计算机

行业快报

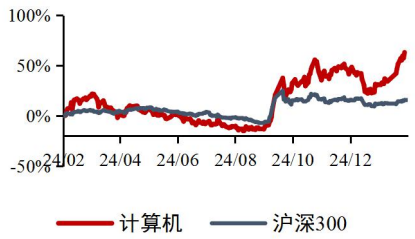
领先大市-A(维持)

UltraMem 架构为推理降本，AI 应用全面落地可期

2025 年 2 月 17 日

行业研究/行业快报

计算机行业近一年市场表现



资料来源：最闻

首选股票		评级
688111.SH	金山办公	买入-A
688041.SH	海光信息	买入-A

相关报告：

【山证计算机】DeepSeek 有望推动 AI 应用生态加速繁荣 2025.2.6

【山证计算机】《人工智能扩散框架》发布，AI 芯片国产化替代持续加速-行业政策点评 2025.1.20

分析师：

方闻千

执业登记编码：S0760524050001

邮箱：fangwenqian@szxq.com

事件描述：

➢ 2 月 12 日，字节豆包大模型团队发布全新的稀疏模型架构 UltraMem，有效解决了当前主流的 MoE 架构在推理时产生的高额访存问题，推理速度较 MoE 架构提升 2-6 倍，同时推理成本最高可降低 83%。

事件点评：

➢ **UltraMem 在 PKM 架构的基础上对模型结构、value 检索方式、稀疏参数进行优化，在保证模型性能的同时大幅提升推理效率。**UltraMem 架构参考 PKM（Product Key Memory）的设计，即 Transformer 层中嵌入大内存层以及推理时以行列路由的方式激活参数，访存效果较 MoE 架构明显改善。同时，UltraMem 对 PKM 架构进行针对性优化以提升模型性能：1）优化模型结构：将 PKM 的单个内存层拆分成多个内存层均匀嵌入 Transformer 层中，使模型能够并行执行访存和 Transformer 层计算操作；2）优化 value 检索方式：在推理时以 TDQKR 的乘法方法替代简单的行列加权方法选出得分最高的多个 value，使模型能够精准检索到与输入相关的 value；3）隐式扩展稀疏参数：引入数倍于 physical memory 的 virtual memory，在不提高模型部署复杂度的情况下提升模型性能。根据实验结果，训练规模达 2000 万 value 的 UltraMem 模型，在同等计算资源下可同时实现业界领先的推理速度和模型性能。

➢ **推理成本持续下降加速应用生态繁荣。**根据 Semianalysis 数据，随着算法持续进步，截至 2024 年底，以 GPT-3 质量的输出为标准，模型推理价格下降了 1200 倍。进入 2025 年，在推理技术优化下，DeepSeek 模型的使用成本不到 o1 模型的 1/25，而字节最新发布的 UltraMem 架构将使主流稀疏模型的推理成本大幅下降。我们认为，模型调用价格是用户选择模型运行应用的重要考量因素，各大模型厂商及科技大厂将持续竞相推动推理成本下降，从而带动上层 AI 应用的加速落地，并有望促进应用从云端场景向端侧场景拓展。

投资建议：UltraMem 架构的模型推理成本大幅下降，将加速 AI 应用落地，并推动应用向端侧渗透，进而刺激推理算力需求，重点关注 1）AI 应用相关标的，包括企业服务领域的金蝶国际、泛微网络、致远互联，用友网络等，办公领域的金山办公、福昕软件等，多模态领域的万兴科技、美图公司等，金融领域的新致软件、同花顺等，教育领域的科大讯飞、佳发教育等，医疗领域的润达医疗、卫宁健康等，以及其他领域的彩讯股份、金桥信息、焦点科技等；2）国产算力芯片厂商，包括海光信息、寒武纪等；3）AI 服务器厂

商，包括四川长虹、神州数码、拓维信息、浪潮信息、中科曙光、华勤技术等；4）算力云厂商，包括青云科技、优刻得、并行科技等；5）端侧硬件厂商，包括美格智能、移远通信、广和通、乐鑫科技、中科蓝讯、恒玄科技等。

风险提示：AI 产品落地不及预期，行业竞争加剧风险，技术研发进展不及预期。

分析师承诺：

本人已在中国证券业协会登记为证券分析师，本人承诺，以勤勉的职业态度，独立、客观地出具本报告。本人对证券研究报告的内容和观点负责，保证信息来源合法合规，研究方法专业审慎，分析结论具有合理依据。本报告清晰准确地反映本人的研究观点。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直接或间接受到任何形式的补偿。本人承诺不利用自己的身份、地位或执业过程中所掌握的信息为自己或他人谋取私利。

投资评级的说明：

以报告发布日后的 6--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；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。

无评级：因无法获取必要的资料，或者公司面临无法预见的结果的重大不确定事件，或者其他原因，致使无法给出明确的投资评级。（新股覆盖、新三板覆盖报告及转债报告默认无评级）

评级体系：

——公司评级

买入： 预计涨幅领先相对基准指数 15%以上；
增持： 预计涨幅领先相对基准指数介于 5%-15%之间；
中性： 预计涨幅领先相对基准指数介于-5%-5%之间；
减持： 预计涨幅落后相对基准指数介于-5%- -15%之间；
卖出： 预计涨幅落后相对基准指数-15%以上。

——行业评级

领先大市： 预计涨幅超越相对基准指数 10%以上；
同步大市： 预计涨幅相对基准指数介于-10%-10%之间；
落后大市： 预计涨幅落后相对基准指数-10%以上。

——风险评级

A： 预计波动率小于等于相对基准指数；
B： 预计波动率大于相对基准指数。

免责声明：

山西证券股份有限公司(以下简称“公司”)具备证券投资咨询业务资格。本报告是基于公司认为可靠的已公开信息，但公司不保证该等信息的准确性和完整性。入市有风险，投资需谨慎。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。在任何情况下，公司不对任何人因使用本报告中的任何内容引致的损失负任何责任。本报告所载的资料、意见及推测仅反映发布当日的判断。在不同时期，公司可发出与本报告所载资料、意见及推测不一致的报告。公司或其关联机构在法律许可的情况下可能持有或交易本报告中提到的上市公司发行的证券或投资标的，还可能为或争取为这些公司提供投资银行或财务顾问服务。客户应当考虑到公司可能存在可能影响本报告客观性的利益冲突。公司在知晓范围内履行披露义务。

依据《发布证券研究报告执业规范》规定特此声明，禁止公司员工将公司证券研究报告私自提供给未经公司授权的任何媒体或机构；禁止任何媒体或机构未经授权私自刊载或转发公司证券研究报告。刊载或转发公司证券研究报告的授权必须通过签署协议约定，且明确由被授权机构承担相关刊载或者转发责任。

依据《发布证券研究报告执业规范》规定特此提示公司证券研究业务客户不得将公司证券研究报告转发给他人，提示公司证券研究业务客户及公众投资者慎重使用公众媒体刊载的证券研究报告。

依据《证券期货经营机构及其工作人员廉洁从业规定》和《证券经营机构及其工作人员廉洁从业实施细则》规定特此告知公司证券研究业务客户遵守廉洁从业规定。

山西证券研究所：

上海
上海市浦东新区滨江大道 5159 号陆家嘴滨江中心 N5 座 3 楼

深圳
广东省深圳市福田区金田路 3086 号大百汇广场 43 层

太原
太原市府西街 69 号国贸中心 A 座 28 层
电话：0351-8686981
http://www.i618.com.cn

北京
北京市丰台区金泽西路 2 号院 1 号楼丽泽平安金融中心 A 座 25 层