

AI Agent 基础设施和模型能力继续迭代

——计算机行业跟踪报告

同步大市 (维持)

2026 年 07 月 13 日

行业核心观点:

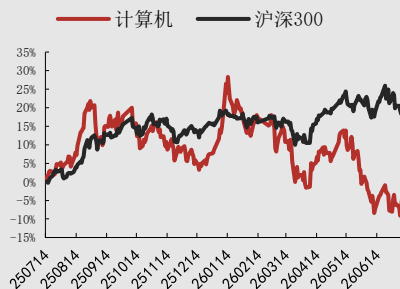
上周, 沪深 300 下跌 1.27%, 创业板指下跌 4.41%, 申万计算机行业指数上涨 2.74%, 分别跑赢沪深 300 和创业板指 4.01 和 7.15 个百分点, 在申万各一级行业中排名第 1 位。上周, 英伟达发布博文指出, 在智能体 AI 时代, CPU 需要承担工具调用、代码执行、数据处理、KV 缓存和结果分析等任务, NVIDIA Vera CPU 在智能体 AI 执行的满载 CPU 工作负载中, 单核持续性能为 x86 的 1.8 倍; 同时, 英伟达 Nemotron 3 Ultra 在 LangChain 的 Deep Agents benchmark 测试中取得开源模型最高准确率, 调优后每次运行的推理成本为领先闭源模型的 1/10。我们认为, 随着 AI Agent 从问答交互走向多步推理、工具调用和复杂任务执行, 算力基础设施的瓶颈正在从单一 GPU 训练能力延伸至 CPU 调度、内存带宽、推理吞吐和系统工程优化等环节, 在此背景下, 能够同时保障低成本、高吞吐、高可靠性的模型与软硬件协同能力, 将成为 AI Agent 规模化落地的重要支撑。SpaceXAI 发布首个专门针对编程和智能体任务训练的 Grok 4.5 模型, 该模型由 SpaceXAI 与 Cursor 联合完成训练, 面向真实工程场景设计, 擅长处理大型代码库以及跨多个仓库、涉及多种工具的长时间任务; 字节跳动发布多模态图像创作模型 Seedream 5.0 Pro, 在复杂信息可视化、交互式精准编辑、真实影像与人像质感、原生多语种输入与生成等方面实现提升。我们认为, 头部厂商正围绕编程、办公、图像创作等高频场景加快模型产品化和应用落地, AI 能力正从通用对话逐步深入具体生产流程, 模型竞争也从单一能力比拼转向场景适配、成本效率的综合竞争, 建议关注具备底层模型能力和应用场景积累的领先厂商。中长期视角下, 继续聚焦 AI 产业、数据产业两大投资主线。

投资要点:

产业动态: (1) **人工智能:** 英伟达 7 月 7 日发布博文, 指出在智能体 AI 时代, NVIDIA Vera CPU 成为“可规模化单线程最强 CPU”, 持续单核性能是 x86 的 1.8 倍。(2) **AI 大模型:** 英伟达 7 月 8 日发布博文, 宣布在 LangChain 的 Deep Agents benchmark (深度智能体基准) 测试中, 旗下 NVIDIA Nemotron 3 Ultra 取得开源模型最高准确率。(3) **AI 大模型:** 7 月 9 日, SpaceXAI 正式发布了其 Grok 4.5 模型, 这是该公司首个专门针对编程和智能体任务训练的模型。据介绍, 该模型由 SpaceXAI 与 Cursor 联合完成训练, 在提供前沿智能水平的同时, 兼具领先的速度与成本效率。(4) **AI 大模型:** 字节跳动 7 月 8 日宣布, 正式发布多模态图像创作模型 Seedream 5.0 Pro。据介绍, 相比之前版本, Seedream 5.0 Pro 在图文匹配、结构合理性、文字渲染与画面美感等基础能力上全面提升。

行业估值: 从估值情况来看, 申万计算机行业 2026 年 7 月 12 日 PE-TTM

行业相对沪深 300 指数表现



数据来源: 聚源, 万联证券研究所

相关研究

AI 应用商业化验证进程加速

全球大模型 token 调用量持续增长, 多款国产大模型位居全球领先梯队

AI 大模型综合能力持续增强, 应用场景有望进一步拓宽拓深

分析师:

