思维链在提升大模型推理能力出现的现象



资料来源：《Towards Reasoning Era: A Survey of Long Chain-of-Thought for Reasoning Large Language Models》-LARG 等，国盛证券研究所绘制

- **智谱采用交错式思考+保留式思考+轮级思考。**交错式思考下，模型会在每次响应和工具调用前都会进行思考，提高了指令遵循和生成质量。保留式思考更侧重编程智能体场景，模型会自动在多轮对话中保留所有思维块，复用现有的推理过程而非从头重新推导。减少了信息丢失和不一致性，适合长期且复杂的任务。而轮级思考支持模型在会话内对每个轮次的推理进行控制，对轻量级请求禁用思维以降低延迟/成本，对复杂任务启用思维以提高准确性和稳定性。

- **MiniMax 推出交错思维链。**显式推理与工具调用之间交替进行，并将推理结果持续带入后续步骤。这一过程能显著提高模型在长程任务中的规划能力、自我修正能力与可靠性。通过将冗长、重度依赖工具的任务转化为稳定的 “计划→行动→反思” 循环，交错思维链有效减少了状态漂移与重复性错误，确保每一步行动都基于最新的证据。

2.1.3 长记忆突破：Engram 与嵌套学习领衔最新进展

模型对上下文的理解与记忆，构成模型在 Agent 及 C 端应用的瓶颈。在 Agent 场景中，模型需要处理的不再只是原始问题，还包括任务所处的阶段、已完成的步骤、待解决的子任务，以及工具执行返回的环境反馈。与此同时，模型往往需要借助外部工具来获取信息或执行操作，这些交互会动态更新上下文，并持续影响后续决策。因此，模型需要在更大的时间尺度上理解上下文。在 AGI-Next 前沿峰会上，腾讯首席 AI 科学家姚顺雨认为上下文环境是模型在 C 端应用的核心瓶颈。

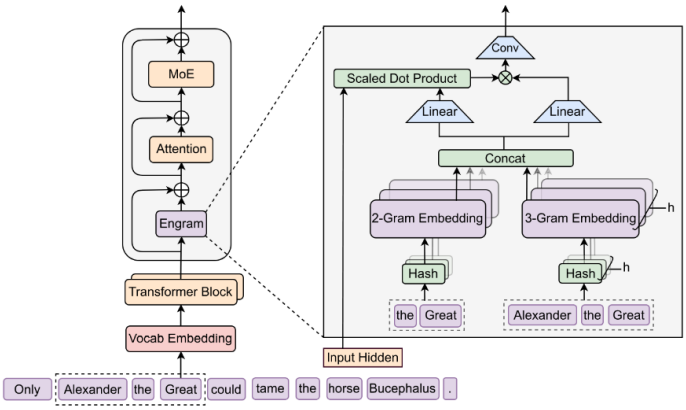
Anthropic 提出 AI Agent 上下文工程的三大核心准则。鉴于大语言模型受限于有限的注意力预算，因此高效的上下文工程需要优化 token 的效用，以持续实现期望的输出结果。具体准则包括：1）系统提示词应采用精简、直观且严谨的语言，追求以最小信息密度触发期望行为；2）工具设计需具备极高的容错性，并确保其功能边界与预期用途高度清晰；3）提示示例应聚焦于构建一组具备多样性与典型性的样本示例，从而精准描绘出 Agent 的期望行为。

在腾讯提出的上下文学习能力的基准测试上，主流模型表现不佳。腾讯混元团队最新提出了一个关于上下文学习的基准测试 CL-bench，检验模型是否能从给定上下文中吸收此前预训练没学过的知识（领域知识、规则系统、复杂流程、从数据归纳的规律），再利用

新知识去完成任务。CL-bench 将上下文学习能力的不足分为忽略、误用以及格式/约束不遵守三大类型，研究结果表示主流模型的错误比例都很高。例如，GPT-5.1 在忽略上下文/误用上下文/格式不遵守情况分别占比 55.3%/61.5%/35.3%。在学术界，通过反思、压缩、记忆系统来解决上下文的问题，在特定场景下有效，但本质上都是在回避核心问题。而 DeepSeek 的显式记忆结构、多通路处理不同类型上下文的方法，或将真正解决模型在上下文的学习问题。

DeepSeek 提出条件记忆，增强上下文能力的同时降低计算成本。条件记忆的重点在 Engram 模块，旨在通过将静态模式存储与动态计算在结构上分离，增强 Transformer 骨干网络。Engram 通常会被插入到 Transformer 的前期层，能够卸载静态模式的重建工作，减轻骨干网络的计算负担。此外，也能保留足够的上下文信息，使得门控机制能够判别所需要激活的记忆类型。Engram 显著扩展了长上下文能力，并在长文本任务中表现突出，例如 Multi-Query NIAH 指标从 84.2 提升到 97.0。

图表22：Engram 的架构



资料来源：《Conditional Memory via Scalable Lookup: A New Axis of Sparsity for Large Language Models》-Deepseek 团队等，国盛证券研究所

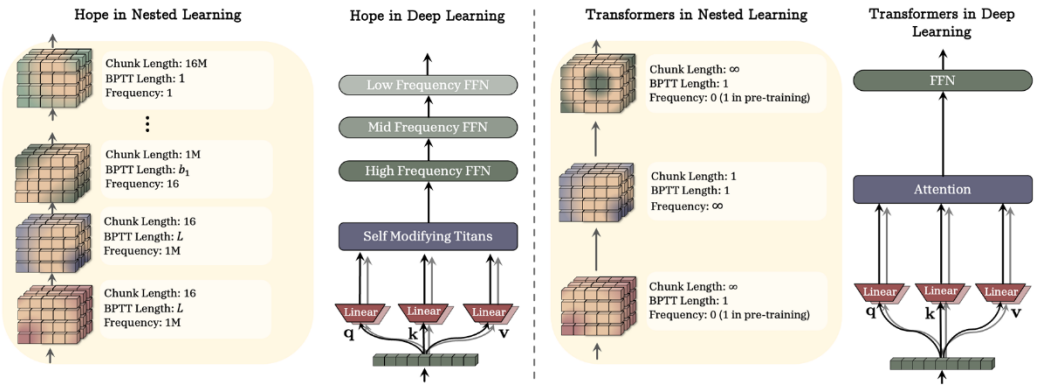
图表23：MoE 与 Engram 的长文本表现对比

Model	RULER (32k)							
	NIAH Accuracy (↑)					Other Tasks (↑)		
	S	MK	MV	MQ	VT	CWE	FWE	QA
MoE-27B *(50k, 1.63)*	**100.0**	88.0	92.7	84.2	77.0	4.5	73.0	34.5
Engram-27B *(41k, 1.66)*	99.6	88.3	93.0	89.5	83.2	3.8	99.6	**44.0**
Engram-27B *(46k, 1.63)*	97.6	89.0	95.5	**97.0**	87.2	4.3	98.6	37.5
Engram-27B *(50k, 1.62)*	99.3	**89.3**	**96.5**	**97.0**	**89.0**	**5.9**	**99.3**	40.5

资料来源：《Conditional Memory via Scalable Lookup: A New Axis of Sparsity for Large Language Models》-Deepseek 团队等，国盛证券研究所

嵌套学习或将成为解锁模型长期记忆以及持续学习的关键。谷歌提出嵌套学习（Nested Learning）的新型学习范式，将深度学习架构重新定义为一套由不同更新频率驱动的嵌套、并行优化问题系统，每个组件本质上都是通过压缩自身上下文流来工作的关联存储模块。基于此理论，谷歌提出了结合自修正 Titans 与持续体记忆系统的 HOPE 架构。实验表明，HOPE 在语言建模、常识推理和长上下文记忆等任务中展现出显著优势，为实现真正具备上下文记忆以及持续学习能力的下一代大模型提供了可行路径。

图表24：Hope 架构与 Transformer 的比较

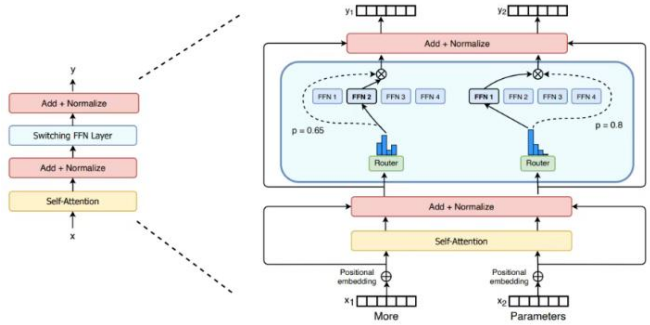


资料来源：《Nested Learning: The Illusion of Deep Learning Architectures》-谷歌团队，国盛证券研究所

2.1.4 计算效率优化：以 MoE 架构与创新注意力机制为主流

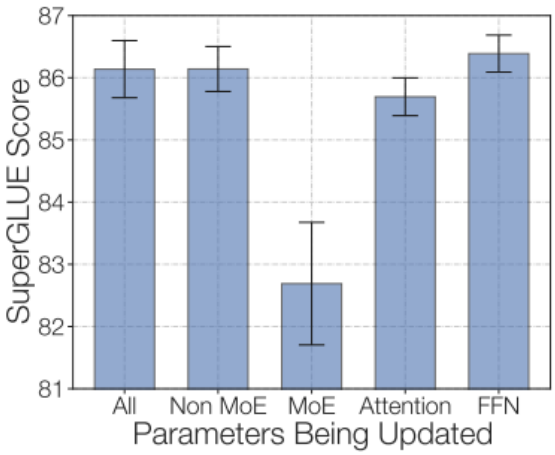
MoE 显著提升模型训练和推理效率。混合专家模型（MoE）由稀疏 MoE 层与门控网络或路由组成，稀疏 MoE 层代替了传统 Transformer 模型中的前馈网络层，包含若干专家，每个专家本身是一个独立的神经网络。而门控网络或路由用于决定 token 被发送到的专家方向。MoE 实现了在相同的计算预算条件下，显著扩大模型或数据集的规模。在预训练阶段，MoE 相较于稠密模型能够更快地达到相同的质量水平。在推理阶段，推理速度比相同参数的模型更快。但 MoE 会占用大量显存，因为所有专家系统都需要加载到内存中。此外，在微调方面也存在诸多挑战（如泛化能力不足、过拟合等）。

图表25：Switch Transformer Layer



资料来源：Hugging Face 公众号，国盛证券研究所

图表26：MoE 与其他架构的 SuperGLUE 分数对比



资料来源：Hugging Face 公众号，国盛证券研究所

新注意力机制旨在提升 Transformer 架构的注意力机制的计算效率。MoE 解决了 FFN 层的计算开销，新的注意力机制则旨在攻克 Transformer 架构的另一个核心瓶颈，自注意力机制（self-attention）与序列长度 L 的二次方计算复杂度 O(L^2)，这会导致长序列处理时计算开销过大。

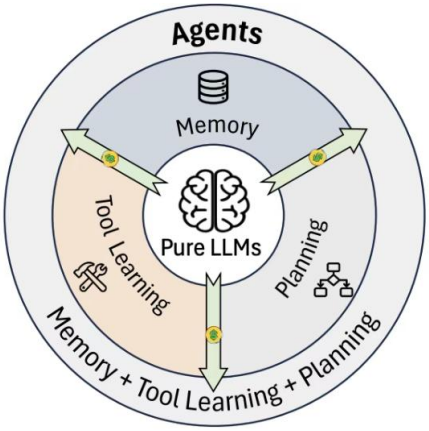
- **DeepSeek 的多头潜在注意力（MLA）：**通过将长序列的 Key 和 Value 向量（即 KV 缓存）压缩成一个单一的、低秩的潜在向量来解决 KV 缓存瓶颈，极大地减少了存储历史信息所需的内存需求，KV 缓存相较于前代模型减少了 93.3%。

