下载日志已记录，仅供内部参考

国联民生证券 GUOLIAN MINSHENG SECURITIES

Token“工厂”深度

打通 AI 上游到应用的关键

- **Token 需求持续扩张，系统级产能成为基础设施投资的核心观察指标。**国家数据局披露，我国日均 Token 调用量由 2024 年初约 1,000 亿增至 2026 年 3 月超过 140 万亿。Token 同时进入模型调用、API 计价与 AI Factory 产出衡量体系；推理基础设施的竞争随之延伸至计算、互连、存储、模型和软件协同形成的有效吞吐与单位 Token 成本。
- **Agent 提高单项任务的 Token 消耗，也抬高对推理系统的要求。**多轮推理、工具调用与长上下文使单项任务可能包含多次模型请求。瓶颈由计算能力延伸至显存、KV Cache、Prefill/Decode 调度和响应时延；缓存复用、PD 分离、智能路由及上下文管理可共同影响单位资源的有效 Token 产出。
- **AI 商业价值的衡量正由 Token 使用量进一步延伸至任务结果。**AI 产业的价值锚正在进一步由 Token 向 Task 与业务结果延伸。Token 衡量模型服务使用量，Action 衡量 Agent 执行的具体业务步骤，而 Resolution/Outcome 则进一步衡量业务问题是否真正得到解决；进入企业场景后，客户最终购买的并非 Token 本身，而是 Token 驱动的数据调用、流程执行和任务结果。企业级 Agent 的竞争壁垒也因此从模型能力进一步延伸至企业数据、业务语义、API、权限治理、工作流和真实生产反馈。我们认为，随着单位 Token 成本持续下降、使用需求释放，AI 产业价值衡量将进一步聚焦 Tokens per Task、Task Success Rate、Cost per Successful Task 以及单项成功 Task 所创造的业务价值。
- **投资建议：**我们判断，随着 Agent 从模型调用走向复杂任务执行，Token 需求有望在调用规模扩大与单任务 Token 密度提升的共同驱动下持续增长，产业链关注重点也将由算力供给进一步延伸至 Token 生产、分发运营、效率优化与企业 Task 变现。沿此主线，建议关注大模型厂商智谱、MiniMax、科大讯飞、昆仑万维；运营商中国电信、中国移动、中国联通；云服务厂商阿里巴巴、金山云、优刻得、青云科技；算力租赁与 Token 工厂运营企业协创数据、宏景科技、软通动力、行云科技等；Token 优化企业星环科技、信测标准、集智股份、并行科技；垂直行业 Token 应用企业金山办公、用友网络、金蝶国际、万兴科技等。
- **风险提示：**AI 应用及 Agent 商业化进展不及预期；Token 调用需求增长不及预期；模型服务降价与行业竞争加剧；推理技术迭代及投入回收不及预期；数据安全与合规风险。

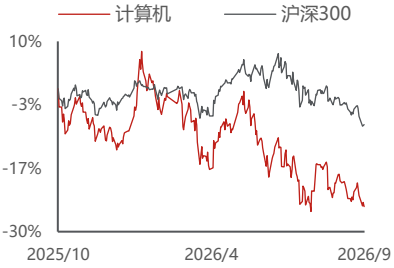
推荐

维持评级



分析师 吕伟
执业证书： S0590525110033
邮箱： lvwei_yj@glms.com.cn
分析师 郭新宇
执业证书： S0590525110034
邮箱： guoxinyu@glms.com.cn

相对走势



相关研究

1. 计算机行业周报 20260919：华为全联接大会全透视-2026/09/19
2. 计算机行业周报 20260913：从浪潮定增看超节点、液冷、存算一体三大机遇-2026/09/13
3. 计算机行业动态报告：昇腾、寒武纪、浪潮与超节点-2026/09/09
4. 计算机行业 2026 年中报分析：AI 依然是业绩重要主线-2026/09/07
5. 计算机行业周报 20260906：AI 投资重新聚焦大模型智能上界-2026/09/06

下载日志已记录，仅供内部参考

投资聚焦

研究背景

随着大模型从生成内容走向调用工具、执行流程，AI 产业的产出不再只体现为模型生成了多少 Token，还体现为这些 Token 能否支持一项任务稳定完成。本报告希望回答两个相互关联的问题：

（1）算力、模型与软件如何共同提高可交付的 Token 产能？

（2）Token 进入企业业务流程后，如何进一步转化为可验证的任务结果？

区别于市场的观点/方法

本报告沿“Token 生产—Token 优化—Task 执行”梳理产业链：基础设施侧关注系统级吞吐与单位 Token 成本，推理服务侧关注满足延迟要求的有效吞吐，应用侧关注任务成功率和单位成功任务成本。报告同时区分 Token、Action 与 Outcome，分别对应模型用量、执行步骤和任务结果，以此识别模型、平台及企业应用环节的不同价值来源。

近期催化

近期产业进展沿模型效率、推理服务和企业数据底座同步展开。9 月，星环科技在 VLDB 2026 展示面向 AI Agent 的 GPU 原生认知数据库，将数据分析、知识检索、长期记忆与上下文供给纳入统一架构；华为云在全联接大会公布灵衢昇腾 950 智算集群服务、Agentic MaaS 及企业级智能体平台的最新进展，并披露其 Token 吞吐性能较上一代算力提升 20%；智谱披露 GLM-5.3-Flash 通过量化、缓存优化及 EPD 分离架构等技术，使端到端服务性能较初始基线提升约 3 倍。我们认为后续应关注技术进展能否转化为模型服务、企业数据调用及可验证任务结果。

结论与建议

建议沿六类公司跟踪 Token 需求的商业化兑现。大模型厂商关注智谱、MiniMax、科大讯飞、昆仑万维；电信运营商关注中国电信、中国移动、中国联通；云服务厂商关注阿里巴巴、金山云、优刻得、青云科技；算力租赁与 Token 工厂运营关注协创数据、宏景科技、软通动力、行云科技、弘信电子、润建股份、东方国信、中国软件国际、彩讯股份、吉宏股份、铜牛信息、超讯通信、南威软件、平治信息、思特奇、亿都国际；Token 优化关注星环科技、信测标准、集智股份、并行科技；垂直行业 Token 应用关注金山办公、用友网络、金蝶国际、汉得信息、拓尔思、万兴科技、创业慧康、新致软件。Token 调用增长最终仍需转化为可控成本下的稳定服务、持续客户使用及可验证的业务结果。

下载日志已记录，仅供内部参考

目录

1 Token 工厂：AI 产业价值传导的新框架 4
1.1 从 Bit 到 Token：AI 产业计量单位的变化 4
1.2 Token 工厂：从基础设施到 AI 应用的完整生产体系 5
1.3 Token 供给：模型、算力与系统协同共同决定智能产能 7
1.4 Token Economics：从算力投入到单位 Token 经济性 9
2 Token 优化：AI 规模化应用的效率基础 12
2.1 Agent 推动 Token 需求从规模增长走向密度提升 12
2.2 Token 生产瓶颈：计算、显存、缓存、调度与延迟 13
2.3 Token 优化：从单点加速走向全栈推理系统 15
2.4 Token 优化如何衡量：从系统性能到成功任务成本 18
2.5 Token 优化走向平台化：推理引擎与 Inference Platform 成为独立中间层 20
3 Token 到 Task：AI 价值从智能产出走向任务结果 23
3.1 从 Token 到 Task：AI 交付与计价单位向业务结果迁移 23
3.2 企业级 Task 壁垒：数据、业务流程与真实反馈 24
3.3 市场空间与价值迁移：Token 需求扩张，Task 成为重要价值锚 26
4 投资建议 28
4.1 行业投资建议 28
4.2 重点公司 28
5 风险提示 32
插图目录 33
表格目录 33

下载日志已记录，仅供内部参考

1 Token 工厂：AI 产业价值传导的新框架

1.1 从 Bit 到 Token：AI 产业计量单位的变化

计算系统长期使用不同指标分别描述数据规模与计算能力。NIST 将 Bit 定义为取值为 0 或 1 的二进制数字，并将 Byte 定义为由 8 个 Bit 组成的序列，Bit 与 Byte 构成数字信息表示和存储的基础单位。IBM 指出，经典计算系统通常使用 FLOPS、内存容量、内存带宽以及处理器时钟频率等指标描述系统能力，其中 FLOPS 表示系统每秒能够完成的浮点运算次数。因此，传统计算体系中，Bit 与 Byte 主要用于描述数字信息规模，FLOPS 等指标则用于刻画计算设备执行运算的能力。

大模型进一步将自然语言、代码等输入转化为 Token 进行处理，并以 Token 序列生成输出。OpenAI 将 Token 定义为模型处理文本时使用的基本单位，一个 Token 可以对应字符、部分单词、完整单词或标点符号，同一段文本产生的 Token 数量也可能因模型、编码方式和语言不同而存在差异。文本发送至模型后首先被拆分为 Token，模型随后处理这些输入 Token 并持续生成新的输出 Token。OpenAI 的 API 结果可以直接记录 input tokens、output tokens 和 total tokens 等使用数据，并进一步区分输入 Token、缓存输入 Token、输出 Token 及推理 Token 等不同类型，使模型实际处理的信息规模能够被持续统计和追踪。

表1：大模型服务中的主要 Token 类型

类别	描述
输入 Token(Input tokens)	在请求中提供给模型的 Token。它们也被称为提示 Token(prompt tokens)。
输出 Token(Output tokens)	由模型生成的 Token。Chat Completions 接口将其称为补全 Token(completion tokens)。
缓存输入 Token(Cached input tokens)	通过提示缓存复用的输入 Token。它们的定价可能与未缓存的输入 Token 不同。
推理 Token(Reasoning tokens)	推理模型在生成其可见答案之前，内部使用的 Token。

资料来源：OpenAI，国联民生证券研究所

Token 已经进一步由模型内部处理单位进入大模型服务的商业计量体系。阿里云百炼对大语言模型调用按照实际 Token 用量计费，并分别设置输入 Token 与输出 Token 价格，部分模型在缓存命中时还采用区别于标准输入 Token 的计费价格。基础设施侧，NVIDIA 进一步将 Token 称为推理模型、Agent 和智能系统的“生产单位”，并将 AI Factory 描述为将能源和数据转化为智能 Token 的生产系统。围绕这一生产过程，NVIDIA 使用 Tokens per Second、Tokens per Watt、Cost per Token、Utilization 和 Uptime 等指标衡量 AI Factory 的产出效率、能耗效率、单位成本以及系统运行情况。Token 由此同时进入模型调用、服务计价以及 AI 基础设施产出衡量体系。

从 FLOPS 等基础算力指标向 Token 指标延伸，AI 推理的衡量对象开始由计

下载日志已记录，仅供内部参考

算资源本身进一步覆盖模型实际交付的输出。NVIDIA 指出，FLOPS per Dollar 衡量单位投入能够获得的原始计算能力，而原始计算能力并不直接等同于现实中的 Token 产出；Cost per Token 则进一步衡量生成实际交付 Token 所对应的综合成本。与此同时，阿里云指出，Token 是模型对输入文本进行切分后的基本单位，不同模型可能采用不同的切分方法，字符数量与 Token 数量并非一一对应；其大模型服务进一步按照实际输入和输出 Token 数量进行计量和计费。因此，模型服务的实际使用成本不仅与每百万 Token 单价有关，也取决于实际调用过程中产生的输入和输出 Token 数量。Token 作为连接底层计算资源与上层模型服务的中间计量单位，其本身的产出需要依赖完整的计算基础设施、模型和软件系统共同实现。

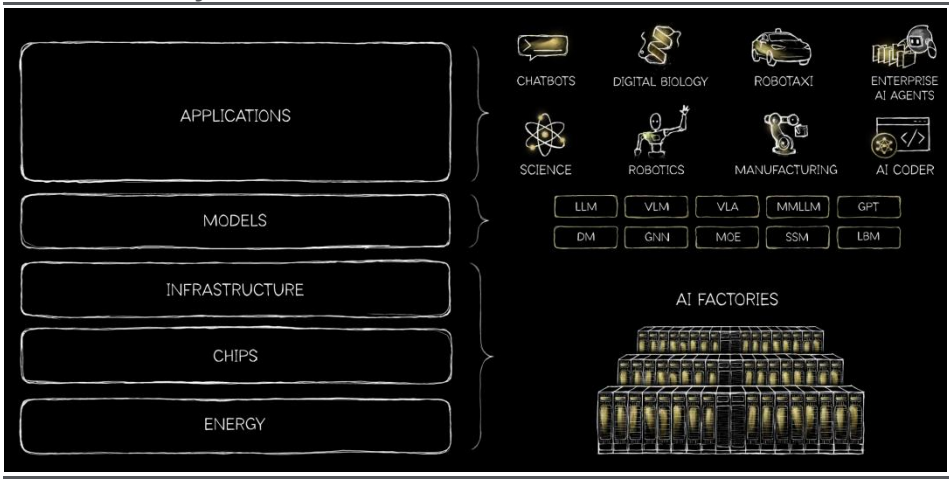
1.2 Token 工厂：从基础设施到 AI 应用的完整生产体系

AI Factory 并非单一 GPU 集群，而是面向 AI 工作负载构建的全栈智能生产体系。NVIDIA 将 AI Factory 描述为一种面向规模化智能生产的新型基础设施。与传统数据中心主要承担通用计算和数据处理不同，AI Factory 通过加速计算基础设施与 AI 软件的紧密结合，覆盖数据摄取、模型训练、微调以及长思考推理等 AI 生命周期环节。NVIDIA 进一步将 Token 视为推理模型、Agent 和智能系统的“生产单位”，并将 AI Factory 描述为将数据与能源转化为智能与 Token 的生产系统。因此，AI Factory 最终交付的不再只是 GPU 算力或理论计算能力，而是由模型持续生成、并可被上层应用进一步调用的智能产出。