夏清莹

执业证书编号:

S0270520050001

电话:

(0755) 8322 3620

邮箱:

xiaqy1@wlzq.com.cn

为 143.13 倍，低于 2023-2025 年历史 PE-TTM 的均值 158.59 倍。

风险提示：中美科技摩擦；地缘政治风险；产业生态建设不及预期；应用落地不及预期，数据隐私安全风险，市场竞争加剧。

正文目录

1 行业周观点.....4

2 市场行情回顾.....4

3 产业动态.....5

 3.1 人工智能：可规模化多线程最强 CPU，英伟达 Vera 在智能体 AI 负载下持续性能是 x86 的 1.8 倍5

 3.2 AI 大模型：推理成本仅为顶级闭源模型的 1/10，英伟达 Nemotron 3 Ultra 获开源模型最高准确率6

 3.3 AI 大模型：马斯克 SpaceXAI 首个编程智能体模型 Grok 4.5 发布，与 Cursor 联合训练，效率翻倍价格减半6

 3.4 AI 大模型：字节发布多模态图像创作模型 Seedream 5.0 Pro，复杂信息可视化、交互式精准编辑6

4 风险提示.....7

图表 1： 沪深 300、创业板指、申万一级行业指数涨跌幅（%）（上周）4

图表 2： 沪深 300、创业板指、申万一级行业指数涨跌幅（%）（2026 年初至 2026 年 7 月 12 日）5

图表 3： 申万计算机行业估值情况（PE-TTM，单位：倍）5

1 行业周观点

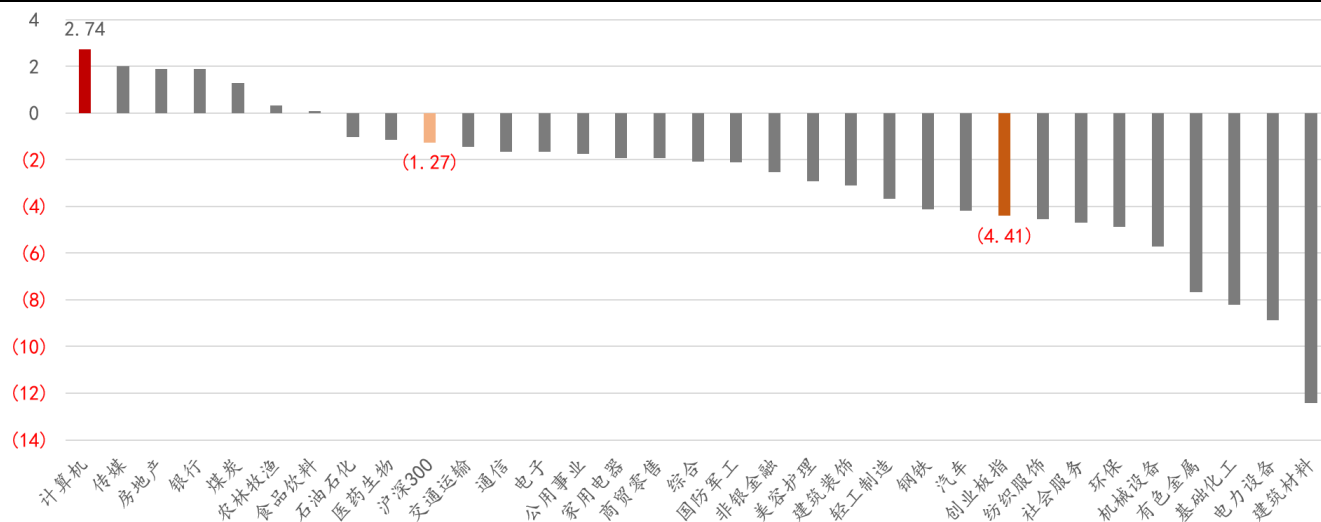
上周，英伟达发布博文指出，在智能体AI时代，CPU需要承担工具调用、代码执行、数据处理、KV缓存和结果分析等任务，NVIDIA Vera CPU在智能体AI执行的满载CPU工作负载中，单核持续性能为x86的1.8倍；同时，英伟达Nemotron 3 Ultra在LangChain的Deep Agents benchmark测试中取得开源模型最高准确率，调优后每次运行的推理成本为领先闭源模型的1/10。我们认为，随着AI Agent从问答交互走向多步推理、工具调用和复杂任务执行，算力基础设施的瓶颈正在从单一GPU训练能力延伸至CPU调度、内存带宽、推理吞吐和系统工程优化等环节，在此背景下，能够同时保障低成本、高吞吐、高可靠性的模型与软硬件协同能力，将成为AI Agent规模化落地的重要支撑。SpaceXAI发布首个专门针对编程和智能体任务训练的Grok 4.5模型，该模型由SpaceXAI与Cursor联合完成训练，面向真实工程场景设计，擅长处理大型代码库以及跨多个仓库、涉及多种工具的长时间任务；字节跳动发布多模态图像创作模型Seedream 5.0 Pro，在复杂信息可视化、交互式精准编辑、真实影像与人像质感、原生多语种输入与生成等方面实现提升。我们认为，头部厂商正围绕编程、办公、图像创作等高频场景加快模型产品化和应用落地，AI能力正从通用对话逐步深入具体生产流程，模型竞争也从单一能力比拼转向场景适配、成本效率的综合竞争，建议关注具备底层模型能力和应用场景积累的领先厂商。