- **MiniMax-M1 的闪电注意力（Lightning Attention）：**将注意力计算分成块内和块间两部分，块内用传统注意力计算，块间用线性注意力的核技巧，避免了累积求和操作拖慢速度。Lightning Attention 还采用了分块技术充分利用 GPU 硬件，让内存使用更高效，训练速度不随序列长度增加而变慢。在架构中，每隔七个使用线性注意力的 Transformer 块插入一个使用标准 softmax 注意力的完整 Transformer 块，理论上可以让推理长度扩展到数十万个 token。

2.2 Agent：将 AI 应用扩展至生产力领域

推理能力的成熟为 AI Agent 的崛起奠定了基础。2024 年被视为“Agent 元年”，因为推理能力的提升使得 Agent 能够有效解决任务规划、错误识别等核心难题。Agent 与传统 LLM 的本质区别在于，它不仅能回答问题，还能通过“感知—规划—行动”这一闭环过程执行任务，具备更强的任务执行力。Agent 通常由四个核心模块组成：大脑（大语言模型）、记忆（上下文持久化）、工具使用（API 调用）和规划（任务分解）。其中，大语言模型作为中心决策引擎，负责理解目标并制定多步执行计划。

图表27：完整 Agent 的组成部分



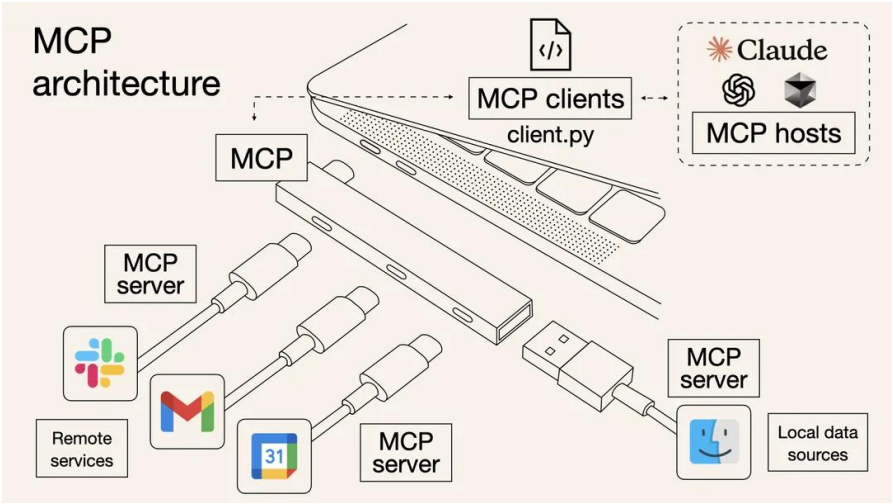
资料来源：《Toward Efficient Agents: A Survey of Memory,Tool learning, and Planning》-上海人工智能实验室等，国盛证券研究所

Anthropic 和智谱为端侧 Agent 技术应用的先行者。Anthropic 的 Claude 3.5 Sonnet 率先在 2024 年 10 月推出“Computer Use”功能，并推出 API。Claude 能够使用标准工具和程序，帮助用户完成简单的重复性任务。此外，Claude 还具备多模态能力，能够理解桌面图像并推理执行任务（如填写表格、导航网页等）。同期，智谱推出了 AutoGLM—全球首个具备“Phone Use”能力的 AI Agent，能够通过文字或语音指令模拟手机操作，并支持 Web 端的检索和信息总结功能。

MCP 协议与 Skills 的推出，共同加速 Agent 的进化。2024 年 11 月，Anthropic 推出了 MCP 协议（Model Context Protocol，模型上下文协议），解决了 Agent 与外部工具连接的“接口”问题，能够将 API 调用转化为具备上下文感知的协作指令集，实现从自然语言指令到任务执行的完整自动化过程。MCP 生态系统包括三个核心角色：MCP Host（负责接收用户需求并调用大语言模型）、MCP Server（提供工具和服务的“货架”）和 MCP Client（负责中间通信协调）。2025 年 10 月，Anthropic 在 Claude 模型中推出 Claude Skills 功能，定位解决通用大模型在垂直场景中“知道但不会做”的问题，提升任务执行的可靠性与一致性。Skills 是一种约定标准，本质上是一个包含元数据、脚本、模板、参考指令等的文件夹，将某个领域沉淀的一套成熟的方法论后实现可复用可管理可分享可迭代。据 FreeBuf 与社区监测数据，公开可查的 Skills 数量已超过 10 万，且保持每周数千新增的速度（如 skills.sh 平台每小时新增 550+技能）。头部 Skill 如 find-skills（用于发

现其他技能）安装量达 194.1K+，成为事实上的“技能搜索引擎”。Skills 正在成为 AI 时代的“AppStore”。

图表28：MCP 架构



资料来源：腾讯技术工程，国盛证券研究所

Agent 能力的提升将 AI 应用扩展至生产力领域，编程场景最为明显。典型的产品如 Cursor 和 Manus: Cursor 是一款基于 VSCode 开发并集成多种顶尖模型（如 Codex 5.3、Claude Sonnet 4.5、Gemini 3 Pro）的 AI 编程工具，能够自动生成代码、优化代码结构并支持在编程软件中实时交互。2025 年 11 月，在完成 D 轮融资后，Cursor 估值已达到 293 亿美元。Manus 是由中国团队推出的一款通用型 AI Agent 产品，通过调用多个第三方模型（如 Claude、Qwen 等）提供智能服务；Manus 采用了 Multi-Agent 架构，在完整的沙盒环境中运行，具备互联网访问、持久文件系统、安装软件、创建自定义工具的能力。这意味着 Manus AI 可以独立进行工作，记住上下文并交付可用于生产的结果。2025 年 12 月，Manus 以 20 亿美元的价格被 Meta 收购，而该月初 Manus 的 ARR 刚突破 1 亿美元。

图表29：Cursor 的应用界面

资料来源：Cursor 官网，国盛证券研究所

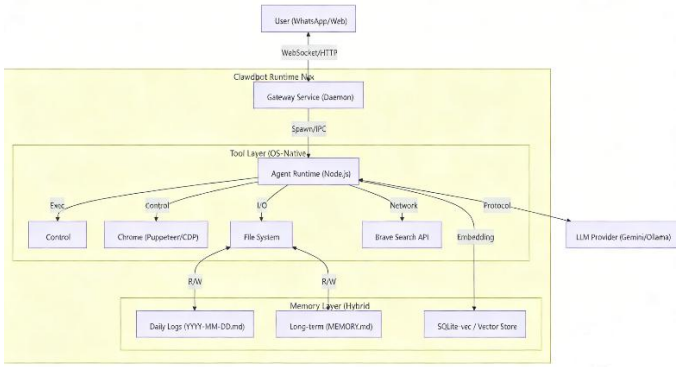
图表30：Manus 官方页面

资料来源：Manus 官网，国盛证券研究所

OpenClaw 为运行在本地设备的 Agent，广受开发者关注。2026 年初，AI Agent 产品 OpenClaw 火爆出圈，其将大模型的能力与用户的本地系统、工具链和通讯软件深度结合。OpenClaw 具有四大技术优势：1）本地特权，区别于沙盒化的 Agent，OpenClaw 可

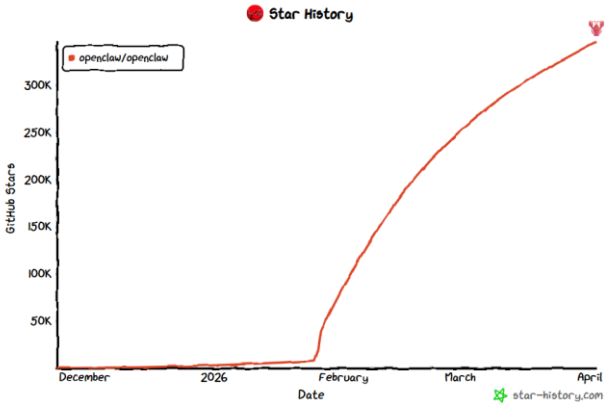
以管理文件、运行脚本、部署代码；2）私有记忆，数据不出域、完全透明，可由用户手动编辑维护；3）多模态交互，支持 WhatsApp 语音、图片等；4）浏览器混合控制，接管用户已打开的 Chrome 实例，复用 Cookie 和登录态。因此，OpenClaw 可以处理跨平台、需要系统级权限且带有自动化性质的任务，如邮件管理、网页自动化、开发者效率工具、实时监控与研究等。截至 2026 年 4 月 11 日，OpenClaw 在 Github 上已累计获得 354k Star 数。

图表31：OpenClaw 核心架构



资料来源：腾讯云开发者，国盛证券研究所

图表32：OpenClaw Star 数量变化趋势



资料来源：Github，国盛证券研究所

国内厂商积极拥抱 OpenClaw，并基于其构建云端 AI 助手产品。目前，国内厂商月之暗面、MiniMax 都已基于 OpenClaw 构建了云端 AI 助手。MiniMax 推出的 MaxClaw 直接集成在 MiniMax Agent 网页端，在云端部署并运行 OpenClaw，无需用户自备服务器或 API Key。此外，用户可通过飞书、钉钉、Telegram、WhatsApp 等渠道与 Agent 内置的 Expert 进行 7×24 小时交互和任务下发，交付内容均可在网页端同步查看，实现云端跨端协作。月之暗面推出的 KimiClaw 支持一键云端部署，拥有 ClawHub 超 5000 个社区插件。用户可以在官网上直接使用，也可配置在飞书群里随时召唤。

3. 解构 MiniMax 产品的驱动力

3.1 人才：扁平灵活的组织构架最大化释放人才能量，实现创新与执行力

公司独特的组织架构推动了公司的可扩展性，助力研发工作的快速推进和高频迭代。自创立以来，公司营造赋能型文化，各层级员工均能各展所长。在公司富有远见的创始人闫俊杰博士的带领下，公司始终站在 AI 行业发展的最前沿，致力于推动 AI 从研究转化为应用。

特意保持扁平灵活的组织架构，CEO 之下最多只有三层，从而实现更快的决策，促进智能大模型迭代。公司以项目制模式组建跨职能团队。打破技术、产品和业务之间的传统隔阂，所有团队都围绕一个目标保持一致：提升公司模型的智能水平，并使其对每个人都能触手可及。

公司通过包容的环境来促进研究和创造力，重视多样化贡献和持续改进。给予年轻人充分授权：许多毕业不久的同事如今已领导关键研发项目。公司不对工作范围设限，积极鼓励个人承担超出其既定角色的更多责任。这在公司的 R&D 团队负责人中得到充分的体现，他们在刚担任负责人时通常年龄在 30 岁以下。公司的激励机制灵活且以绩效为导向，包括持续的薪资调整和项目奖金。

研发人均薪酬远高于计算机行业平均水平，高薪酬激励锁定高精尖技术人才。根据 MiniMax 招股书披露数据计算可知，公司 2025 年年化研发人均薪酬约 94 万元，相比之下申万计算机板块 2024 年平均人均薪酬仅为 25 万元。公司给到研发的薪酬水平远高于行业平均，以高薪酬激励锁定高精尖技术人才。

图表33：Minimax 研发人员年化薪酬估算

MiniMax 2025Q3 研发人员薪酬（千美元）	28,531
2025 年化研发人员薪酬（万人民币）	26,629
2025Q3 研发人员人数（人）	284
2025 年年化研发人均薪酬（万元）	94
申万计算机板块 2024 年平均人均薪酬（万元）	25

*资料来源：MiniMax 招股书、Wind，国盛证券研究所测算（原始数据来自于招股书，数据截至 2025Q3，按/3*4 做年化计算，假设 1 美元兑 7 人民币的汇率）*

3.2 技术：坚持第一性原理，多模态融合前沿

公司在模型及产品演进方向上，始终锚定 AGI。创始人闫俊杰基于在上一代 AI（图像与视觉）领域的深厚经验，判断传统专用模型过于依赖场景定制，导致模型复用率低，边际成本难以下降，从而制约了 AI 能力的规模化扩展。因此，MiniMax 始终聚焦于提升模型的通用能力，并将通用人工智能（AGI）定义为接近通过图灵测试的智能体。公司成立之初便同步布局语言、视觉和声音三大模态的大模型体系，旨在通过多模态融合实现 AGI。在技术演进上，依据 OpenAI 提出的 L1–L5 能力路线图，当前大模型整体已接近 L3 水平，对应 MiniMax 从 Talkie/星野等早期产品逐步演进至现阶段的智能体产品形态。

