# GTC 大会重磅发布，腾讯加码 AI 投入，看好 AI 基建与应用景气度

## ——人工智能动态跟踪 2025 年 3 月第 2 期

| 行业评级 | 看好（维持） |
|---|---|
| 国家/地区 | 中国 |
| 行业 | 传媒行业 |
| 报告发布日期 | 2025 年 03 月 24 日 |


## 核心观点

- **英伟达 GTC 大会发布多款重磅 AI 软硬件产品：**英伟达 2025 年 GTC 大会发布多款重磅产品和技术，硬件、基模与算法侧均有明显优化，AI 算力规模与利用效率持续优化，大模型的 Scaling law 有望推动 AI 大模型持续实现突破。此外黄仁勋于 GTC 大会发布 Agentic AI 时代主题演讲，宣布 AI 将由生成式 AI 时代步入 Agentic AI 时代，届时算力需求将会迎来百倍增长。我们认为，一方面 Deepseek 的高效低成本的强大推理能力与本身 MaaS 服务的高毛利验证了推理类 AI 的商业模式可行性，另一方面如 Manus、新夸克等 AI agent 的出现首次展现了 AI agent 产品的范式，如 GTC 大会所述，算力将成为 AI 时代不可或缺的资源。因此我们持续看好国内云计算龙头、算力基模领先的阿里巴巴与用户基础扎实、AI 应用前瞻布局的腾讯。
- **腾讯加码 AI 投入 CAPEX 大幅增长，国内 AI 产业链迈入景气周期。**本周腾讯公布 2024Q4 财报，2024 年资本开支同比两倍以上增长，参考 Bloomberg 预期计算，假设 2025 年 CAPEX 占收入比例为 12%-15%，则对应腾讯 2025 年 CAPEX 为 864-1080 亿。结合前期阿里对未来 3 年投资 3800 亿资本开支用于云和 AI 硬件基础设施建设的宏伟战略，且字节跳动、百度等互联网大厂及联通、移动等运营商对 AI 投入态度均趋于积极乐观，我们认为本轮由 Deepseek 带动 AI 产业链热潮已推动行业步入景气周期，持续时间和力度均有强力资本投入支撑，上游算力侧的投入将有力转化为下游 AI 算力及大模型产品供给的增长，算力效率提升带动的降本将明显提升 AI 需求及 B/C 端渗透率，云计算及 AI 应用服务有望迎来高增，从而带动相关互联网龙头公司的云及 AI 业务重估。
- **AI 基模应用领先，看好腾讯 AI ToC 应用龙头地位。**腾讯 AI 大模型布局完善，拥有混元系列大语言模型与多模态大模型混元 video 与混元生图等多款行业领先大模型，新发布深度思考大模型混元 T1 达到行业领先水平，AI 基模技术国内领先。AI 应用侧，元宝于 2 月登顶 app store 免费榜首，DAU 增长近 20 倍，此外还有微信这一国民级应用作为用户支撑。集团内部广告、游戏及微信多业务可依靠 AI 赋能，我们认为腾讯 AI 应用侧场景、用户优势明显，有望取得 AI ToC 应用龙头地位。
- **本周 AI 行业重点动态：1）产业政策端：**上海市政府发布《促进服务业创新发展若干措施》明确提到支持 AI 产业发展，并颁布产业基金及技术人才扶持等政策，有望带动 AI 产业迎来系统性政策红利。**2）大模型端：**昆仑万维发布首个多模态深度推理大模型 R1V，首次将推理范式引入多模态大模型。**3）算法端：**METR 团队发现大模型平均每 7 个月可持续任务时长翻倍且准确率明显提升，AI 智能化持续加速。Zoom 团队开发新草稿链推理范式，有效提升 AI 大模型对简单问题的解决效率并显著降低算力消耗。**4）AI 应用端：**字节跳动团队发布开源类 Manus 综合 AI agent 产品 TARS，在多个基准测试中表现出色。建议关注后续 AI agent 产品动态。