中长期视角下，继续聚焦AI产业、数据产业两大投资主线。1) AI产业：主要关注国内AI大模型的技术升级和产品迭代、头部AI大模型厂商的生态建设、智能驾驶、具身智能等新型智能终端及智能体的应用落地进程，以及人工智能治理的生态建设。2) 数据产业：主要关注数据基础设施建设、高质量数据集等数据要素资源的开发利用、数据要素的价值释放及应用落地以及数据隐私安全。

2 市场行情回顾

申万计算机行业跑赢沪深300和创业板指，周涨跌幅位于申万31个一级行业的上游。上周，沪深300下跌1.27%，创业板指下跌4.41%，申万计算机行业指数上涨2.74%，分别跑赢沪深300和创业板指4.01和7.15个百分点，在申万各一级行业中排名第1位。

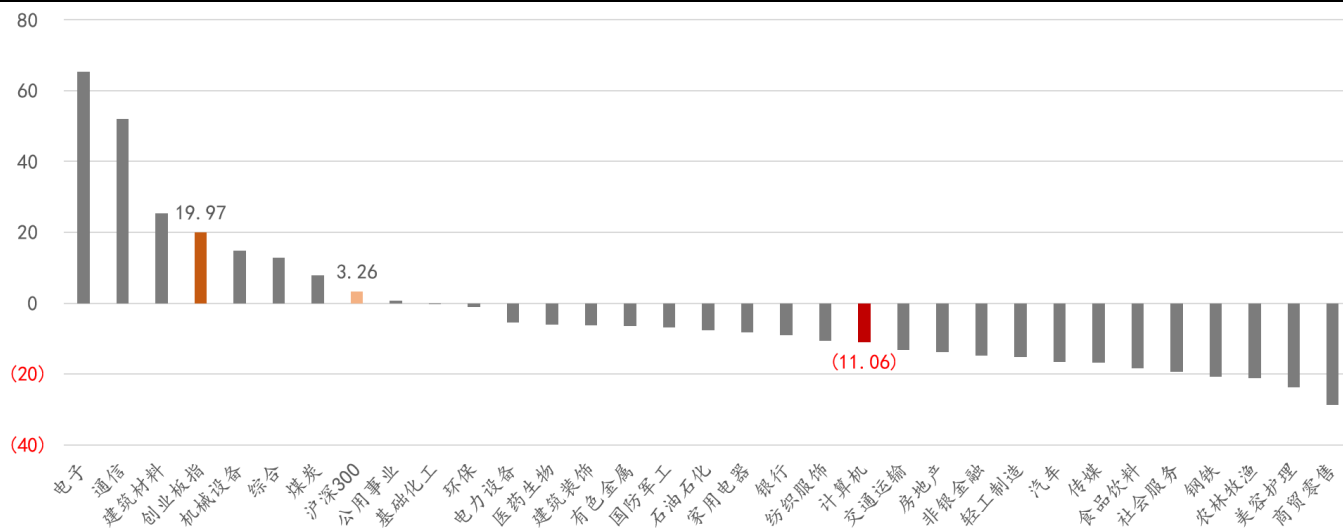
图表1：沪深300、创业板指、申万一级行业指数涨跌幅（%）（上周）



资料来源：iFinD，万联证券研究所

申万计算机行业2026年初至7月12日累计下跌11.06%，位于申万31个一级行业的中下游。2026年初至7月12日，沪深300上涨3.26%，创业板指上涨19.97%，申万计算机行业指数下跌11.06%，分别跑输沪深300和创业板指14.31和31.02个百分点，在申万各一级行业中排名第19位。

