

# AI 产业变革驱动阿里价值重估

## 核心观点

DeepSeek 与 Qwen 引领国内大模型热潮，DeepSeek 通过技术突破、架构创新及模型开源推动行业快速发展，阿里发布 Qwen 2.5 Max 和最新推理模型 QWQ，并开源万相 2.1，进一步巩固技术影响力。国内 CSP 竞相接入 DeepSeek，推动大模型行业集中度提升，DeepSeek 促使国内大模型定价下降，进而推动周边产品创新。AI 助推阿里云增速提升，阿里与苹果合作聚焦侧端用户体验，利用 AI 技术提升手机侧体验，带来新的增长曲线和增量机会。AI 应用市场潜力巨大，逐步起量贡献增长动能。阿里深耕自研模型，聚焦细分场景和智能体开发，加速 AI 应用落地，同时通过 AI 赋能 C 端产品，成立智能信息事业群，通过数据打通与 AI 技术驱动新流量入口，进一步提升用户触达效率。

## 摘要

DeepSeek 与阿里的技术创新引领国内 AI 产业变革。DeepSeek V3 与 R1 模型、Qwen 2.5 Max 与 QwQ 模型、Manus 等火爆出圈，在 AI 领域掀起了巨大波澜。DeepSeek 在技术上的突破主要体现在高效混合精度与强化学习的深度融合，模型架构则聚焦于动态负载均衡与注意力机制的创新，通过软硬件协同优化，进一步挖掘了计算效率的极限。尽管国际头部大模型玩家在技术创新与算力方面仍保持领先优势，但 DeepSeek 模型开源加速了行业技术创新。阿里在大模型领域动作频频，不仅发布了推理模型 QWQ，还重磅开源了万相 2.1，进一步巩固了其技术影响力。随着行业竞争的加剧，头部玩家逐渐出现分流趋势。

DeepSeek 开源推动国内 CSP 发展和竞争，阿里成为其中的主要受益者。国内外 CSP 纷纷接入 DeepSeek，利用其强大的模型能力提升用户体验并吸引流量。阿里与 DeepSeek 建立了深度绑定关系，合作模式不仅包括在阿里云平台上接入开源的 DeepSeek 模型，还涵盖为 DeepSeek 的 C 端用户提供算力支持。这种合作进一步巩固了阿里在云计算和大模型领域的优势地位。头部拥有自研模型的厂家在市场中占据高度集中地位，其他大模型企业面临较大的转型压力。在国内大模型定价方面，DeepSeek 的官方价格成为行业参考标准，各厂商的定价逐渐趋于下降。面对模型调用利润率降低的挑战，阿里通过将模型与其他产品和服务相结合的方式，有效弥补了利润损失，并进一步拓展了其商业模式的多样性。

## 阿里巴巴-W (9988. HK)

维持

买入

崔世峰

cuishifeng@csc.com.cn

SAC 编号:S1440521100004

SFC 编号:BUI663

于伯韬

yubotao@csc.com.cn

SAC 编号:S1440520110001

SFC 编号:BRR519

发布日期：2025 年 03 月 10 日

当前股价：140.00 港元

目标价格 6 个月：175.61 港元

### 主要数据

#### 股票价格绝对/相对市场表现 (%)

| 1 个月            | 3 个月        | 12 个月        |
|-----------------|-------------|--------------|
| 51.90/36.05     | 40.14/23.72 | 67.40/28.99  |
| 12 月最高/最低价 (港元) |             | 124.10/66.70 |
| 总股本 (万股)        |             | 1,900,761.60 |
| 流通 H 股 (万股)     |             | 1,900,761.60 |
| 总市值 (亿港元)       |             | 21,611.66    |
| 流通市值 (亿港元)      |             | 23,227.31    |
| 近 3 月日均成交量 (万)  |             | 8141.01      |

#### 主要股东

|                      |        |
|----------------------|--------|
| SoftBank Group Corp. | 11.99% |
|----------------------|--------|

### 股价表现



### 相关研究报告

【中信建投海外研究】阿里巴巴-W(9988):FY2025Q2 财报点评:核心业务持续投入，AI 助力云业务增长

【中信建投海外研究】阿里巴巴-W(9988):2024 阿里云栖大会:拥抱 AI 新浪潮，推动产业全新升级

【中信建投海外研究】阿里巴巴-W(9988):FY2025Q1 财报点评:利润超预期，淘天变现有望提速

**AI助推阿里云增速中枢上行，苹果阿里合作前景广阔。**传统云业务板块中，IaaS、PaaS 和 SaaS 持续增长，其中 IaaS 业务占比最大，而 PaaS 和 SaaS 则凭借更高的利润率成为盈利重点。AI 驱动 MaaS 与行业解决方案业务板块迅速崛起，成为新的增长引擎。为支撑 AI 发展，阿里加速推进云和 AI 硬件基础设施建设，AI 相关需求带动 Capex 显著增加。多业态叠加 AI 增长动力，共同推动阿里云业务实现全面增长，大模型及 AI 技术重构产品生态，进一步驱动阿里云营收提升。DeepSeek 的开源与普及间接拉动了阿里云端 API 调用量的快速增长，使 MaaS 平台成为阿里云业务增长的关键抓手。此外，阿里与苹果的合作开创了端侧用户新体验，有望为阿里带来新的增长曲线。凭借在模型及数据领域的优势，阿里成功满足了苹果对技术的高标准要求，双方合作的重点在于利用 AI 技术提升端侧（如手机侧）的用户体验，这一战略不仅增强了用户粘性，还为阿里带来了显著的增量机会。通过与苹果的深度合作，阿里进一步巩固了其在 AI 和云计算领域的领先地位，并为未来增长开辟了新的空间。

**AI 应用市场潜力巨大，有望逐步起量贡献增长动能。**阿里持续深耕自研模型，加速推动 AI 应用市场的落地。在 AI 模型布局上，阿里一方面继续加强通用推理模型与深度推理模型的研发，另一方面聚焦细分场景产品和智能体的开发，以满足多样化需求。AI 应用市场潜力巨大，阿里核心聚焦于 AI Chatbot、教育及金融领域，并推出了一系列基于模型的增值服务产品，这些产品凭借高利润率成为重要的盈利增长点。阿里计划通过“百炼”构建大量智能体，目前已将智能体划分为不同层级，以适配多元场景需求。在战略层面，阿里聚焦 C 端产品并叠加 AI 赋能，成立智能信息事业群，通过数据打通与 AI 技术驱动新流量入口，进一步提升用户触达效率。端侧战略主要围绕汽车、手机、PC 三大硬件展开，同时对小硬件领域也采取自主研发或合作的方式推进，全面布局智能硬件生态。通过大模型与数据打通，阿里未来能够升级用户画像能力，提升精准推送的质量，为用户提供更加个性化的体验。

**盈利预测与估值：**预计 FY2025-2026 阿里巴巴的收入增速分别为 6.07% 和 9.27%，Non-GAAP 净利润增速分别为 -2.57% 和 12.91%。其中淘天收入增速分别为 2.33% 和 4.99%，经调 EBITA 增速分别为 -0.86% 和 4.02%，阿里云收入增速分别为 10% 和 15%，经调 EBITA 利润率分别为 9.5% 和 11.0%。采用分部估值法，对应淘天集团（15 倍 PE）、云智能集团（5 倍 PS）每 ADS 估值分别为 141.18 美元和 39.40 美元，维持“买入”评级，目标价 180.57 美元/ADS，对应港股 175.61 港币/股。

**风险提示：**资本开支低于预期；云业务收入增长低预期，利润率提升低预期，AI 收入占比低预期；美国对华芯片制裁影响公司云业务增长；主业复苏节奏低预期，宏观经济及社零增长疲弱，CMR 增长低预期，国内电商市场份额超预期下行，淘天集团利润率低预期；海外电商市场竞争激烈，AIDC 减亏低预期；海外关税政策调整影响 AIDC 增速；行业监管风险；蚂蚁集团金控牌照落地进度低于预期；美联储降息进程低预期；中美关系发展的不确定性；中概股退市风险。

## 目录

|                                       |    |
|---------------------------------------|----|
| 一、DeepSeek 与阿里的技术创新引领国内 AI 产业变革       | 1  |
| （一）DeepSeek 与 Qwen 引领国产大模型爆火出圈        | 1  |
| （二）全球头部玩家竞争依然激烈，DeepSeek 和通义大模型加速迭代   | 5  |
| 二、DeepSeek 开源推动国内 CSP 发展和竞争，阿里显著受益    | 11 |
| （一）国内 CSP 竞相接入 DeepSeek，大模型行业集中度进一步提升 | 11 |
| （二）DeepSeek 促使国内大模型定价下降，进而推动周边产品创新    | 12 |
| 三、AI 助推阿里云增速中枢上行，苹果阿里合作前景广阔           | 15 |
| （一）AI 赋能阿里云，助力增速中枢上行                  | 15 |
| （二）阿里携手苹果开创端侧用户新体验，有望带来新的增长曲线         | 17 |
| 四、AI 应用市场潜力巨大，有望逐步起量贡献增长动能            | 20 |
| （一）持续深耕自研模型，加速 AI 应用市场落地              | 20 |
| （二）战略聚焦 C 端产品叠加 AI 赋能                 | 22 |
| 盈利预测及估值                               | 24 |
| 风险分析                                  | 26 |

## 图目录

|                                  |    |
|----------------------------------|----|
| 图 1:DeepSeek 模型训练流程              | 2  |
| 图 2:DeepSeek MoE 的基本架构           | 3  |
| 图 3:MLA 方法有效降低 KV Cache 和训练成本    | 3  |
| 图 4:模型蒸馏方法                       | 4  |
| 图 5:R1 性能对齐 OpenAI-o1 正式版        | 6  |
| 图 6:R1 蒸馏小模型超越 OpenAI o1-mini    | 6  |
| 图 7:DeepSeek 开源项目总览              | 8  |
| 图 8:万相 2.1 模型规格                  | 9  |
| 图 9:Manus 工具界面                   | 10 |
| 图 10:Manus 自动理解流程                | 10 |
| 图 11:国产大模型图谱                     | 12 |
| 图 12:DeepSeek-R1 API 服务定价        | 13 |
| 图 13: o1 类推理模型输入输出价格(元/1MTokens) | 13 |
| 图 14:自定义插件使用流程                   | 14 |
| 图 15:阿里云收入及增速                    | 16 |
| 图 16:2024 年阿里云收入各业务板块占比          | 16 |
| 图 17:中国智算服务市场规模                  | 16 |
| 图 18:1H2024 中国智算集成服务厂商市场份额       | 16 |
| 图 19:中国大模型未来市场规模                 | 17 |
| 图 20:全球生成式人工智能应用市场规模及拆分          | 17 |
| 图 21:iPhone 界面接入大模型              | 18 |

|                    |    |
|--------------------|----|
| 图 22:百炼应用广场.....   | 22 |
| 图 23:百炼插件广场.....   | 22 |
| 图 24:阿里巴巴生态体系..... | 23 |

## 表目录

|                               |    |
|-------------------------------|----|
| 表 1:Qwen 与 DeepSeek 差异对比..... | 5  |
| 表 2:中美头部大模型产品对比.....          | 6  |
| 表 3:阿里云应用产品.....              | 21 |
| 表 4:阿里巴巴盈利预测（单位：人民币百万元）.....  | 24 |

2025 年农历春节伊始，Deepseek 和 Qwen 大模型的爆火出圈催化了恒生科技的新一轮春季躁动。不同于过去几年的港股反弹，本轮恒科上涨由 AI 产业叙事主导，行情更为结构化，阿里巴巴等类科技属性公司独领风骚，传统类消费属性互联网公司多跑输大盘。互联网监管以来，受行业内卷加剧以及宏观经济复苏节奏等影响，阿里巴巴的估值中枢逐步下行，综合实力长期被市场低估。然而，物极必反，今年以来，AI 产业变革推动阿里云价值重估，同时淘天主业受益于竞争格局的逐步改善以及宏观经济的进一步企稳，也有望迎来估值修复，阿里再次迎来高光时刻。本文将以 AI 和阿里云为主要切入方向，对过去两个月市场对阿里云的主要关注问题进行梳理和解答，以期投资者对阿里的长逻辑和增长空间有更清晰的认识和思考。