## 投资建议与投资标的

- 看好 AI 新周期带动算力-算法-应用生态三端依次持续推进，我们建议增加港股互联网板块配置仓位，核心推荐处于产业链前端，云消费弹性明显+算法具备优势的阿里巴巴-W(09988，买入)，多模态视频生成模型技术全球领先的快手-W(01024，买入)，以及卡位最佳社交场景，具备数据+应用生态优势的腾讯控股(00700，买入)。

## 风险提示

行业发展及 AI 应用落地不及预期，假设条件变化影响测算结果，宏观经济风险，海外政策风险。

# 目 录

# 图表目录

# 1 本周 AI 专题跟踪：英伟达召开 GTC 大会，腾讯加码 AI 投入

## 1.1 聚焦英伟达 GTC 大会：AI 算力效率持续优化，Agentic AI 时代有望带来百倍算力增长

北京时间 3 月 19 日凌晨，英伟达召开 2025 年度 GTC 大会。本次 GTC 大会重磅发布了 Blackwell 芯片与下一代 GPU 架构 Rubin、Groot N1 开源人形机器人基础模型与用于机器人仿真的开源物理引擎 Newton、Dynamo 推理操作系统与 CUDA-X 全栈加速库等新产品/服务，并对 AI 算力、AI agent 及具身智能、量子计算等前沿技术的发展趋势进行展望。具体看 AI 相关核心内容如下：

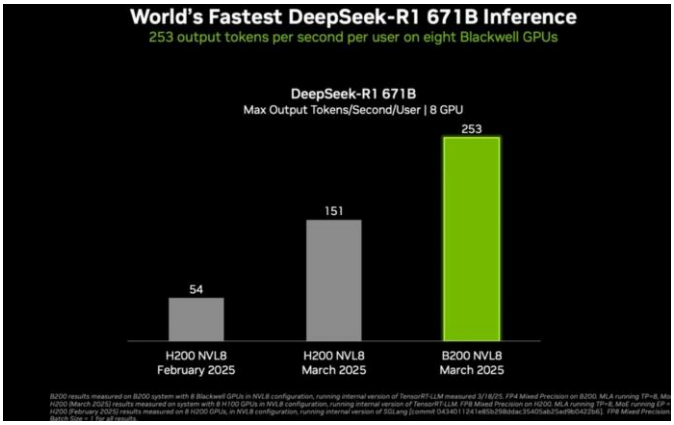
**1）AI 硬件：**Blackwell 投产，可显著提升 AI 大模型推理效率。基于 Blackwell 架构的升级版芯片 Blackwell Ultra 已投入生产，据官方数据，新一代芯片性能较上一代提升 1.5 倍，搭载 288GB HBM3e 存储，FP4 推理性能显著增强。全新 GB300 NVL72 机架系统整合 72 个 Blackwell Ultra GPU 与 36 颗基于 Arm Neoverse 的 Grace CPU，专为复杂 AI 推理而生。相比前代 GB200 NVL72，推理速度飙升 1.5 倍，LLM 推理效率较 Hopper 架构提升 11 倍，Deepseek-R1671B 模型响应时间从 H100 的 90 秒缩短至 10 秒，可显著提升 token 吞吐量及吞吐速率。

**表 1：Blackwell Ultra NVL72 和 GB200 NVL72 硬件参数对比**

| | Blackwell Ultra NVL72 | GB200 NVL72 |
|---|---|---|
| Blackwell 芯片\|Grace 芯片 | **72\|36** | 72\|36 |
| FP4 | **1.1ExaFLOPS** | 1440PetaFLOPS |
| 快速内存 | **40TB** | 30TB |
| 内存 | **20TB** | 14TB |
| 内存带宽 | **576TB/S** | 576TB/S |
| NVLink 带宽 | **130TB/S** | 130TB/S |