AI Factory 需要能源、芯片、基础设施、模型和应用多个层级协同运行，而非依靠单一技术环节完成 Token 生产。NVIDIA 目前将 AI 体系划分为 Energy、Chips、Infrastructure、Models 和 Applications 五个关键层级，并指出 AI Factory 将这五个层级整合为面向 Agentic AI、Physical AI 和高性能计算的完整系统。NVIDIA 进一步指出，五个层级需要协同扩展，能源支撑芯片运行，芯片构成计算基础，基础设施承载模型运行，模型再向上支撑具体 AI 应用。在基础设施内部，AI Factory 还需要同时整合加速计算集群、高速网络、存储和基础设施软件，以支持训练、推理及其他生产级 AI 工作负载。

下载日志已记录，仅供内部参考

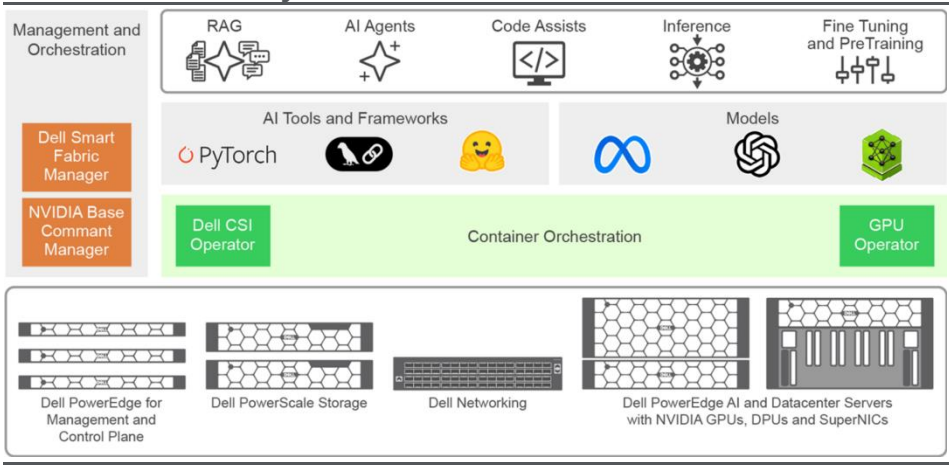
图1：AI Factory 由能源、芯片、基础设施、模型与应用五层共同构成



资料来源：NVIDIA 官网，国联民生证券研究所

企业级 AI Factory 进一步体现出从底层硬件向数据、软件和模型延伸的全栈特征。Dell 将其与 NVIDIA 共同构建的 AI Factory 定义为端到端企业 AI 解决方案，将 Dell 的计算、存储、客户端设备、软件和服务能力与 NVIDIA AI 基础设施及软件体系进行整合，并通过高速网络连接各类组件。Dell 最新参考架构进一步将计算、网络、存储、编排和 AI 软件能力整合至统一平台，可支持 RAG、AI Agent、代码助手、模型微调和推理等不同 AI 工作负载。在软件与模型侧，平台可进一步接入 NVIDIA AI Enterprise、模型服务运行时及相关 AI 框架，使企业能够在统一基础设施上完成模型构建、调整、部署和推理。

图2：企业级 AI Factory 整合计算、存储、网络、AI 软件与上层应用



资料来源：Dell Technologies，国联民生证券研究所

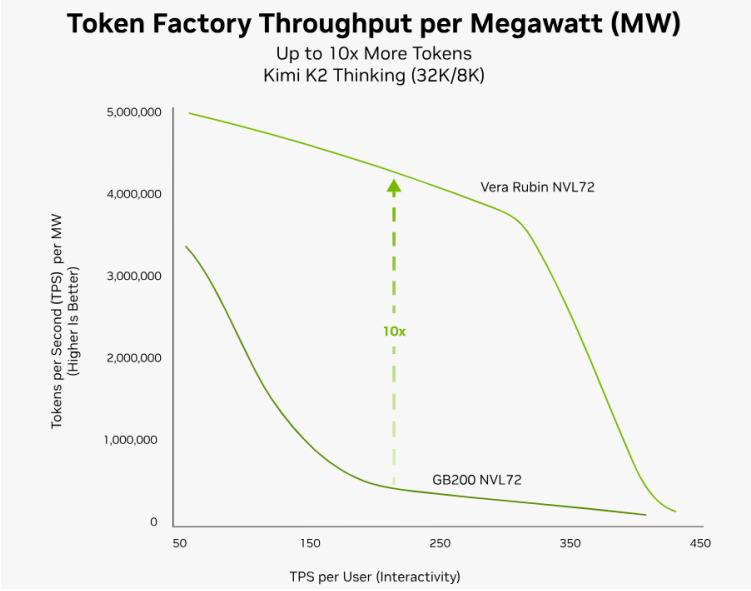
下载日志已记录，仅供内部参考

Token 贯穿上述技术栈，并成为不同系统组件共同参与生产的对象。NVIDIA 指出，当用户向 AI 系统提交 Prompt 后，AI Factory 首先将输入转换为 Token，再由运行于 GPU 之上的 AI 模型处理这些 Token 并生成响应，而 GPU 之间的并行计算依赖高速网络和互连共同完成。NVIDIA 将 AI Factory 的组成进一步概括为加速计算、网络、软件、存储、系统以及工具和服务，说明实际 Token 生产是完整技术栈协同运行的结果。我们认为，基于上述全栈体系，可以进一步将本文关注的 Token 产业链抽象为“算力供给—模型与 Token 优化—应用与任务执行”三个层级，以此作为后续分析不同环节技术作用与价值来源的产业地图。

1.3 Token 供给：模型、算力与系统协同共同决定智能产能

NVIDIA 对 AI Factory 的衡量正由单芯片性能进一步延伸至系统级 Token 产能。NVIDIA Vera Rubin NVL72 将 72 颗 Rubin GPU、36 颗 Vera CPU 以及高速互连与网络能力集成于机架级系统，面向大规模 Agentic AI 与推理工作负载。NVIDIA 进一步以 Tokens per Second per MW 衡量 AI Factory 在功率约束下的 Token 吞吐效率；根据其披露的 KimiK2-Thinking 性能口径，Vera Rubin NVL72 的每兆瓦 Token 吞吐最高可达 GB200 NVL72 的约 10 倍。我们认为，这表明 AI 基础设施的竞争正在从单颗 GPU 峰值算力，进一步转向计算、互连、网络与软件协同后的系统级 Token 产能。

图3：Vera Rubin NVL72 单位功率 Token 吞吐最高达到 GB200 NVL72 的 10 倍

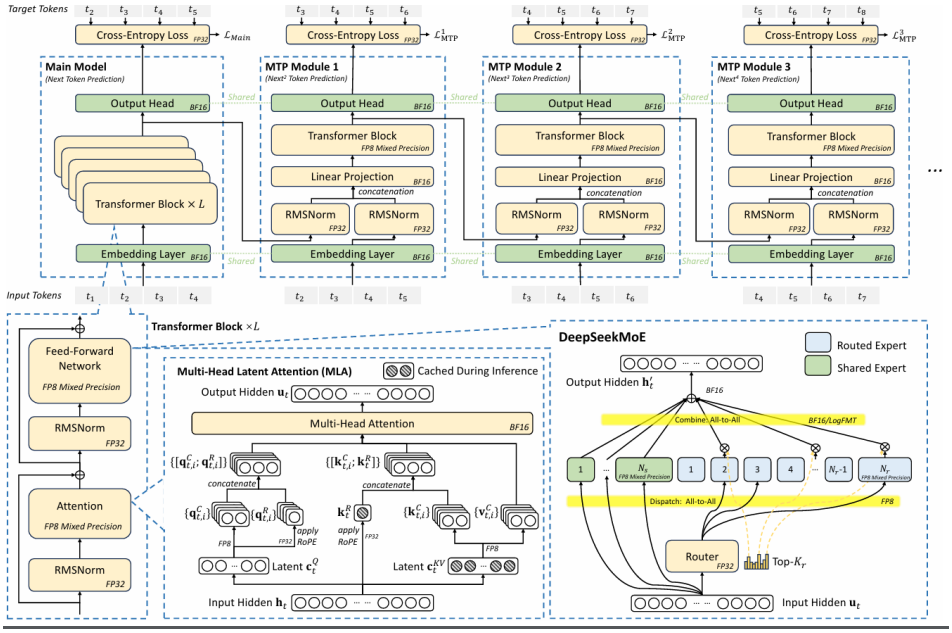


资料来源：NVIDIA 官网，国联民生证券研究所

下载日志已记录，仅供内部参考

DeepSeek-V3 的案例显示，模型架构可以直接改变模型运行中的计算与存储需求。DeepSeek-V3 采用 671B 参数的 MoE 架构，每个 Token 仅激活 37B 参数，通过稀疏激活降低每个 Token 对应的计算需求。与此同时，DeepSeek-V3 采用 Multi-head Latent Attention（MLA），将 KV 表示压缩至较小的潜在向量，并在推理阶段仅缓存该潜在表示，从而降低 KV Cache 的存储开销。DeepSeek 团队进一步测算，在 BF16 精度下，DeepSeek-V3 的 KV Cache 约为 70.3KB/Token，而 Qwen2.5-72B 和 Llama3.1-405B 分别约为 327.7KB/Token 和 516.1KB/Token。由此可见，MoE 与 MLA 等模型架构设计能够分别改变计算量与 KV Cache 占用，进而影响模型对底层计算和存储资源的需求。

图4：DeepSeek-V3 基本架构

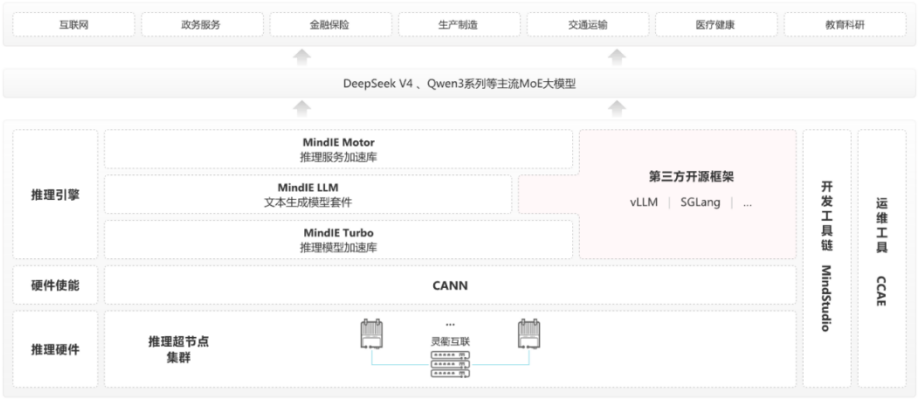


资料来源：DeepSeek《Insights into DeepSeek-V3:Scaling Challenges and Reflections on Hardware for AI Architectures》，国联民生证券研究所

模型能力向实际 Token 产能的转化，还取决于模型架构与底层算力平台的适配。华为昇腾大规模专家并行集群推理解决方案基于昇腾 AI 基础软硬件构建，深度适配 DeepSeek-V3/R1 及主流 MoE 模型，并以相同算力下更大吞吐、更高并发和更低时延为目标；在中心训推算力平台中，华为进一步将 DeepSeek、Qwen 等 MoE 模型与大规模专家并行、PD 分离等推理机制结合，官方披露吞吐可提升 3 倍以上。2025 年华为全联接大会上，华为还披露，面向 MoE 推理场景，大规模专家并行方案相较常规服务器堆叠可实现约 4 倍的单卡吞吐提升，并降低 50%的系统时延。我们认为，这表明模型结构、并行策略与底层软硬件协同共同影响算力向 Token 吞吐的转化效率，评价 Token 供给能力不宜仅依据单颗芯片的峰值 FLOPS。