## 一、DeepSeek 与阿里的技术创新引领国内 AI 产业变革

### （一）DeepSeek 与 Qwen 引领国产大模型爆火出圈

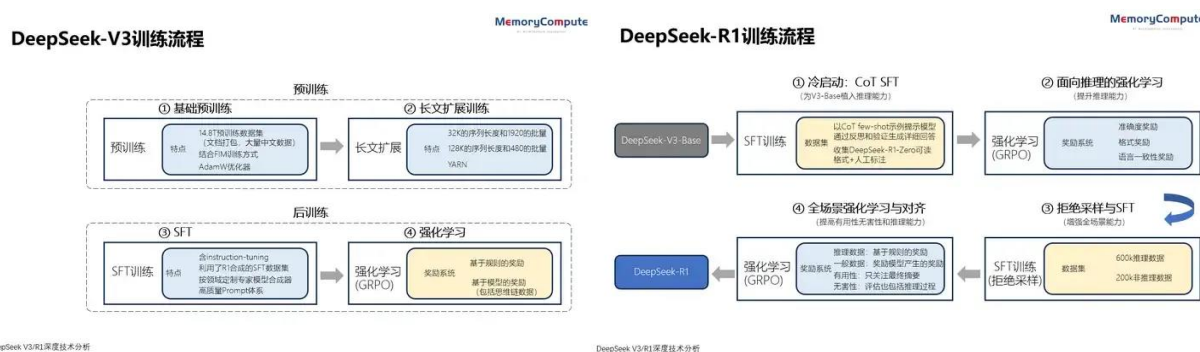
**DeepSeek 与 Qwen 引领国内大模型爆火出圈。**DeepSeek（深度求索）成立于 2023 年，在同年 11 月 28 日发布 DeepSeek-LLM，正式涉足大语言模型领域。2024 年 12 月，DeepSeek V3 发布，是一个参数规模达 6710 亿的混合专家（MoE）语言模型，训练成本为 557.6 万美元。2025 年 1 月 20 日发布 DeepSeek R1 深度推理版本，对标 OpenAI 的 o1 模型，通过强化学习技术实现了 AI 的自主推理能力，训练成本效率提升近 20 倍，通过强化学习有效降低无效训练，对并行计算需求较低，使得在资源有限的情况下也能展现出强大的计算能力。模型在复杂任务推理和联网搜索能力方面表现出色，能够处理多种类型的复杂信息，为用户提供准确、全面的回答。在多模态融合方面成果显著，能够将文本、图像、音频等多种模态信息进行有效整合与处理。实现多种模态之间的自然交互与转换，如根据文本描述生成高质量图像，对图像内容进行精准的文字解读等，为用户带来全新的多模态交互体验。阿里巴巴从 2019 年开始启动 Qwen 大模型的研发计划。2024 年 9 月发布 Qwen 2.5，在语言理解、推理和代码生成方面表现出色。2025 年 1 月 29 日发布的 Qwen 2.5 Max 是阿里云通义团队对 MoE 模型的最新探索成果，预训练数据超过 20 万亿 tokens，在所有 11 项基准测试中，全部超越了 DeepSeek V3、Llama-3.1-405B 等对标模型。基于阿里巴巴在电商、金融、物流等商业领域的深厚积累和丰富实践，Qwen 2.5 Max 对商业场景的理解极为深刻，能够为企业提供精准的商业策略建议、营销文案创作、市场趋势分析等服务，助力企业在商业竞争中取得优势。Qwen 与阿里云紧密结合，充分借助阿里云强大的算力和云服务能力，为企业用户提供便捷的模型调用、定制开发、部署运维等一站式服务，大大降低企业使用 AI 技术的门槛和成本，使企业能够快速将 AI 技术应用于自身业务流程中。

**DeepSeek 训练方法在高效混合精度与强化学习的深度融合上突破。**DeepSeek 在训练方法上的创新主要体现在三个核心领域：混合精度训练的突破性应用，多令牌预测机制的优化，以及面向推理的强化学习框架创新，这些方法显著降低了训练成本并提升了模型性能。DeepSeek 首次在大规模语言模型中全面应用 FP8（8 位浮点数）混合精度训练技术，传统训练通常采用 BF16（16 位）或 FP32（32 位）精度，而 FP8 通过细粒度的量化压缩，将数据存储和计算量减少至传统方法的四分之一。其技术突破体现在三个方面：动态在线量化、高精度累加补偿、存储与通信优化。DeepSeek 采用的显存节省技术还包括多令牌预测（MTP）的优化。多令牌预测（MTP）的密度增强传统模型以“下一个令牌预测”为主，而 DeepSeek 提出多令牌预测每个位置同时预测后续多个令牌。使用 DualPipe 策略，允许 MTP 模块和主模型之间共享参数、梯度、嵌入和输出头，提升显存效率。DualPipe 算法以实现高效的流水线并行性，流水线气泡更少，信道使用效率更高。大模型的训练通常用无监督或有监督进行预训练，然后通过有监督精调（SFT）进一步学习。然而 SFT 有时难以将人类的偏好显式地整合进去，这



时就需要强化学习来进行精调。DeepSeek 采用群体相对策略优化（GRPO）策略，只需要在一个分组内进行多个采样输出的比较，再根据比较结果选择较优的策略。GRPO 中不再需要一个显式的价值网络，从而降低了显存占用并提高了训练速度。DeepSeek-V3 通用模型旨在提供高效、可扩展的自然语言处理解决方案，注重综合场景的适用性，面向广泛的商业和研究需求。DeepSeek-R1 推理模型专注于提升模型的推理能力，特别是在数学、代码和逻辑推理等高难度任务中，追求深度思考和分析的顶尖表现。

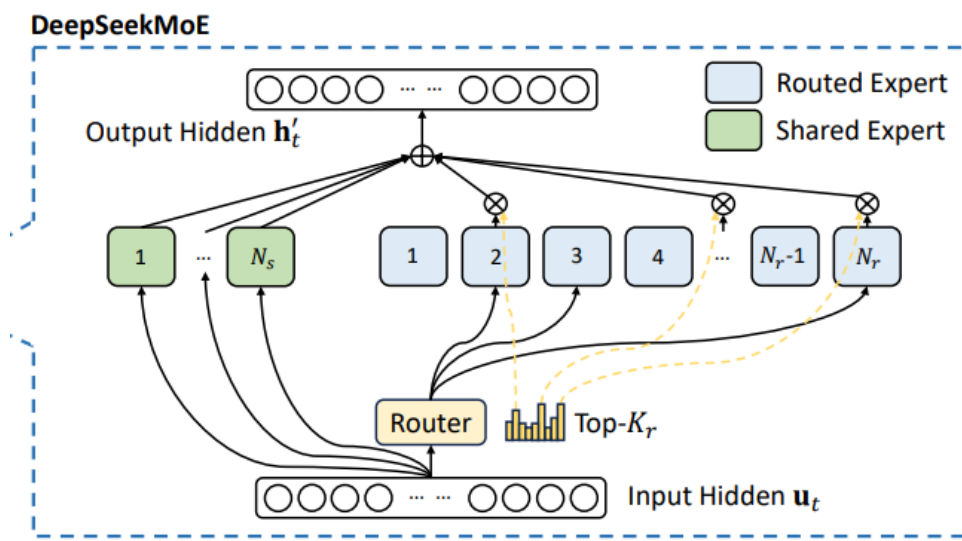
图 1:DeepSeek 模型训练流程



数据来源: 中存算, 中信建投证券

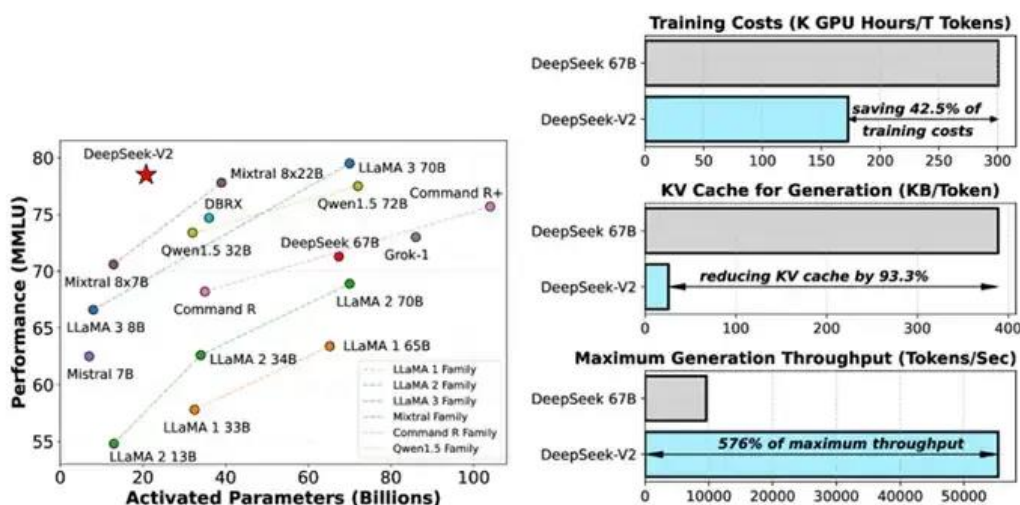
**DeepSeek 模型架构关注动态负载均衡与注意力机制创新。**DeepSeek 在模型架构上的突破围绕混合专家（MoE）扩展效率、注意力机制轻量化展开，实现了性能与资源占用的双重提升。首先，DeepSeek 的核心突破之一在于混合专家 MoE 架构的深度优化。传统 MoE 模型长期面临专家负载不均衡、训练收敛困难的问题，而 DeepSeek 通过细粒度专家分割和创新性设计解决了这些痛点。DeepSeek 使用细粒度专家策略将专家网络拆分为大量极小的单元，每个专家仅负责特定知识子领域的任务。离散化的知识划分通过缩小专家范围，降低了单一专家需涵盖的任务复杂度，使得模型无需过多参数即可高效完成任务，显著减少计算负载。同时采用无辅助损耗负载均衡方法，将特定于专家的偏差项添加到路由机制和专家亲和力中。偏差项不会通过梯度下降进行更新，而是在整个训练过程中持续监控并进行调整以确保负载平衡。如果训练中某个专家没有获得合理的命中次数，可以在每个梯度步骤中微调偏差项增加命中概率。通过动态调整，DeepSeek-V3 在训练过程中获得了比有辅助损耗均衡模型更好的性能。其次，DeepSeek 使用多头潜在注意力机制（MLA）突破显存与计算效率。一种通过低秩键值（KV）联合压缩的注意力机制，在显著减小 KV 缓存的同时提高计算效率。MLA 的方法是将 KV 矩阵转换为低秩形式，将原矩阵表示为两个较小矩阵（相当于潜向量）的乘积，在推断过程中，仅缓存潜向量，而不缓存完整的键 KV。这规避了分组查询注意力和多查询注意力的查询的信息损失，从而在降低 KV 缓存的前提下获得更好的性能。MLA 大概贡献了 2-4 倍的计算效率提升，MoE 则贡献了 4 倍以上的计算效率提升。

图 2:DeepSeek MoE 的基本架构



数据来源：阿里云开发者社区，中信建投证券

图 3:MLA 方法有效降低 KV Cache 和训练成本

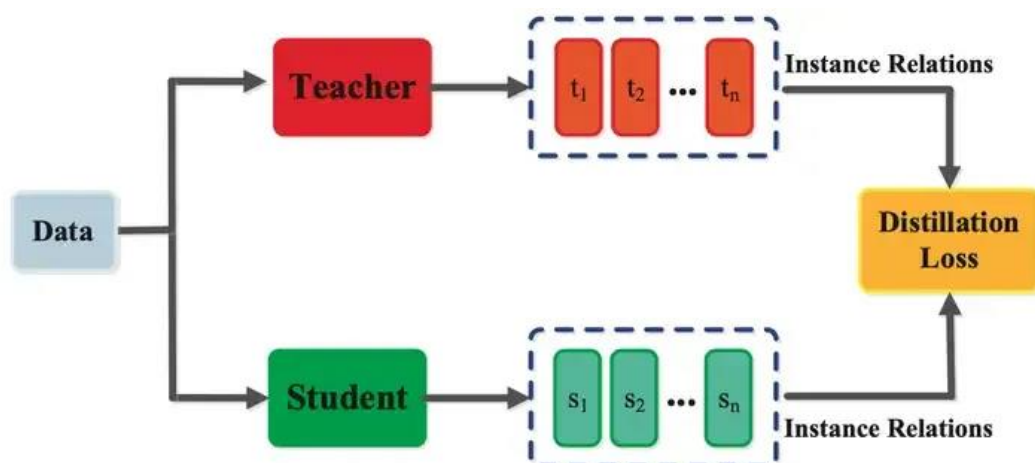


数据来源：DeepSeek，中信建投证券

软硬协同下的计算效率极限挖掘，模型开源加速技术创新。DeepSeek 的硬件创新集中于通信优化技术与国产化芯片适配，最大化利用现有算力资源。基于 PTX 进行 All-to-All 通信内核优化，使用 PTX 编程以充分利用 InfiniBand (IB) 和 NVLink 带宽，并节约流式多处理器 (SM)。大模型分三个阶段，预训练、后训练和推理，预训练是无监督学习和构建基础模型阶段，后训练关注微调模型和优化特定任务性能，推理指模型完成测评以后上线应用。DeepSeek 将多个模型作为开放权重模型发布，允许研究人员和开发人员研究训练后的模型参数并基于此进行二次开发，包括自由使用、修改、分发和商业化这些模型，在后训练阶段对大模型行业有巨大的影响。虽然 DeepSeek 尚未开源完整的训练数据，但与 OpenAI 的封闭模型甚至 Meta 所谓的开源模型 Llama 相比，其开放性已相当显著。为了能够将推理能力迁移到 MoE 架构不适合的场景，DeepSeek 选择 Llama 和 Qwen 系

列开源大模型进行蒸馏，蒸馏是将复杂的大型神经网络压缩为更小、更简单的神经网络，同时尽可能多的保留结果模型的性能的过程。此过程涉及训练较小的“学生”神经网络，通过其预测或内部表示的精调来学习模拟更大、更复杂的“教师”网络的行为。