图表34：OpenAI 定义的 L1-L5 五级线路图

级别	名称	描述
L1	Chatbots	可对话的 AI
L2	Reasoners	具备像人类一样解决问题能力的 AI
L3	Agents	能行动的 AI
L4	Innovators	能辅助发明的 AI
L5	Organizations	能像组织一样运作的 AI

资料来源：OpenAI，国盛证券研究所

战略原点：以智能水平为核心变量，构建长期竞争优势。大模型行业的市场空间呈现出由模型智能水平这一核心变量驱动的非线性增长特征。模型能力每一次关键性涌现，都会解锁全新的应用与商业场景，例如 AI 从对话能力延展至科学研究等高复杂度逻辑推理领域。创始人闫俊杰提出，AI 时代的核心资源在于智能本身。围绕这一认知，MiniMax 将提升模型智能水平构筑其竞争力，并以技术进步作为公司长期发展的主要驱动力。具体来看，公司已逐步构建起较为完善的多模态模型矩阵，通过“模型性能提升带动商业化收入增长，收入反哺持续研发”的正向循环机制，推动企业实现长期增长。

架构演进：第一性原理下的工程理性。创始人闫俊杰基于“效率与效果可相互转化”的第一性原理，带领团队在架构选型上展现出极高的战略定力与灵活性：

- **MoE 架构的先行者：**在 2023 年，创始人闫俊杰便判断竞争焦点将从参数规模转向单位成本优化，率先在国内落地 MoE（混合专家）架构。公司上线国内首个 MoE 大模型 abab 6，并进一步推出 abab 6.5 系列，实现 MoE 架构在中国市场的首次稳定商用部署，验证了其技术判断的可行性与工程落地能力。

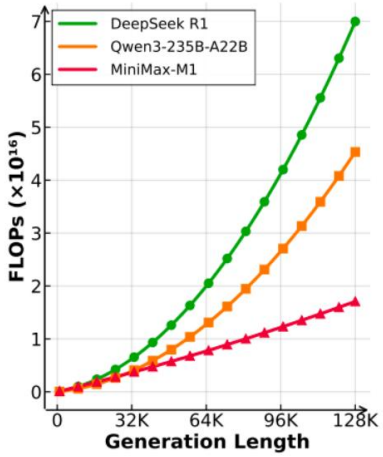
●坚持相关共识的研究：

1）闪电注意力（Lightning Attention）旨在解决传统 Transformer 模型中 Softmax 注意力存在的 $O(n^2)$计算复杂度问题，尤其是长文本处理的瓶颈。通过采用线性注意力变体，计算复杂度被降低至 O(n)，同时提升了模型在处理大量输入输出文本时的效率。

2）交错思维链（Interleaved Thinking）指在显式推理与工具调用之间交替进行，并将推理结果持续带入后续步骤。这一过程能显著提高模型在长程任务中的规划能力、自我修正能力与可靠性。通过将冗长、重度依赖工具的任务转化为稳定的“计划→行动→反思”循环，交错思维链有效减少了状态漂移与重复性错误，确保每一步行动都基于最新的证据。

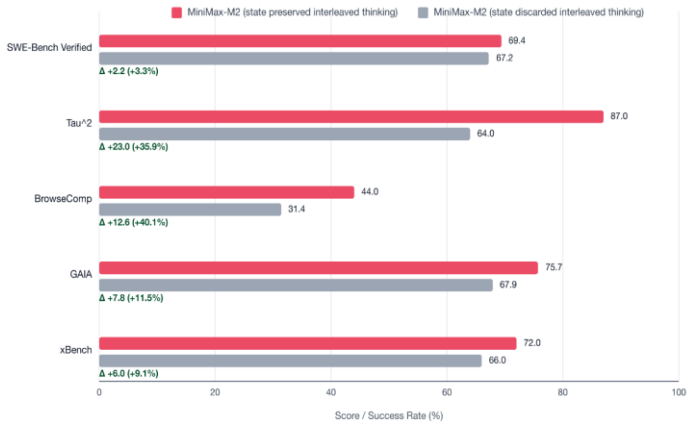
- **M1 至 M2 的理性迭代：Attention 机制的取舍。**M1 系列采用 Lightning Attention 与 Softmax 混合架构，在长文本（100k+Token）场景下，FLOPs 仅为传统架构的 25%-50%，显著降低长上下文的推理成本。但随着模型向代码、数学、Agent、多模态、长 CoT 与 RL 扩展，线性注意力在数值稳定性、低精度计算与工程复杂度上的约束逐步显现。M2 系列最终放弃线性注意力，回归全注意力，并结合 MoE 架构与交错思维链，以换取更强的工程稳定性与推理一致性。但 MiniMax 并未否定线性注意力的长期价值，而是在当前算力与应用约束下优先保障模型的可用性与商业落地能力。这种在“前沿探索”与“工程实现”间的平衡能力，也是 MiniMax 的核心竞争力之一。

图表35：MiniMax 与 Deepseek、Qwen 的算力消耗对比



资料来源：《MiniMax-M1: Scaling Test-Time Compute Efficiently with Lightning Attention》-MiniMax 团队，国盛证券研究所

图表36：MiniMax-M2 模型有无交错思维链的 benchmark 对比



资料来源：MiniMax 官方公众号，国盛证券研究所

算法与数据：重塑后训练效率边界。在通用能力的构建上，公司通过算法创新解决了传统强化学习和数据匮乏的痛点：

- 突破传统强化学习算法：CISPO（Clipped Importance Sampling Policy Optimization）算法从底层机制上改造了 RL 稳定性控制逻辑，在不牺牲训练数据利用率的前提下显著提升收敛效率，是 MiniMax 降低大模型训练门槛的关键技术之一。传统 RL 方法通过 Token Clipping 维持训练稳定，但容易误删高价值学习信号。CISPO 转而裁剪重要性抽样权值，而非直接裁剪 Token，使所有样本仍参与核心优化过程。实验数据表明，在 M1 架构全量 RL 训练中，仅耗时 3 周、花费 512 张 H800，计算成本控制在 53.5 万美元左右，显著优于 Deepseek 的 GRPO 与字节跳动的 DAPO。

- 逻辑推理的合成数据引擎：SynLogic 针对高质量逻辑推理训练数据匮乏的问题，构建 SynLogic 合成数据框架。实验证实，基于 SynLogic 的训练不仅提升了数学与代码能力，还能有效泛化到数学和代码以外的未曾训练的领域，为开发更强的通用推理模型提供了重要方向。

多模态能力：语言、视频、语音模型均位居行业前列。MiniMax 成立之初便同步布局语言、视觉和声音三大模态的大模型体系，在多模态领域已建立起深厚的技术壁垒。

- 语音方面：MiniMax-Speech 建立高保真零样本标准。MiniMax 通过引入可学习说话人的编码器并提出 Flow-VAE 架构，率先实现高保真的内生零样本语音克隆，显著提升 TTS 在内容创作与情感交互场景中的商业可用性。在 2026 年 3 月 6 日的 Artificial Arena TTS 排行榜中，MiniMax 在全球厂商中位列第三，ELO 评分达到 1107，仅次于 OpenAI 与 InWorld。

- 视觉方面：Minimax 提出 VTP 框架，解决了视觉 Tokenizer 的缩放悖论。在不改变下游 DiT 模型训练配置的前提下，仅通过提升 Tokenizer 预训练阶段的 FLOPs，即可带来 FID 指标 65.8%的相对提升，验证了视觉 Tokenizer 具备独立且显著的缩放价值。

- 多模态强化学习：Minimax 提出了 V-Triune 系统，探索在统一 RL 框架下同时提升视觉语言模型的复杂推理与精细感知能力。该系统通过引入 Dynamic IoU 机制解决感知任务奖励设计困难的问题，并通过冻结 ViT 主干缓解多模态联合训练的不稳定性，为构建高阶多模态模型提供了可行工程路径。

图表37：MiniMax 核心技术布局与工程化路线图

架构演进	• MoE架构先行者：abab 6是国内首个MoE大模型。 • M1系列：采用Lightning Attention与Softmax混合架构，显著降低长上下文的推理成本。 • M2系列：理性迭代。回归全注意力，并结合MoE架构与交错思维链，以换取更强的工程稳定性与推理一致性。
算法与数据	• CISPO：裁剪重要性抽样权值，显著提升收敛效率。在M1架构全量RL训练中，仅耗时3周、花费512张H800，计算成本约53.5万美元。 • SynLogic：合成数据框架。提升数学与代码能力并展现出了泛化能力。
多模态能力	• 语音：引入可学习说话人的编码器并提出Flow-VAE架构，率先实现高保真的内生零样本语音克隆。 • 视觉：提出VTP框架，在不改变下游DiT模型训练配置的前提下，仅通过提升Tokenizer预训练阶段的FLOPs。 • 多模态强化学习：引入Dynamic IoU机制解决感知任务奖励设计困难的问题，并通过冻结ViT主干缓解多模态联合训练的不稳定性，为构建高阶多模态模型提供了可行工程路径。

资料来源：来自 MiniMax 团队的《MiniMax-M1: Scaling Test-Time Compute Efficiently with Lightning Attention》、《SynLogic: Synthesizing Verifiable Reasoning Data at Scale for Learning Logical Reasoning and Beyond》、《MiniMax-Speech: Intrinsic Zero-Shot Text-to-Speech with a Learnable Speaker Encoder》、《Towards Scalable Pre-training of Visual Tokenizers for Generation》、《One RL to See Them All: Visual Triple Unified Reinforcement Learning》，国盛证券研究所绘制

纵观 MiniMax 的技术演进，其核心竞争力并非源于单一技术范式的固守，而是依托第一性原理驱动的工程实用主义。公司通过探索线性注意力机制，始终致力于寻找计算效率的新摩尔定律，M2 系列模型在架构上的调整展现了团队卓越的自我迭代与纠错能力。这种不依赖路径依赖、以最终场景效果为导向的决策机制，是大模型公司保持长期竞争力的关键。在多模态技术方面，MiniMax 通过在编码阶段最大化信息的语义密度，显著降低了下游生成模型的学习门槛与计算负载。这种系统级的全局优化能力，构成了 MiniMax 在大模型赛道上的独特壁垒。

3.3 工程化能力和算力效率：软硬协同极致优化，算力利用率高

效率也是模型能力的边界，MiniMax 在效率维度具备优势。CEO 闫俊杰认为，在算力资源受限的客观约束下，效率直接决定了模型能力的上限。若算力资源仅为同业的 1/10，唯有通过极致的系统优化释放潜力，才能在模型能力上追平甚至反超。因此，公司并未采取“先冲击能力边界，后优化效率”的模式，而是采用了两者同步推进的研发策略。通过在算力架构、算法设计与系统优化等多个维度的提升，增强模型能力。在衡量推理模型算力使用效率的 MFU 指标上，MiniMax 模型的 MFU 超过 75%，显著优于行业平均的 40%-50%水平。高 MFU 水平意味着计算资源得到了更充分的释放，这不仅是推理成本降低与性能增益的核心驱动力，更为系统的大规模扩展奠定了坚实的基础。

自研基础设施构筑成本优势。公司成立之初即组建内部基建团队，自主研发 AI 基础设施，以支持全面灵活的模型训练和推理，具体包括训练及推理框架、并行和可扩展能力以及自动化的运维支持等。AI 基础设施已形成多维度的工程优势，具体包括：

- **系统级框架优化（算子层）：**自研训练与推理框架。针对模型训练与推理所涉及的基础计算单元进行深度优化，通过并行计算策略、多级键值（KV）缓存机制及分布式专家并行推理架构，实现了计算堆栈的系统级优化，显著降低延迟并提升了算力利用率，完美适配自研模型。