下载日志已记录，仅供内部参考

图5：昇腾大规模专家并行推理方案适配 DeepSeek V4、Qwen3 系列等主流 MoE 大模型

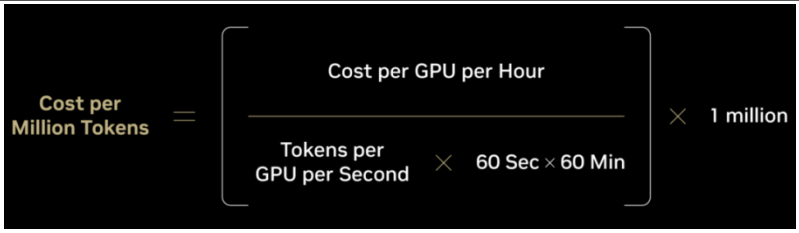


资料来源：华为企业业务官网，国联民生证券研究所

1.4 Token Economics：从算力投入到单位 Token 经济性

NVIDIA 在 AI 推理基础设施的 TCO 评价中，进一步从计算资源投入关注实际 Token 产出。NVIDIA 指出，仅比较 Compute Cost 或 FLOPS per Dollar 并不能完整反映推理 TCO：前者反映企业购买或租赁 AI 基础设施的成本，后者衡量单位投入对应的理论计算能力，而实际计算能力与真实 Token 产出并不等同，因此 NVIDIA 进一步将 Cost per Token 作为衡量推理经济性的重要指标。NVIDIA 给出的计算框架为：Cost per Million Tokens = Cost per GPU per Hour÷（Tokens per GPU per Second×3,600）×1,000,000；其中云部署的 GPU 小时成本对应云服务商的小时费率，自建部署则可采用自有基础设施摊销后的有效小时成本，而实际 Token 吞吐位于公式分母。我们认为，该框架意味着推理经济性不仅取决于单位算力成本，也取决于单位时间内实际交付的 Token 数量，因此评价推理基础设施时需要同时观察资源成本与实际 Token 吞吐。

图6：单位 Token 成本由 GPU 小时成本与实际 Token 吞吐共同决定

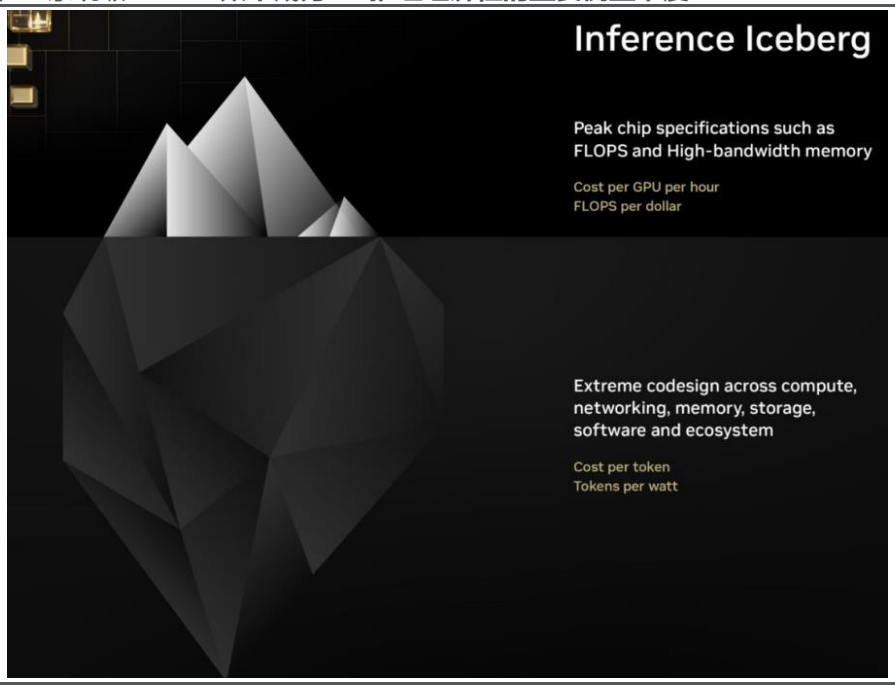


资料来源：NVIDIA Blog，国联民生证券研究所

下载日志已记录，仅供内部参考

Cost per Token 能够进一步反映软硬件协同后的实际 Token 产出效率。以 DeepSeek-R1 为例，NVIDIA 基于自身分析及 SemiAnalysis InferenceX v2 数据披露，HGX H200 与 GB300 NVL72 的 GPU 小时成本分别约为 1.41 美元和 2.65 美元，后者约为前者的 2 倍；但对应 Tokens per Second per GPU 由约 90 提升至 6,000，Tokens per Second per MW 由约 5.4 万提升至 280 万，单位百万 Token 成本则由约 4.20 美元下降至 0.12 美元。NVIDIA 同时指出，实际 Token 产出受到计算、网络、内存、存储及软件等全栈优化的共同影响。我们认为，上述 DeepSeek-R1 案例表明，硬件投入与实际 Token 产出之间并非简单对应关系，评价推理经济性还需要结合系统实际交付的 Token 吞吐与单位 Token 成本。

图7：系统级 Token 效率成为 AI 推理经济性的重要衡量维度

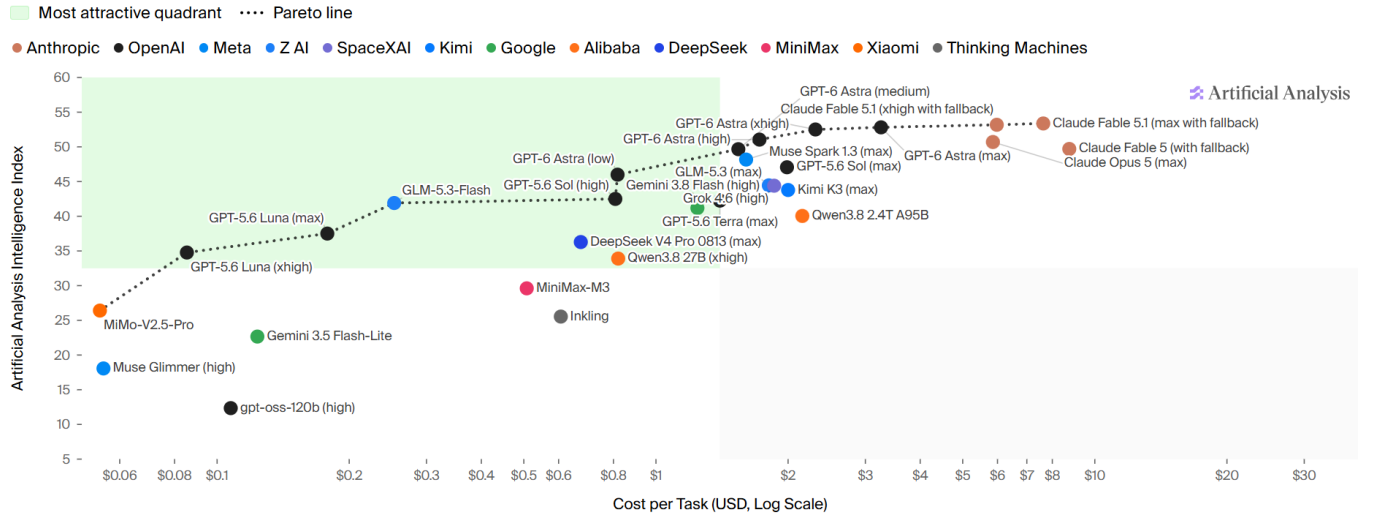


资料来源：NVIDIA Blog，国联民生证券研究所

模型能力与推理成本并非简单对应，单独比较 API 的 Token 价格难以完整反映模型完成任务的实际成本。Artificial Analysis 于 2026 年 9 月发布的 Intelligence Index v4.3 显示，GPT-6 Astra（max）与 Claude Fable5.1（max with fallback）的 Intelligence Index 均为 53 分，但平均 Cost per Intelligence Index Task 分别约为 3.26 美元和 7.63 美元；GLM-5.3-Flash 与 GPT-5.6Terra（max）均为 42 分，对应任务成本则分别约为 0.25 美元和 1.40 美元。Artificial Analysis 的 Cost per Intelligence Index Task 按照各项评测实际涉及的输入、缓存命中、缓存写入、Reasoning 和回答 Token 价格计算，并按任务数量及 Intelligence Index 权重进行加权，因此相近的模型能力得分可以对应不同的任务成本。

下载日志已记录，仅供内部参考

图8：相近智能水平的模型可以对应明显不同的单任务成本



资料来源：Artificial Analysis，国联民生证券研究所

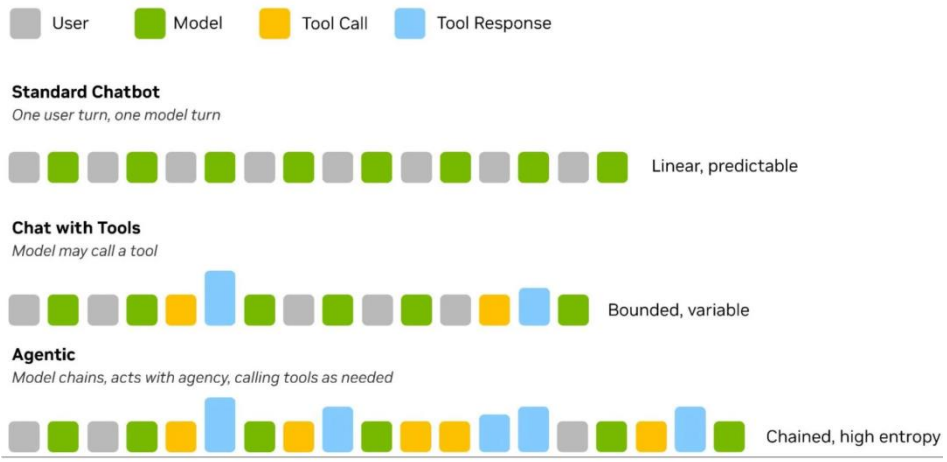
下载日志已记录，仅供内部参考

2 Token 优化：AI 规模化应用的效率基础

2.1 Agent 推动 Token 需求从规模增长走向密度提升

Agentic 工作负载正在成为 Token 需求增长的重要来源。OpenRouter 基于 2026 年 1 月 1 日至 6 月 14 日超过 450 万亿 Token 的真实请求日志，将工作负载划分为 Agentic、Mixed 和 Human 三类，并显示 Agentic 工作负载的 Token 使用量约在 2026 年 2 月初超过 Human 工作负载。Agent 不仅增加调用次数，也显著提高单次请求的 Token 密度，根据 OpenRouter 披露，Agentic 工作负载平均每次请求消耗的 Token 约为普通 Human 工作负载的 15 倍。NVIDIA 指出，与传统“人类发送请求、模型进行响应”的生成式 AI 交互不同，Agent 并不遵循预先确定的动作序列，而是可以调用工具、生成子 Agent、在 Memory 中保留信息并管理自身的 Context Window。

图9：从 Chatbot 到 Agent，模型交互由线性请求演进为多步骤调用链

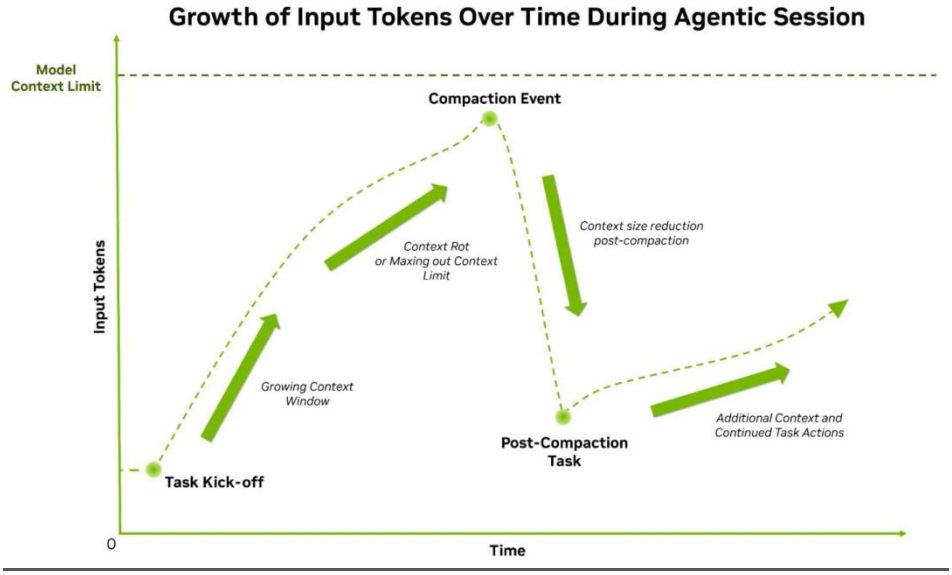


资料来源：NVIDIA Developer Blog，国联民生证券研究所

长周期 Agent 任务进一步放大上下文累积带来的 Token 消耗。NVIDIA 记录的一次真实 Claude Code 编程任务持续约 33 分钟，共发生 283 次模型请求，其中包括 58 次主 Agent 请求和 225 次子 Agent 请求。在该任务中，主 Agent 上下文由约 1.5 万 Token 逐步增长至 15.6 万 Token，在触发 Context Compaction 后降至约 2 万 Token，随后随着任务继续执行重新累积。主 Agent 会在连续推理过程中快速累积输入上下文，而上下文压缩和子 Agent 分工能够控制长周期任务中的输入 Token 处理规模。