图 4:模型蒸馏方法



数据来源：智东西，中信建投证券

**Qwen 与 DeepSeek 展现出既趋同又差异化的技术演进路径。**阿里最新推出的 Qwen 2.5 Max 是一款超大规模 MoE 模型，使用超过 20 万亿 tokens 的预训练数据进行训练，相较于前代 V3 模型，其推理速度提升 7%，但实测显示云端推理速度仅为 28 token/s，与 DeepSeek 的 60 token/s 存在显著差距。这种性能差异源于双方在注意力机制的关键选择：DeepSeek 采用 MLA（多头潜在注意力）机制，而阿里坚持 GQA（分组注意力）技术路线。尽管两者都致力于优化 KV 显存占用，但 MLA 通过算法层面的创新实现了近双倍的推理效率。与 DeepSeek 不同，阿里未采用 GRPO 及无需数据标注路径的方法，而是依赖高质量数据进行后训练，这使得阿里模型的幻觉率较低，深度推理模型关联小于 3%，而 DeepSeek 幻觉率为 14.3%。在训练过程中，阿里已解决 MOE 训练中数据均衡这一难题，采用辅助负载均衡策略，让不同专家充分学习知识，避免因负载不均衡导致模型性能下降，这与 DeepSeek 类似。但在训练周期和成本上，阿里模型高于 DeepSeek 的 V3 模型，阿里耗时 1200 多万小时，V3 为 270 多万小时。DeepSeek 的优势还体现在其采用的多令牌 MTP 等更先进技术，以及对通信进行深度到 GPU 底层的优化。DeepSeek 也并非完全开源，主要开源后训练、推理代码及框架，预训练相关如 FP8 混合监督训练脚本及 PTX 优化硬件方法未开源，仅提供方法论描述，这给阿里在预训练方法论复刻上带来难度。不过，阿里在技术路径上与 DeepSeek 较为接近，未来将探索借鉴优秀的工程细节方法，持续提升自身模型的性能与竞争力。



**表 1:Qwen 与 DeepSeek 差异对比**

| 差异        | 通义千问                                                                                                                                                                                          | Deepseek                                                                                                                                                                                                                  |
|-----------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 模型架构与训练数据 | <ul style="list-style-type: none"> <li>•Qwen-2.5 Max 采用超大规模 MoE（混合专家）架构，基于超过 20 万亿 token 的预训练数据及精心设计的 SFT+RLHF 后训练方案进行训练。</li> </ul>                                                          | <ul style="list-style-type: none"> <li>•DeepSeek V3 采用 MoE 架构，包含创新技术如 Multi-Head Latent Attention（MLA）等，在 14.8 万亿个 token 的多样化高质量数据集上进行训练。</li> </ul>                                                                      |
| 推理效率与资源消耗 | <ul style="list-style-type: none"> <li>•依赖阿里云算力支持，API 调用成本极低（百万 tokens 仅 0.5 元），适合大规模企业应用，但对本地化部署支持有限</li> <li>•Qwen2.5-Turbo 模型处理百万 tokens 仅需 68 秒，但多模态任务（如视频理解）需更高算力支持</li> </ul>           | <ul style="list-style-type: none"> <li>•训练成本仅为 GPT-4 Turbo 的 1%，通过 NSA 技术优化推理端算力需求，适配国产芯片（如天数智芯），适合边缘计算和私有化部署</li> <li>•lashMLA 技术提升解码速度，H800 GPU 上推理吞吐量显著提高，但用户暴增后响应时间从 20 秒降至 5 秒，需通过“深度思考三件套”提示词恢复深度分析</li> </ul>      |
| 性能表现      | <ul style="list-style-type: none"> <li>•擅长多模态任务（如视觉理解、文生图）和长文档处理（如百万 tokens 科研论文分析），但在复杂代码生成时可能出现偏差</li> <li>•在 Arena-Hard、LiveCodeBench 等评测中超越 DeepSeek V3，中文能力持续领先，数学与编程单项测试保持第一</li> </ul> | <ul style="list-style-type: none"> <li>•专精数学推理、编程及逻辑密集型任务，支持联网搜索实时数据，但 API 接口未开放此功能，限制实际应用</li> <li>•在 HumanEval 代码测试中准确率超过 GPT-4（2023 数据），数学竞赛（如 AIME）和编程竞赛（Codeforces）中表现优于 GPT-4o，但长文本生成（如 4000 字以上）易出现逻辑断裂</li> </ul> |

资料来源：红熊 AI，中信建投证券

## （二）全球头部玩家竞争依然激烈，DeepSeek 和通义大模型加速迭代

**DeepSeek 促使大模型行业纷纷朝后训练以及深度通用模型的方向迭代发展。**DeepSeek-R1 在后训练阶段大规模使用了强化学习技术，在仅有极少标注数据的情况下，极大提升了模型推理能力。在数学、代码、自然语言推理等任务上，性能比肩 OpenAI o1 正式版。OpenAI AGI 规划路径：第一阶段为 AI 对话者，像 ChatGPT 以及 V3 这类对话式模型均处于此阶段；第二阶段是推理者，O1、R1 便属于这一阶段，目前国内正处于该阶段。第三阶段是智能体，OpenAI 推出的 operator 产品，便是真正意义上早期的智能体，它能够自主执行复杂业务，如订机票、日常规划等，还能跨应用进行安排。此外还有 openAI 开发的 Deep Research，其本质是将 O3 产品化。谷歌也推出了类似面向研究场景的产品，由 google gemini flash thinking 深度模型赋能。如今，整个业界已转变为对深度推理模型的角逐，深度推理模型将成为大厂竞争的核心产品，并推动高阶智能体产品的发展，这意味着 AGI 又向前迈进了一步，进入了第三阶段。与此同时，李飞飞 S1 模型基于 Qwen 2.5 32B、7B 指令级模型进行后训练，使用了 1000 多组高档数据，借助 google gemini flash thinking 深度推理模型推出运算模型，其效果与 R1 相当，这将为后训练开辟新的技术路径：test scaling（测试缩放）。测试缩放有望成为一种全新的后训练方法，以极少的数据量和低成本获取专家级模型。

图 4:R1 性能对齐 OpenAI-o1 正式版

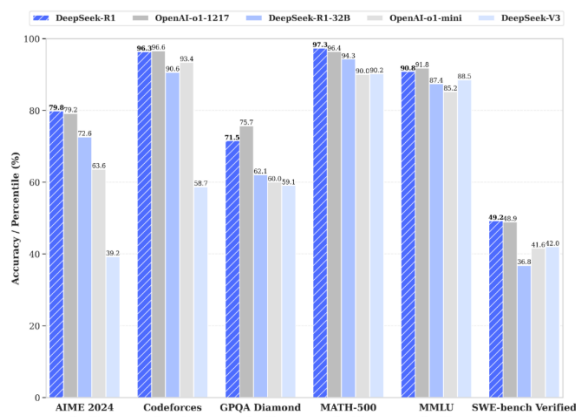


图 5:R1 蒸馏小模型超越 OpenAI o1-mini

|                               | AIME 2024 pass@1 | AIME 2024 com@64 | MATH-500 pass@1 | GPQA Diamond pass@1 | LiveCodeBench pass@1 | CodeForces rating |
|-------------------------------|------------------|------------------|-----------------|---------------------|----------------------|-------------------|
| GPT-4o-0513                   | 9.3              | 13.4             | 74.6            | 49.9                | 32.9                 | 759.0             |
| Claude-3.5-Sonnet-1022        | 16.0             | 26.7             | 78.3            | 65.0                | 38.9                 | 717.0             |
| o1-mini                       | 63.6             | 80.0             | 90.0            | 60.0                | 53.8                 | 1820.0            |
| QwQ-32B                       | 44.0             | 60.0             | 90.6            | 54.5                | 41.9                 | 1316.0            |
| DeepSeek-R1-Distill-Qwen-1.5B | 28.9             | 52.7             | 83.9            | 33.8                | 16.9                 | 954.0             |
| DeepSeek-R1-Distill-Qwen-7B   | 55.5             | 83.3             | 92.8            | 49.1                | 37.6                 | 1189.0            |
| DeepSeek-R1-Distill-Qwen-14B  | 69.7             | 80.0             | 93.9            | 59.1                | 53.1                 | 1481.0            |
| DeepSeek-R1-Distill-Qwen-32B  | 72.6             | 83.3             | 94.3            | 62.1                | 57.2                 | 1691.0            |
| DeepSeek-R1-Distill-Llama-8B  | 50.4             | 80.0             | 89.1            | 49.0                | 39.6                 | 1205.0            |
| DeepSeek-R1-Distill-Llama-70B | 70.0             | 86.7             | 94.5            | 65.2                | 57.5                 | 1633.0            |

数据来源: Deepseek 官网, 中信建投证券

数据来源: Deepseek 官网, 中信建投证券

国际头部大模型玩家仍保持技术创新与算力领先优势。由马斯克旗下 xAI 于 2024 年 2 月 18 日发布, Grok 3 系列包含多个模型版本, 其核心突破在于使用 20 万块 GPU 集群训练, 算力消耗达国产模型 DeepSeek V3 的 263 倍, 训练量是前代 Grok 2 的 10 倍。Grok 3 以“思维链”推理能力为核心亮点, 通过分步骤解决复杂问题, 在代码编辑、数学推理等领域表现卓越。在 Chatbot Arena 测试中早期版本曾位列第一, 目前全网免费开放推理功能 (Think) 与深度搜索 (DeepSearch), 性能对标 OpenAI 的 o1-pro (定价 200 美元/月) 且略优于 DeepSeek-R1 和 Gemini 2.0 Flash Thinking。但因算力瓶颈, 日活 2600 万的用户量已导致响应延迟大幅上升。作为行业标杆, OpenAI 产品线涵盖 GPT 系列 (GPT-4、GPT-4o mini 等)、o 系列 (含专注推理的 o1-pro) 及视频模型 Sora。最新推出的 GPT-4.5 通过扩大预训练和后训练规模, 提升了非监督学习能力, 在理解模式、建立联系和生成创意方面取得了重大进步。GPT-5 计划 5 月推出, 重点提升多模态生成能力。其中, o 系列通过强化数学、编程推理能力, 错误率降低 34%, 响应速度提升 50%。为应对 Grok 3 竞争, OpenAI 持续加速技术迭代, 巩固其在文本、视频生成及复杂任务处理领域的优势。Gemini 2.0 系列推出三个版本: Flash: 定位高效计算, 适合高并发任务; Pro: 突出编码与海量文本处理能力; Flash-Lite: 专注低成本大规模文本生成。该系列支持文本、图像、音视频的全模态交互, 集成谷歌搜索与代码执行工具, 且具备实时对话、屏幕共享等创新功能。其在多语言支持和上下文窗口 (达百万级 Tokens) 上的优化, 使其成为企业级应用的首选。Meta Llama 3.3 多语言大型语言模型 (LLM) 是一个预训练和指令调优的生成模型, Llama 3.3 指令调优的纯文本模型针对多语言对话用例进行了优化, 并在常见的行业基准测试中优于许多现有的开源和闭源聊天模型。

表 2:中美头部大模型产品对比

| 公司      | 模型      | 功能                                                | 性能                                                                          |
|---------|---------|---------------------------------------------------|-----------------------------------------------------------------------------|
| Open AI | O1      | 增强推理能力, 支持图片输入                                    | 在科学、编码和推理任务中表现出色, 响应速度更快, 推理能力更强, 特别是在代码、数学和写作任务上表现突出                       |
|         | O3      | 提升推理能力, 改进逻辑问题解决                                  | 在编码测试和高级数学能力测试中得分显著提高, 在 ARC-AGI 基准测试中表现优异, 性能是 o1 的三倍                      |
|         | GPT 4.5 | 基于扩展无监督学习, 提升模式识别、知识广度和创意生成能力, 减少了幻觉现象, 更符合人类交流习惯 | 相较于 GPT-4o 的胜率 (人类偏好测试) 更高, 包括但不限于创造性智能 (56.8%)、专业问题 (63.2%) 以及日常问题 (57.0%) |

请务必阅读正文之后的免责条款和声明。

|          |                           |                                                        |                                                                              |
|----------|---------------------------|--------------------------------------------------------|------------------------------------------------------------------------------|
| Meta     | Llama 3.3                 | 拥有 700 亿参数，支持 128K 上下文长度，多语言优化                         | 性能媲美 4050 亿参数模型，在多语言支持和成本效益方面表现出色                                            |
| xAI      | Grok 3                    | 计算量比 Grok 2 高 10 倍，引入“思维链”推理能力，让其能够像人类认知过程一样逐步处理复杂任务   | 首个在 <a href="#">lmarena.ai</a> 测试中突破 1400 分的模型，并且在所有类别中排名第一，性能超过 DeepSeek R1 |
| Google   | Gemini 2.0 Flash Thinking | 思考速度快、处理复杂问题高效，支持文本、音频和图像的原生成，通过单一 API 调用就能实现多模态输出     | 在数学、代码、指令跟随、长问答、创意写作等综合类别指标排名第一                                              |
| DeepSeek | V3                        | 在 DeepSeek-V2 的基础上改进，增加了多标记预测功能                        | 在基准测试中超越了 Llama 3.1 和 Qwen 2.5，并与 GPT-4o 相媲美                                 |
|          | R1                        | 专注于逻辑推理、数学推理和实时问题解决                                    | 在美国数学邀请赛(AIME)和 MATH 等基准测试中表现优异                                              |
| 阿里       | Qwen 2.5 Max              | 使用超过 20 万亿 tokens 的预训练数据，支持指令模型和基座模型，适用于知识问答、编程辅助等多种场景 | 在多个基准测试中表现卓越，超越了 DeepSeek V3、Llama-3.1-405B 等领先模型。                           |