- **资源的统一动态调度（集群层）：**打破了训练与推理的资源物理隔离。公司部署的智能调度系统可将数据处理、模型评估等离线任务设为“可抢占作业”。当高优先级的训练或推理任务启动时，系统自动回收低优先级任务的算力资源，既保障了时效性需求，又极大降低了集群闲置率。

- **跨集群的高可用架构（网络层）：**为支持规模化部署及企业级应用需求，在基础设施设计阶段即引入多计算集群架构。结合实时负载监测与反馈调节，实现了跨集群的动态调度与资源切换。该策略体系能够在不同集群之间分配推理请求和计算资源，降低了单一集群故障对整体服务影响，并支持算力的弹性扩展，确保了服务的高可用性、容错性及规模化部署的运行要求。

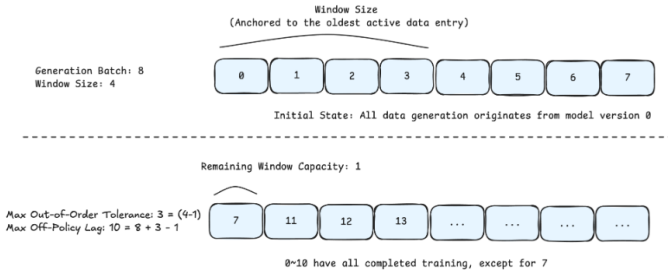
图表38：AI 基础设施的主要亮点

资料来源：MiniMax 招股书，国盛证券研究所绘制

Agent RL 架构加速模型训练过程，打破“不可能三角”。 MiniMax-M2.5 采用的原生 AgentRL 系统，在吞吐量、训练稳定性与 Agent 灵活性这三者之间取得平衡。构建了标准化的 Agent 与 LLM 交互协议，使不同类型的 Agent 脚手架能够被统一纳入训练体系，从而支撑超大规模强化学习的稳定运行。在此基础上，配合高度工程化的系统优化以及成熟的算法与奖励设计，整体训练效率与可控性显著提升。

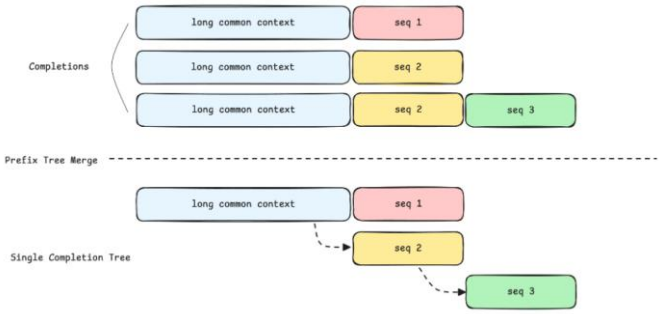
在工程实现层面，主要创新体现在三个方面：1）引入 Windowed FIFO 调度策略，在保障系统吞吐能力的同时有效约束样本的 off-policyness，提升训练稳定性。2）通过 Prefix Tree Merging，将原本线性的训练样本重构为树形结构，实现约 40 倍的训练加速，并显著降低显存占用。3）结合 Dynamic MTP、Rollout 阶段的 PD 分离、全局 L3 KV Cache Pool，对 LLM 推理链路进行针对性优化。

图表39：Windowed FIFO 原理



资料来源：MiniMax 官网，国盛证券研究所

图表40：Prefix Tree Merging 原理



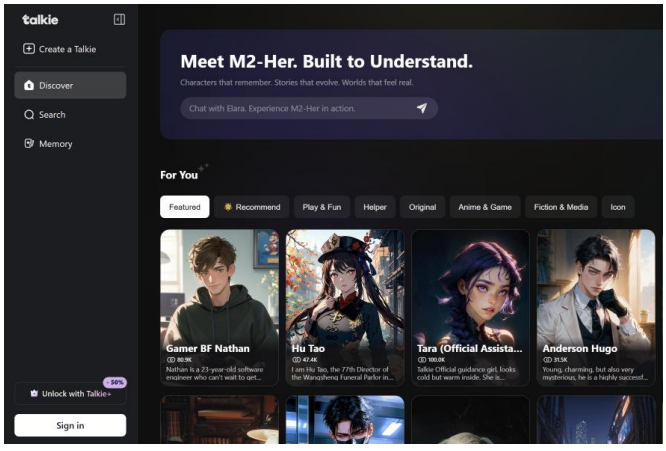
资料来源：MiniMax 官网，国盛证券研究所

3.4 产品力：C 端产品布局 AI 娱乐与虚拟陪伴，产品全球认可度持续提升

娱乐是生产力之外的 AI 第二大赛道。全球千万的年轻用户在其中创建自己的个性化 AI agent 并与之对话互动，下游场景极其长尾，包括互动、生活陪伴和广泛的日常问答等。娱乐赛道的竞争趋于稳定，头部代表产品为 Character AI、Talkie/星野，共性是通过优化模型能力提升用户体验、通过丰富的多模态创作工具激发用户想象力创作，从而做到高用户时长、用户黏性、深度互动体验。

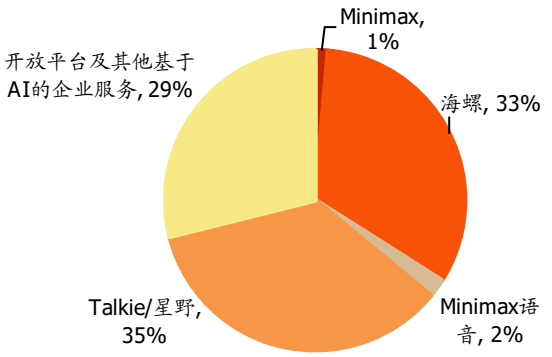
基于多模态的全栈模型能力布局，公司推出 C 端 AI 陪伴产品 talkie/星野。公司模型布局涵盖文本、语音、视频、图像与音乐五大方向，全栈的多模态能力使得公司得以在 AI 娱乐方向具备竞争力。基于此，公司推出 Talkie 和星野：Talkie（面向国际市场）/星野（面向中国本土市场）是全球公认的 AI 原生全模态交互平台。依托公司的自研 AI 模型支持，用户可与 AI 智能体或虚拟角色互动产生情绪链接。截止 2025 年前三季度，这两款产品合计贡献公司收入的 35%。

图表41：Talkie 官网



资料来源：Talkie 官网，国盛证券研究所

图表42：2025Q1-Q3 公司收入占比情况



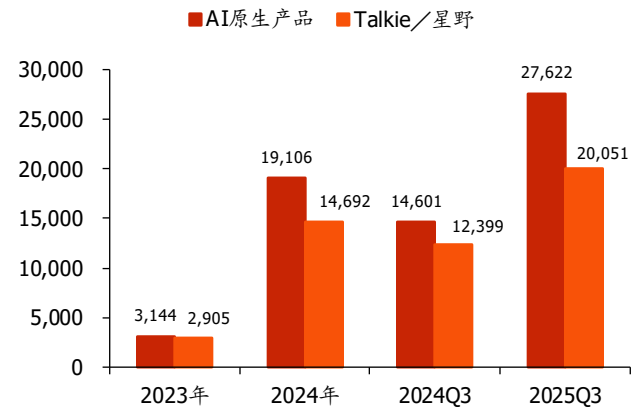
资料来源：MiniMax 招股书，国盛证券研究所

月活用户数、付费用户数、付费率持续提升，体现出公司产品在全球市场的认可度。截至 2025 年 9 月 30 日，公司的 AI 原生产品累计为来自超过 200 个国家及地区的逾 2 亿

名个人用户，以及来自超过 100 个国家及地区的超过 10 万家企业以及开发者提供服务。通过 MiniMax 招股书披露数据可知，公司 C 端产品平均月活用户数、付费用户数与付费率持续提升，印证了公司 C 端产品在全球市场的认可度。

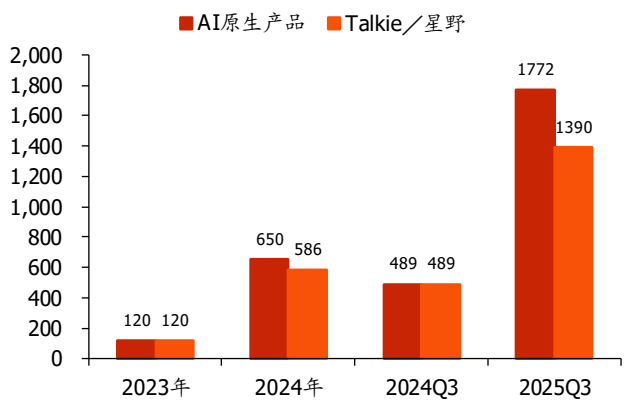
用户留存能力提升，品牌知名度和市场接受度持续改善。虽然公司战略性调整以减少营销及推广开支，并更加注重自然用户获取及用户质量，将重点转向 MiniMax 的变现阶段；然而尽管营销开支水平降低，但新用户下降幅度远小于相关营销预算的削减幅度，反映 MiniMax 及 Talkie／星野的用户留存能力提升、品牌知名度提高以及市场接受度持续改善。

图表43：公司平均月活用户数（千用户）



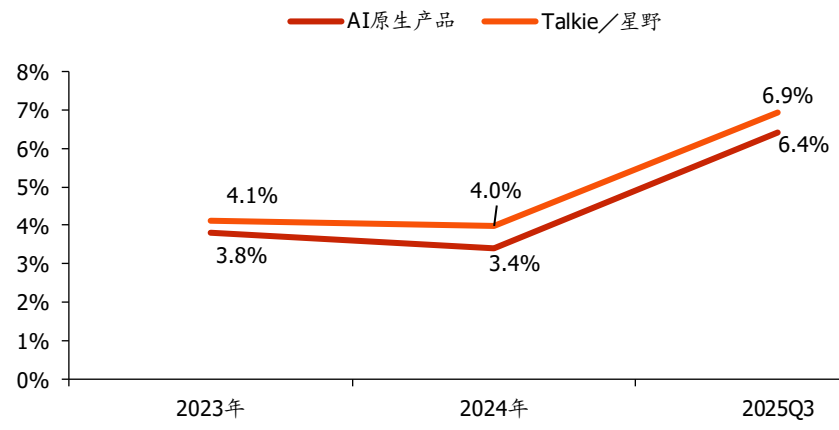
资料来源：MiniMax 招股书，国盛证券研究所

图表44：公司付费用户数（千用户）



资料来源：MiniMax 招股书，国盛证券研究所

图表45：公司 C 端产品付费率（按付费用户/平均月活用户数计算）



资料来源：MiniMax 招股书，国盛证券研究所

4. 财务预测及估值

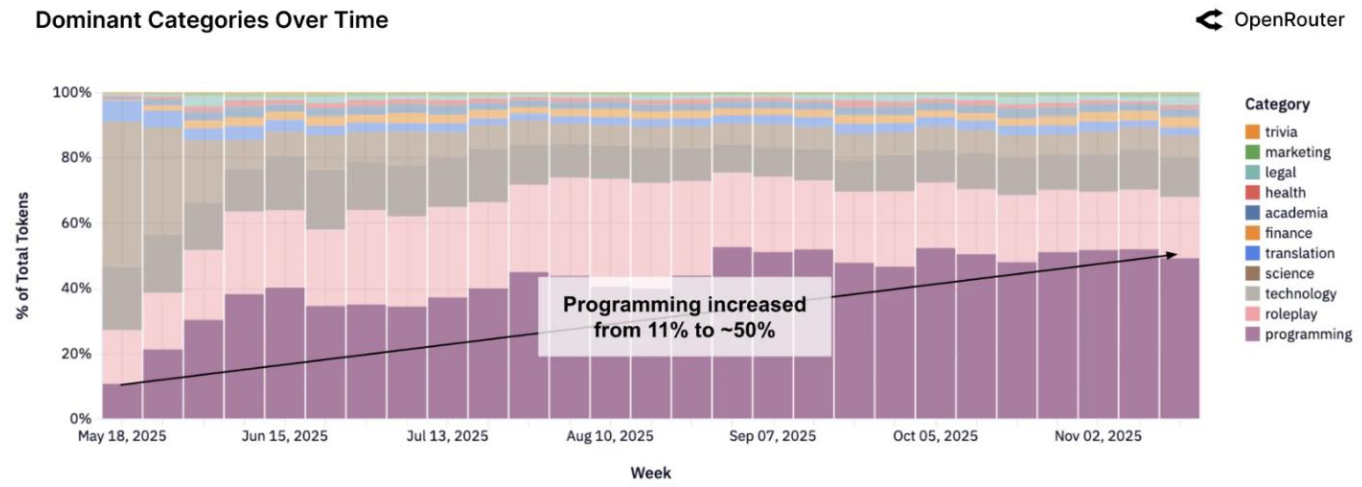
4.1 核心商业化方向：编程、办公与多模态创作

4.1.1 编程：M 系列模型兼具高智能与高性价比

LLM 嵌入开发者工作流，Coding 占 AI 使用的比重已超 50%。在 2025 年 12 月 OpenRouter 发布的 100 万亿 Token 实证研究中，编程是目前 AI 使用的最大场景。编程类请求的占比已从 2025 年初的 11%飙升至 50%以上，反映了 AI 从探索、对话的使用向应用任务（如代码生成、调试和数据脚本）的转变。随着 LLM 嵌入到开发人员工作流中，AI 作为编程工具的角色正成为开发者的标配。