下载日志已记录，仅供内部参考

图10：真实 Agent 会话中的 Input Sequence Length 随请求持续变化



资料来源：NVIDIA Developer Blog，国联民生证券研究所

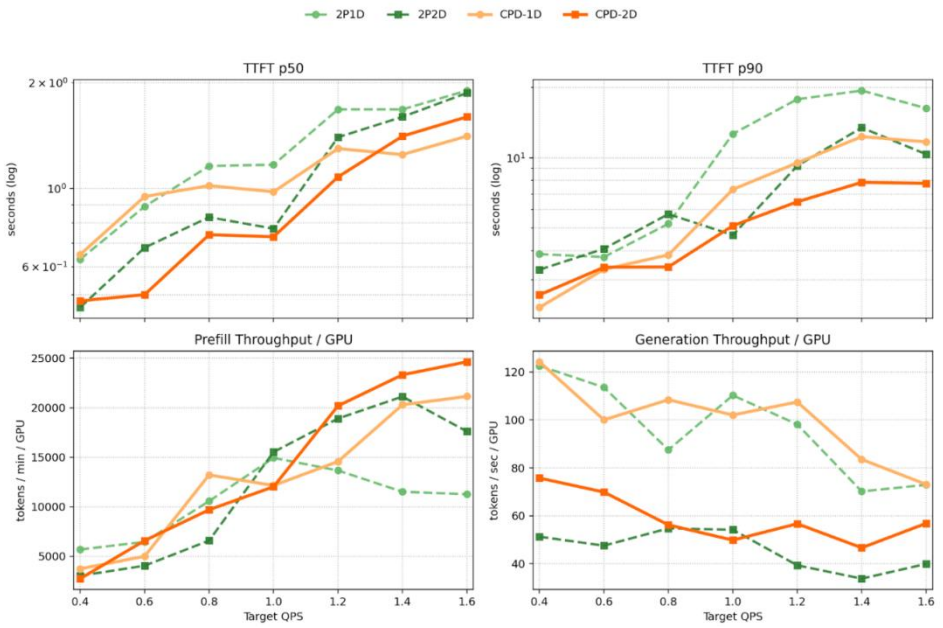
2.2 Token 生产瓶颈：计算、显存、缓存、调度与延迟

大模型推理中的 Prefill 与 Decode 具有不同的资源特征。NVIDIA 指出，在 Aggregated Serving 中，Prefill 与 Decode 运行于同一 GPU；其中 Prefill 通常以密集矩阵计算为主，更偏向 Compute-bound，而 Decode 需要加载 KV Cache，更偏向 Memory-bound。在 SGLang 默认调度策略下，Prefill 与 Decode 阶段严格串行执行，可能导致 Decode 阶段 GPU 计算单元利用不足，同时 Prefill 阶段的内存带宽利用不足。

长上下文与冷热请求混合进一步放大了 Prefill 侧的排队压力。Together AI 指出，多轮对话、Coding Copilot、Agent Memory 和 RAG 等应用正在推动长 Prompt 成为常态；在高负载环境中，大规模 Cold Prompt 需要执行完整 Prefill 并可能占用 Prefill 资源数秒，使原本可以通过 Cache 复用快速处理的 Warm Request 也需要在同一队列等待。在 Together AI 基于 Coding Agent 场景的测试中，传统 2P1D 架构的实际 QPS 在 0.75—0.8QPS/GPU 附近开始趋于饱和，此后排队延迟快速增加。

下载日志已记录，仅供内部参考

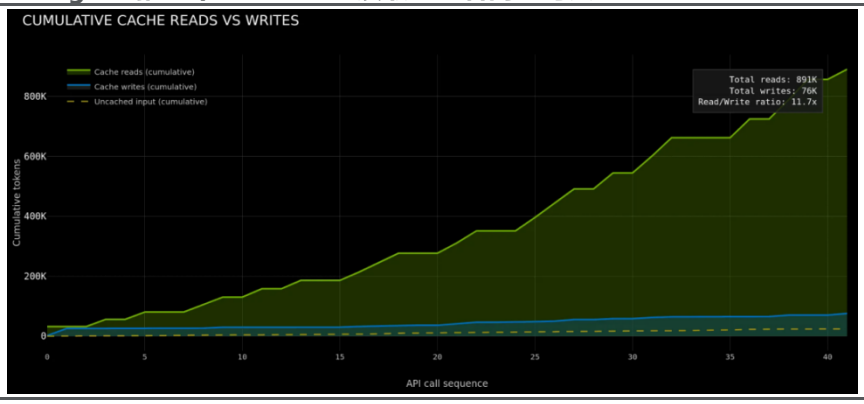
图11：长上下文混合负载下，Prefill 容量饱和将快速推高 TTFT



资料来源：Together AI，国联民生证券研究所

Agent 推理呈现出显著的 KV Cache 重复复用特征。NVIDIA 以 Claude Code 为例指出，首次 API 调用将 Conversation Prefix 写入 KV Cache 后，后续请求在相同 Worker 上的 Cache Hit Rate 达到 85%—97%；在 4 个 Opus Teammates 组成的 Agent 团队中，整体 Cache Hit Rate 达到 97.2%。同一数据中，KV Cache 的 Read/Write Ratio 达到 11.7 倍，即系统每写入一份 Cache 数据，后续需要进行接近 12 次读取。NVIDIA 将这一访问特征概括为“Write-once-read-many”，并指出提高跨 Worker 的 Cache 复用率、保持高价值 KV Block 可复用，是 Agent 推理的重要优化目标。

图12：Agent 推理中 KV Cache 读取量显著高于写入量



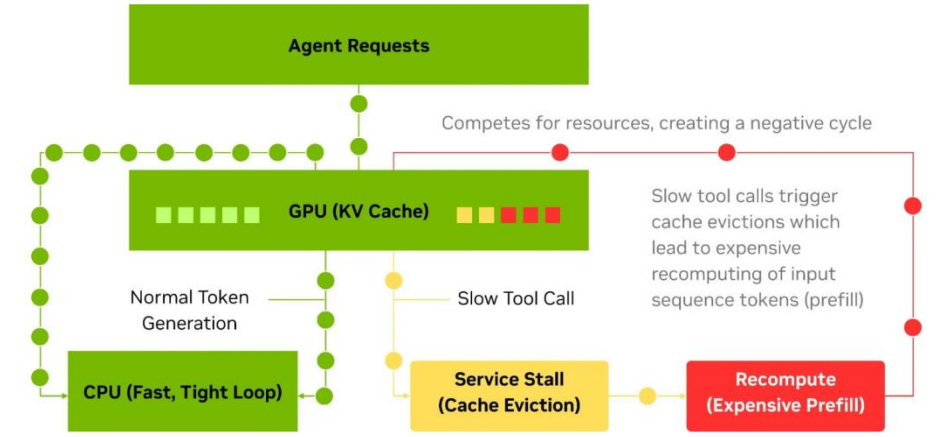
资料来源：NVIDIA Developer Blog，国联民生证券研究所

工具调用进一步增加了 Agent 推理中的缓存管理难度。在高流量环境中，Agent 等待外部工具返回的 2—30 秒间隔可能使原有 KV Block 在默认 LRU 机制下淘汰；当 Agent 恢复执行时，此前的完整 Prefix 可能需要重新计算。NVIDIA

下载日志已记录，仅供内部参考

因此提出，Agent 推理中的 Cache 管理需要进一步区分不同 KV Block 的复用价值，并控制其保留位置和保留时间。我们认为，Agent 工作负载正在使推理瓶颈由单纯计算能力约束，扩展为计算、显存、缓存与调度共同作用的系统问题，因此 Token 优化也需要由单点加速进一步演进为覆盖完整 Serving 链路的系统工程。

图13：较慢的工具调用可能触发 KV Cache 淘汰并导致高成本重计算



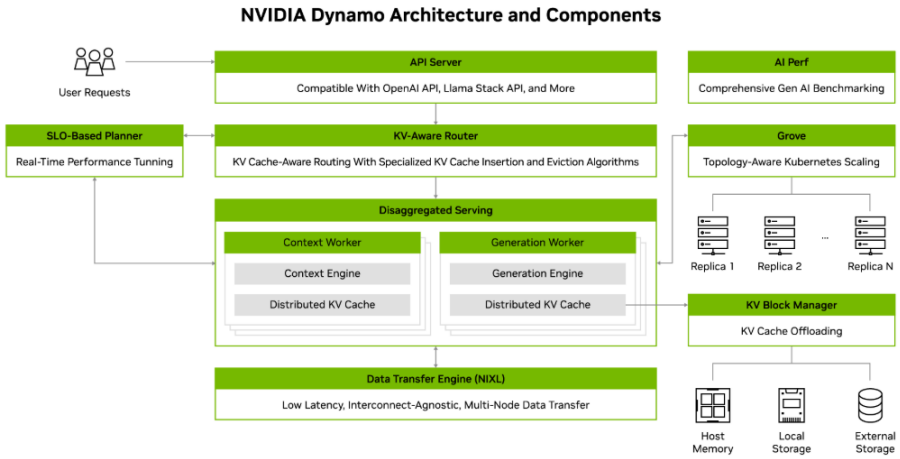
资料来源：NVIDIA Developer Blog，国联民生证券研究所

2.3 Token 优化：从单点加速走向全栈推理系统

Token 优化贯穿 KV Cache 管理、请求调度、算子执行与分布式并行等多个环节。vLLM 通过 PagedAttention 提升 KV Cache 内存管理效率，并支持 Continuous Batching、Chunked Prefill、Prefix Caching、量化、优化 Attention/GEMM/MoE Kernel、Speculative Decoding，以及 Tensor、Pipeline、Data、Expert 和 Context Parallelism 等能力；同时支持将 Prefill 与 Decode 分离部署，通过为两个阶段配置不同的并行策略，分别优化首 Token 时延与 Token 间时延。在分布式推理服务层面 NVIDIA Dynamo 进一步从集群层面协调推理资源。其 SLO Planner 根据容量与 Prefill 负载动态调整 GPU 资源，KV-aware Router 根据 KV Cache 分布进行请求路由，KV Block Manager 支持 KV Cache 在不同内存层级间迁移，NIXL 则用于加速 GPU 及异构内存、存储之间的数据的传输。