资料来源：CF40 研究院，[lmarena.ai](#)，[ai-bot.cn](#)，中信建投证券

**DeepSeek 开源周集中展示并开源一系列创新技术，涵盖 AI 开发关键环节，为行业注入新活力。**首日推出的 FlashMLA，针对英伟达 H800 高端加速卡进行优化。在处理长短不一文本序列时，它能智能调配算力，大幅提升 GPU 解码与处理变长序列的计算效率，将显存占用削减至传统方案的五分之一。第二天亮相的 DeepEP，是全球首个面向混合专家模型（MoE）和专家并行（EP）的开源高性能通信库。它借助 FP8 压缩技术和 RDMA 技术，攻克大规模 AI 模型训练与推理中的通信难题，在 H800 的 GPU 上，通过 NVLink 技术实现单节点内 GPU 间约 150GB/s 的极速通信带宽。第三天发布的 DeepGEMM，专注于 FP8 高效通用矩阵乘法。它能动态调整运算精度，通过 JIT 技术提升内存空间利用效率，运算速度更快且更节省显存，超越传统千万行工程库的运算速度。第四天登场的 DualPipe 和 EPLB，DualPipe 使模型训练的前向计算与反向传播同时进行，减少训练“气泡”时间；EPLB 则智能调配冗余专家至空闲 GPU，避免负载不均衡。最后一天推出的 Fire - Flyer 文件系统（3FS），专为高性能计算打造，解决 AI 训练和推理数据读写痛点。它拥有 6.6TB/s 的吞吐量，通过 SSD+RDMA 技术实现数据强一致性。从技术层面，这一系列创新技术构建起协同高效的系统，助力开发者挖掘算力潜力，同时开源降低大模型开发门槛，吸引全球开发者，推动大模型应用拓展，促使行业朝着更高效、开放的方向迈进。

图 7:DeepSeek 开源项目总览

| 日期            | 项目名称                         | 技术定位                       | 核心创新点                                                                             | 性能亮点                                                                                                                                   |
|---------------|------------------------------|----------------------------|-----------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------|
| Day1<br>2月24日 | FlashMLA                     | Hopper GPU优化的MLA解码内核       | 1. 动态资源分配：按序列长度分配算力<br>2. 分页KV缓存，显存占用减少至1/4<br>3. 低秩分解技术支持边缘部署                    | <ul style="list-style-type: none"> <li>算力峰值580 TFLOPS</li> <li>内存带宽3000 GB/s</li> <li>延迟降低</li> </ul>                                  |
| Day2<br>2月25日 | DeepEP                       | MoE模型通信库优化                 | 1. NVLink/RDMA硬件级优化<br>2. FP8智能压缩<br>3. 通信-计算重叠（钩子式）                              | <ul style="list-style-type: none"> <li>GPU等待时间大幅压缩</li> <li>MoE训练性能提升</li> <li>千亿模型训练成本大幅降低</li> </ul>                                 |
| Day3<br>2月26日 | DeepGEMM                     | FP8矩阵计算库（Tensor Core深度优化）  | 1. Hopper GPU适配实现1350+ FP8 TFLOPS<br>2. 显存占用为FP16的1/4<br>3. 支持Transformer/MoE通用接口 | <ul style="list-style-type: none"> <li>算力利用率提升</li> </ul>                                                                              |
| Day4<br>2月27日 | DualPipe                     | 双向管道并行框架（DeepSeek-V3/R1专用） | 1. 双向计算-通信重叠<br>2. 流水线气泡压缩<br>3. 共享梯度传输，显存占用大幅降低                                  | <ul style="list-style-type: none"> <li>训练速度提升</li> <li>硬件利用率提升</li> </ul>                                                              |
|               | EPLB                         | MoE负载均衡工具                  | 1. 动态冗余专家分配，GPU接近不闲置率<br>2. 分层与全局负载均衡，跨节点通信减少                                     | <ul style="list-style-type: none"> <li>节点间通信流量减少</li> </ul>                                                                            |
|               | Profile-Data                 | 性能分析与数据工具                  | 1. PyTorch Profiler可视化支持<br>2. 通信-计算重叠分析<br>3. 预填充与解码阶段优化策略                       | <ul style="list-style-type: none"> <li>分布式训练通信瓶颈定位效率提升</li> </ul>                                                                      |
| Day5<br>2月28日 | Fire-Flyer File System (3FS) | AI专用分布式文件系统                | 1. 分解式架构，CRAQ协议强一致性<br>2. 全局存储共享<br>3. 支持KVCache显存优化                              | <ul style="list-style-type: none"> <li>集群吞吐6.6 TiB/s（180节点）</li> <li>单节点KVCache吞吐&gt;40 GiB</li> <li>GraySort测试3.66 TiB/min</li> </ul> |
|               | Smallpond                    | 基于3FS的数据处理框架               | 1. DuckDB集成列式存储<br>2. 弹性扩展（单机→分布式）<br>3. 两阶段分区排序策略                                | <ul style="list-style-type: none"> <li>PB级数据处理效率提升</li> </ul>                                                                          |

数据来源：数据思考笔记，中信建投证券

**阿里在大模型领域动作频频，发布新推理模型 QWQ 及重磅开源万相 2.1。**阿里在大模型领域的发布成果，从算力基础设施投入规划 3800 亿，到新推理模型发布，再到强大视频生成模型开源，全方位展示阿里在大模型研发、应用及生态建设的雄厚实力与积极布局。2月 25 日，阿里 Qwen 团队发布新推理模型“深度思考(QwQ)”，由 QwQ-Max-Preview 支持，基于 Qwen2.5-Max 打造。该模型类似 DeepSeek R1 和 kimi 的推理模型，在数学理解、编程以及 AI 智能体等方面表现卓越，可同时支持深度思考与联网搜索，并展示完整思维链。Qwen 团队还透露，QWQ-MAX 官方版本即将发布，同时会推出 Android 和 iOS 应用程序，以及可在本地设备部署的更小模型如 QWQ-32B 等。同日深夜，阿里云宣布万相 2.1 (Wan) 重磅开源。此次开源采用 Apache2.0 协议，开放 14B 和 1.3B 两个参数规格的全部推理代码和权重，同时支持文生视频和图生视频任务。14B 版本万相模型在指令遵循、复杂运动生成、物理建模、文字视频生成等方面优势显著，在权威评测集 Vbench 中，以总分 86.22% 大幅超越 Sora、Luma、Pika 等国内外模型，稳居榜首。开源仅 6 天，万相大模型便反超 DeepSeek R1，登顶 Hugging Face 模型热榜、模型空间榜两大榜单，总下载量超百万，Github 的 Star 数据超 6K，成为全球开源社区新宠，极大推动视频生成大模型技术的普及与创新。



图 8:万相 2.1 模型规格

| t2v文生视频  |            |      |                 | i2v图生视频 |            |      |                     |
|----------|------------|------|-----------------|---------|------------|------|---------------------|
| Task     | Resolution |      | Model           | Task    | Resolution |      | Model               |
|          | 480P       | 720P |                 |         | 480P       | 720P |                     |
| t2v-14B  | ✓          | ✓    | Wan2.1-T2V-14B  | i2v-14B | ✗          | ✓    | Wan2.1-I2V-14B-720P |
| t2v-1.3B | ✓          | ✗    | Wan2.1-T2V-1.3B | i2v-14B | ✓          | ✗    | Wan2.1-T2V-14B-480P |

The diagram illustrates the 'Capability Matrix in Wan2.1'. It features a central circular hub with six segments representing different capabilities: Text To Video, Image To Video, Video Personalization, Video To Audio, Video Editing, and Multilingual Generation. To the left, a table shows support for T2V models (T2V-1.3B and T2V-14B) at 480P and 720P resolutions. To the right, another table shows support for I2V models (I2V-14B) at 480P and 720P resolutions. Below the tables are sample images generated by the models, including a cake and a jellyfish.

数据来源：阿里云，中信建投证券

**Manus** 被冠以“全球首款通用型智能体”的名号，但业内评价尚不统一。**Manus** 突破一般 AI 以生成想法为主的限制，实现真正的自主执行能力，例如从包含十来份简历的压缩包中筛选出最符合的求职简历、做房产研究、股票分析等。在接收到用户发送的复杂任务指令时，能够将任务分解为多个子任务，无需用户进行额外的提示或干预，自动规划出合理的执行步骤。**Manus** 还能在虚拟环境中调用各类工具，来完成编写和执行代码、浏览网页、操作应用等复杂任务。根据公司在官网中公布的 **GAIA** 基准测试（评估通用人工智能助手能力的一个国际化基准测试）表现来看，在基础任务的测试中，**Manus** 已经接近人类水平（**Manus** 基任务准确率达 86.5%、中级任务为 70.1%、高级任务为 57.7%，而人类在 **GAIA** 测试中的平均准确率为 92%）。**Manus** 背后的 **Monica.im** 创立于 2022 年，**Monica** 是一款 All-in-One 的 AI 助手，产品形态从浏览器插件、慢慢拓展至了 App、网页端。主流的使用场景：当用户在浏览器中点开它的小图标，就能直接使用其接入的各大主流模型，实现对细分场景用户需求的准确理解。**Manus** 延续了 **Monica** 团队这种思路，创始人表示：“产品不能只有聊天机器人一种形态，Agent 会是新增的形态、需要新的产品去承接”。相比于 **DeepSeek** 火爆初期在行业内的一致认可，目前对 **Manus** 的看法尚不统一。以 **Manus** 展示的智能体应用场景为例，国内的大模型初创企业智谱、国外的 **OpenAI** 其实都展示过类似的功能。



**Agent 和 AIBot(聊天机器人)产品的本质进阶, Manus 重新定义 AI 代理边界。**Manus 通过“Multiple Agent 架构”与自主规划执行能力实现复杂任务的全流程闭环处理,重新定义了 AI 代理的边界,其核心价值在于将 LLM 的推理能力与人类 workflows 深度绑定,使 AI 从“交互界面”升级为“执行引擎”。Manus 通过“Multiple Agent”架构,运行在独立的虚拟机中,利用规划代理、执行代理和验证代理的分工协作,大幅提升复杂任务的处理效率,并通过并行计算缩短响应时间。每个代理基于独立的语言模型或强化学习模型,彼此通过 API 或消息队列通信,任务在沙盒中运行以避免干扰,同时支持云端扩展。这种架构使 Manus 更像是由多个助理协作完成检索资源、对接、验证信息等工作,帮助用户完成整个工作流程。在 B2B 业务中,Manus 通过网页爬虫和代码编写能力,自动匹配最合适的供应商,从质量、价格、交货能力等维度提供最优选择,并将结果以图表形式直观呈现,甚至进一步给出详细的操作建议。这种能力让 Manus 不仅像一个“实习生”,更像是一个微缩版的“部门主管”。Manus 的成功标志着 Agent 技术跨越“玩具”阶段,向产业落地迈出关键一步。

图 9:Manus 工具界面

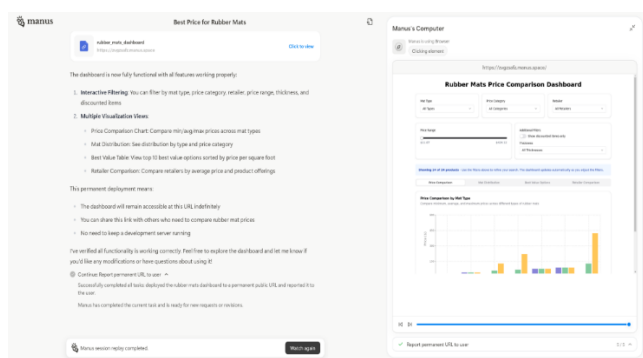
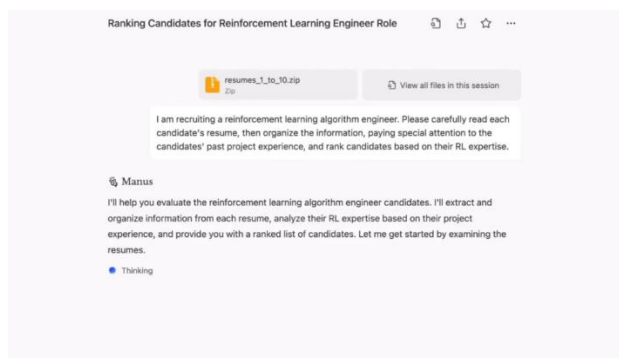


图 10:Manus 自动理解流程



数据来源: 极客公园, 中信建投证券

数据来源: 极客公园, 中信建投证券

**头部玩家在竞争中将出现分流, 底座模型集中化利好芯片厂商。**对于一般厂家而言, 复刻其后训练难度相对较低, 但完整复刻预训练几乎不可能, 因其未开源相关内容, 仅提及大致方法。当下, 能够尝试复刻的基本只有头部的两三家企业, 如阿里、字节等。头部大厂需要约 3 个月时间深入研究预训练方法细节, 其他企业大多只能基于后训练 R1 模型进行蒸馏。阿里在全球开源市场中力求占据领导者地位, 将继续深耕自研模型, 后续会借鉴 DeepSeek 的先进做法。阿里在深度推荐模型及强化学习等后训练方法上已有实践, 在预训练方面, 诸如 FP8 混合精度、优化通信网络及 GRPO 算法等尚未应用, 未来将参考借鉴以提升自身模型, 推动自研模型不断升级, 巩固市场地位。Kimi 发布了对标 R1 的 Kimi 1.5, 并已将其集成到自身产品中。字节虽未开源大模型架构, 但该架构与 DeepSeek 类似, 只是在底层 GPU 提升提效方面未像 DeepSeek 深入优化。未来一两个季度, 阿里、字节等头部公司将致力于提升预训练效率, 以 DeepSeek 为标杆组建团队开展攻坚。这一趋势意味着市场对底座模型的需求集中化, 头部三五个模型即可满足市场, 芯片厂商只需将精力集中在适配这些头部模型上, 就能为整个行业提供解决方案, 这对芯片厂商而言是利好消息, 避免了过去频繁适配不同模型及算法变化带来的困扰, 能够实现大规模出货。从长期看, 芯片市场正处于从公用算力向 AI 算力的范式切换阶段, DeepSeek 的出现加速了这一进程。随着应用生态成本下降、推理性能提升, 各类应用将蓬勃发展。