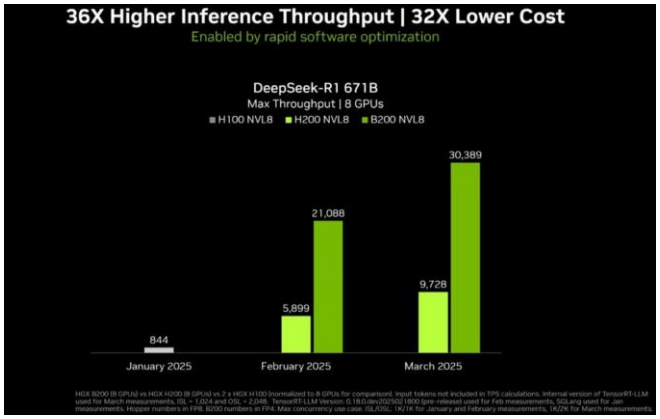
数据来源：腾讯科技

**图 1：使用 Blackwell 芯片后 Deepseek R1 吞吐速率明显增长**




数据来源：AI 寒武纪

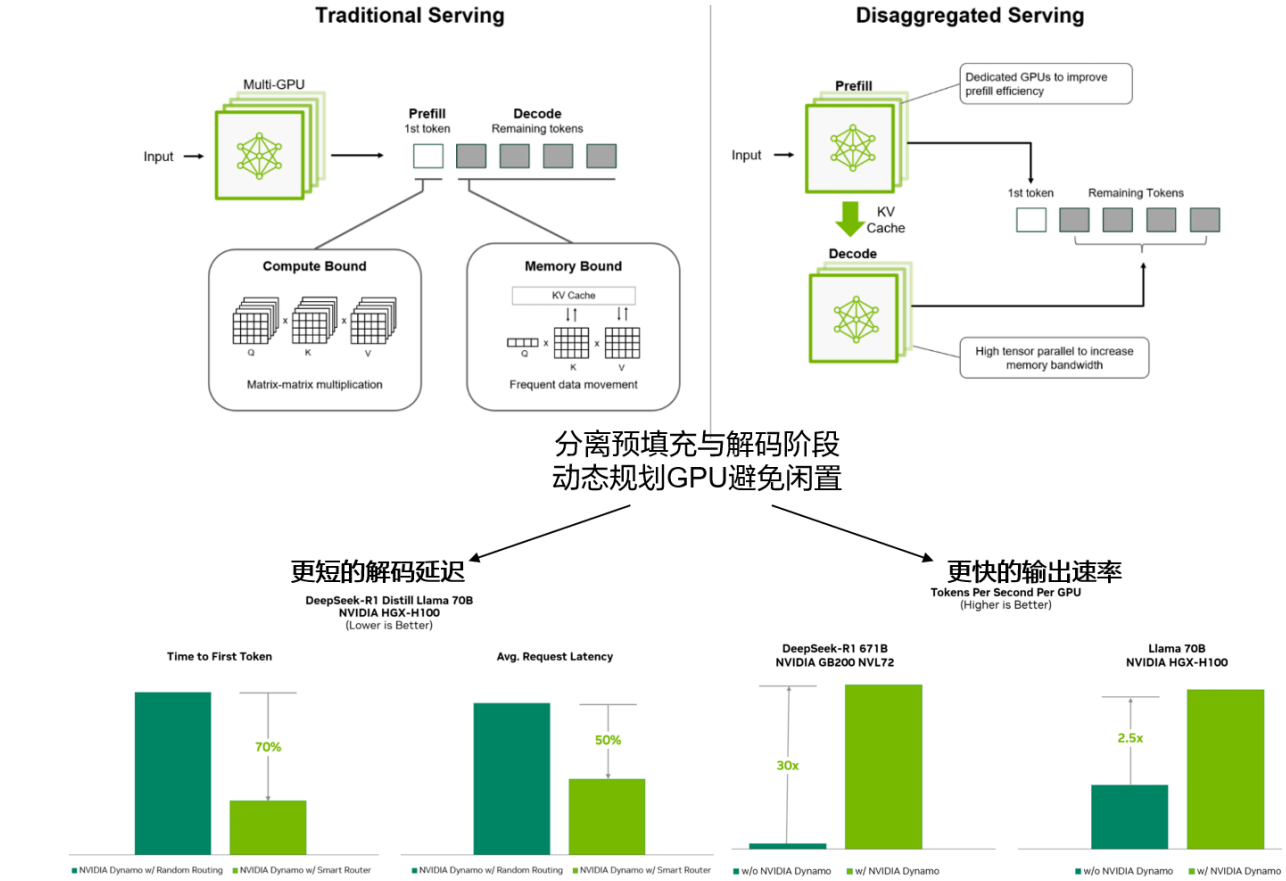
**图 2：使用 Blackwell 芯片后 Deepseek R1 吞吐量提升至 36 倍，token 成本下降 32 倍**