下载日志已记录，仅供内部参考

图14：NVIDIA Dynamo 分布式推理架构：SLO 调度、KV 感知路由、PD 分离与 NIXL 数据的传输



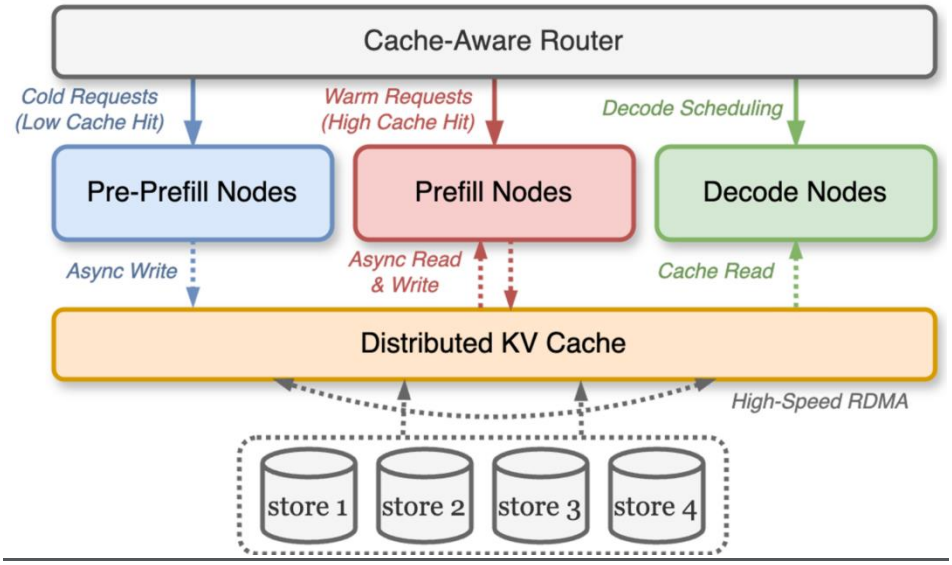
资料来源：NVIDIA Developer Blog，国联民生证券研究所

底层推理优化通过改进计算执行、请求批处理与多 GPU 并行，提高硬件资源利用效率。FlashAttention-2 通过减少非矩阵乘法运算、改进并行方式及工作分配，提高注意力计算效率。NVIDIA 指出，权重量化能够降低模型的内存占用，进一步量化中间激活值，则可以利用低位宽下吞吐能力更高的专用 Tensor Core 硬件加快推理。在推测解码中，轻量级草稿机制快速提出多个候选 Token，目标模型在一次前向计算中对其进行验证。Continuous Batching 在每个生成步骤动态调整批次，已有请求完成后，新请求立即加入，无需等待整个批次完成。Chunked Prefill 将较长的预填充任务拆分为小块，与解码请求共同组批，通过平衡计算密集的预填充与内存密集的解码操作，改善吞吐和延迟。多 GPU 部署中，张量并行在模型层内拆分参数，流水线并行将模型层分布到多个 GPU，专家并行将不同专家网络分布到多个 GPU，数据并行则复制整个模型，并行处理不同请求批次。

缓存感知的 PD 分离通过减少重复的输入 Token 计算，优化长上下文请求的 Token 生产效率。Together AI 在常规 Prefill 与 Decode 分离的基础上，按照上下文缓存命中情况分流请求：低复用的冷请求由 Pre-Prefill 节点计算新上下文，并将生成的 KV Cache 写入分布式缓存；高复用的热请求由 Prefill 节点读取已有 KV Cache，跳过可复用前缀的重复计算。Decode 节点保持以延迟为中心，并隔离 Prefill 干扰。在 Together AI 披露的冷、热长上下文混合负载测试中，该架构相较常规 PD 分离方案将每 GPU 可持续处理的请求量提高约 35%—40%；这一结果反映了特定负载下缓存复用和请求分流带来的系统收益。

下载日志已记录，仅供内部参考

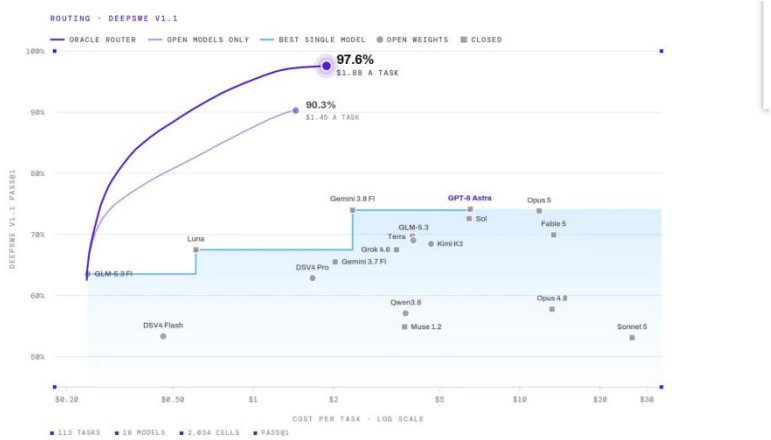
图15：Cache-aware P/D 分离分离通过请求分流与 KV Cache 复用减少重复 Token 计算



资料来源：Together AI，国联民生证券研究所

智能路由则进一步在模型层面对单任务成本进行优化。Fireworks 指出，智能路由的基本思路是将简单任务交给成本效率更高的模型，而将高难度任务保留给能力更强的模型，从而避免所有任务均使用高成本模型。Fireworks Nexus 中的 Fire Router 会对每个请求的任务难度进行评分，将常规任务路由至 Fireworks 提供的高成本效率开放模型，将困难任务转发至原有模型供应商，并根据任务复杂度和缓存优化动态进行路由；Fireworks 称，该机制通常可带来约 3—5 倍的成本降低。Fireworks 披露的 Factory 案例显示，其客户任务大致呈现三等分：约三分之一的任务需要最新 Frontier Model，约三分之一属于可使用成本效率最高模型的常规任务，其余约三分之一介于两者之间；Factory Router 通过将每项任务匹配至能够达到可靠性门槛的成本效率最优模型，使平均单任务成本降低约 30%—40%。

图16：智能路由相较单一模型实现更优任务成本效率

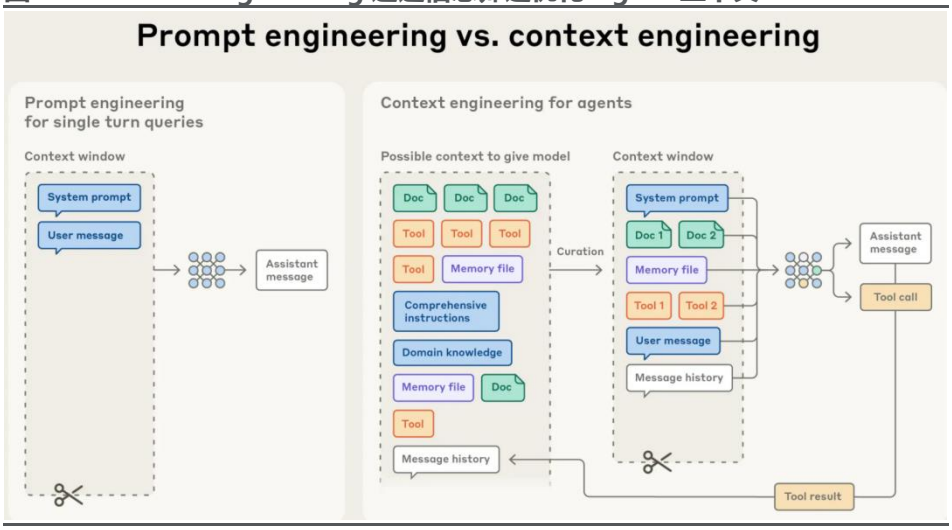


资料来源：Fireworks，国联民生证券研究所

下载日志已记录，仅供内部参考

Token 优化的边界还在进一步由“如何更高效计算 Token”扩展至“哪些 Token 值得进入上下文”。Anthropic 将 Context Engineering 定义为在模型推理过程中持续筛选和维护最优 Token 集合，并指出在有限的 Attention Budget 下，应尽可能保留能够支撑目标结果的最小高信息量 Token 集合。对长周期 Agent，Anthropic 进一步提出 Compaction、Structured Note-taking 与子代理架构等机制：Compaction 通过压缩接近上限的历史上下文并保留关键内容，Structured Note-taking 将任务状态持久化至上下文窗口之外，而子代理架构则通过将聚焦任务分配给具有独立上下文窗口的专业子代理，由主代理负责高层计划协调，子代理进行深入技术工作或信息搜索，并仅返回浓缩、提炼后的工作总结，从而将详细搜索上下文与主代理的综合分析过程相互隔离。OpenAI 同样已在 Responses API 中加入 Native Compaction，当上下文持续增长时，将此前关键状态压缩为更节省 Token 的表示并带入后续上下文。由此，Token 效率开始由底层计算效率进一步延伸至上下文效率，即完成单项任务究竟需要在 Context Window 中持续保留和处理多少有效 Token。

图17：Context Engineering 通过信息筛选优化 Agent 上下文



资料来源：Anthropic，国联民生证券研究所

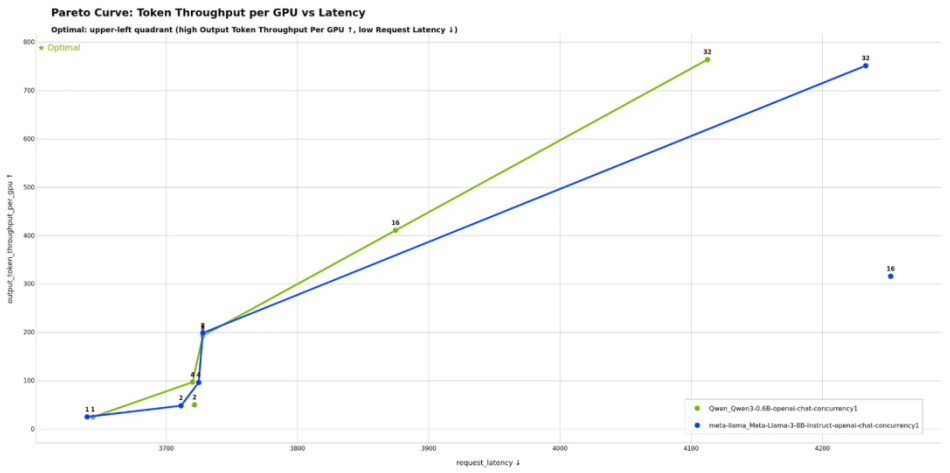
2.4 Token 优化如何衡量：从系统性能到成功任务成本

Token 优化首先需要通过资源利用与延迟指标识别推理系统中的效率瓶颈。NVIDIA AIPerf 指标包括 TTFT、TTFO、ITL 和 Goodput 等，从响应速度、生成速度与有效吞吐等不同维度评价推理服务性能。其中，TTFT 衡量请求发出至收到首个 Token 的时间，并包含网络、排队、Prompt 处理和首 Token 生成过程；ITL 则衡量连续输出 Token 之间的时间间隔。因此，不同指标分别反映请求排队、Prefill、Decode 和缓存等不同环节的运行状态，单独观察峰值 Tokens/s 难以完整描述生产级推理性能。

下载日志已记录，仅供内部参考

生产环境中的评价重点进一步由峰值 Throughput 转向满足服务质量约束的 Goodput。NVIDIA AIPerf 将 Goodput 定义为满足用户设定 Service Level Objective（SLO）的请求吞吐率，只有满足全部 SLO 约束的请求才计入 Good Request Count。与此同时，AIPerf 通过 Token Through put per GPU 与 Latency 的 Pareto Curve 展示推理系统中 GPU 效率与延迟之间的权衡，不同配置在更高单位 GPU 吞吐和更低 Latency 之间形成不同效率前沿。我们认为，相较于单纯追求更高峰值 Throughput，生产级推理更应关注在既定服务质量约束下能够实现的有效吞吐。

图18：单位 GPU Token 吞吐与推理延迟形成 Pareto 效率前沿



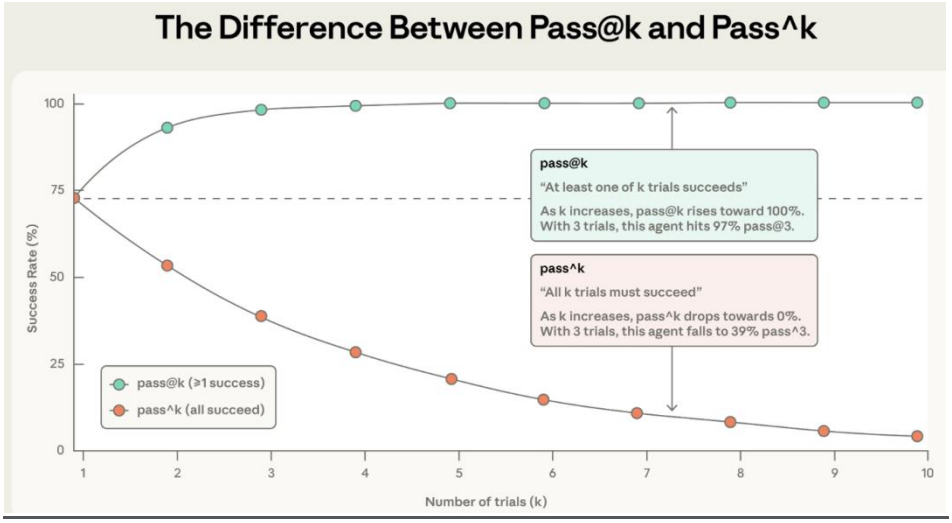
资料来源：NVIDIA AIPerf，国联民生证券研究所

当评价对象由单次模型调用进一步扩展至完整任务时，AI 经济性还需要同时考虑任务是否成功。OpenAI 指出，仅观察 Cost per Token 不足以衡量 AI 投入产出，因为更低价的模型虽然 Token 更便宜，但可能需要更多尝试、更长时间或更多人工审核；而能力更强、单 Token 价格更高的模型，则可能一次即可完成同样任务。因此，更关键的指标是 Cost per Successful Task。OpenAI 指出，模型层面的成功任务成本取决于模型价格、实际使用的计算量以及得到正确结果的概率；对于企业而言，完整成本还需要进一步计入员工时间、人工审核、重试和返工。OpenAI 据此提出，应将完成工作的全部成本除以达到质量要求的成功任务数量，从而衡量单个成功任务的实际成本，因此更低的 Cost per Token 并不必然对应更低的 Cost per Successful Task。