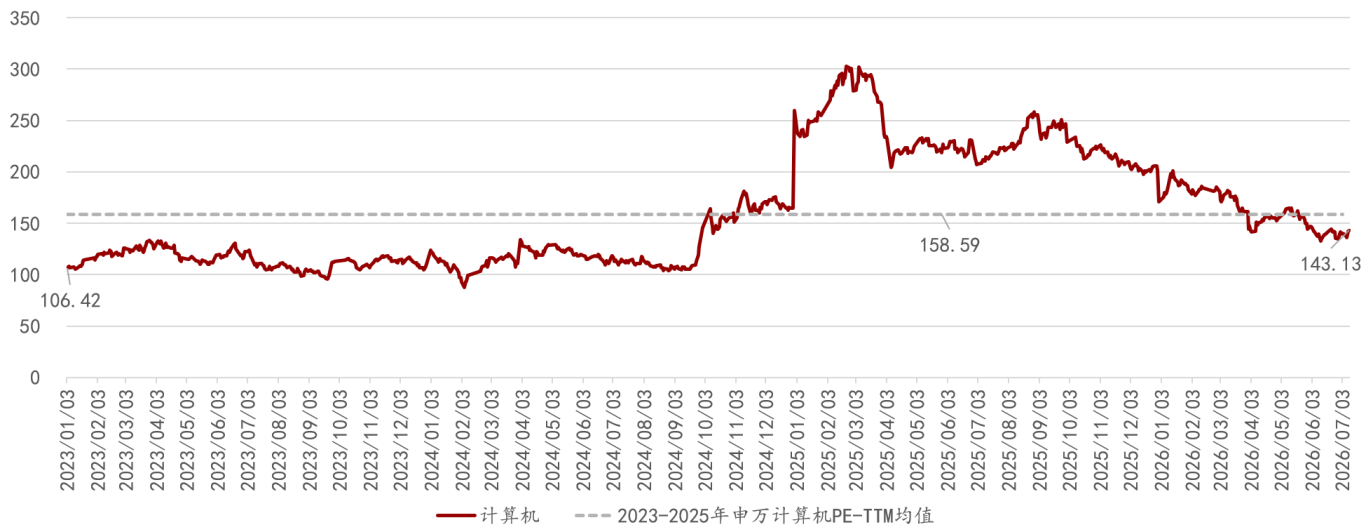
图表2：沪深300、创业板指、申万一级行业指数涨跌幅（%）（2026年初至2026年7月12日）



资料来源：iFinD，万联证券研究所

行业估值水平低于历史中枢水平。从估值情况来看，申万计算机行业2026年7月12日PE-TTM为143.13倍，低于2023-2025年历史PE-TTM的均值158.59倍。

图表3：申万计算机行业估值情况（PE-TTM，单位：倍）



资料来源：iFinD，万联证券研究所

3 产业动态

3.1 人工智能：可规模化单线程最强 CPU，英伟达 Vera 在智能体 AI 负载下持续性能是 x86 的 1.8 倍

英伟达7月7日发布博文，指出在智能体 NVIDIA Vera CPU成为“可规模化单线程最强CPU”，持续单核性能是x86的1.8倍。在博文中，英伟达提出“可规模化单线程最强CPU”（Max single-threaded CPUs at scale）概念，指出在智能体AI推理、响应和学习过程中，CPU需要承担工具调用、代码执行、数据处理、KV缓存和结

