国金证券 SINOLINK SECURITIES

DeepSeek V4 发布，国产算力加速计算机行业研究

2026 年 04 月 25 日

买入（维持评级）

行业周报

证券研究报告

计算机组
分析师：李可夫（执业 S1130525120009） 分析师：刘高畅（执业 S1130525120005）
likefu@gjzq.com.cn liugaochang@gjzq.com.cn

DeepSeek V4 发布，国产算力加速

DeepSeek-V4 重磅发布——百万上下文普惠时代正式启幕

摘要：2026 年 4 月 24 日，DeepSeek 正式上线并开源 DeepSeek-V4 预览版，同步发布完整技术报告，推出 DeepSeek-V4-Pro 与 DeepSeek-V4-Flash 两个版本，分别具备 1.6 万亿总参数（49B 激活参数）与 2840 亿总参数（13B 激活参数）。该系列模型通过架构创新，将最大上下文长度提升至 100 万 Token，大幅降低计算与内存成本，其中 V4-Pro、V4-Flash 在百万 Token 场景下的单 Token 推理 FLOPs 及 KV Cache 占用较 V3.2 大幅下降。同时，V4-Pro-Max 模式在 Agent 能力、世界知识、推理性能上表现优异，接近世界顶级闭源模型水平，当前已应用于公司内部 Agentic Coding 场景；成本方面官方预计下半年昇腾点批量上市后将大幅下调价格推动技术普惠。

架构革新赋能效率跃升——从算法到基建的全链路创新

DeepSeek-V4 的高效能得益于多项架构创新，形成从算法到底层基建的全链路优化体系。算法层面，采用 CSA 与 HCA 融合的混合注意力机制，搭配流形约束超连接（mHC）与 Muon 优化器，既提升长上下文运算效率，又强化模型建模能力、加快训练收敛速度。底层基建方面，通过 MoE 模块一体化融合内核、领域专用语言 TileLang、FP4 量化感知训练等多项优化，实现计算、通信与内存访问的高效协同，降低内存占用与计算开销；训练与推理框架层面，通过自动微分模块拓展、异构 KV 缓存架构设计等策略，全方位提升训练与推理效率。这些创新不仅支撑了 DeepSeek-V4 的性能突破，其开源特性也为国内其他国产模型提供了可借鉴的技术范式，助力国内大模型整体水平提升。

国产算力协同适配——Day 0 双首发，助力产业自主可控

DeepSeek-V4 发布当日即实现寒武纪、华为昇腾两大国产芯片的 Day 0 适配，彰显了国产大模型与国产算力的深度协同能力。寒武纪基于 vLLM 推理框架完成适配并开源代码，通过自研融合算子库、高性能编程语言及多维度推理框架优化，深度挖掘硬件特性，充分释放模型推理潜能。华为昇腾超节点全系列产品全面支持该模型，昇腾 950 通过融合 kernel、多流并行等技术及底层架构升级，实现高吞吐、低时延部署；昇腾 A3 超节点也完成适配并提供训练参考实现。体现了 DeepSeek 与国产芯片厂商在技术预研、软硬协同上的深入合作，对推动国内大模型软硬件产业协同发展、加速 AI 算力生态自主可控具有重要意义。

投资建议

相关标的：

国内算力：寒武纪、东阳光、海光信息、利通电子、协创数据、浪潮信息、华勤技术、网宿科技、芯原股份、华丰科技、亿田智能、豫能控股、星环科技、首都在线、神州数码、百度集团、中芯国际、华虹半导体、中科曙光、润泽科技、大位科技、润建股份、奥飞数据、云赛智联、瑞晟智能、科华数据、潍柴重机、金山云、欧陆通、杰创智能。

海外算力/存储：胜宏科技、中际旭创、东山精密、欧科亿、天孚通信、天岳先进、新易盛、工业富联、兆易创新、大普微、源杰科技、景旺电子、英维克、唯科科技、领益智造等；Lumentum、闪迪、博通、marvell、铠侠、美光、SK 海力士、中微公司、北方华创、拓荆科技、长川科技。

CPU：海光信息、中科曙光、澜起科技、禾盛新材、中国长城、龙芯中科、兴森科技、深南电路、宏和科技、广合科技。

AI 应用：1）大模型&自定义 Agent：智谱、Minimax、腾讯控股、阿里巴巴、科大讯飞。2）星环科技、德才股份、美年健康、真爱美家、中控技术、金蝶国际、迪普科技、云知声、多点数智、聚水潭、迈富时、阜博集团、范式智能、汇量科技等 AI INFRA&高景气&高壁垒。其他：空天时代、具身智能等。