Agent 任务还需要通过重复执行衡量可靠性。Anthropic 指出，Agent 行为在不同运行之间存在差异，同一 Task 在一次 Eval 中成功、下一次执行时仍可能失败，因此单次运行结果不能完整反映任务可靠性。Anthropic 进一步区分 pass@k 与 pass^k：前者衡量 k 次尝试中至少成功一次的概率，后者衡量 k 次执行全部成功的概率，分别反映 Agent 获得成功解的能力和重复执行的一致性。

下载日志已记录，仅供内部参考

图19：Agent 任务成功率需同时考虑单次能力与重复执行可靠性



资料来源：Anthropic，国联民生证券研究所

我们认为，Token 优化的评价体系正在由底层系统性能逐步延伸至服务质量和任务经济性：GPU 利用率、KV Cache 及 Prefill/Decode 耗时用于识别底层资源瓶颈，TTFT、TPOT 与 Goodput 用于评价生产 Serving 能力，Cost per Token 进一步刻画 Token 经济性，而进入 Agent 阶段后，还需要结合 Task Success Rate 与 Cost per Successful Task 评价完整任务的资源投入与最终有效产出。Token 优化的最终目标也由“更快地产生 Token”进一步转向“在质量、延迟和可靠性约束下，以更低成本稳定完成更多成功任务”。

2.5 Token 优化走向平台化：推理引擎与 Inference Platform 成为独立中间层

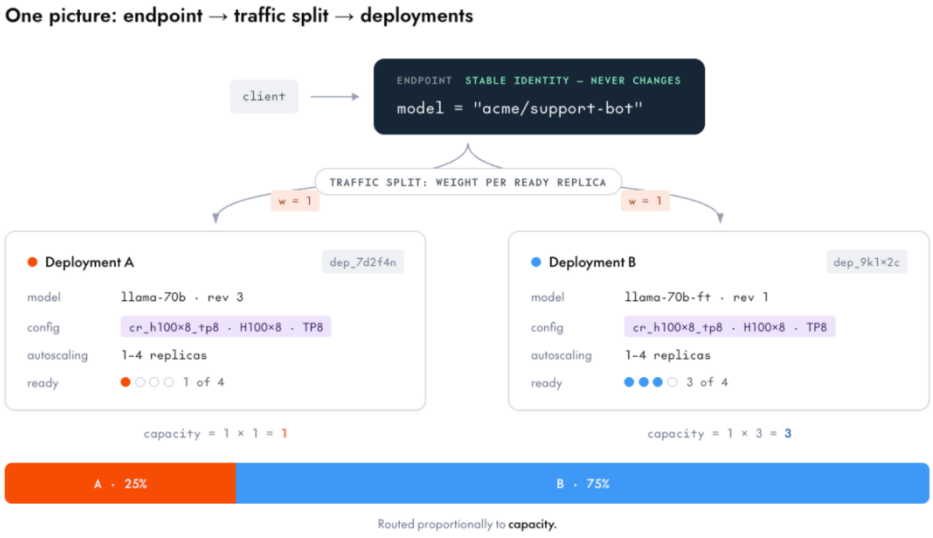
随着推理优化涉及的技术环节增加，推理引擎开始将多项底层优化能力集成到统一 Serving 体系中。vLLM 目前集成 Paged Attention、Continuous Batching、Chunked Prefill、Prefix Caching、Speculative Decoding 以及多种并行方式，并支持 Prefill 与 Decode 解耦部署。由此，应用开发者可以通过统一推理引擎使用缓存、调度、并行和模型执行优化，而无需分别实现上述底层能力。

Inference Platform 进一步将模型运行配置、部署实例与应用调用接口分层管理。Together AI 的 Dedicated Model Inference 由 Config、Deployment 和 Endpoint 三类对象组成：Config 用于指定推理引擎、GPU 类型及数量、并行方式，以及 Latency、Throughput 或 Balanced 等优化目标；Deployment 将特定模型版本与 Config 绑定，并配置 Auto scaling Policy、运行具体 Replica；Endpoint 则作为应用调用时保持不变的固定接口，通过 Traffic Split 将请求路由至一个或多个 Deployment。Together AI 通过 capacity-aware traffic 连接这三

下载日志已记录，仅供内部参考

类对象，在这一架构下，平台还可支持 Roll out、A/B Test、Shadow Experiment 和 Zero-downtime Change 等生产操作。

图20：Together AI 通过 Endpoint—Deployment—Config 架构解耦应用接口与底层推理资源



资料来源：Together AI，国联民生证券研究所

国内推理平台也在通过算子、推理与模型三个层级优化，将异构算力进一步转化为高性能模型服务。硅基流动在某大型央企航空集团项目中，将推理加速框架概括为三层技术突破：算子优化侧，对多款主流国产芯片进行适配与优化，并支持多品牌、多架构算力资源的统一纳管与智能调度；推理优化侧，通过集群化大 EP 与 PD 分离等方式优化 MoE 模型的执行效率、显存利用效率和分布式通信效率；模型优化侧，则基于公有云 MaaS 平台的模型适配经验，从模型、机制、框架和算子等层面进行联合优化。图 21 进一步展示了硅基流动高性能推理引擎由底层多元算力资源，向上经过资源优化、执行优化、通信优化、精度优化、缓存优化和调度优化，并进一步形成 Dense 模型优化、MoE 模型优化以及统一模型 Serving，最终连接 Agent、RAG 等上层应用的分层架构。

下载日志已记录，仅供内部参考

图21：硅基流动通过分层推理架构连接异构算力、模型优化与上层 AI 应用



资料来源：硅基流动官网，国联民生证券研究所

下载日志已记录，仅供内部参考

3 Token 到 Task：AI 价值从智能产出走向任务结果

3.1 从 Token 到 Task：AI 交付与计价单位向业务结果迁移

大模型服务首先可以按照 Token 使用量进行计价。OpenAI API 按照每百万 Token 列示模型价格，并分别区分 Input、Cached Input 和 Output 等 Token 类型，使模型调用量能够对应标准化的服务使用成本。我们认为，Token 由此构成模型服务的重要计量单位，但进入企业 Agent 场景后，客户最终需要获得的并非 Token 本身，而是 Token 进一步驱动的业务动作和任务结果。

Agent 进入业务流程后，计量单位进一步由 Token 向具体 Action 延伸。Salesforce 将 Flex Credits 定义为 Agentforce 的使用计量方式，并明确按照 Action 进行计量；其中 Action 是 AI Agent 在平台上执行的一项具体功能，例如更新记录、总结复杂案例、回答产品问题或执行自定义 Prompt 和 Flow。Salesforce 同时明确，每个 Action 均单独计量，标准 Agentforce Action 消耗 20 个 Flex Credits。

Salesforce 的业务案例显示，一项完整业务请求可以包含多个 Action。在“Where is My Order”示例中，Agent 依次执行“Authenticate Customer from Customer Email”、“Retrieve Customer Order Details”两步，因此该 Use Case 对应 2 个 Action、40 个 Flex Credits、0.20 美元。与 Token 相比，Action 已经开始将 AI 使用量与具体业务执行步骤建立更直接的对应关系。

图22：Agentforce 将完整业务流程拆解为可计量的 Agent Action



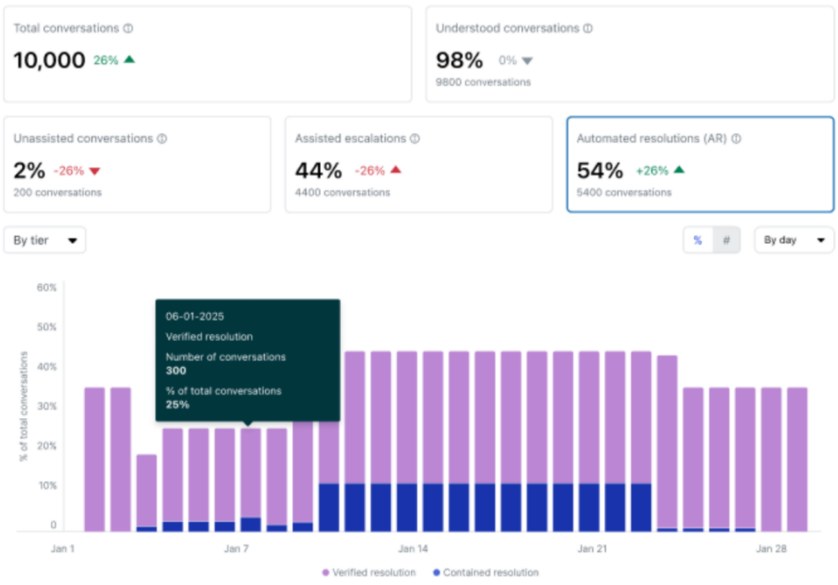
资料来源：Salesforce，国联民生证券研究所

应用层的计量对象还在进一步由业务动作向可验证的任务结果延伸。Zendesk 在 2026 年扩展 Outcome-based Pricing，并明确提出，其收费对象是

下载日志已记录，仅供内部参考

能够被验证的 Resolution：交互需要由 AI Agent 端到端完成，同时由独立的 AI Evaluation Model 进行确认，客户仅为真正得到确认解决的 Outcome 付费。在结果评价层面，Zendesk 进一步通过 Resolution Tier 区分不同任务处理状态，其中 Contained Resolution 表示 AI Agent 将交互处理至结束且客户没有进一步请求帮助，Verified Resolution 则表示 AI Agent 成功解决该次交互。我们认为，相较于 Token 和 Action，Resolution 已经更加接近客户最终关心的“业务问题是否得到解决”。

图23：Zendesk 通过 Contained 与 Verified Resolution 衡量 AI Agent 任务完成情况



资料来源：Zendesk，国联民生证券研究所

我们认为，Token、Action 和 Outcome 并不是相互替代的计量单位，而是对应 AI 价值传导过程中的不同环节：Token 衡量模型服务的使用量，Action 衡量 Agent 执行的具体业务步骤，Resolution/Outcome 则进一步衡量业务问题是否得到解决。随着 AI 由模型调用进一步进入企业任务执行，价值衡量对象也由模型生成量逐步接近最终业务结果。

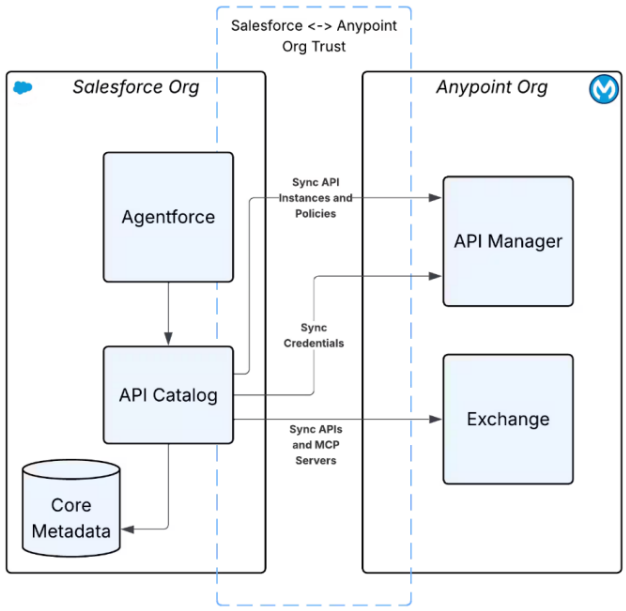
3.2 企业级 Task 壁垒：数据、业务流程与真实反馈

当 AI 由生成内容进入企业任务执行，竞争壁垒进一步由模型能力延伸至企业数据、业务语义、权限体系和工作流。Salesforce Agentforce 通过 Data360 连接 Salesforce 及外部系统中的结构化和非结构化数据，并结合 Salesforce Metadata 为 Agent 提供业务上下文，使 Agent 的 Action 能够与企业自身的业务逻辑保持一致。同时，Salesforce 通过 Mule Soft Agent Fabric、Flow 以及 MCP、A2A 等开放协议编排连接企业应用、API 和其他 Agent 的多步骤工作流，