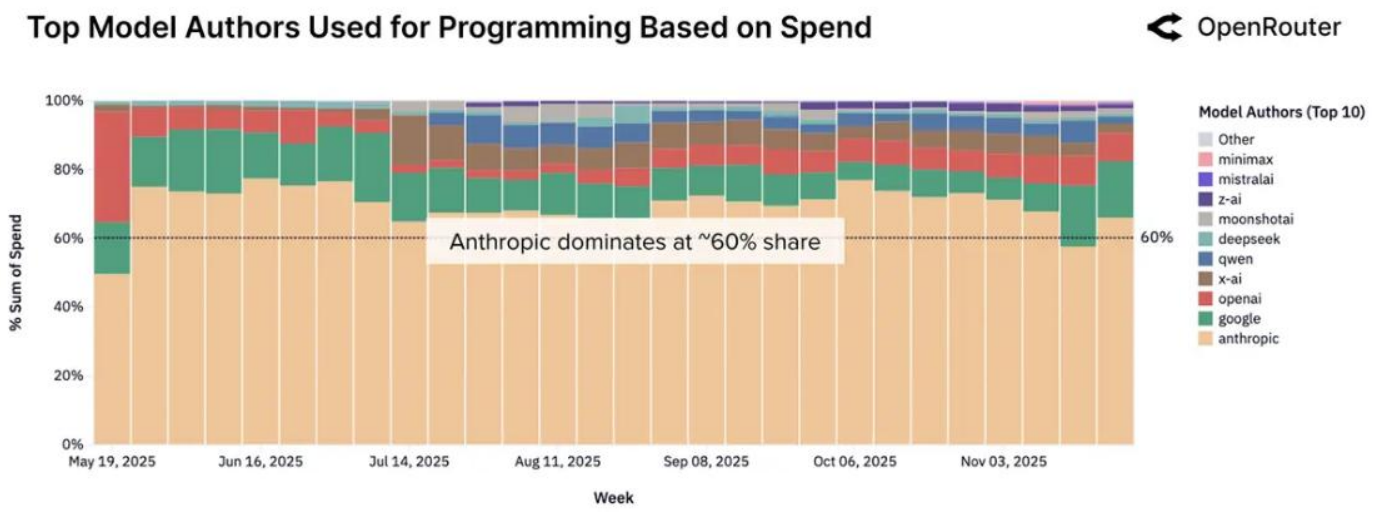
Anthropic 为编程领域的王者，国内模型厂商逐步得到市场认可。在细分 Coding 领域的主要参与者当中，Anthropic 的 Claude 系列始终占据主导地位。在 2025 年 5 月至 12 月期间，Anthropic 的市占率大都维持在 60%以上。谷歌的 Gemini 模型份额稳定在 15%左右，而 OpenAI 从年初的 2%上升至 8%。中端市场也在加速，中国厂商如智谱、阿里巴巴、MiniMax、Deepseek、月之暗面也持续得到市场关注。

图表46：全球大模型能力的需求转向编程领域



资料来源：OpenRouter，国盛证券研究所

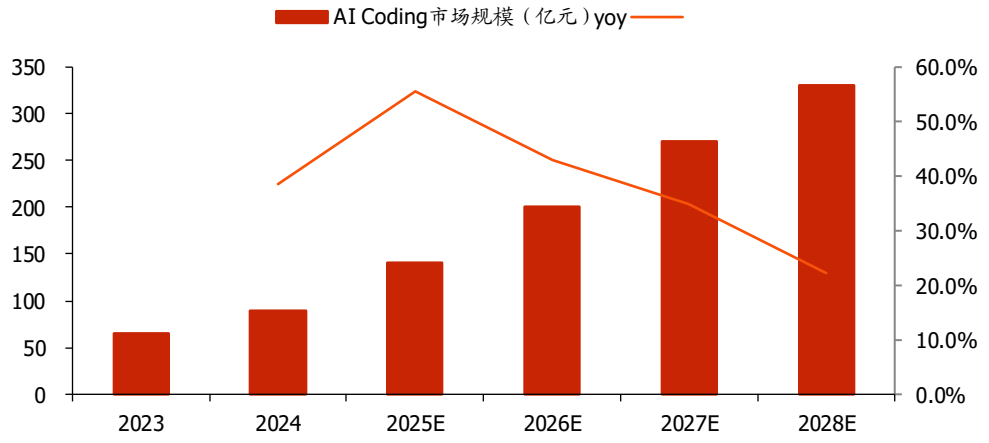
图表47：基于模型支出衡量的头部模型厂商份额



资料来源：OpenRouter，国盛证券研究所

AI Coding 已成企业级生产力工具：微软内部已有 30%的代码由 AI 自动生成。Meta 也预计 AI 很快将承担内部 50%的编程工作量。AI Coding 在国内企业的渗透速度同样显著，腾讯云代码助手在客户侧的采纳率达到 30%、单测执行率提升 18%、代码评审覆盖率提升 20%。在 MiniMax 内部真实业务场景中，整体任务的 30%由 M2.5 自主完成，覆盖研发、产品、销售、HR、财务等职能，且渗透率仍在持续上升。其中，在编程场景表现尤为突出，M2.5 生成的代码已占新提交代码的 80%。根据亿欧智库数据，全球 AI Coding 市场预计在 2031 年达 244.6 亿美元。中国 2024 年 AI Coding 市场规模为 90 亿元，预计 2028 年将突破 330 亿元，年复合增速达 38.4%，市场发展潜力可观。

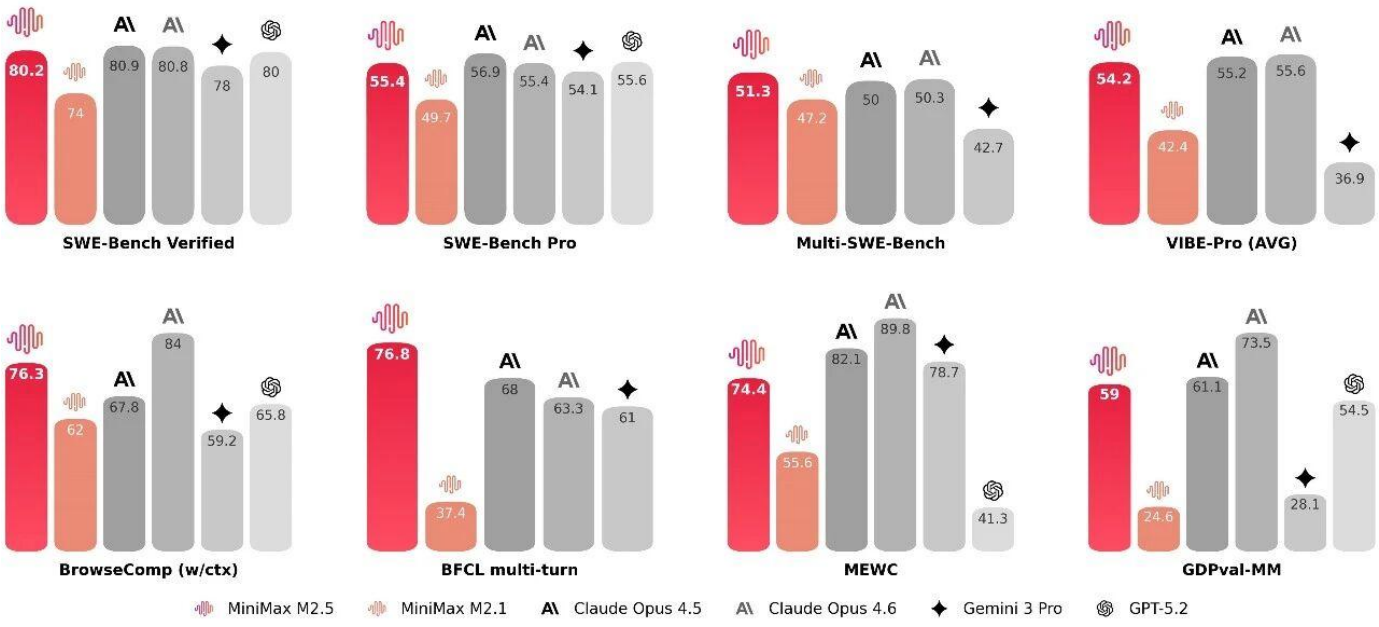
图表48：中国 AI Coding 市场规模（亿元）



资料来源：亿欧智库，国盛证券研究所

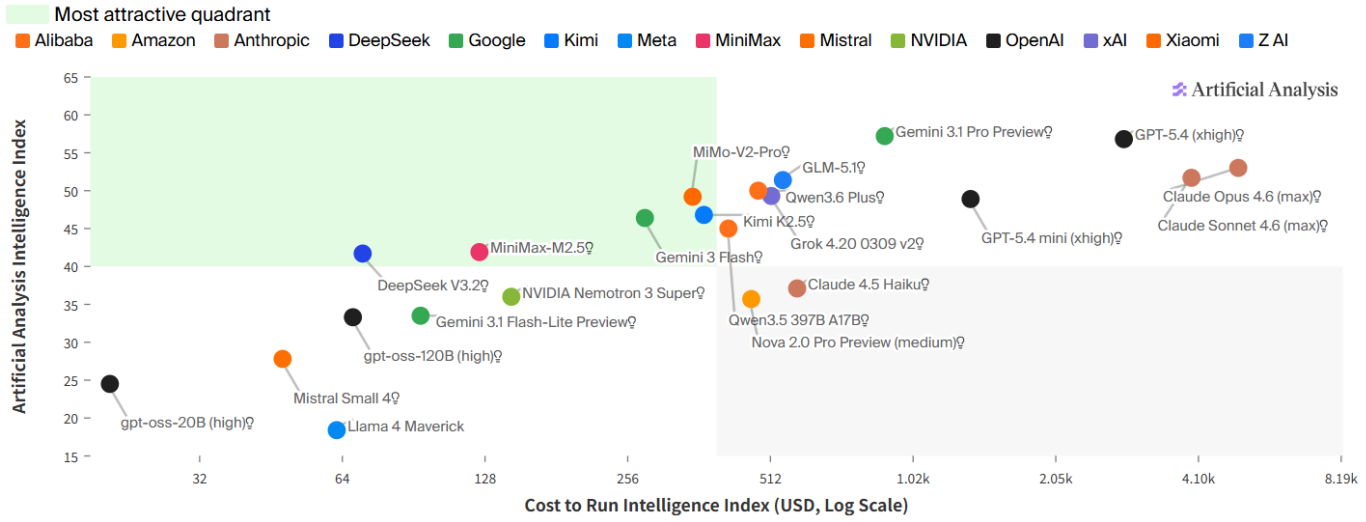
MiniMax M2 系列专为代码和 Agent 而设计。2025 年 10 月发布的 MiniMax M2 通过优化的 MoE 架构，能够在保证模型高性能的同时降低成本与优化推理速度，在 Agent 和代码能力上表现优异。2026 年 2 月发布的 MiniMax M2.5 在编程、工具调用和搜索、办公等生产力场景都达到或刷新了行业的 SOTA。在复杂 Agent 场景下，M2.5 能够拆解复杂任务效率提升、速度更快。同时，M2.5 拥有极高性价比，在每秒输出 100 个 token 的情况下，工作一小时仅需 1 美元。综合模型智能与成本来看，MiniMax-M2.5 已进入最佳模型象限，验证了其“高智能、高性价比”的工程化路径。