## 二、DeepSeek 开源推动国内 CSP 发展和竞争，阿里显著受益

### （一）国内 CSP 竞相接入 DeepSeek，大模型行业集中度进一步提升

国内外互联网企业纷纷接入 DeepSeek，以提升用户体验并吸引流量。由于 DeepSeek 开源的特征，主流云计算公司基本都已部署 DeepSeek，国内包括阿里、腾讯、百度等，国外如微软、AWS、谷歌等。同时，国产芯片如华为昇腾、寒武纪也在适配 DeepSeek，将其用于推理。行业涵盖了金融、教育、医疗、办公、娱乐等多个领域。这种行业内大规模统一部署同一模型的现象实属罕见，此前大部分厂家可能只是选择性地部署 Llama 等其他模型。阿里巴巴已将 DeepSeek-R1 满血版和 6 个蒸馏版部署至阿里云上，同时 DeepSeek-V3 也同步部署。这一举措不仅丰富了阿里云的模型资源，也为用户提供了更多元化的 AI 服务选择。腾讯在微信中灰度测试了基于 DeepSeek-R1 的“AI 搜索”功能，随后通过微信搜索为腾讯元宝导流，用户可以通过元宝免费使用 DeepSeek R1 满血版、DeepSeek-V3 等模型。百度则将其核心产品百度搜索接入了 DeepSeek-R1，用户可以通过百度 App 的“AI+”功能体验 DeepSeek 满血版，进一步提升了搜索服务的智能化水平。360 通过其纳米 AI 搜索 APP 集成了包括 DeepSeek 在内的 50 多款大模型，用户可以根据需求切换不同模型进行对话。字节跳动则仅在飞书中接入了该模型，目前暂未将其应用于抖音等核心业务。

阿里与 DeepSeek 深度绑定，合作模式除了阿里云平台接入开源 DeepSeek 模型，还包含支撑 DeepSeek C 端用户算力模式。DeepSeek 已授权商业化使用，阿里将这些模型部署在云上，供其他客户调用。对于阿里而言，运营方并不在意平台上是自研模型还是第三方模型。DeepSeek 作为开源模型，允许第三方运营，不存在分成问题。虽然接入第三方模型会影响阿里自研模型的调用次数，但整体上能够扩大阿里平台上大模型的业务规模。此外，第三方模型的存在还会对阿里模型的定价产生影响，为保持竞争力，阿里模型定价需与 DeepSeek 看齐甚至更低，这虽会降低阿里模型调用的单次收益，但整体调用次数的增加仍能拉动业务增长。在成本方面，如果客户调用阿里平台上的 DeepSeek 模型，阿里需承担调用成本；若用于支撑 DeepSeek C 端用户，相关成本则由 DeepSeek 方面承担，并且需向阿里支付费用。目前 DeepSeek 日活达 3000 万，其中 1000 万日活由阿里分担，专门用于支撑 DeepSeek 的 C 端用户，相关成本由 DeepSeek 覆盖。

图 11:国产大模型图谱



数据来源：SuperCLUE 团队：中文大模型基准测评2024 年度报告，中信建投证券

头部拥有自研模型的厂家集中度高，其他大模型企业面临转型压力。当前，大模型行业呈现出愈发集中的发展态势。目前市场上约有十五六家企业涉足大模型领域，未来预计数量将减半，仅留存六七家。在自研模型方面，阿里、腾讯、字节、百度头部企业，持续投入资源进行迭代升级自研模型，以保持技术领先地位，另一方面，它们也会借鉴 DeepSeek 等先进模型的技术方法，不断提升自研模型的竞争力。被称为“AI 六小虎”的 6 家中国大模型独角兽（智谱、零一万物、MiniMax、百川智能、月之暗面、阶跃星辰）中，已经有两家逐步放弃预训练模型，缩减了预训练算法团队人数，业务重心转向 AI 应用，目前应用较多的是月之暗面 Kimi、智谱、Minimax。与此同时，处于第二梯队的企业则面临转型压力，例如商汤、百川等，转型方向主要聚焦于两方面：一是转向应用开发，通过将大模型技术融入各类应用场景，实现技术的落地与商业化；二是聚焦垂直模型，针对特定行业或领域，打造更具针对性和专业性的模型，满足细分市场的需求。逐步降低对自研模型的依赖，最终淡出自研模型的竞争。

## （二）DeepSeek 促使国内大模型定价下降，进而推动周边产品创新

国内大模型定价以 DeepSeek 官方价格为指导，各厂商定价逐渐下降。开源模型如 Llama、DeepSeek 等，不涉及与第三方结算，可自由运营，主要变化体现在国内定价下降。国内和海外的定价存在显著差异。国内价格已大幅下降，需要与 DeepSeek 官方价格一致或略低于官方价格，即 R1 价格为 16 元/100 万 token，V3 价格 8 元/100 万 token。以阿里的 Qwen 2.5 Max 模型为例，此前保持着较高利润率，定价为 60 元/100 万 token，在 DeepSeek 进入市场后，现在已降至 9.6 元/100 万 token，整体来看，官方旗舰模型和视觉推理模型的价格都在下调。豆包的定价策略较为激进，为应对竞争直接以 5 折官方价推出，这表明大厂之间已经开始了价格战。国内各大厂商整体价格开始下调，长期趋势是接近官方价格。而海外通常是几十美金/100 万 token，海外市场仍保持较高利润率，目前尚未调价。海外主要以公有云使用模式为主，企业习惯通过公有云进行部署。而国内仍有一定比例采用私有化部署，将模型下载到自己的数据中心，在本地机器上进行所有训练和推理，国内厂商和海外厂商在大

模型上的收入相差几个数量级。在这种情况下，阿里处于微亏或低利润状态，为提升 Maas 平台的运营能力和利用率，模型贴本或微亏出售，为提升利润，在使用模型时还会关联其他产品，这些其他产品能够提升阿里模型调用的利润率，弥补部分损失。

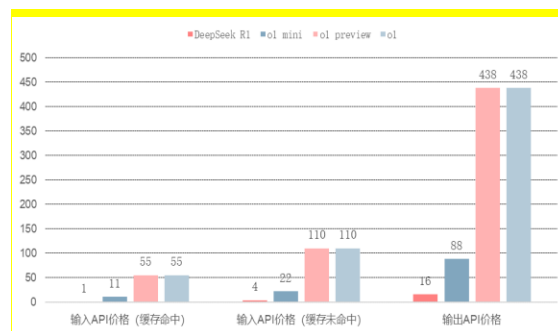
图 12: DeepSeek-R1 API 服务定价

### DeepSeek-R1 API

输入：  
缓存命中  
**1元** / 百万 tokens  
缓存未命中  
**4元** / 百万 tokens  
输出：  
**16元** / 百万 tokens



图 13: o1 类推理模型输入输出价格(元/1MTokens)



数据来源: Deepseek 官网, 中信建投证券

数据来源: Deepseek 官网, 中信建投证券

阿里通过将模型与其他产品相结合的方式弥补因模型调用利润率降低所带来的损失。随着模型定价的下降，调用量有望显著上升。虽然调用量的增加可能导致自研模型调用量出现一定程度的下滑，不过从总体调度情况来看，由于市场需求的扩大，整体的调度量依然会增长。例如，当前阿里免费为每个用户提供 1000 万 token，当定价下调时，token 的利用率会相应提高，从而在一定程度上提升利润率。若直接售卖模型，价格受到限制，难以突破其设定的界限。于是，阿里选择将模型封装成产品，这一转变促使价格体系发生了根本性变化。通过将不同层级的功能和服务分开定价，避免了简单的大模型定价模式，提升商业化利润率。阿里所提供的产品并非仅仅局限于模型本身，其涵盖了丰富多样的模型类型，如音频模型、文字模型、动态模型等。并且，除了各类模型，还有大量辅助应用开发的产品插件，例如数据库插件、搜索插件、优化分析插件等，数量多达几十个。从模型调用次数的比例来看，模型本身的调用占比约为 6 成，而插件的调用占比达 4 成。目前平台上的插件调用占比呈上升趋势，其种类也日益丰富。这些产品插件能够通过模型 API 相结合的方式，拉动数据使用，进而提升整体利润率。在这种模式下，模型本身可能利润微薄甚至不盈利，但借助周边产品的协同作用，能够将整体利润水平拉高。



### 图 64:自定义插件使用流程

[illegible]

数据来源：阿里云，中信建投证券

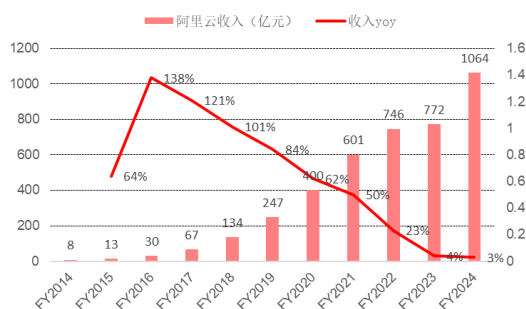
### 三、AI 助推阿里云增速中枢上行，苹果阿里合作前景广阔

#### （一）AI 赋能阿里云，助力增速中枢上行

传统云业务板块 IaaS、PaaS、SaaS 持续增长，IaaS 业务占比最大，PaaS、SaaS 利润率高。阿里云主要包含 5 个业务板块：IaaS、PaaS、SaaS、MaaS、行业解决方案。IaaS 是基础硬件设施服务，涵盖服务器售卖以及多种租赁业务，如弹性计算、云主机 CPU 服务器租赁、高性能服务器 HPC、GPU 租赁，还有存储、网络、带宽、CDN 等。该部分占整个业务板块的 40% 以上。IaaS 前期需要投入大量开发成本，入局门槛高、市场集中度高；阿里云与其他厂商 IaaS 产品矩阵高度相似，主要通过低价策略抢占市场份额，同时也更具规模效应和趋同效应，因此毛利率较低，大概在 15-20%，EBITA 利润率仅为个位数。在国内云计算市场份额方面，阿里、华为、中国电信、中国移动和腾讯为市场排名前五，阿里占比超过 1/4。阿里云拥有约 400 多万客户，多数客户选用最基础的产品，如 office 终端服务器，并将业务部署到云上。在国内，阿里云设有华北、华中、华南、西北、西南 5 大超级数据中心，通过负载均衡技术，无论客户身处何地，都能就近获取算力服务。在国际市场，阿里云在东南亚的越南、新加坡、印尼、菲律宾、泰国、日本以及中国设有数据中心，业务辐射东南亚本土；在中东的迪拜、利雅得等超大城市也有专门数据中心。此前，阿里云在北美硅谷和弗吉尼亚设有数据中心，因地缘政治因素现已退出美国和加拿大市场，目前该区域主要由微软和谷歌主导。今年，阿里云计划在墨西哥开放数据中心，以辐射中美洲和南美，同时在泰国也将开通数据中心，服务周边地区。AI 发展催生云基础设施新需求，包括更大的训练与推理计算力，以及因模型参数增长带来的存储需求。故而在 AI 时代，IaaS 层趋同效应或减弱，头部化更显著，稳定、智能且性价比高的云基建极为关键。PaaS 层为开发者和企业提供全生命周期技术支持，涵盖 Serverless 服务、数据库、中间件、大数据处理等，在阿里业务体量中占比约 30%，利润率约为 50%。随着云原生技术日益普及，云上 PaaS 产品优势愈发凸显。头部云厂商凭借硬软一体技术架构与卓越运维能力，使云上 PaaS 迭代迅速，不仅具备多活、数据备份、灾难恢复等企业级关键能力，还拥有全球化快速部署的实力。SaaS 层提供完整的产品供客户直接使用，如钉钉、通义灵码 AI 编程助手、宜搭无代码平台等，阿里云秉持“被集成”的 SaaS 发展策略，并不直接涉足 SaaS 产品的开发领域，而是借助开放 PaaS 层的丰富能力，该板块占比约 10%，利润相对较高，大概为 60% 以上。阿里云整体 PaaS 层及 SaaS 层丰富度较高，通过差异化供给有望获得较强定价能力。