下载日志已记录，仅供内部参考

使 Agent 能够进一步跨系统执行任务。我们认为，企业 Agent 的生产部署因此不仅依赖模型能力，也需要与企业数据、业务系统和既有治理体系深度结合。

图24：企业 Agent 任务执行依赖数据、API 与权限治理体系



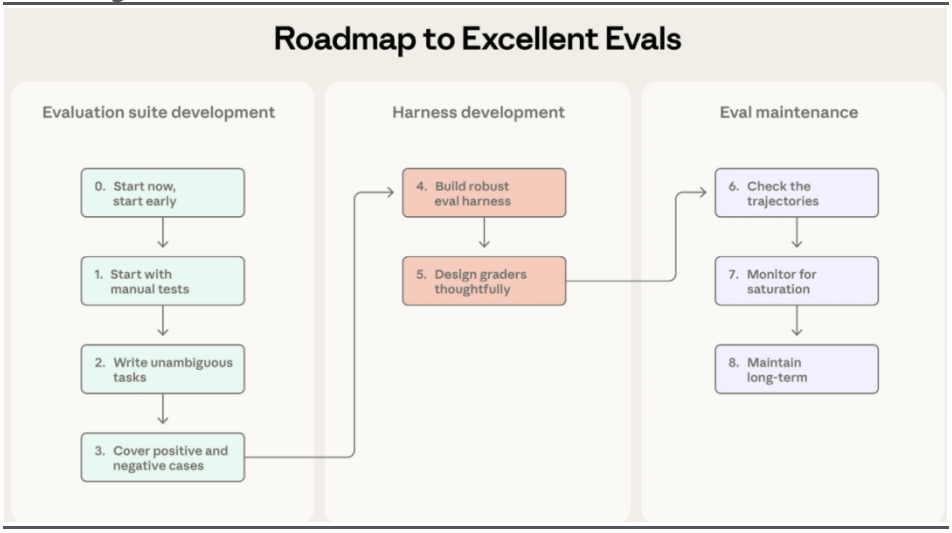
资料来源：Salesforce，国联民生证券研究所

在获得企业业务上下文后，Agent 还需要进一步连接实际业务系统，并将模型推理转化为跨系统业务执行。Salesforce 指出，Agentforce 可以编排连接应用、API 和其他 Agent 的多步骤 Workflow。Agentforce Gateway 则作为集中式治理层，对 Agent force 发起的外部流量实施策略控制，包括工具限制、基于属性的访问控制和配额限制，并管理身份验证与授权。我们认为，数据决定 Agent “知道什么”，工作流和工具决定 Agent “能够做什么”，而权限治理进一步决定 Agent “被允许做什么”，三者共同构成企业 Task 进入生产环境的基础。

企业任务执行还需要通过真实评估与生产反馈形成持续优化闭环。Anthropic 指出，Evals 可与 Production Monitoring、A/B Tests、User Research 等方法结合，为 Agent 在规模化应用过程中的持续改进提供反馈信号。对于已经进入生产环境的 Agent，可进一步从 Bug Tracker 和 Support Queue 中提取用户报告的真实失败案例，并将其转化为 Test Case，使 Eval Suite 更贴近实际使用；随着评估体系逐步建立，企业还可基于固定任务形成基准与回归测试，持续跟踪 Agent 的时延、Token 使用量、单任务成本及错误率等指标。我们认为，企业级 Agent 的应用壁垒不仅在于接入数据与业务系统的能力，也在于能否持续积累真实任务反馈，并将其转化为更稳定的任务执行能力。

下载日志已记录，仅供内部参考

图25：Agent 通过任务评价与生产反馈形成持续优化闭环



资料来源：Anthropic，国联民生证券研究所

3.3 市场空间与价值迁移：Token 需求扩张，Task 成为重要价值锚

Token 调用量正在快速扩张，并成为观察 AI 应用规模化落地的重要量化指标。国家数据局披露，我国日均 Token 调用量由 2024 年初约 1,000 亿增长至 2025 年底 100 万亿，2026 年 3 月进一步突破 140 万亿，两年增长超过 1,000 倍；《全国数据资源调查报告（2025 年）》显示，2025 年我国全年 Token 调用量约为 21,100 万亿。与此同时，2025 年用于人工智能训练和推理的数据总量达到 199.48EB，其中推理数据量达到 101.34EB，首次超过训练数据量 98.14EB。国家数据局专家解读指出，推理数据超过训练反映 AI 开始进入推理阶段并嵌入真实业务场景。

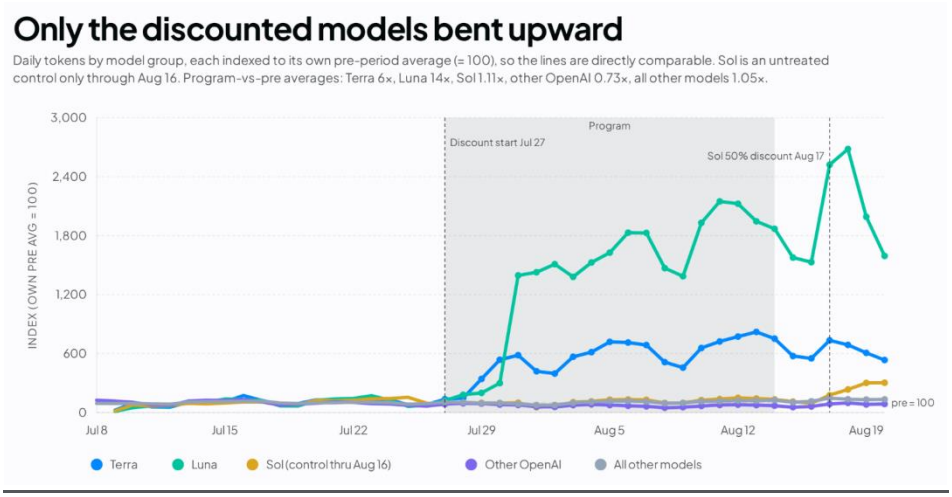
Agent 进一步使 AI 市场空间的衡量由 Token 调用规模向成功任务数量延伸。单项 Task 通常包含多次模型调用和多个业务 Action，不同任务对应的 Tokens per Task 存在差异，同样规模的 Token 投入也会因 Task Success Rate 不同形成不同数量的成功任务。我们认为，Token 调用量主要描述底层智能计算需求，而应用层市场空间还需要进一步观察每项任务消耗多少 Token、任务能否成功完成，以及成功 Task 最终能够创造多少业务价值。

单位 Token 价格下降并不必然意味着 AI 产业价值同步下降，较低价格可能伴随 Token 使用需求进一步释放。OpenRouter 对 2026 年 7 月 27 日至 8 月 14 日 GPT-5.6 Terra 和 Luna 的折扣活动进行复盘，并以同期维持原价的 Sol 作为阶段性参照；活动期间，Terra 日均 Token 使用量提升至活动前的 5.6 倍，Luna 升至 13.8 倍，而按原价销售的 Sol 模型仅小幅增长至 1.1 倍。7 月 30 日 OpenAI

下载日志已记录，仅供内部参考

进一步调整模型标价后，叠加 OpenRouter 折扣，Luna 和 Terra 相对活动前价格的有效折扣分别达到约 90%和 60%；同期 Terra 和 Luna 合计 Token 份额由活动前的 0.7%提升至 7.8%，OpenAI 全系模型 Token 份额由 7.1%提升至 12.4%。

图26：OpenRouter 折扣实验显示模型价格下降后 Token 调用量显著提升



资料来源：OpenRouter，国联民生证券研究所

我们认为，单位 Token 价格下降与 AI 产业价值增长并不构成简单矛盾。一方面，更低的单位 Token 价格可以降低模型和 Agent 的使用成本，并可能释放更多 Token 调用需求；另一方面，Agent 执行完整 Task 通常需要多轮模型调用和多个业务 Action，使 AI 应用的价值衡量进一步由单次模型调用转向完整任务。随着 AI 产业由 Token 生产逐步走向 Task 执行，核心变量也将由“每个 Token 多少钱”进一步延伸至“每项 Task 需要多少 Token、能够完成多少成功 Task，以及成功 Task 能够创造多少业务价值”，Task 由此成为连接底层 Token 生产与上层企业价值的重要价值锚。

下载日志已记录，仅供内部参考

4 投资建议

4.1 行业投资建议

我们认为，Token 需求扩张与企业 Agent 落地构成主要投资线索。底层看可持续交付的 Token 吞吐和单位 Token 成本，中间看模型服务及推理平台的有效调用与运营效率，应用层看企业任务成功率、客户持续使用和实际业务价值。建议围绕模型 API 服务、Token 运营平台与企业任务执行筛选标的，避免仅按算力规模或名义 Token 单价判断经营成果。

4.2 重点公司

围绕“Token 生产—Token 优化—Task 执行”的产业链框架，我们将重点公司划分为大模型、电信运营商、云服务、算力租赁与 Token 工厂运营、Token 优化及垂直行业 Token 应用六类，分别对应模型侧 Token 生成、Token 分发、模型服务承载、Token 产能运营、推理效率优化及 Task 执行与业务价值变现六个环节。随着产业链由算力建设向 Token 生产、运营、优化及 Task 执行延伸，后续应重点跟踪各环节实际 Token 调用量、单位成本、客户持续使用及任务结果等指标。

表2：部分上市公司 Token 工厂业务梳理

一、大模型厂商（自研模型，直接对外售卖 API 词元）	
智谱	2026 年上半年开放平台及 API 业务收入 8.25 亿元、同比增长超 27 倍，占总收入 86.5%。
MiniMax	2026 年上半年开放平台及其他 AI 企业服务收入增至 7390 万美元、占总收入 63.4%，Token 计划迅速获采用。
科大讯飞	AI 开放平台推出星辰 TokenPlan（统一额度池、多模型共享、500 万 TPM 吞吐），2026 年上半年大模型 API 及 MaaS 平台服务收入同比增长约 70%。
昆仑万维	自研天工大模型，AI 软件技术收入中的 API 调用服务按用户消耗的 Tokens 数量及对应单价确认收入。
二、电信运营商（Token 经营与算力网络）	
中国电信	以 Token 经营为主线重塑经营发展模式，TokenHub 汇聚了 170 个大模型，在行业内率先推出全国 Token 套餐。
中国移动	抢抓 Token 发展新机遇，组建 Token 办公室统筹 Token 全栈创新，升级移动模型服务平台 MoMA 并打造 Token 工厂。

下载日志已记录，仅供内部参考

中国联通	以“全国一张算力大网、一个算网调度平台、一个模型和 Token 分发平台”搭建“三个一”集约化 Token 运营体系，元景 MaaS 平台汇聚 200 余款大模型并建设规模化“Token 超市”。
三、云服务厂商（MaaS/词元服务）	
阿里巴巴	成立 Alibaba Token Hub（ATH）事业群，以“创造 Token、输送 Token、应用 Token”为目标。
金山云	2026Q2 AI 云服务业务总收入同比增长 82%，占公有云收入的 56%。
优刻得	推出私有化 Token 运营平台 UToken，依托乌兰察布、上海青浦双智算中心构建“词元工厂”基础设施，并新设全资子公司贵州词元启智。
青云科技	青云 AI 智算平台完成 Token 资源购买与管理能力建设，支持主流大模型服务上线及专属云实例部署，并建立模型服务流控和资源管理能力，为后续用户套餐体系、Token 计费等商业化运营能力建设提供基础支撑。
四、算力租赁与 Token 工厂运营（算力转化为 Token 供给）	
协创数据	全资设立天津词元智算科技有限公司，依托 FCloud 算力与模型底座，布局四大业务方向，构建“基础设施—平台—应用”服务链条。
宏景科技	2025 年年度报告提出发力 Token 工厂，以标准化、可计量、高性价比的 Token 产能为核心产品，系统性构建“Token 工厂”业务模式。
软通动力	“词元（Token）工厂计划”首个标杆项目“北京壹号词元工厂”在京点亮，与韶关词元工厂形成一南一北双核心布局。
行云科技	算力服务以 AI 推理调用量（算力 Token）为计量单位按需弹性计费，打造 TaaS（Token 即服务）模式，并新设全资子公司杭州行云词元科技。
弘信电子	以 Token 工厂模式为核心推动算力业务进入规模化运营阶段，2026 年 5 月燧弘华创首个华为昇腾 384 超节点算力集群落地。
润建股份	以“Token 为核、绿算协同、智维百业”为业务发展战略并设立 Token 事业群，五象云谷智算中心转型“Token 工厂”模式，2026 年上半年算力网络业务收入同比增长超 50%。
东方国信	聚焦 Token 工厂运营、算力集群租赁、算力零售服务三大业务板块，具备 Token 规模化生产与市场化销售能力，并成功中标中国电信 Token 工厂项目候选人。
中国软件国际	构建从智算基建到高价值 Token 运营的完整商业链条，算力业务采用算力硬件转售、算力租赁、Token 销售三种模式。
彩讯股份	依托自研 Rich AICloud 完善智能体感知、决策、执行能力，助力客户词元（Token）经营战略落地。
吉宏股份	将携手阿里云、火山引擎、谷歌、AWS，结合跨境社交电商全链路业务场景与生态伙伴算力供应链能力，探索大模型、智能体、跨境业务、Token 出海分销的全球化服务全链条协同。

