

韩国科技行业

本文内容由 Alex Wu (吴俊豪) 提供

中信证券财富管理(香港)

产品及投资方案部

电话: (852) 2237 9250 / 电邮: wminvestmentsolutions@citics.com.hk

Deepseek 新模型效率提升

摘要

中信证券财富管理与中信里昂研究观点一致。根据中信里昂研究在 2026 年 9 月 13 日发布的题为《Deepseek new model efficiency gains》的报告，DeepSeek 于 9 月 10 日发布的最新模型 DeepSeek-V4.1-Flash，将基于 HBM 的 KV 缓存需求降低了 75%（至 890 bytes/token），并将 eSSD 缓存需求降低了 87.5%，这似乎对内存股构成了压力。需要注意的是，这一改进仅适用于 KV 缓存效率，并不代表 GPU 总内存或存储需求的削减。历史表明，对内存效率提升会降低内存需求的担忧是没有根据的。AI 模型一直在持续提高效率，DeepSeek 从 V1 到 V4 将每 token 内存效率提升了约 110x。然而，随着更低成本推动更广泛使用，内存需求仍在持续增长。随着 token 量激增（8 月 token 需求翻倍），模型效率的提升以及内存/计算性能的进步预计将继续推动整体内存需求。中信里昂维持对内存股的正面看法。

- 75%的降幅具体适用于推理期间的 KV 缓存，而不适用于模型权重、训练内存或更广泛的 HBM 占用。如果推理变得更便宜，开发者就会运行更多智能体、更长上下文以及更多推理工作负载。处理的 token 数量上升的速度可能远快于每 token 所需内存下降的速度。即使缓存压缩降低了容量需求，AI 模型仍然需要巨大的带宽在 GPU 与内存之间传输数据。因此，较低的容量强度并不一定会按比例转化为更低的 HBM 需求。
- 然而，这再次证明算法效率可以缓解限制超长上下文和智能体工作负载的内存容量瓶颈。压缩 KV 缓存占用意味着数据中心能够承载大得多的批处理规模、支持更长的上下文窗口，并服务显著更多的并发用户。
- 就像 Google 的 Turbo quant（2026 年 3 月）和 Deepseek 的早期版本（2025 年 1 月）一样，这降低了每次查询或每位用户的内存负担，但它往往会触发杰文斯悖论（Jevons Paradox）：随着推理变得更便宜且更高效，整体使用量、采用率和 API 调用量激增，这可以在很大程度上抵消单位硬件节省。
- 分析预计，软件创新将继续推动内存效率提升，投资者应预期会周期性看到宣称内存使用下降的新闻标题。然而，算力需求正呈指数级增长，上下文长度持续扩展，而智能体 AI 正大幅增加推理量。最终结果是，即使每 token 内存消耗下降，总内存需求也可能强劲上升。

免责声明

本文由中信证券财富管理（香港）（“**中信证券财富管理（香港）**”）撰写。中信证券财富管理（香港）是中信证券经纪（香港）有限公司（“**中信证券经纪（香港）**”）的财富管理品牌。本文为一般性资讯，仅供参考之用。本文具有教学性质，但不被视为对任何司法管辖区的任何特定投资产品、策略、计划特征或其他目的的建议或推荐，中信证券财富管理（香港）或其任何子公司也不承诺会参与任何此处提及的交易。任何示例都是通用的、假设的并且仅用于说明目的。本文尚未经香港证券及期货事务监察委员会审核。

本文是由投资／产品专员而非分析师撰写的文章汇编。它不构成研究报告，也不应被解释为研究报告，也不旨在提供专业、投资或任何其他类型的建议或推荐。本文件所载观点为投资／产品专员个人观点，或投资／产品专员对公司观点的理解，可能与中信证券财富管理（香港）的官方观点和利益（包括投资研究部门的观点）存在差异，也可能不存在差异。

适合性、风险及适当性

本文不含足够资讯以支持投资决策，并且不得依赖本文来评估投资任何证券或产品的优劣。本文仅提供一般资讯，未针对任何接收者量身定制，亦未考虑任何接收者的投资目标、投资组合持仓、策略、财务状况或需求。本文具有教育性质，并非中信证券经纪（香港）或其附属公司参与任何交易的建议、推荐或承诺。任何范例均为通用、假设及仅供参考之用。

读者应自行独立评估法律、监管、税务、信用及会计方面的影响，并与其财务专业人员共同决定本文提及的任何投资是否适合其个人目标。投资者应确保在进行任何投资前获取所有相关资讯。本文中所涉及的任何预测、数据、观点、技术或策略仅供参考，基于某些假设及当前市场状况，可能会在未通知的情况下发生变更。

信息来源被认为是可靠的，但不保证其准确性，也不对任何错误或遗漏承担责任。请注意，投资涉及风险，投资的价值和收益可能会根据市场情况和税收协议而波动，投资者可能无法收回投资的全部金额。过去的表现和收益率都不是当前和未来结果的可靠指标。

本文不应被解释为在任何司法管辖区向任何不合法接受此类邀请或招揽的人发出购买或出售任何证券的邀请或招揽。此处包含的信息均未在任何司法管辖区获得批准。

责任限制

中信证券财富管理（香港）对于您或任何第三方在使用此材料或其内容时遭受的任何损失或损害（包括但不限于直接损失、间接损失、或由此导致的间接损失、利润损失或机会损失）不承担任何责任。此材料是根据您的请求提供给您的，并适用于熟悉相关市场且愿意并能够承担投资相关风险的个人。在您充分了解所有涉及风险并独立确定该投资适合您之前，请勿进行任何交易。请注意，我们并非您指定的投资顾问。

法律主体声明

香港：本资料由中信证券经纪（香港）有限公司（中央编号 AAE879）在香港发布及分发，该公司受香港证券及期货事务监察委员会监管。

附属披露

本文可能包含先前发布的信息，包括由 CLSA Limited（中央编号 AAB893）发布的信息，该公司受香港证券及期货事务监察委员会监管。中信证券经纪（香港）与 CLSA Limited 是关联实体，均属于中信证券国际有限公司（“**CSI**”）的共同所有或控制之下。

中信证券经纪（香港）与 CSI 及其子公司（“**CSI 集团**”）内的其他实体之间可能存在潜在的利益冲突，这些冲突可能涉及多个领域，包括投资产品、企业咨询服务、证券交易以及提供给本文提及公司（们）的其他金融服务。此外，CSI 集团可能在本文讨论的公司（们）中持有利益或头寸。

中信证券经纪（香港）遵循一项整个集团的政策（可能会不时更新），该政策阐明了 CSI 集团实体如何管理或消除任何实际或潜在的利益冲突。此政策旨在确保 CSI 集团发布的任何研究报告或评论保持公正。

针对不同司法管辖区的声明

香港：根据任何时间下所适用的法律和法规，中信证券财富管理（香港）、其各自的关联公司或与中信证券财富管理（香港）有关联的公司或个人可能在发布本资料前使用了此处包含的信息，并且可能不时购买、出售或在所提及的任何证券或相关证券中拥有重大利益，或在目前或将来可能产生或曾经有业务或财务关系，或可能或已经向本文提及的实体、其顾问和／或任何其他关联方提供投资银行、资本市场和／或其他服务。因此，投资者应注意中信证券财富管理（香港）和／或其各自的关联公司或公司、或此等个人可能存在一项或多项利益冲突。

未经中信证券经纪（香港）有限公司事先书面授权，任何人不得以任何目的复制、分发或销售本文。

版权所有 2026。中信证券经纪（香港）有限公司。保留所有权利。