图表49：MiniMax M2.5 与海外顶尖模型的 Benchmark 对比



资料来源：MiniMax 稀宇科技公众号，国盛证券研究所

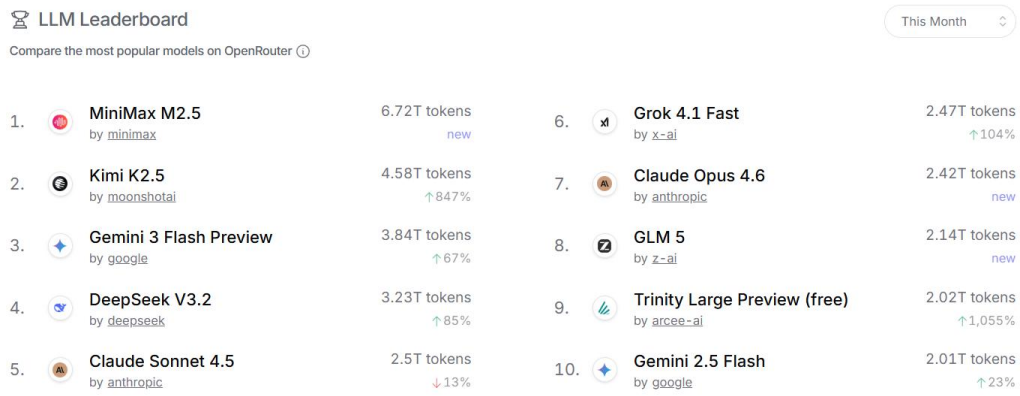
图表50：Intelligence vs. Cost to Run Artificial Analysis Intelligence Index（截至 2026 年 4 月 13 日）



资料来源：Artificial Analysis，国盛证券研究所

主流开发平台的接入，是模型综合性能的重要验证。目前，国内外已有微软、谷歌、OpenRouter、Cline、TRAE 等超 20 家知名开发平台接入了 MiniMax M2。2026 年 1 月，硅谷明星公司 Kilo 在最新编程产品 Kilo for Slack 的发布中，宣布将其默认模型切换为 MiniMax M 系列模型。Kilo Code 的联合创始人兼 CEO 表示，在开发者直接评判的真实编码工作流程中，M2.1 的表现已能够与全球顶尖模型相媲美。自 MiniMax M2.5 模型发布以来，模型在 OpenRouter 平台 3 月份的 Token 总消耗量跃升至全球第一，超越谷歌的 Gemini 3 与 Claude Opus 4.6。

图表51：MiniMax-M2.5 的 Token 消耗量（2026 年 3 月）

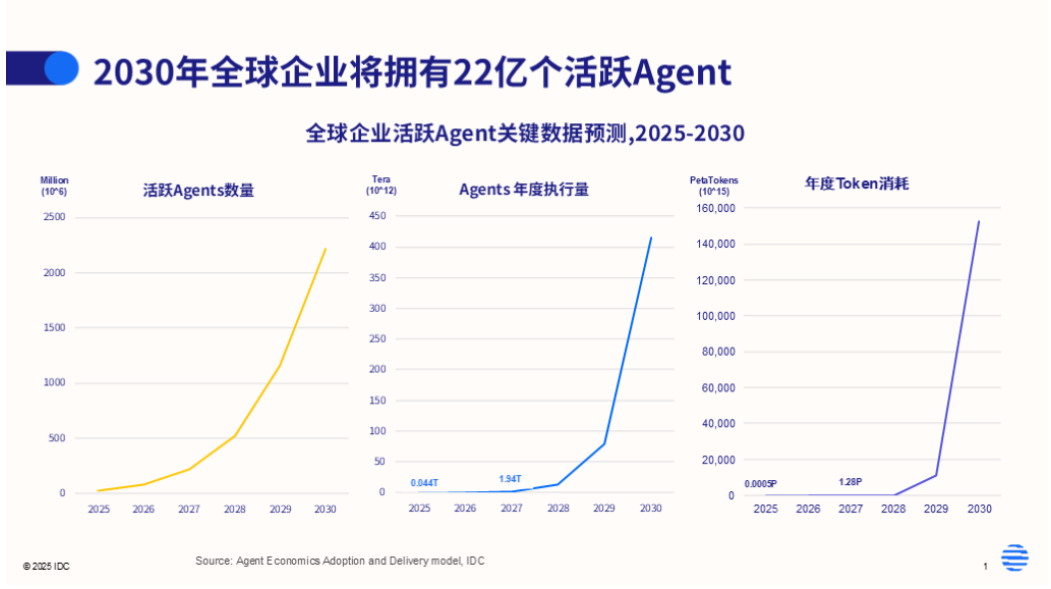


资料来源：OpenRouter，国盛证券研究所

4.1.2 办公领域：MiniMax Agent 逐步实现多场景自动化

Agent 正驱动生产力组织方式的改变。企业内的工作正由一批可调用、可协作、可复用、可交付结果的 Agent 来智能化处理，IDC 预测，1）活跃 Agent 的数量将从 2025 年的约 2860 万，快速攀升至 2030 年的 22.16 亿，年复合增长率达 139%。2）Agent 将深度嵌入进企业的业务流中，年执行任务数将从 2025 年的 440 亿次暴涨至 2030 年的 415 万亿次，年复合增长率高达 524%。3）Agent 处理任务的复杂程度加深，驱动底层 Token 消耗出现数量级跃迁。预计年度 Token 消耗将从 2025 年的 0.0005 Peta Tokens 暴增至 2030 年的 152,667 Peta Tokens，年复合增长率高达 3418%。

图表52：2025-2030 年 Agent 数量、任务执行量及年度 Token 消耗量



资料来源：IDC，国盛证券研究所

Agent 助力 MiniMax 进行 AI 原生组织变革。在推进模型与产品研发的同时，MiniMax 也在持续向 AI 原生组织演进。内部的 Agent 实习生已经覆盖了近 90%的员工，应用场景包括编程开发、数据分析、运维管理、人力招聘及市场销售等多个职能维度。MiniMax

将其内部环境视为 AI 原生能力演化的试验场，并将该组织形式的变革视作其决定未来研发效率的关键变量。

能力产品化落地，MiniMax 完善 Agent 生态布局。2026 年 1 月，MiniMax 正式发布桌面端及专家 Agent，标志着其 Agent 能力从云端交互迈向本地工作场景。该系列产品的核心技术演进与应用优势体现在以下四个维度：

- 长程复杂任务：Agent 实现了跨源搜索、复杂逻辑理解与多条件匹配的深度融合。在处理大规模跨源信息检索时，Agent 能够自主执行网页遍历、并行浏览及细节提取，最终实现非结构化数据的结构化整理。
- 本地化办公环境渗透：通过桌面端入口，Agent 能够直接进入用户工作流。其具备处理海量本地文件及执行跨模态任务的能力，协助用户完成文件分类、信息梳理等高频办公场景任务。
- 专业场景：针对法务、金融等专业行业，Agent 能够对庞杂的行业数据进行深度梳理，并输出具有专业深度的结构化洞察。
- 专家 Agent：通过将特定领域的专业知识、工作流程、输出标准等预先注入到 Agent 中。用户仅需通过自然语言指令，即可驱动 Agent 自动展开完整的专业级工作。用户可以在专家社区内直接使用内置的专家，也可以创建个性化专家 Agent。

图表53：MiniMax 专家 Agent 官方界面

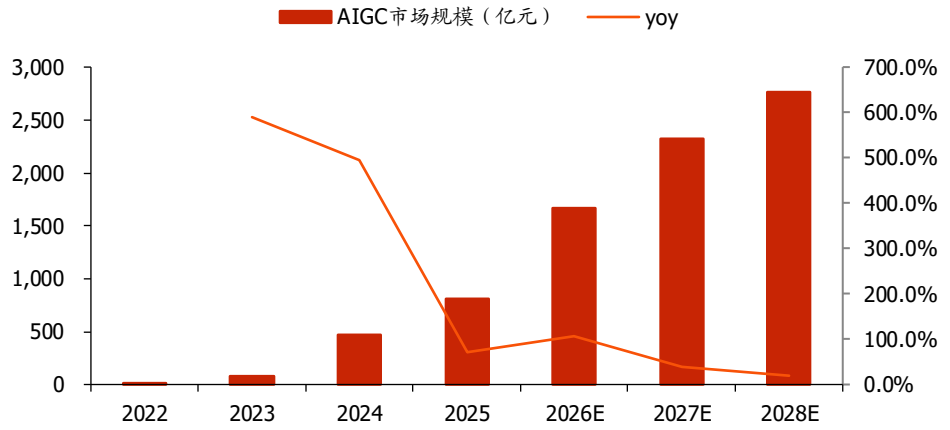
资料来源：MiniMax Agent 官网，国盛证券研究所

4.1.3 多模态：视频与语音模型性能领先

AI 视频模型性能驱动其渗透率的提升，打开 AIGC 市场的长期空间。从消费端来看，视频是用户消费时间最长的内容形态，有丰富的应用的场景和大型内容分发平台，长期或有诞生超级应用的机会。随着 AI 视频生成的能力不断提升，AI 生成视频占视频消费内容

的比例将不断提升，推动内容供给端变革逐步渗透视频消费市场。艾媒咨询数据显示，2025 年中国 AIGC 行业核心市场规模为 805.8 亿元，同比增长 70.8%，预计 2028 年将达 2767.4 亿元，AIGC 行业市场规模未来仍保持高速增长。

图表54：中国 AIGC 市场规模（亿元）



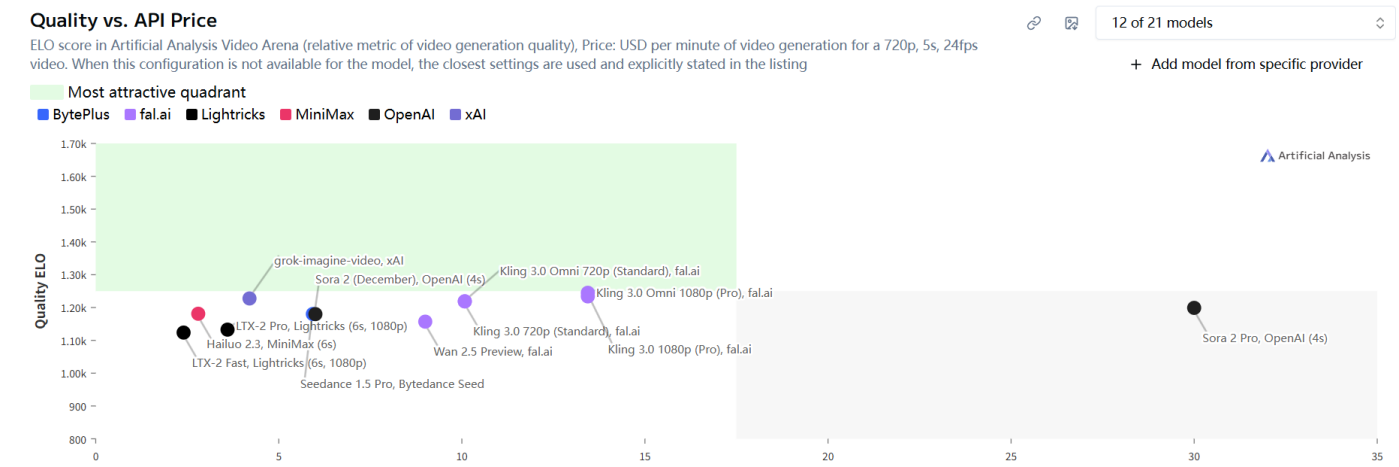
资料来源：艾媒咨询，国盛证券研究所

视频模型性能与成本持续迭代，未来将解锁更多应用场景。目前视频模型存在一定缺陷：生成视频时长短（0-15s）、抽卡属性强、指令遵循能力不足等，所以现阶段 AIGC 市场规模的增长是由模型的性能主导。典型的案例为字节跳动的 Seedance 2.0，该模型在指令遵循、物理规律、智能分镜等能力上显著增强，已成为国内出品的现象级 AI 视频模型。当模型的能力稳定后，AI 视频推理成本的降低将解锁更多专业的商业场景（电影、电视剧等），成为推动行业增长的主导因素。