风险提示

行业竞争加剧的风险；技术研发进度不及预期的风险；特定行业下游资本开支周期性波动的风险。

内容目录

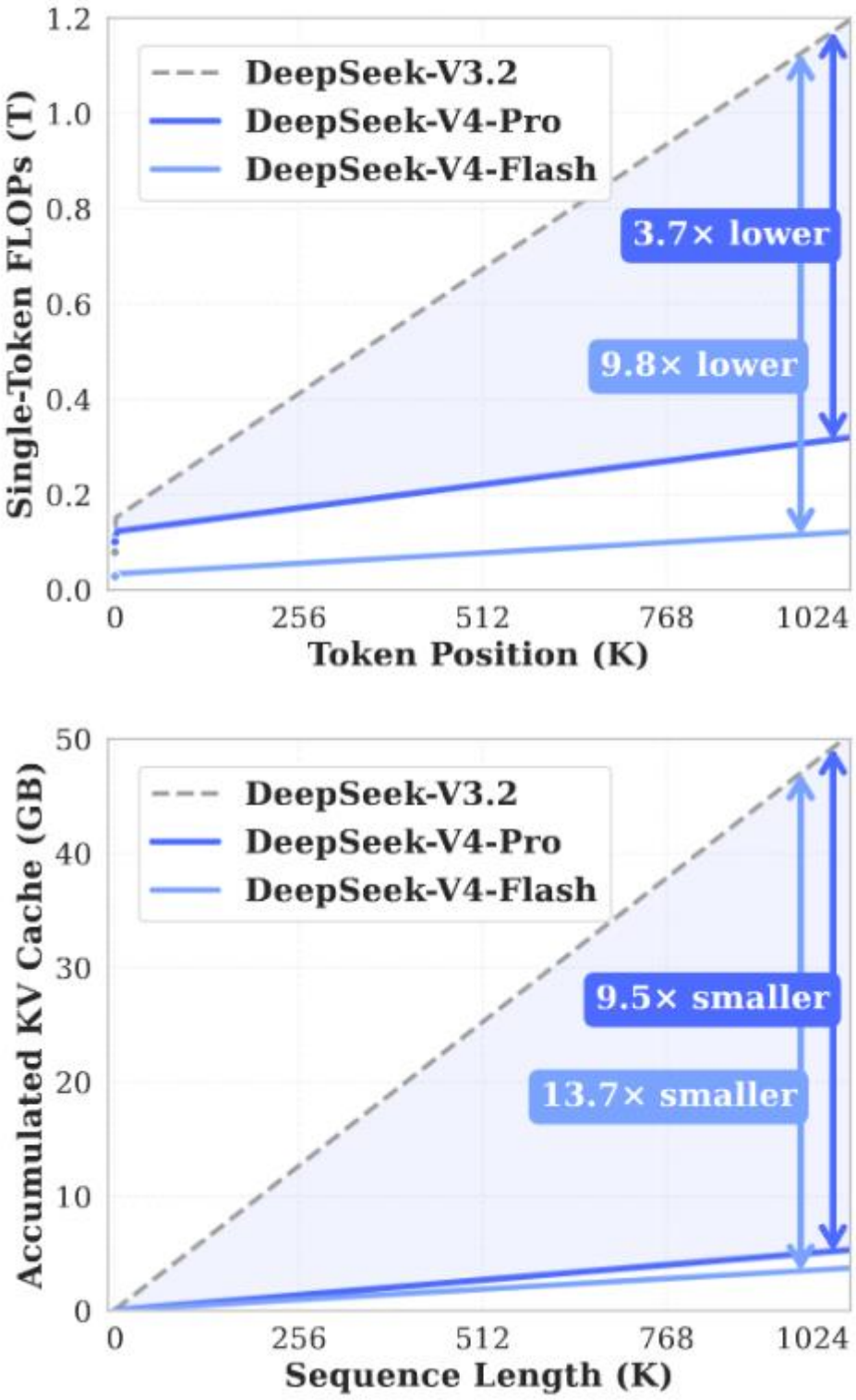
图表目录

DeepSeek-V4 概览：百万上下文普惠时代开启

2026 年 4 月 24 日，DeepSeek 正式上线并开源 DeepSeek-V4 预览版，同步发布完整技术报告。此次发布分为两个版本：DeepSeek-V4-Pro（1.6 万亿总参数，49B 激活参数）和 DeepSeek-V4-Flash（2840 亿总参数，13B 激活参数）

DeepSeek 通过架构创新大幅降低了计算和内存成本，V4-Pro 与 V4-Flash 最大上下文长度为 1M，技术报告数据显示，在 100 万 Token 场景下，相比 V3.2，V4-Pro 单 Token 推理 FLOPs 相比 V3.2 降低 3.7 倍，KV Cache 降低 9.5 倍；V4-Flash 进一步降低至 FLOPs 的 1/9.8、KV Cache 的 1/13.7。这意味着处理同等长度上下文的硬件成本大幅下降，使百万 Token 推理在商业环境中具备实际可行性

图表1：DeepSeek-V4 和 DeepSeek-V3.2 的计算量和显存容量随上下文长度的变化



来源：deepseek 技术报告，国金证券研究所

DeepSeek-V4-Pro-Max（最高推理强度模式）在多个维度的核心评测表现如下：

Agent 能力大幅提高：相比前代模型，DeepSeek-V4-Pro 的 Agent 能力显著增强。在 Agentic Coding 评测中，V4-Pro 已达到当前开源模型最佳水平，并在其他 Agent 相关评测中同样表现优异。目前 DeepSeek-V4 已成为公司内部员工使用的 Agentic Coding