果分析等任务，而单线程是衡量速度的一个关键指标。英伟达在官方博文中指出，NVIDIA Vera就是此类新型CPU的代表，面向智能体循环中的串行步骤设计，相比较NVIDIA Grace，其自研Olympus CPU核心每周指令数高50%。在硬件参数方面，Vera配备最高1.2TB/s的LPDDR5X内存带宽，内存功耗低于40瓦；同时采用单片计算裸片设计，核心间带宽为3.4TB/s。在性能对比方面，英伟达称，在智能体AI执行的满载CPU工作负载中，Vera的单核持续性能为x86的1.8倍。Perplexity测试一项真实编码流程，即克隆代码仓库并在沙箱中运行测试套件，结果显示Vera完成任务速度约为x86的1.5倍，并发沙箱启动速度最高为x86的1.9倍。在数据处理类负载中，合作伙伴测试显示，Vera运行Starburst（分布式SQL分析平台）的大规模SQL分析速度可达领先x86服务器CPU的3倍；在Redpanda（流式数据平台）实时流处理场景中，延迟最高可降低至1/6。（资料来源：IT之家）

3.2 AI大模型：推理成本仅为顶级闭源模型的1/10，英伟达 Nemotron 3 Ultra 获开源模型最高准确率

英伟达7月8日发布博文，宣布在LangChain的Deep Agents benchmark（深度智能体基准）测试中，旗下NVIDIA Nemotron 3 Ultra取得开源模型最高准确率。英伟达表示本次测试的核心在于成本和吞吐量，调优后的Nemotron 3 Ultra每次运行的推理成本为领先闭源模型的1/10，在开源模型中实现了最高的准确率，同时以更高的吞吐量完成了更多任务。在业务任务能力方面，英伟达Nemotron 3 Ultra模型通过优化模型周围环境，达到了顶尖闭源模型的水平。LangChain 联合创始人兼首席执行官Harrison Chase表示：“构建更优秀智能体的关键在于不断改进模型周围的系统。当团队能够协同调整内存、工具使用、评估和模型行为时，这些因素都会产生叠加效应。我们与NVIDIA的合作表明，企业可以利用开放技术栈获得强大的性能，同时还能掌控他们正在构建的智能体系统。”（资料来源：IT之家）

3.3 AI大模型：马斯克 SpaceXAI 首个编程智能体模型 Grok 4.5 发布，与 Cursor 联合训练，效率翻倍价格减半

7月9日，SpaceXAI正式发布了其Grok 4.5模型，这是该公司首个专门针对编程和智能体任务训练的模型。据介绍，该模型由SpaceXAI与Cursor联合完成训练，在提供前沿智能水平的同时，兼具领先的速度与成本效率。马斯克将其称为“Opus级模型”。Grok 4.5面向真实工程场景设计，擅长处理大型代码库以及跨多个仓库、涉及数百种技能和多种工具的长时间任务。模型在数万个NVIDIA GB300 GPU上完成训练，团队在数据过滤与筛选上投入了大量精力，通过去重、质量评分和领域聚焦筛选来保证数据混合的高覆盖与高信号。在强化学习阶段，训练覆盖了数十万个任务，重点围绕多步软件工程及其他技术工作，采用自动化和基于模型的评分。据介绍，Grok 4.5单个任务消耗的Token仅为同级领先模型的一半，每秒输出Token数达80TPS，且单次任务步骤数少一半。定价为每百万输入Token2美元（现汇率约合13.6元人民币）、每百万输出Token6美元（现汇率约合40.8元人民币），成本不到同类可比模型的一半。除了代码生成，Grok 4.5还展现出Office工作场景下的细致能力。在Grok Build中，该模型可以构建包含网络研究、多工作表公式的复杂Excel模型，并留下便签以供后续参考；在PowerPoint中能利用原生形状绘制复杂图表并设计直观的幻灯片内容；在Word里则可撰写清晰的文章。模型还具备根据简单提示直接构建端到端功能应用的能力，在Rust、C/C++等挑战性任务中均有高效表现。（资料来源：IT之家）