数据来源：AI 寒武纪

**2）AI 软件：**软件层面，NVIDIA 发布了分布式推理服务库 NVIDIA Dynamo。NVIDIA Dynamo 是一个高吞吐量、低延迟的开源推理服务框架，用于在大规模分布式环境中部署生成式 AI 和推理模型。在 NVIDIA Blackwell 上运行开源 Deepseek-R1 模型时将 token 输出速率提高了 30 倍，且解码延迟降低了 50%以上。

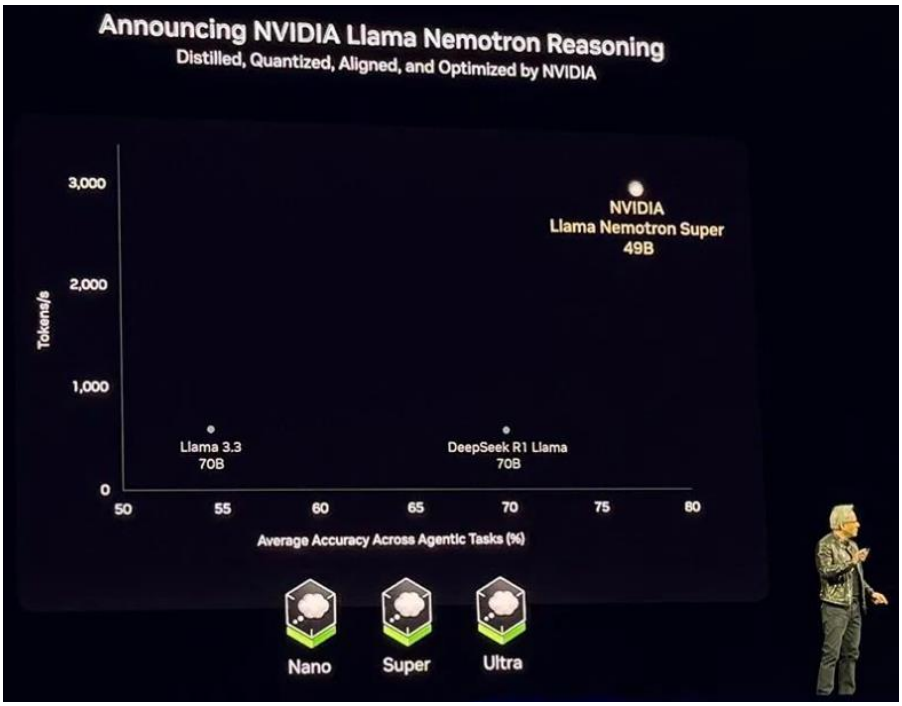
图 3：NVIDIA Dynamo 作用原理及效果




数据来源：NVIDIA developer

**3）AI 大模型：**英伟达基于 Llama 3.1 开发了全新 Llama Nemotron 推理模型系列，分别针对不同的部署需求提供 Nano、Super、Ultra 版本。通过后训练阶段的增强有效提升了多步骤数学计算、编码、推理和复杂决策的能力，使得基于该模型开发的 Agent 能够以更高的准确率处理复杂的自动化任务，增强决策能力。其准确性比基础模型提高了 20%，其中 Super 49B 版本在生成速度和 AI 智能体任务的准确性两个维度超过 Deepseek-R1，吞吐量达到 Llama 3.3 70B、Deepseek R1 Llama 70B 的 5 倍。

图 4：英伟达 Llama Nemotron 推理模型效果领先 Deepseek 与 Llama 基础模型

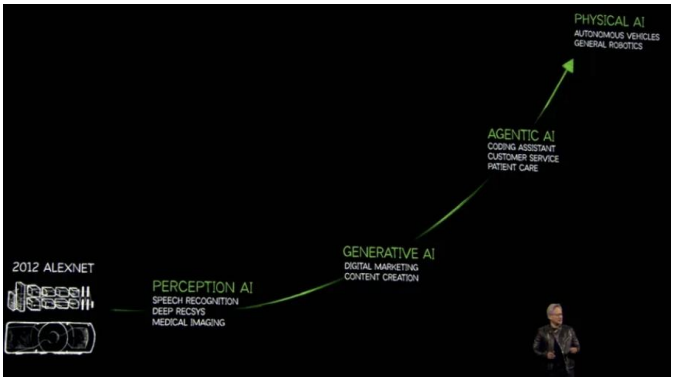