根据我们测算，视频模型生成一条 1 分钟的视频成本大约在 103.2-350.4 元。具体来看，以 Seedance 2.0 为例，一条 10s 视频需要花费 80 积分，基础会员至高级会员的积分价格区间在 4.3-7.3 元/100 积分，假设生成一段符合要求的视频需要生成 5-10 次，则单条 1min 视频所需的价格在 103.2-350.4 元。而本土动画电影/好莱坞电影的每分钟制作成本为 10 万/200 万美元，因此 AI 视频模型在成本侧具备显著的优势。

Hailuo 02 为出圈最早的视频模型，在综合性价比层面拥有较强的竞争力。MiniMax 视频模型最早出圈的是 2025 年 6 月发布的 Hailuo 02 模型，该模型是全球首个能够稳定生成体操等高难度复杂动作的视频模型。一段“猫猫跳水”视频在 TikTok 发布后 5 小时内播放量突破 150 万次，累计播放量超过 7000 万次，点赞量超过 310 万，并进一步衍生出各类“动物运动会”题材内容。2025 年 12 月，Hailuo 再次引发关注。模型将《三体》原著中不存在的“丁仪砸向水滴”进行视频化演绎，在 B 站获得超过 300 万次播放和 17 万次点赞，进一步验证了其在视频画面的表现能力。根据 Artificial Analysis 的数据，Hailuo-2.3 在性能位于第一梯队的同时，定价处于行业低位。

图表55：视频模型综合性价比对比（截至 2026 年 3 月 6 日）

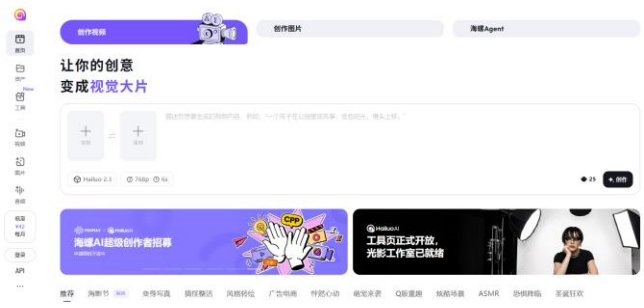


资料来源：Artificial Analysis，国盛证券研究所

海螺 AI 平台以 Hailuo 模型为核心，提供全方位的视频创作解决方案。MiniMax 推出的海螺 AI 平台不仅支持高保真、高质量的视频内容生成，还针对专业级项目提供进阶的镜头运动控制功能，模拟电影导演常用的镜头语言和表现手法。此外，海螺 AI 平台还整合了图像和音频生成模块，为用户提供了完整的视频创作素材支持。图像模型方面，不仅支持自研的 Image 模型，还接入了多家主流厂商的模型，包括阿里巴巴的 Qwen Image、字节跳动的 Seedream 5.0 Lite 等，进一步丰富了用户的图像选择。

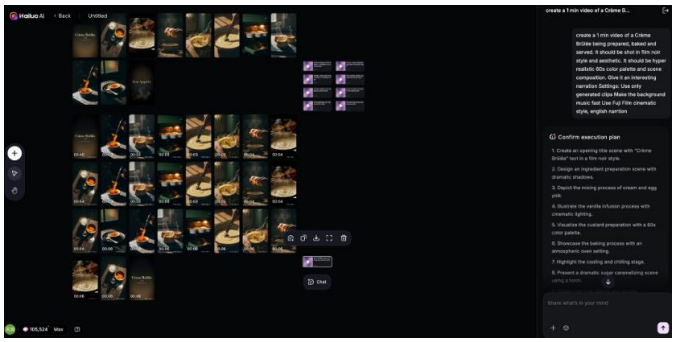
Media Agent 能够同时满足普通用户与专业用户的创作需求。2025 年 10 月推出全模态创作能力的 Media Agent，将 MiniMax 旗下的的视频、语音与音乐模型封装成统一的 Agent。用户只需输入希望呈现的内容，即可完成从素材生成到成片输出的全过程，实现“一键成片”。对于有更高创作要求的用户，还可以在 Agent 工作流中对具体素材片段进行精细化调整，从而在效率与可控性之间取得平衡。

图表56：海螺 AI 网页端界面



资料来源：海螺 AI 官网，国盛证券研究所

图表57：Media Agent 的工作流画面



资料来源：海螺 AI 公众号，国盛证券研究所

TTS 市场高速增长，创新企业与科技巨头并驱争先。文本转语音模型（TTS）通过将数字文本转换为语音，增强了文本内容的可访问性。目前 TTS 主要应用在移动设备和物联网设备（智能音箱、车载系统等）领域。此外，内容创作（如播客、有声读物和新闻播报）领域的公司也正逐步采用 TTS，创造新的收入来源。根据 Global Market Insights 的数据，2025 年全球 TTS 市场规模为 48 亿美元，预计 2035 年将达到 353 亿美元，未来十年的年复合增速达 22.4%。从 TTS 市场的参与者的角度，国外市场上有 Inworld AI、OpenAI、ElevenLabs 等，国内为 MiniMax、阿里巴巴、字节跳动等。

图表58：国内外文本转语音模型及特点

	公司	模型系列	模型特点
国外	Inworld AI	TTS-1.5	首音时间 P90 延迟：Max 模型低于 250ms，Mini 模型低于 130ms。TTS-1.5 表现力、词错误率、幻觉、截断等方面都得到提升，支持超 14 种语言。
	OpenAI	GPT-4o-mini-TTS	开发者可以预设多种语音风格，还能根据指令调整语音风格，比如口音、情感范围、语调、语速等，支持 57 种语言。
	ElevenLabs	Eleven v3	聚焦情感表达与表演，通过音频标签实现全方位情感控制，有声读物、角色配音和高度创造性的内容。支持超 70 种语言。
国内	MiniMax	Speech-2.6	端到端延迟低于 250ms，支持多种语言的网址、邮箱、电话号码、日期及金额等非标准文本格式的直接转换，Fluent LoRA 提升音韵自然度。支持超 40 种语言。
	阿里巴巴	Qwen3-TTS	支持音色克隆、音色创造、超高质量拟人化语音生成，以及基于自然语言描述的语音控制。端到端合成延迟低至 97ms，10 种语言与 9 个精品音色。
	字节跳动	Doubao-Seed-TTS 2.0	解锁深度语义理解和上下文理解能力，针对教育场景专项优化，使得全科复杂公式符号的合成平均准确率高达 90%左右。模型在互动拟人感、情感演绎与指令遵循能力较好。支持中英日和西班牙 4 种语言以及中国 7 种方言口音。

资料来源：InWorld AI 官网、ElevenLabs 官网、火山引擎官网、OpenAI 官网、MiniMax 稀宇科技公众号、通义实验室公众号、火山引擎公众号，国盛证券研究所

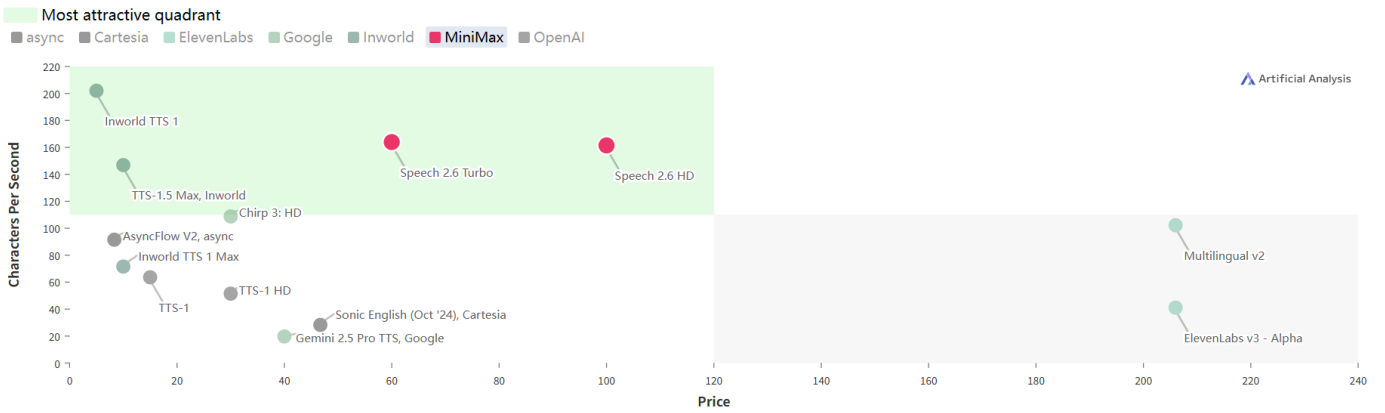
Speech-02 性能位居全球第三。2025 年 5 月发布的 Speech-02 模型，在智能语音合成能力上实现了显著跃升。Speech-02 系列已迭代至 Speech-2.6-HD 与 Speech-2.6-Turbo 两个版本，分别面向高保真音质与快速、实时处理场景进行优化。截至 2026 年 2 月，MiniMax 的 Speech-02-Turbo 在 Artificial Analysis 的语音榜单上位列全球第三，仅次于 Inworld 与 OpenAI。在保持高性能的同时，Speech 模型还具备价格优势，处在全球语言模型行业中的最佳性价比区间内。

图表59：全球文生语音领域排行榜（截至 2026 年 3 月 6 日）

	Creator	Model	ELO	95% CI	Samples	Released
1	Inworld	Inworld TTS 1 Max	1,162	-15/15	2,164	Jun 2025
2	OpenAI	TTS-1	1,111	-8/8	6,913	Nov 2023
3	MiniMax	Speech-02-Turbo	1,107	-11/11	3,592	Mar 2025
4	ElevenLabs	Multilingual v2	1,105	-7/7	10,206	Aug 2023
5	StepFun	Step TTS 2	1,075	-20/20	1,213	Dec 2025
6	Fish Audio	OpenAudio S1	1,070	-9/9	5,108	Jun 2025
7	Kokoro	Kokoro 82M v1.0 Open Weights	1,060	-9/9	5,746	Jan 2025
8	Amazon	Polly Generative	1,057	-10/10	4,350	May 2024
9	Cartesia	Sonic 3	1,054	-14/14	2,270	Oct 2025
10	Hume AI	Octave 2	1,051	-14/14	2,444	Oct 2025

资料来源：Artificial Analysis，国盛证券研究所

图表60：语音模型综合性价比对比（截至 2026 年 3 月 6 日）



资料来源：Artificial Analysis，国盛证券研究所

Speech 模型一经上线，迅速获得了多家知名企业的合作。2023 年 11 月，公司推出首个 Speech-01 模型。模型上线九个月，累计服务超 2,000 家企业用户，为语言学习、PC 语音助手、语音声聊唱聊、情感配音等十余种场景提供落地解决方案，与 Haivivi、阅文起点中文网、呱呱有声、高途教育等国内知名公司达成合作。随着 Speech 模型性能的迭代，其已成为全球语音智能领域的核心基础设施之一。多个知名平台和技术框架已将 Speech 模型作为底层技术引擎，包括支撑 ChatGPT 高级语音模式的 LiveKit、GitHub 热门开源框架 Pipecat、YC 孵化的语音平台 Vapi 等。在智能硬件领域，Fuzozo 和 Rokid Glasses 等新兴产品也选用 Speech 模型，以提升其自然语音交互体验。