3.4 AI大模型：字节发布多模态图像创作模型 Seedream 5.0 Pro，复杂信息可视化、交互式精准编辑

字节跳动7月8日宣布，正式发布多模态图像创作模型Seedream 5.0 Pro。据介绍，相

比之前版本，Seedream 5.0 Pro在图文匹配、结构合理性、文字渲染与画面美感等基础能力上全面提升：

复杂信息可视化：将数据、概念与密集文字准确转化为专业排版，直接用于高信息密度的内容生产。

交互式精准编辑：基于对空间位置与区域语义的理解，支持点选、圈选、草图渲染、色彩与材质替换、图层分离、多图融合的自由组合，实现可控编辑。

真实的影像与人像质感：还原现实世界的光影、材质与皮肤肌理，兼顾CG表现与摄影质感，画面更具生命力。

原生多语种输入与生成：支持十余种世界常用语言的直接输入与高质量渲染，准确传递本地化特征。

目前，Seedream 5.0 Pro已上线火山方舟体验中心，也将陆续在豆包、即梦上线。（资料来源：IT之家）

4 风险提示

中美科技摩擦；地缘政治风险；产业生态建设不及预期；应用落地不及预期，数据隐私安全风险，市场竞争加剧。

行业投资评级

强于大市：未来6个月内行业指数相对大盘涨幅10%以上；

同步大市：未来6个月内行业指数相对大盘涨幅10%至-10%之间；

弱于大市：未来6个月内行业指数相对大盘跌幅10%以上。

公司投资评级

买入：未来6个月内公司相对大盘涨幅15%以上；

增持：未来6个月内公司相对大盘涨幅5%至15%；

观望：未来6个月内公司相对大盘涨幅-5%至5%；

卖出：未来6个月内公司相对大盘跌幅5%以上。

基准指数：沪深300指数

风险提示

我们在此提醒您，不同证券研究机构采用不同的评级术语及评级标准。我们采用的是相对评级体系，表示投资的相对比重建议；投资者买入或者卖出证券的决定取决于个人的实际情况，比如当前的持仓结构以及其他需要考虑的因素。投资者应阅读整篇报告，以获取比较完整的观点与信息，不应仅仅依靠投资评级来推断结论。

证券分析师承诺

本人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证券分析师，以勤勉的执业态度，独立、客观地出具本报告。本报告清晰准确地反映了本人的研究观点。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。

免责声明

万联证券股份有限公司（以下简称“本公司”）是一家覆盖证券经纪、投资银行、投资管理和证券咨询等多项业务的全国性综合类证券公司。本公司具有中国证监会许可的证券投资咨询业务资格。

本报告仅供本公司的客户使用。本公司不会因接收人收到本报告而视其为客户。在任何情况下，本报告中的信息或所表述的意见并不构成对任何人的投资建议。本报告中的信息或所表述的意见并未考虑到个别投资者的具体投资目的、财务状况以及特定需求。客户应自主作出投资决策并自行承担投资风险。本公司不对任何人因使用本报告中的内容所导致的损失负任何责任。在法律许可情况下，本公司或其关联机构可能会持有报告中提到的公司所发行的证券头寸并进行交易，还可能为这些公司提供或争取提供投资银行、财务顾问或类似的金融服务。

市场有风险，投资需谨慎。本报告是基于本公司认为可靠且已公开的信息撰写，本公司力求但不保证这些信息的准确性及完整性，也不保证文中的观点或陈述不会发生任何变更。在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的报告。分析师任何形式的分享证券投资收益或者分担证券投资损失的书面或口头承诺均为无效。

本报告的版权仅为本公司所有，未经书面许可任何机构和个人不得以任何形式翻版、复制、刊登、发表和引用。未经我方许可而引用、刊发或转载的引起法律后果和造成我公司经济损失的概由对方承担，我公司保留追究的权利。

万联证券股份有限公司 研究所

上海浦东新区世纪大道1528号陆家嘴基金大厦

北京西城区平安里西大街28号中海国际中心

深圳福田区深南大道2007号金地中心

广州天河区珠江东路11号高德置地广场