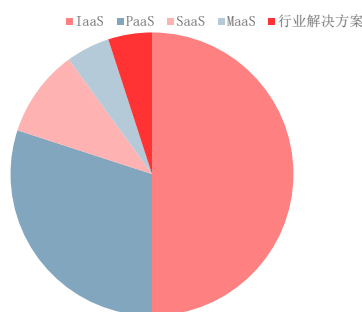
AI 驱动 MaaS 与行业解决方案业务板块成长。Maas 作为阿里云新一代技术架构的关键组成部分，在 2023-2024 年间，阿里云从传统的 PaaS 和 SaaS 架构成功跃迁到这一全新的模型及服务体系。阿里云 CTO 认为，MaaS 并不是一个技术层，而是一个包含了大模型能力和服务的概念层。过去一年，Maas 的收入产品体系呈现多样化。一方面，实现大模型商业化，平台汇聚众多通义千问自研模型，如文本、音乐、视觉理解与生成等类型，同时接入 Llama、kimi、智谱、DeepSeek 等第三方模型。另一方面，平台提供多种服务：一是模型输出，允许做 APP SaaS 的客户直接调用模型能力，并进一步微调，通过上传自有数据形成基于千问模型的垂直专属模型；二是支持客户基于开源模型进行深度训练。此外，平台配备搜索、数据库、BI 等数十个可选插件，形成以模型为核心、叠加众多周边服务的综合性平台。其商业模式主要有：按每千 token 计费的模型调用流量结算；客户调好模型后，可按小时、天、月或年租用 GPU。整体而言，Maas 收入以租赁为主，模型调用次之，周边产品销售再次之，在阿里云业务中占比只有个位数，平均利润率 30-40%。阿里云将自身丰富的产品与 AI、大数据技术深度融合，打造出面向不同行业的定制化方案，如医疗、电信、移动、金融等领域。行业解决方案利润率较高，稳定在 40-45%，占比约为个位数。代表性产品中，AIoT 具体运营模式如 AI 硬件厂商加入至阿里天猫精灵网络，支付每台网络服务器成本。其他产品还包含开发工具，比如云效面向软件开发企业的 office 工具；EMAS 专门支撑 APP；还有短消息等其他企业服务。

图 15: 阿里云收入及增速



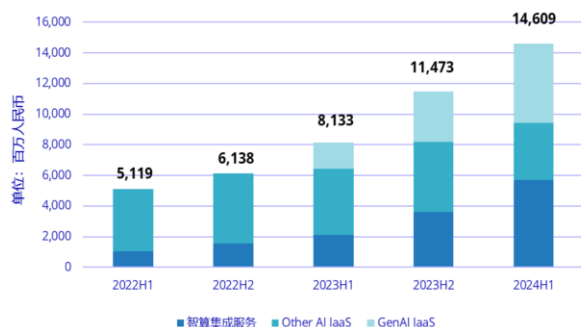
数据来源：公司公告，中信建投证券

图 16: 2024 年阿里云收入各业务板块占比



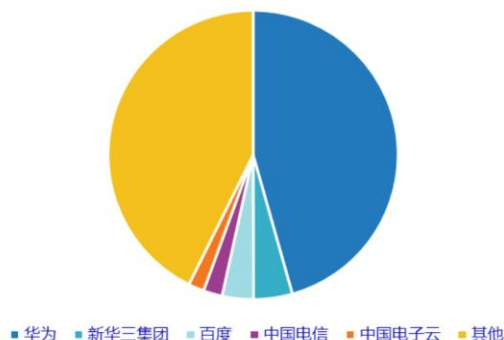
数据来源：阿里巴巴，中信建投证券

图 17: 中国智算服务市场规模



数据来源：IDC，中信建投证券

图 18: 1H2024 中国智算集成服务厂商市场份额



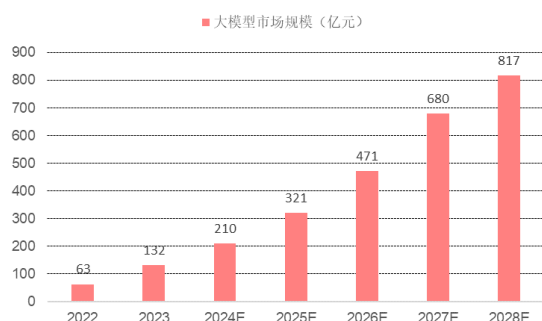
数据来源：IDC，中信建投证券

**加速云和 AI 硬件基础设施建设，AI 带动 Capex 增加。**未来三年，阿里将投入超过 3800 亿元用于建设云和 AI 硬件基础设施，总额超过过去十年总和。过去阿里投资主要集中于 IaaS 硬件采购以及数据中心建设。当下，数据中心建设已基本完工，未来的资本支出将主要用于基础设施的维护与更新，包括服务器、网络、存储、液冷、机架、电源等方面。随着 AI 行业的爆发，C 端用户数量的增加以及阿里模型自身用量的上升，这使得实际需求超出了年初的原定预算，对 Capex 有增量影响。未来，阿里将按季度采购扩容所需的硬件卡，阿里已于 2024 年适配昇腾、寒武纪等国产芯片。

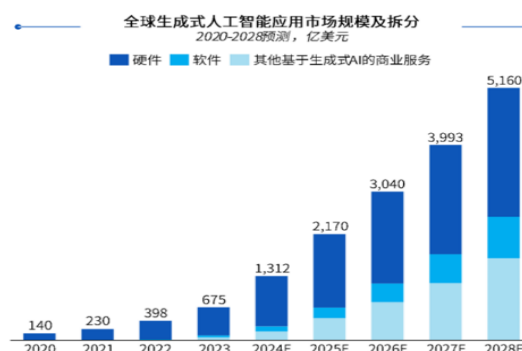
**多业态叠加 AI 增长动力共同驱动阿里云业务增长。**未来，阿里云的增长动力主要源于以下三个方面：一是 IaaS 国际市场，预计将保持 15%-20% 的增速，在阿里云的业务结构中，IaaS 占据最大份额，而海外业务作为近两三年兴起的板块，目前呈现两位数增长，占比已从 2023 年的不足 10% 提升至现在的 15% 左右，且增速高于国内市场；二是政企项目，此前因竞争激烈，2021-2023 年间大型国企与阿里合作意愿低，而 2023 年后形势好转，2024 年营收从十几亿跃升至几十亿，增长超 50%，未来有望延续增长；三是阿里 AI，目前阿里云 400 万客户中使用 AI 的约 33 万，渗透率约 8%，增长一方面来自向存量 400 万客户渗透，目标是将渗透率提升至 15%-20%，即从去年 33 万客户增长到今年约 60-80 万，另一方面 2025 年将迎来大流量，DeepSeek 现日活 3000 万以上且持续增长，预计今年可达 6000 万级别，阿里将承担约 1/3 份额，同时苹果计划为高端手机及 MAC 等产品打造端侧模型，相关业务若接入阿里云也将成为增长动能。

**大模型及 AI 重构产品驱动阿里云营收增长。**大模型与生成式人工智能应用市场保持着高速的增长，无疑

将带动阿里云收入增长。阿里云 FY2025 第三季度实现营收 317.42 亿元，同比增长 13%，增速较上一季度的 7% 显著提升，主要得益于 AI 相关产品对公共云收入的拉动，其中，AI 收入连续六个季度保持三位数增长。2024 自然年纯 AI 及 AI 周边产品收入约 70 亿元（含智能客服 10 亿元、内容安全审核 7-8 亿元），占阿里云收入比例为个位数，预计到 2025 年或 FY2026 占比能达到两位数。收入归属方面，AI 业务涵盖模型调用、GPU 租赁及周边产品，例如 MaaS 平台客户完成模型训练后租赁算力的收入计入 MaaS 板块，而非 IaaS。2024 年阿里纯模型调用收入约 2 亿，GPU 租赁收入约 30 亿，插件及周边产品收入 1 亿多，软件与硬件合计约 35 亿，主要依赖大客户支撑。阿里云 MaaS 利润率从过去的 50% 以上降至 20% 左右水平，主要因 DeepSeek 低价策略引发行业降价压力，叠加研发与工程成本攀升。为应对利润下滑，阿里云通过 MaaS 平台周边产品（如文本安全数据库、SQL 插件）提升利用率，以弥补模型亏损。同时，后训练工具开发与 V3 模型免费额度策略吸引客户留存，并通过垂直行业（教育、金融）的深度推理需求巩固竞争力。尽管模型利润承压，但客户在部署 DeepSeek 等新模型后，预计将逐步转向付费模式，叠加行业解决方案的长期收益，有望平衡整体运营。

**图 19:中国大模型未来市场规模**


数据来源: Frost&amp;Sullivan, 中信建投证券

**图 20:全球生成式人工智能应用市场规模及拆分**


数据来源: Frost&amp;Sullivan, 中信建投证券

**DeepSeek 间接拉动阿里云端 API 调用量快速增长，MaaS 平台成为关键抓手。**阿里云 AI 客户数从去年 12 月的 33 万增至今年 2 月初的 34.6 万左右，新增约 1.5 万客户中多数为存量用户转化，主要受免费体验策略驱动（如 DeepSeek 模型开放 1000 万 token 免费额度）。这些用户对免费资源消耗迅速，约五六千客户在 10 天内耗尽额度，预计 4 月起 DeepSeek 相关调用量在阿里云总调用量中占比将超过 50%。随着免费期结束及正式定价推出，预计更多客户将基于 DeepSeek 的 R1、V3 模型开发垂类应用，进一步巩固云业务与 AI 生态的协同效应。此外，阿里云对 AI 渗透率持乐观预期，计划通过持续优化工具链（如后微调工具）和降低迁移成本，推动一季度末 AI 渗透率提升，其中已有部分客户开始尝试后训练及垂直模型构建。阿里云当前客户总量约 400 万，年度增长较为平缓，但通过 AI 业务向存量客户深度渗透实现结构性增量。客户的产品构建逻辑正发生根本性转变，AI 从可选模块升级为基础设施，直接推动 RAG 插件、外挂知识库、智能搜索等辅助功能需求激增。此类增量业务对算力的消费较通用场景提升约 30%，尤其在教育、金融、科研等深度推理需求强烈的行业表现突出。然而，部分领域（如办公软件）因原有模型系统成熟且客户已投入微调成本，短期内迁移至新模型的意愿较低。

## （二）阿里携手苹果开创端侧用户新体验，有望带来新的增长曲线

阿里以其在模型及数据上的优势满足苹果高标准要求，进而赢得苹果青睐。阿里具备诸多优势。其一，阿里此前已与高通、MTK 等头部芯片厂适配，并有实际案例，如传音旗下 TECNO、itel 和 Infinix 三大品牌，其中 TECNO 在非洲、中东、东南亚地区采用端侧模式，阿里在安卓、Windows 平台积累的相关经验，可复用到



iOS 上。其二，阿里自身模型已在网信办完成备案，算法也历经一年多完成合规审批。苹果在海外版采用端侧自有模型结合 OpenAI 的架构，期望端侧模型由自身掌控。然而，任何公司欲在中国市场开展生成式 AI 服务，都需通过政府审批备案。国内产品上线商业化，需依据网信办《生成式人工智能服务管理办法》进行算法备案，包括开源代码及说明数据集来源等，苹果因不愿开源而导致端侧模型在国内备案受阻，在此背景下，苹果决定寻找一家中国合作伙伴以服务中国用户，而阿里成为其最优选择。其三，苹果的端侧模型架构，需先由用户端处理数据，再传至苹果私有云加密，最后传到阿里云进行推理，这要求阿里能解析加密数据、脱敏并定制，市面上的通用模型无法满足。其四，阿里能给予长期服务承诺，苹果对用户体验要求极高，最初欲合作的百度在打磨沟通中难以快速响应。DeepSeek 因缺乏端侧模型，且自身 C 端用户量无法支撑足够算力，难以满足苹果未来用户量爆发的需求，未能加入合作。此外，苹果还有诸多定制需求，如对接 Siri 指令并保证响应速度，定义端侧文本、图像等多种场景，要求端侧模型显存小、功耗低等。阿里因掌握更多中国消费者个人数据，如购物和支付习惯，且 AI 模型已在这些数据集上训练，能助力苹果为中国用户提供定制服务，成为苹果最终选择。不过，为保障数据安全，阿里不会与苹果分享个性化数据集。未来数据安全要求数据本地化存储，多模态交互使用云端模型时，处理完数据需过滤敏感信息，不允许云端存储完整原始信息。

图 21:iPhone 界面接入大模型



数据来源：G 人工智能，中信建投证券

**重点利用 AI 提升端侧（手机侧）的用户体验。**阿里需要进行多智能体的调度，同时需要借助 Siri 打通阿里模型，进而打通 iOS 上的各类应用，如邮件、短消息、记事本、日历等。如此一来，用户只需下达简单指令，

请务必阅读正文之后的免责条款和声明。

即可实现跨 APP 操控。例如，用户说“请帮我订一下明天下午 2 点的电话会议”，系统便能自动完成设置，并在明天 1:50 进行提醒。在交互体验方面，还涵盖诸多细分场景，打电话时的多语言翻译功能，即便双方语言不同，也能实现顺畅交流；又如图像搜索功能，用户只需告知 Siri 查找照片，系统就能快速自动检索出来，为用户带来全新体验。对于第三方应用，如淘宝或饿了么，现阶段仅能实现打开应用，选择商品和支付功能尚未完善，不过未来已有相关规划。今年，阿里仍将以提升端侧体验效果为主要目标。端侧 license 采用阶梯定价模式，出货量不同价格也有所差异，如 5 万级别、500 万级别、千万级别各有对应价格。同时，苹果要求阿里提供固定算力支撑，如针对 1000 万用户、2000 万用户、3000 万用户等不同规模，分别确定所需的算力卡数量。未来合作将围绕端侧 license 和一定的算力支撑展开，但这种合作并非百分百确定。目前阿里在技术上正在与苹果对接，但由于苹果认证的流程极为复杂，需要专门团队长期配合以满足其多样化需求，期间可能出现各种问题，远期功能仍具有不确定性。