图表61：Speech 模型的客户合作案例

公司	场景	解决方案
Haivivi BubblePal	智能玩具	MiniMax 先进的语音合成、文本模型的技术实力，使每一次互动都更适合儿童对话场景，让 BubblePal 随时准备回应儿童的好奇问题，充分激发孩子们的好奇心和想象力。
阅文起点有声书	听书	MiniMax 凭借其在长文本语音生成和超长文本语音生成的技术优势，能快速理解上下文整体语境，在长篇小说的有声读物制作中保持情感的一致性，同时准确解析角色情绪，进行风格化演绎。
呱呱有声	合成音色辅助创作	MiniMax 为“呱呱有声”提供了领先的超拟人语音合成技术，涵盖 30 多种丰富音色，支持汉语、英语、西语、韩语、日语在内的十种语言。
高途	AI 教育	精准还原吴彦祖的声线特点，如美式发音的连读技巧和语调韵律，能基于对话场景实时调整情绪表达。

资料来源：MiniMax 稀宇科技公众号、MiniMax 官网，国盛证券研究所

4.2 财务预测

收入方面，我们预计 2026-2028 年收入为 7.0、13.2 亿美元。以长期视角，我们预计 MiniMax 业务会由“Agent+Generative 开发平台及企业服务”三大结构组成，背后反映的是公司模型与产品一体化能力。

- 开发平台及企业服务：核心在于模型的智能、价格与速度。MiniMax 坚持通过自研基础设施、模型算法等工程化方式，实现训练和推理成本的极致优化。根据 Artificial Analysis 的数据，MiniMax-M2.5 是行业内为数不多位于最佳性价比区间内的模型，

且生成速度快，让复杂 Agent 的运行在经济成本上变得可无限扩展。在 2026 年 2 月，M2 系列文本模型平均单日 token 消耗量已增长至 2025 年 12 月的超过 6 倍。我们认为，当模型智能提升打开新应用场景后，智能、价格与速度”将成为 开发平台及企业服务业务重要的驱动因素。

- Agent：包含目前的 MiniMax Agent 与星野/talkie 等。两者对应了两大不同场景：MiniMax Agent 对应生产力场景，而星野/talkie 则对应生活陪伴场景。MiniMax 在 2025 年业绩公告上提到，内部的 Agent 实习生已经覆盖了近 90%的工作人员，包括编程开发、数据分析、运维管理等职位内容，并且预计 2026 年办公领域 AI Agent 的交付能力和渗透率会显著提升。星野/talkie 已经积累了大量寻求虚拟情感陪伴的原始用户，在生活陪伴场景上具备产品、数据等优势。我们认为，模型智能的每一次突破都将带来新的应用场景以及原有产品的升级，MiniMax Agent 与星野/talkie 或将随着模型能力的升级进化为不同的 Agent 产品形态。

- Generative Media：包含海螺 AI、语音、音乐业务等。MiniMax 在语言、视频、语音、音乐等主要模态均拥有了具备全球竞争力的模型，未来将通过将各部分模型融合训练，打造出具备强大生成与理解能力的多模态模型。目前，国内尚未有厂家能够实现该一目标，而 MiniMax 已占据先发优势。

毛利率方面，影响因素主要包括模型智能水平、收入结构、定价策略以及推理成本效率等。随着 MiniMax 模型的智能水平、模型及系统效率、基础设施资源配置优化持续提升，我们预计 2026-2028 年毛利率为 27.1%、30.6%、34.6%。

费用方面，销售及营销、行政及研发占比会随着规模效应而逐步稀释。但在大模型行业中，模型迭代速度与研发能力的积累是公司的重要壁垒，因此我们预计研发投入仍会保持高增速的态势，预计 2026-2028 年研发开支为 5.3、7.3、8.5 亿美元，占收入的比重为 196.3%、103.4%、64.0%。

在归母净利润方面，由于销售、行政、研发等开支会随着公司规模的扩大会持续投入，但增速小于整体收入的增速。我们预计 2026-2028 年公司的归母净亏损分别为 6.6、8.8、7.8 亿美元，Non-GAAP 归母净亏损为 6.3、8.1、6.4 亿美元。

图表62：MiniMax 核心财务预测：年度

	2024	2025	2026E	2027E	2028E
总收入（百万美元）	31	79	269	703	1321
AI 原生产品	22	53	142	401	757
开放平台等企业服务	9	26	127	302	564
yoy	782%	159%	241%	161%	88%
AI 原生产品	2777%	143%	168%	182%	89%
开放平台等企业服务	223%	198%	390%	137%	87%
毛利额（百万美元）	3.7	20.1	73.1	215.2	456.8
GPM	12%	25%	27%	31%	35%
销售费用率	-285%	-66%	-40%	-29%	-18%
行政费用率	-9%	-47%	-40%	-24%	-11%
研发费用率	-619%	-320%	-196%	-103%	-64%
归母净利润（百万美元）	-465	-1872	-656	-883	-776
NPM	-1524%	-2368%	-244%	-126%	-59%
调整后归母净利（百万美元）	-244	-251	-629	-812	-644
Adjusted NPM	-800%	-317%	-234%	-116%	-49%

资料来源：公司公告、Wind，国盛证券研究所

图表63：MiniMax 核心财务预测：半年度

	25Q4	26H1e	26H2e	27H1e	27H2e	28H1e	28H2e
总收入（百万美元）	26	84	185	289	414	568	753
AI 原生产品	15	46	96	162	239	328	429
开放平台等企业服务	11	38	89	127	175	240	324
yoy	131%	58%	623%	242%	124%	97%	82%
AI 原生产品	82%	22%	535%	249%	150%	103%	80%
开放平台等企业服务	278%	146%	747%	234%	96%	89%	85%
毛利额（百万美元）		22	51	85	130	190	267
GPM	30%	26%	27%	29%	31%	33%	35%
销售费用率	-49%	-44%	-38%	-33%	-27%	-22%	-16%
行政费用率	-58%	-45%	-37%	-29%	-21%	-13%	-10%
研发费用率	-283%	-231%	-181%	-132%	-84%	-71%	-59%
归母净利润（百万美元）	-1360	-241	-416	-470	-413	-408	-369
NPM	-5311%	-285%	-225%	-163%	-100%	-72%	-49%
调整后归母净利（百万美元）	-61	-232	-397	-441	-371	-351	-293
Adjusted NPM	-238%	-275%	-215%	-153%	-90%	-62%	-39%

资料来源：公司公告、Wind，国盛证券研究所

4.3 估值与投资建议

我们预计 MiniMax 2026-2028 年收入为 2.7、7.0、13.2 亿美元，同比增长 240.8%、161.0%、88.0%；Non-GAAP 归母净亏损为 6.3、8.1、6.4 亿美元。

由于大模型行业尚处于跑马圈地阶段，公司需要保持高强度的研发投入以获得一定的市场份额。模型智能的每次突破，都将给行业带来新的应用场景与收入增量，因此我们采用 P/S 的方法对大模型公司进行估值。我们选取了与公司同属于大模型行业、在港股上市，AI 业务存在一定可比性的公司智谱作为核心可比公司。此外，我们还参考了利用 AI+数据辅助企业决策的软件公司 Palantir 的估值水平。考虑到 MiniMax 模型性能/价格/速度全球领先、全模态模型布局完善以及细分市场成长空间广阔的因素，我们给予公司 2027 年 75 倍 P/S，目标价 1317 港元。

根据我们测算，预计公司未来 Non-GAAP 归母净利润将收窄。主要假设：

- **收入侧：**1）Agent 应用从情感陪伴、Coding、基础的办公场景扩展至更高价值的金融、法律、科学等领域，新场景的扩展能够给公司收入带来增量。同时，Agent 也将衍生出如按结果付费、作为入口收取服务费等新型商业化方式。2）当前视频模型难以同时兼备强大的理解与生成能力，顶尖的语言、视频、语音等多模态模型能力使得 MiniMax 已具备解决该难题的基础能力。此外，多模态模型的训练成本远高于文

本模型，产品定价更高带来多模态业务的高毛利。3）开发平台及企业服务核心在于模型的智能、价格与速度，与 Agent、多模态模型的增长逻辑保持一致。

- **成本侧：** MiniMax 持续通过自研基础设施、算力架构、算法设计等多个维度提升模型训练和推理效率，驱动其模型的计算资源使用效率提升，降低推理成本。MFU 指标已超 75%，显著优于行业平均的 40%-50%水平。2026 年 2 月 M2 系列模型单位推理算力成本环比 2025 年 12 月已降低超 50%，海螺视频生成模型推理延迟环比下降超过 30%。

倘若公司远期年收入能够达到 56 亿美元，稳态毛利率为 60%、各项费用和开支占收入比例为 25%，则稳态净利率为 35%、稳态净利润为 19.6 亿美元，目标市值对应远期 PE 约 27 倍。

- Agent 收入：我们假设 MiniMax Agent 用户数量稳定在 1.5 亿、MAU 占比为 25%、月活付费率为 10%、单用户月度 ARPU 为 40 美元，则 MiniMax 长期稳态收入约为 18.0 亿美元。星野/talkie 方面，根据 Precedence Research 的数据，2035 年全球 AI 伴侣市场规模预计将达到 5524.9 亿美元，2025 年星野/talkie 在市场上估算占比 0.06%，推算 2035 年收入将达到 3.5 亿美元。考虑到产品定位会根据模型的迭代变化、外部市场竞争格局的变化，给予星野/talkie 85%的折价后收入为 3.0 亿美元。综合来看，Agent 远期收入预计能够达到 21.0 亿美元。

- Generative Media 收入：根据 Grandview Research 数据，全球 AI 视频生成的市场规模预计将从 2025 年的 45 亿美元增长至 2033 年的 423 亿美元，复合增长率为 32.2%。2025 年海螺 AI 收入预计达 0.26 亿美元，市占率约为 0.6%。考虑到海螺视频领先的市场地位以及即将推出的理解能力更强的 hailuo-03 模型，叠加语音与音乐模型，假设长期市占率增长至 3%，长期收入至少能达 12.7 亿美元。

- 开发平台及企业服务：未来增长的核心在于模型智能突破带来新的应用场景以及现有场景渗透率的提升，收入由文本模型、视频模型、语音模型、音乐模型、Coding Plan 等构成，区别于 Agent 和 Generative Media 的是用户群体的差异。假设长期 B 端收入与 C 端收入比例为 2：3，则收入预计能够达到 22.4 亿美元。

首次覆盖，给予“买入”评级。

图表64：可比公司估值

代码	股票简称	最新股价	市值：亿	收入：亿			P/S		
				2026E	2027E	2028E	2026E	2027E	2028E
2513.HK	智谱	976	3865	28	67	157	120	50	21
PLTR.O	Palantir	135.7	3244	72	104	149	45	31	22
平均值							82	41	22

资料来源：Bloomberg 一致预期、Wind，国盛证券研究所 注：智谱股价与市值单位为港元，收入单位为人民币；Palantir 股价、市值及收入的单位均为美元。汇率换算采用 1 港元兑人民币 0.9 元，数据日期截至 2026 年 04 月 15 日

风险提示

模型迭代不及预期的风险。在大语言模型行业，模型的每一次迭代意味着模型性能便会提升一个台阶，且当前头部模型的迭代速度较快。若迭代速度不及预期，导致模型处于行业第二、第三梯队，则公司的收入会存在较大的不确定性。

行业竞争激烈的风险。除了独立模型供应商，MiniMax 还面临着头部互联网科技公司的竞争，这类企业的主业资金雄厚、人才密度高，模型水平与独立供应商接近。诸如字节跳动视频模型 Seedance 2.0 的爆火，可能会对公司核心业务造成挤压。

盈利改善不及预期的风险。MiniMax 当前仍面临较大的亏损，且需要投入大量人力和算力研发新模型。若收入侧定价不能有所上升或者成本侧下降放缓，则会持续消耗企业现金流。

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。

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：北京市东城区永定门西滨河路 8 号院 7 楼中海地产广场东塔 7 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