数据来源：智东西

除此以外，本次 GTC 大会上英伟达 CEO 黄仁勋发表了 Agentic AI 专题演讲：

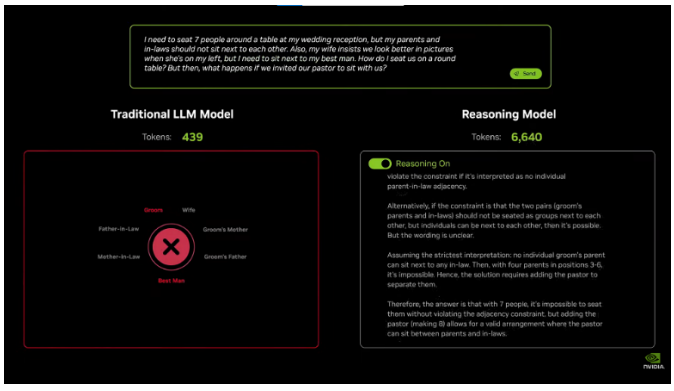
**从生成式 AI 到 AI agent，Agentic AI 时代来临。**AI 的发展史可以分为四个阶段：1. 感知式 AI（语音识别、图像识别）2. 生成式 AI（文本/图像生成）3. Agentic AI（自主推理、规划和行动）4. Physical AI（机器人、自动驾驶等）。当前感知式 AI 与生成式 AI 的发展接近完善，随之而来的多模态识别及数据检索/分析功能为下一阶段的 AI agent 打好了基础。近年 AI agent 开始出现突破，逐渐产生了自主推理并执行任务的能力。现场演示生成婚礼座位的安排，以 Deepseek 为例，对比传统 LLM 大模型可以考虑到宾客的人际关系从而做出更合理的座次安排。

图 5：AI 发展的四个阶段




数据来源：techradar

图 6：Deepseek 更擅长解决推理问题




数据来源：智东西

**Scaling law 并未失效，算力仍是 AI agent 时代的核心生产要素。**Agentic 时代的 AI 不再像 LLM 一样是对下一个 token 进行预测，而是基于训练范式的思维链生成一系列表示推理步骤的 token 序列，从而可以实现推理分析和解决方案的自我规划。而在这一过程中的 token 生成量将是生成式