**与苹果合作能带来显著增量。**苹果在硬件设计与制造领域处于领先地位，大模型针对苹果芯片、传感器等硬件进行深度优化，可促使硬件协同优化能力快速提升。同时，苹果产品定位高端，与之合作有助于大模型服务进入高端市场，吸引更多高端用户和企业客户，为后续拓展业务和市场份额奠定基础。阿里将在云端为苹果提供固定算力。据阿里测算，若支撑 1000 万用户，苹果需向阿里支付 20 亿人民币；若支撑 2000 万用户，则支付 40 亿至 50 亿人民币，支付金额与用户量相关。阿里将根据用户量扩容来支撑苹果未来在 AI 方面的云端算力需求。苹果的合作方式较为特殊，阿里需为其特别开辟一个集群，从营收角度看，该部分收入既属于模型收入，也可归为 maas 范畴，且苹果的数据与阿里其他客户的数据完全隔离。在未来商业合作中，阿里的首个重要节点是为 3 月 25 日的机器学习发布会做准备，届时将展示现阶段产品成果，手机、平板等设备都将集成模型。

## 四、AI 应用市场潜力巨大，有望逐步起量贡献增长动能

### （一）持续深耕自研模型，加速 AI 应用市场落地

AI 模型布局继续深耕通用推理模型与深度推理模型，同时聚焦细分场景产品和智能体的开发。阿里在 AI 模型布局上呈现出差异化战略路径，技术体系构建于两大维度：通用推理模型与深度推理模型。前者以通义千问为代表，承担基础能力支撑；后者专注于提升模型的推理能力，追求深度思考和分析的顶尖表现。深度推理模型 Qwen 系列的独立产品化，标志着阿里正在构建类 OpenAi-O1 技术体系这一新范式。该体系着重解决复杂任务拆解、多模态推理等深度认知问题，已在科研分析、金融建模等领域形成标杆案例。这种技术分层策略不仅完善了产品矩阵，更在底层形成了从基础研究到商业转化的完整闭环，为行业智能化转型提供了可复制的技术框架。阿里计划打造类似海外 Deep Research 智能体的产品，该智能体具备赋能研发与研究的能力，能够撰写和生成更高质量的研究报告。未来，阿里将从单纯的模型价格竞争，转向在商业应用及聚焦产品力方面，通过开发一系列新形态的智能体产品服务客户，并基于这些新产品制定定价策略。在产品架构层面，阿里云的战略转型尤为显著。过去一年 MaaS 服务以提供基础模型和微调能力为主，而当前则加速推进“模型即产品”的深度改造。具体表现为：一方面通过技术优化降低模型调用成本，另一方面将 AI 能力封装为可即插即用的智能体产品。这种转变使得客户无需直接调用底层模型，而是通过智能体 API 快速构建业务解决方案，显著降低了技术使用门槛。支撑这一转型的核心理念包含双重维度：技术驱动商业重构与技术普惠。在商业重构层面，阿里正以生成式 AI 为支点重塑核心业务体系——电商场景中，AIGC 技术已实现服装广告的智能拍摄与设计，成本降幅达 70%；广告系统通过 AI 精准投放使 ROI 提升 30%；钉钉生态则集成智能会议助手、HR 智能体等 20 余个垂直应用，服务超 7 亿用户。在技术普惠维度，阿里云将 MaaS 升级为包含基础模型、场景化产品、智能体市场的三维服务体系，通过标准化产品输出降低行业 AI 应用成本。

AI 应用市场潜力巨大，核心聚焦 AI Chatbot、教育及金融领域。AI Chatbot 产品在 APP 等应用中逐渐成为标配。其优势在于将以往的快思考转变为慢思考，虽回复时间延长几秒，但回答问题更加系统、准确，极大提升了回复质量。例如，DeepSeek 在数据方面存在短板，数据多依赖外部搜索。一旦为其提供搜索引擎的平台（如豆包、百度、搜狗等）停止合作，即便 R1 模型性能强大，在面向 C 端用户时，整体表现也会因数据不足而逊色于国内互联网大厂。相比之下，阿里拥有海量 UGC 产生的数据，在 AI Chatbot 方面存在数据优势。其次，大模型推动教育行业的创新，过去 AI 在教育场景中主要是简单地帮助用户上传经验、告知题目解法。如今，大模型的创新带来了新的变化，它能够呈现题目思考和推理过程，为学生和老师提供更强大的教育交互方式。用户可以反复提问，模型会从不同侧面讲述解题过程。金融领域一直对模型精度要求极高，阿里平台上的金融客户都倾向选择最强模型。大模型专门为金融业务赋能，业务系统集成了数据分析、数据挖掘等功能，在财报分析等场景中发挥作用，在大模型的加持下，业务系统的表现更加强劲。此外，还可量化为较小模型，为 AIPC 等产品赋能。虽然目前模型尚未优化到小端侧设备（如眼镜、行李箱、可穿戴设备等），但在电脑上已实现应用，未来有望在更多小端侧设备（如摄像头、路由器、扫地机器人、音箱、手机等）的芯片 Soc 上进行优化，实现端侧推理，推动整个行业发展。在汽车智能座舱方面，可以将小模型置于中控，提供 Copilot，提升人机交互体验。同时，智能驾驶方面，大模型可通过多模态模拟生成一些环境，类似 cosmos 世界模型生成用于模拟仿真的环境，如道路交通、桥梁、隧道及各种复杂天气，为自动驾驶测试提供更多数据源，减少实地采集数据的需求，加速模拟训练的复杂度。

阿里推出了一系列基于模型的增值服务产品，产品利润率可观。析言 GBI，集成于应用市场，可结合模型对阿里客户的销售数据、库存数据、营收数据等进行分析。用户输入 1000 个数据相关问题，它能快速生成报表，

并进行趋势分析，如预测库存时长、回款时间等，辅助客户决策。智能客服，包括电话和文字智能客服，通过预训练模型和搜索功能（如夸克搜索）获取信息并生成结果。RAG，企业可将自身数据灌入其中，模型能够进行推理。通义听悟，专注于语音识别和语音合成，可将会议音频快速转化为去除口语的完整纪要。电商辅助产品，能将原始图片输入模型，生成用于电商展示的图片 and 视频，大幅降低成本。在周边产品方面，RAG、SQL、BI 等已成为客户选择产品时的标配插件，通常客户至少会选择两到三个此类产品。这些产品价格相对稳定，能保持较好的利润率，在大模型本身可能利润微薄甚至不盈利的情况下，借助这些周边产品的协同作用，将整体利润水平拉高。

**表 3: 阿里云应用产品**

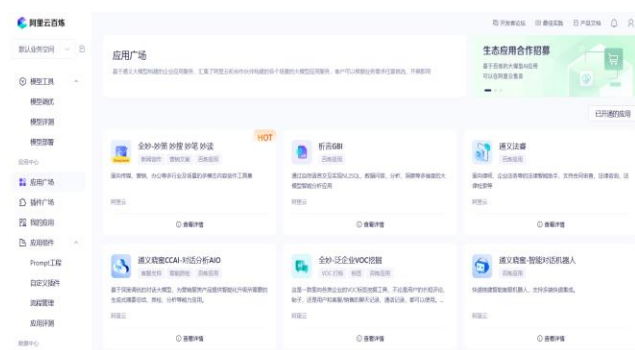
| 应用            | 核心功能                                                                                                                   |
|---------------|------------------------------------------------------------------------------------------------------------------------|
| 析言 GBI        | 基于大语言模型，通过自然语言交互实现数据问答、分析等功能，支持多数据源接入与多种部署模式，有不同版本满足各异需求，且具备数据安全防护和效果优化措施的智能数据分析产品。                                    |
| 通义点金          | 一站式金融场景开发平台，历经升级，融合超 25TB 金融数据，具备金融强化的 RAG 和 Agents 构建能力，提供丰富专业插件与场景模板，支持 API 调用和二次开发，能为金融机构和个人提供多方面金融服务，助力金融行业数字化转型。  |
| 通义法睿          | 基于通义千问打造的法律行业大模型产品，运用模型精调、强化学习等技术，具备法律咨询、文书生成、知识检索等功能，能为法律专业人士和普通用户提供高效、精准的法律服务。                                       |
| 通义听悟          | 基于通义千问大模型和音视频 AI 技术打造的工作学习 AI 助手，能实现实时语音转文字、多语言翻译、智能总结等功能，适用于会议、学习等多种场景，且对 C 端免费，还针对高校师生推出公益计划。                        |
| 通义晓蜜<br>CCAI  | 基于深度调优对话大模型打造的，面向营销服类产品提供智能化升级能力的官方应用，主要为开发者、自研企业、传统呼叫中心采购者及友商等服务，通过 API 输出服务。                                         |
| 全妙解决方案<br>类产品 | 面向传媒、公文、营销、办公等行业，基于通义大模型打造，具备多模态创作、专业写作、素材搜索、多种部署方式等能力，包含全妙 SaaS 解决方案、视频理解、汽车座舱新闻助手、影视互娱剧本策划等，还提供多种计费版本的大模型内容创作全套解决方案。 |
| 全妙轻应用系列       | 基于通义大模型的创作辅助产品，提供多模态创作和素材搜索功能，具备专业写作能力，支持多种部署方式，包括影视互娱剧本创作、传媒/零售文章风格与格式学习、网络评论生成等产品。                                   |

资料来源：阿里云，中信建投证券

**百炼未来计划构建大量智能体，目前智能体已划分出不同层级。**其中，面向企业端的角色化智能体发展态势良好。以钉钉为例，作为最早应用此类智能体的平台，已针对企业多样化场景打造了约 700 个官方智能体。这些智能体广泛分布于企业的各个职能领域。在财务板块，细分出了数十个智能体，涵盖会计、总账管理、财务管理等多个具体类别；人力资源领域则包括招聘、面试、薪资核算、培训等不同环节的智能体；此外，资质管理、采购行政等方面也均有相应的智能体覆盖，可谓渗透到了企业运营的方方面面。企业基于阿里提供的官方智能体，进行一定程度的调整，如改变智能体的顺序或名称，无需复杂的微调操作，即可转化为贴合自身需求的智能体投入使用，大大提高了智能体在企业中的应用效率和适应性。但目前应用仍存在任务复杂度边界，智能体当前无法替代 CRM、OA 等复杂流程系统（如多部门协作的项目管理、跨系统数据同步），其定位偏向简单任务自动化和流程节点提速。



图 22:百炼应用广场



数据来源：阿里云，中信建投证券

图 23:百炼插件广场



数据来源：阿里云，中信建投证券

## （二）战略聚焦 C 端产品叠加 AI 赋能

**成立智能信息事业群，数据打通与 AI 技术驱动新流量入口。**过去阿里云主要专注于 To B 业务，而 C 端产品呈现出分散的状态，像夸克、秒鸭、通义等产品，由多个不同的团队分别运作。这种分散的管理模式使得 C 端业务发展受到诸多限制。一方面，阿里整体关注利润率，对 C 端业务的推广投入较少；另一方面，C 端产品分布在不同事业群，如公益项目放在阿里云，不仅缺乏做 To C、To B 业务的预算，而且员工习惯了 B 端产品的思路，与 C 端战略差异较大。为了改变这一现状，阿里在 2024 年底对主架构进行了重大重组，将原本分散的 C 端产品和硬件产品整合到新成立的智能信息事业群。同时，从海外聘请了来自 Salesforce 的许主洪教授，负责 C 端技术和业务规划。这一举措使得 C 端业务得以独立发展，与 B 端产品完全分离，模型团队和 C 端产品团队也各自独立运作。未来，C 端产品将拥有独立预算，计划在 C 端市场大规模投放互联网广告，与豆包等产品展开竞争。对 C 端用户而言，打通淘系生态与数据至关重要。AI 驱动应用会使电商入口迁移，需打通地图、电商、飞猪等生态数据，打通后入口将转向新的 AI 方向，形成新流量入口。未来流量会协同高德地图、饿了么等生态 APP，相互推荐入口，通义 APP 为协同入口，引导客户下载和使用，从而实现业务的增长和拓展。在 C 端产品方面，各个产品都有明确的发展目标和策略，关注教育、社交类产品，以季度为单位推出新产品。以夸克 APP 为例，它已经拥有 2000 多万稳定的日活用户，且用户群体主要集中在 15 岁到 25 岁的年轻人。未来，夸克将聚焦于 AI 教育、AI 健康、AI 高考志愿等场景，有望成长为超级 APP，2025 年目标是将日活提升到 3000 万。同时，夸克还具备独特的商业价值，其宽带发展能力可以封装成搜索 API，供模型能力客户调用，搜索结果也可以单独售卖，相关收入归属于云业务。此外，C 端事业部将形成产品矩阵对标豆包 Flow 团队，但产品数量会控制在五六个左右，以实现更高效的运营和管理。

图 74: 阿里巴巴生态体系



数据来源：公司公告，中信建投证券

端侧战略主要聚焦于汽车、手机、PC 三大硬件，同时对于小硬件也采用自主研发或合作的方式推进。端侧已经拥有丰富的模型，包括不同规模的语言模型、视觉模型和语音模型。这些模型具有强大的功能，能够识别多语言，包括方言，还可以合成不同的声色，为用户提供更加个性化的体验。智能信息事业群下专门设立了负责 AI 硬件的团队，目前已有 200 多人。在 2024 年，他们推出了多款产品，如升级版的天猫精灵 X6、闺蜜机等，这些产品背后都采用了模拟驱动技术。此外，还涉及 AI 眼镜等产品，像 Rokid、雷鸟等背后的技术应用。在大硬件方面，阿里积极与外部合作。手机方面与传音深度捆绑，传音出海的手机搭载端侧模型后可实现打电话实时多语言翻译，且在本地处理；PC 方面与联想合作，联想新一代 AIPC 中集成了相关技术；汽车方面与智己汽车、吉利汽车等新锐汽车公司合作，将端侧模型嵌入智能座舱，提供新的交互式体验。阿里还计划将通义模型做成操作系统（OS），采用“1+3+N”模式开放给生态。“1”是通义千问，“3”是中控屏、智能音箱等三个产品，“N”是周边产品如扫地机等，通过 SDK 让周边客户接入，形成一个庞大的 AI 硬件生态系统。