下载日志已记录，仅供内部参考

铜牛信息	北京数据集团控股上市平台，集团旗下京算公司发布北京首家国资词元工厂“京算 Token 工厂”。
超讯通信	前瞻性布局 Token 算力运营平台，打通“算力基础设施供给—Token 标准化算力调度服务—行业 AI 综合解决方案输出”的完整价值链。
南威软件	建设北京七星园智算中心，拟打造每分钟产出超 10 亿 Token 的超级智能算力工厂；数字员工平台支持线上 Token 套餐付费与线下 AI 定制两种销售模式。
平治信息	平治智算云平台推出“Token 工厂”模块，用于集中管理多家 AI 模型供应商的接口调用，采用按量计费或按次计费并支持令牌分组管理。
思特奇	以 Token 为统一计量标准搭建集计费、结算、鉴权于一体的运营体系，并推出面向 AI 运营的计费支撑系统 TokenBoss。
亿都国际	旗下全资子公司算丰信息正式进驻香港数码港，重点启动国产 Token-as-a-Service（“TaaS”）业务布局，公司涵盖算力产业投资、运营与技术开发三大业务板块、部署有“丰收”等国产算力集群，并与香港宽频、天仪智科正式签署战略合作协议，深度参与平台从算力底座搭建、国产 GPU 资源整合到全场景方案落地的全流程。
五、Token 优化（提升单位算力 Token 产出效率）	
星环科技	Sophon LLMOps 新增提供 Token 工厂能力，通过大模型推理优化加速、KV Cache 优化、PD 分离和调度提升运行效率，并提供 Token 计量计费、SLA 感知智能路由与跨供应商容灾能力。
信测标准	参股存算一体企业上海锋行致远并新设深圳信测存储技术有限公司，布局面向大模型推理的存算加速模组及整机。
集智股份	下属子公司计划通过自持 Arm 架构 CPU 服务器与 GPU 算力服务器，重点生产满足 Agentic 时代复杂推理任务的 Token 产品，首批 Arm 架构高密度核心 CPU 服务器已完成部署。
并行科技	发布算力普惠 MaaS 平台，实现算力券与平台 Token 互通；智算云业务针对推理场景以应用运行特征分析推动算力选型与优化，并通过参股清程极智、趋境科技构建“算力资源+推理优化”的闭环服务体系。
六、垂直行业 Token 应用（企业级 AI 平台，Token 应用与变现出口）	
金山办公	自研分布式推理系统已在数字员工、代码审核、文档生成场景落地，模型成本较主流 Token 厂商价格低约 50%。
用友网络	2026 年上半年 AI 相关合同签约金额 9.06 亿元、实现 AI 相关收入 4.64 亿元；子公司畅捷通采用“MCP+TOKEN”综合计费的增值服务模式。
金蝶国际	发布企业智能体操作系统“灵基”，按“席位数+灵基点数”收费（灵基点数对应 AI 调用、智能体运行的资源消耗），2026 年上半年 AI 原生产品收入 2.96 亿元、同比增长约 189%。
汉得信息	构建“得灵”企业级 AI 能力体系，AI 智能化应用业务加速商用化落地。
拓尔思	以 Token 为核心重构数据价值，构建规模超 200 万亿 Token 的高质量通用语料池，其中行业高价值语料超 10 万亿 Token。

下载日志已记录，仅供内部参考

万兴科技	强化算力资源调度与 Token 管理能力，第三方大模型 API 最快可实现“T+0”快速接入，推动 AI 能力实现产品价值提升与 Token 消耗的商业闭环。
创业慧康	AI 全栈技术底座持续完善模型统一接入、算力运营、Token 管理与成本核算能力，打通从算力调度到智能体编排的技术链路。
新致软件	商业模式由销售模型能力（Token 付费）迭代至销售智能体与工作结果，保险行业平台依托超 30 亿条行业数据及千亿 Token 语料。

资料来源：各公司官网/公众号/公告、中国网财经、全景网、中华网投资等，国联民生证券研究所整理

下载日志已记录，仅供内部参考

5 风险提示

1）AI 应用及 Agent 商业化进展不及预期。若企业客户仍停留在试点阶段，智能体任务成功率、可靠性或可验证收益不足，模型调用增长可能难以转化为持续付费和稳定收入。

2）Token 调用需求增长不及预期。若模型能力提升、客户预算收缩或应用部署周期拉长，推理负载和 Token 消耗可能低于预期，影响基础设施、平台和应用各环节收入。

3）模型服务降价与行业竞争加剧。若模型厂商、云服务商持续通过降价、折扣或低价模型争夺调用量和客户，可能加剧模型服务及推理平台的价格竞争。若单位 Token 价格下降快于调用量和任务规模增长，相关企业收入及毛利空间可能承压。

4）推理技术迭代及投入回收不及预期。模型架构、硬件平台或推理优化方案快速变化，可能使现有基础设施利用率、Token 吞吐或单位成本改善不及预期。

5）数据安全与合规风险。企业 Agent 逐步接入客户信息、经营数据、内部知识库及跨系统工具后，数据访问范围和自动执行权限相应扩大。若权限控制、敏感数据保护、数据使用边界或操作审计机制不完善，可能出现数据泄露、越权调用或操作不可追溯等问题，进而增加企业客户的安全审查和合规评估要求，延长产品部署及采购周期。

下载日志已记录，仅供内部参考

插图目录

图 1： AI Factory 由能源、芯片、基础设施、模型与应用五层共同构成 6
图 2： 企业级 AI Factory 整合计算、存储、网络、AI 软件与上层应用 6
图 3： Vera Rubin NVL72 单位功率 Token 吞吐最高达到 GB200 NVL72 的 10 倍 7
图 4： DeepSeek-V3 基本架构 8
图 5： 昇腾大规模专家并行推理方案适配 DeepSeek V4、Qwen3 系列等主流 MoE 大模型 9
图 6： 单位 Token 成本由 GPU 小时成本与实际 Token 吞吐共同决定 9
图 7： 系统级 Token 效率成为 AI 推理经济性的重要衡量维度 10
图 8： 相近智能水平的模型可以对应明显不同的单任务成本 11
图 9： 从 Chatbot 到 Agent，模型交互由线性请求演进为多步骤调用链 12
图 10： 真实 Agent 会话中的 Input Sequence Length 随请求持续变化 13
图 11： 长上下文混合负载下，Prefill 容量饱和将快速推高 TTFT 14
图 12： Agent 推理中 KV Cache 读取量显著高于写入量 14
图 13： 较慢的工具调用可能触发 KV Cache 淘汰并导致高成本重计算 15
图 14： NVIDIA Dynamo 分布式推理架构：SLO 调度、KV 感知路由、PD 分离与 NIXL 数据的传输 16
图 15： Cache-aware P/D 分离分离通过请求分流与 KV Cache 复用减少重复 Token 计算 17
图 16： 智能路由相较单一模型实现更优任务成本效率 17
图 17： Context Engineering 通过信息筛选优化 Agent 上下文 18
图 18： 单位 GPU Token 吞吐与推理延迟形成 Pareto 效率前沿 19
图 19： Agent 任务成功率需同时考虑单次能力与重复执行可靠性 20
图 20： Together AI 通过 Endpoint—Deployment—Config 架构解耦应用接口与底层推理资源 21
图 21： 硅基流动通过分层推理架构连接异构算力、模型优化与上层 AI 应用 22
图 22： Agentforce 将完整业务流程拆解为可计量的 Agent Action 23
图 23： Zendesk 通过 Contained 与 Verified Resolution 衡量 AI Agent 任务完成情况 24
图 24： 企业 Agent 任务执行依赖数据、API 与权限治理体系 25
图 25： Agent 通过任务评价与生产反馈形成持续优化闭环 26
图 26： OpenRouter 折扣实验显示模型价格下降后 Token 调用量显著提升 27

表格目录

表 1： 大模型服务中的主要 Token 类型 4
表 2： 部分上市公司 Token 工厂业务梳理 28

下载日志已记录，仅供内部参考

分析师承诺

本报告署名分析师具有中国证券业协会授予的证券投资咨询执业资格并登记为注册分析师，基于认真审慎的工作态度、专业严谨的研究方法与分析逻辑得出研究结论，独立、客观地出具本报告，并对本报告的内容和观点负责。本报告清晰准确地反映了研究人员的研究观点，结论不受任何第三方的授意、影响，研究人员不曾因、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利益。

评级说明

投资建议评级标准	评级		说明
以报告发布日后的 12 个月内公司股价（或行业指数）相对同期基准指数的涨跌幅为基准。其中：A 股以沪深 300 指数为基准；新三板以三板成指或三板做市指数为基准；北交所以北证 50 指数为基准；港股以恒生指数为基准；美股以纳斯达克综合指数或标普 500 指数为基准。	公司评级	推荐	相对基准指数涨幅 15%以上
		谨慎推荐	相对基准指数涨幅 5%～15%之间
		中性	相对基准指数涨幅-5%～5%之间
		回避	相对基准指数跌幅 5%以上
	行业评级	推荐	相对基准指数涨幅 5%以上
		中性	相对基准指数涨幅-5%～5%之间
		回避	相对基准指数跌幅 5%以上

免责声明

本报告由国联民生证券股份有限公司或其关联机构制作。国联民生证券股份有限公司具有中国证监会许可的证券投资咨询业务资格。本报告的分销依据不同国家、地区的法律、法规和监管要求由国联民生证券于该国家或地区的具有相关合法合规经营资质的子公司/经营机构完成。在遵守适用的法律法规情况下，本报告亦可能由国联证券国际金融有限公司在香港地区发行。国联证券国际金融有限公司具备香港证监会批复的就证券提供意见（4 号牌照）的牌照，接受香港证监会监管，负责本报告于中国香港地区的发行与分销。

本报告仅供本公司授权之机构及个人使用，本公司不会因任何人收到本报告而视其为客户。本报告仅为参考之用，并不构成对任何人的操作建议或任何保证，不应被视为买卖任何证券、金融工具的要约或要约邀请。本报告所包含的观点及建议并未考虑获取本报告的机构及个人的具体投资目的、财务状况、特殊状况、目标或需要，客户应当充分考虑自身特定状况，进行独立评估，并应同时考量自身的投资目的、财务状况和特定需求，必要时就法律、商业、财务、税收等方面咨询专家的意见，不应单纯依靠本报告所载的内容而取代自身的独立判断。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而导致的任何可能的损失负任何责任。

本报告是基于已公开信息撰写，但本公司不保证该等信息的准确性或完整性。本报告所载的资料、意见及预测仅反映本公司于发布本报告当日的判断，且预测方法及结果存在一定程度局限性。在不同时期，本公司可发出与本报告所刊载的意见、预测不一致的报告，但本公司没有义务和责任及时更新本报告所涉及的内容并通知客户。

在法律允许的情况下，本公司及其附属机构可能持有报告中提及的公司所发行证券的头寸并进行交易，也可能为这些公司提供或正在争取提供投资银行、财务顾问、咨询服务等相关服务，本公司的员工可能担任本报告所提及的公司的董事；本公司自营部门及其他投资业务部门可能独立做出与本报告中的意见或建议不一致的投资决策。客户应充分考虑可能存在的利益冲突，勿将本报告作为投资决策的唯一参考依据。

若本公司以外的金融机构发送本报告，则由该金融机构独自为此发送行为负责。该机构的客户应联系该机构以交易本报告提及的证券或要求获悉更详细的信息。本报告不构成本公司向发送本报告金融机构之客户提供的投资建议。本公司不会因任何机构或个人从其他机构获得本报告而将其视为本公司客户。提示客户及公众投资者慎重使用未经授权刊载或者转发的本公司证券研究报告，慎重使用公众媒体刊载的证券研究报告。

本报告的版权仅归本公司所有，未经书面许可，任何机构或个人不得以任何形式、任何目的进行翻版、转载、公开传播、篡改或引用，不得将报告内容作为诉讼、仲裁、传媒所引用之证明或依据，不得用于营利或用于未经允许的其它用途。所有在本报告中使用的商标、服务标识及标记，除非另有说明，均为本公司的商标、服务标识及标记。本公司版权所有并保留一切权利。

无锡 江苏省无锡市金融一街 8 号国联金融大厦 8 楼

上海 上海市虹口区杨树浦路 188 号星立方大厦 B 座 7 层

北京 北京市西城区丰盛胡同 20 号丰铭国际大厦 B 座 5F

深圳 深圳市福田区中心四路 1 号嘉里建设广场 1 座 10 层 01 室