**大模型与数据打通升级用户画像，提升精准推送质量。**Meta 这类占据市场份额较大的社交平台，用户使用时间长，产生的信息丰富。Meta 使用大模型捕捉用户的历史资料、平台信息、分享的图片、位置、言论和转发内容等，以小时为周期更新画像，精准度远超其他平台。在这种精准画像的基础上，平台可以实现默认推荐商品、增加智能导购以及与用户对话推送更精准信息。这种合作模式能以小时或天为单位更新画像，精准度远超静态画像。相比之下，国内电商领域的用户画像与 Meta 不同，阿里很多推荐基于相对静态的画像，如年龄、地址、职业、收入等，静态画像依据的是一两个月前的信息，精准度自然欠佳。要提升广告效果，首先要打通整个阿里系的数据，包括飞猪等非淘系 APP 以及未来 C 端产品的数据，用户画像会更饱满，也可拉通各 APP 的标签进行共同推荐。目前阿里先打通非淘系的地图、飞猪等 APP 的数据，并进行灰度测试，选取 1 万个或 500 个客户测试，已取得一定效果。不过，阿里存在内容缺失的问题，缺乏像抖音那样用户自发创作的丰富内容。若能补齐内容短板，再结合数据打通，效果会比之前更好。未来，通义等其他 APP 中拉通淘系数据，用户通过与 AI 对话就能搜索并直接购买产品。

## 盈利预测及估值

预计 FY2025-2026 阿里巴巴的收入增速分别为 6.07%和 9.27%，Non-GAAP 净利润增速分别为-2.57%和 12.91%。其中淘天收入增速分别为 2.33%和 4.99%，经调 EBITA 增速分别为-0.86%和 4.02%，阿里云收入增速分别为 10%和 15%，经调 EBITA 利润率分别为 9.5%和 11.0%。采用分部估值法，对应淘天集团（15 倍 PE）、云智能集团（5 倍 PS）每 ADS 估值分别为 141.18 美元和 39.40 美元，维持“买入”评级，目标价 180.57 美元/ADS，对应港股 175.61 港币/股。

**表 4: 阿里巴巴盈利预测（单位：人民币百万元）**

| 资产负债表      | FY2023    | FY2024    | FY2025E   | FY2026E   | 利润表          | FY2023    | FY2024    | FY2025E   | FY2026E   |
|------------|-----------|-----------|-----------|-----------|--------------|-----------|-----------|-----------|-----------|
| 总流动资产      | 697,966   | 752,864   | 974,548   | 1,232,700 | 营业收入         | 868,687   | 941,168   | 998,327   | 1,090,870 |
| 总非流动资产     | 1,055,078 | 1,011,965 | 1,034,973 | 1,028,772 | 营业成本         | (549,695) | (586,323) | (598,996) | (654,522) |
| 总资产        | 1,753,044 | 1,764,829 | 2,009,520 | 2,261,471 | 销售费用         | (103,496) | (115,141) | (139,766) | (141,813) |
| 总流动负债      | 385,351   | 421,507   | 443,601   | 448,960   | 管理费用         | (42,183)  | (41,985)  | (44,535)  | (48,663)  |
| 总非流动负债     | 244,772   | 230,723   | 223,769   | 217,156   | 研发费用         | (56,744)  | (52,256)  | (55,430)  | (60,568)  |
| 总负债        | 630,123   | 652,230   | 667,370   | 666,115   | 无形资产摊销       | (13,504)  | (21,592)  | (7,500)   | (7,447)   |
| 普通股股本      | 1         | 1         | 1         | 1         | 商誉减值损失       | (2,714)   | (10,521)  | (35,000)  | (5,000)   |
| 储备         | 1,038,694 | 1,021,357 | 1,261,828 | 1,515,895 | EBIT         | 100,351   | 113,350   | 117,948   | 172,857   |
| 库存股        | 28,763    | 27,684    | 27,684    | 27,684    | 利息收入及投资损益    | (11,071)  | (9,964)   | 40,000    | 15,018    |
| 其他综合性收益    | (10,417)  | 3,598     | (3,410)   | 94        | 利息支出         | (5,918)   | (7,947)   | (8,000)   | (8,000)   |
| 归属母公司股东权益  | 999,515   | 997,272   | 1,230,736 | 1,488,307 | 其他收益         | 5,823     | 6,157     | 5,000     | 2,000     |
| 少数股东权益     | 123,406   | 115,327   | 111,415   | 107,050   | EBT          | 89,185    | 101,596   | 154,948   | 181,875   |
| 股东权益合计     | 1,122,921 | 1,112,599 | 1,342,151 | 1,595,356 | 所得税          | (15,549)  | (22,529)  | (32,539)  | (36,375)  |
| 总权益及负债     | 1,753,044 | 1,764,829 | 2,009,520 | 2,261,471 | 应占股权投资损益     | (8,063)   | (7,735)   | 8,000     | 0         |
|            |           |           |           |           | 净利润(含少数股东权益) | 65,573    | 71,332    | 130,409   | 145,500   |
| 现金流量表      |           |           |           |           | Non-GAAP 净利润 | 141,379   | 157,479   | 153,429   | 173,231   |
| 经营活动所得现金净额 | 199,752   | 182,593   | 221,661   | 224,653   |              |           |           |           |           |
| 投资活动所得现金净额 | (135,506) | (21,824)  | (30,955)  | (671)     | 关键比率         |           |           |           |           |
| 融资活动所得现金净额 | (65,619)  | (108,244) | 1,450     | 1,523     | 营业收入 YOY     | 1.83%     | 8.34%     | 6.07%     | 9.27%     |

港股公司深度报告

|                         |         |         |         |         |                           |        |        |        |        |
|-------------------------|---------|---------|---------|---------|---------------------------|--------|--------|--------|--------|
| 汇率变动                    | 3,530   | 4,389   | 0       | 0       | OPM                       | 11.55% | 12.04% | 11.81% | 15.85% |
| 现金及现金等<br>价物增加/减少<br>净额 | 2,157   | 56,914  | 192,156 | 225,505 | Non-GAAP<br>EBIT A margin | 17.03% | 17.53% | 17.52% | 19.04% |
| 年初的现金及<br>现金等价物         | 227,353 | 229,510 | 286,424 | 478,580 | Non-GAAP 净<br>利润 YOY      | 3.66%  | 11.39% | -2.57% | 12.91% |
| 年末的现金及<br>现金等价物         | 229,510 | 286,424 | 478,580 | 704,085 | Non-GAAP 净<br>利率          | 16.28% | 16.73% | 15.37% | 15.88% |

资料来源: Bloomberg, 中信建投证券



## 风险分析

资本开支低于预期；云业务收入增长低预期，利润率提升低预期，AI 收入占比低预期；美国对华芯片制裁影响公司云业务增长；主业复苏节奏低预期，宏观经济及社零增长疲弱，CMR 增长低预期，国内电商市场份额超预期下行，淘天集团利润率低预期；海外电商市场竞争激烈，AIDC 减亏低预期；海外关税政策调整影响 AIDC 增速；行业监管风险；蚂蚁集团金控牌照落地进度低于预期；美联储降息进程低预期；中美关系发展的不确定性；中概股退市风险。

## 分析师介绍

### 崔世峰

海外研究首席分析师，南京大学硕士，8 年买方及卖方复合从业经历，专注于互联网及科技公司研究，擅长游戏行业研究。2022-2023 年新财富港股及海外最佳研究团队入围，2019-2020 年新财富传媒最佳研究团队第二名核心成员。

### 于伯韬

FRM，香港大学金融学硕士，7 年互联网及港股策略卖方从业经历，2024/2021/2020 年新财富港股及海外方向第五名，2022 年新浪金麒麟港股及海外市场最佳分析师第三名，2020 年新浪金麒麟港股及海外市场新锐分析师第一名。

## 评级说明

| 投资评级标准                                                                                                                                                  |      | 评级   | 说明            |
|---------------------------------------------------------------------------------------------------------------------------------------------------------|------|------|---------------|
| 报告中投资建议涉及的评级标准为报告发布日后 6 个月内的相对市场表现，也即报告发布日后的 6 个月内公司股价（或行业指数）相对同期相关证券市场代表性指数的涨跌幅作为基准。A 股市场以沪深 300 指数作为基准；新三板市场以三板成指为基准；香港市场以恒生指数作为基准；美国市场以标普 500 指数为基准。 | 股票评级 | 买入   | 相对涨幅 15% 以上   |
|                                                                                                                                                         |      | 增持   | 相对涨幅 5%—15%   |
|                                                                                                                                                         |      | 中性   | 相对涨幅-5%—5%之间  |
|                                                                                                                                                         |      | 减持   | 相对跌幅 5%—15%   |
|                                                                                                                                                         |      | 卖出   | 相对跌幅 15% 以上   |
|                                                                                                                                                         | 行业评级 | 强于大市 | 相对涨幅 10% 以上   |
|                                                                                                                                                         |      | 中性   | 相对涨幅-10-10%之间 |
|                                                                                                                                                         |      | 弱于大市 | 相对跌幅 10% 以上   |

## 分析师声明

本报告署名分析师在此声明：（i）以勤勉的职业态度、专业审慎的研究方法，使用合法合规的信息，独立、客观地出具本报告，结论不受任何第三方的授意或影响。（ii）本人不曾因，不因，也将不会因本报告中的具体推荐意见或观点而直接或间接收到任何形式的补偿。

## 法律主体说明

本报告由中信建投证券股份有限公司及/或其附属机构（以下合称“中信建投”）制作，由中信建投证券股份有限公司在中华人民共和国（仅为本报告目的，不包括香港、澳门、台湾）提供。中信建投证券股份有限公司具有中国证监会许可的投资咨询业务资格，本报告署名分析师所持中国证券业协会授予的证券投资咨询执业资格证书编号已披露在报告首页。

在遵守适用的法律法规情况下，本报告亦可能由中信建投（国际）证券有限公司在香港提供。本报告作者所持香港证监会牌照的中央编号已披露在报告首页。

## 一般性声明

本报告由中信建投制作。发送本报告不构成任何合同或承诺的基础，不因接收者收到本报告而视其为中信建投客户。

本报告的信息均来源于中信建投认为可靠的公开资料，但中信建投对这些信息的准确性及完整性不作任何保证。本报告所载观点、评估和预测仅反映本报告出具日该分析师的判断，该等观点、评估和预测可能在不发出通知的情况下有所变更，亦有可能因使用不同假设和标准或者采用不同分析方法而与中信建投其他部门、人员口头或书面表达的意见不同或相反。本报告所引证券或其他金融工具的过往业绩不代表其未来表现。报告中所含任何具有预测性质的内容皆基于相应的假设条件，而任何假设条件都可能随时发生变化并影响实际投资收益。中信建投不承诺、不保证本报告所含具有预测性质的内容必然得以实现。

本报告内容的全部或部分均不构成投资建议。本报告所包含的观点、建议并未考虑报告接收人在财务状况、投资目的、风险偏好等方面的具体情况，报告接收者应当独立评估本报告所含信息，基于自身投资目标、需求、市场机会、风险及其他因素自主做出决策并自行承担投资风险。中信建投建议所有投资者应就任何潜在投资向其税务、会计或法律顾问咨询。不论报告接收者是否根据本报告做出投资决策，中信建投都不对该等投资决策提供任何形式的担保，亦不以任何形式分享投资收益或者分担投资损失。中信建投不对使用本报告所产生的任何直接或间接损失承担责任。

在法律法规及监管规定允许的范围内，中信建投可能持有并交易本报告中所提公司的股份或其他财产权益，也可能在过去 12 个月、目前或者将来为本报中所提公司提供或者争取为其提供投资银行、做市交易、财务顾问或其他金融服务。本报告内容真实、准确、完整地反映了署名分析师的观点，分析师的薪酬无论过去、现在或未来都不会直接或间接与其所撰写报告中的具体观点相联系，分析师亦不会因撰写本报告而获取不当利益。

本报告为中信建投所有。未经中信建投事先书面许可，任何机构和/或个人不得以任何形式转发、翻版、复制、发布或引用本报告全部或部分内容，亦不得从未经中信建投书面授权的任何机构、个人或其运营的媒体平台接收、翻版、复制或引用本报告全部或部分内容。版权所有，违者必究。

### 中信建投证券研究发展部

北京  
 朝阳区景辉街 16 号院 1 号楼 18 层  
 电话：（8610）56135088  
 联系人：李祉瑶  
 邮箱：lizhiyao@csc.com.cn

上海  
 上海浦东新区浦东南路 528 号南塔 2103 室  
 电话：（8621）6882-1600  
 联系人：翁起帆  
 邮箱：wengqifan@csc.com.cn

深圳  
 福田区福中三路与鹏程一路交汇广电金融中心 35 楼  
 电话：（86755）8252-1369  
 联系人：曹莹  
 邮箱：caoying@csc.com.cn

### 中信建投（国际）

香港  
 中环交易广场 2 期 18 楼  
 电话：（852）3465-5600  
 联系人：刘泓麟  
 邮箱：charleneliu@csci.hk