模型，据评测反馈使用体验优于 Sonnet 4.5，交付质量接近 Opus 4.6 非思考模式，但仍与 Opus 4.6 思考模式存在一定差距。

丰富的世界知识：DeepSeek-V4-Pro 在世界知识测评中，大幅领先其他开源模型，仅稍逊于顶尖闭源模型 Gemini-Pro-3.1。

世界顶级推理性能：在数学、STEM、竞赛型代码的测评中，DeepSeek-V4-Pro 超越当前所有已公开评测的开源模型，取得了比肩世界顶级闭源模型的优异成绩。

成本方面，DeepSeek-V4-Pro 定价输入百万 token 输入成本为 1 元（缓存命中）/12（缓存未命中），百万 token 输出成本为 24 元。deepseek 官方表示：受限于高端算力，目前 Pro 的服务吞吐十分有限，预计下半年昇腾 950 超节点批量上市后，Pro 的价格会大幅下调。

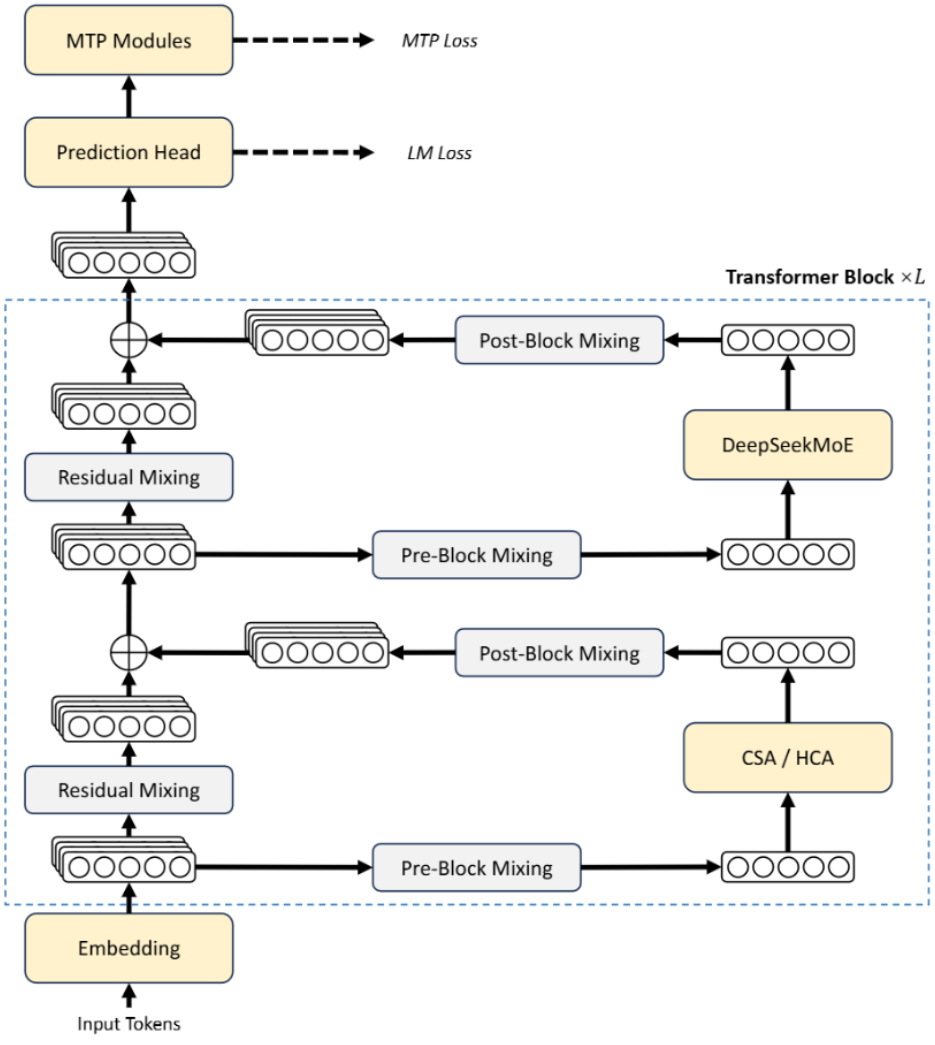
架构革新赋能效率跃升——从算法到基建的全链路创新

2.1 混合注意力机制：CSA + HCA

为提升长上下文运算效率，团队设计融合压缩稀疏注意力（CSA）与高强度压缩注意力（HCA）的混合注意力机制：CSA 沿序列维度压缩 KV 缓存，再执行 DeepSeek 稀疏注意力（DSA）；HCA 则采用更高压缩率处理 KV 缓存，同时保留稠密注意力计算。

为强化模型建模能力，引入流形约束超连接（mHC），对传统残差连接完成全面升级。此外，训练环节引入 Muon 优化器，有效加快收敛速度、提升训练稳定性。

图表2：模型架构示意图



来源：deepseek 技术报告，国金证券研究所

2.2 多项底层基建优化：

为 MoE 模块设计一体化融合内核，实现计算、通信与内存访问的完全重叠；

采用领域专用语言 TileLang，平衡开发效率与运行时性能；

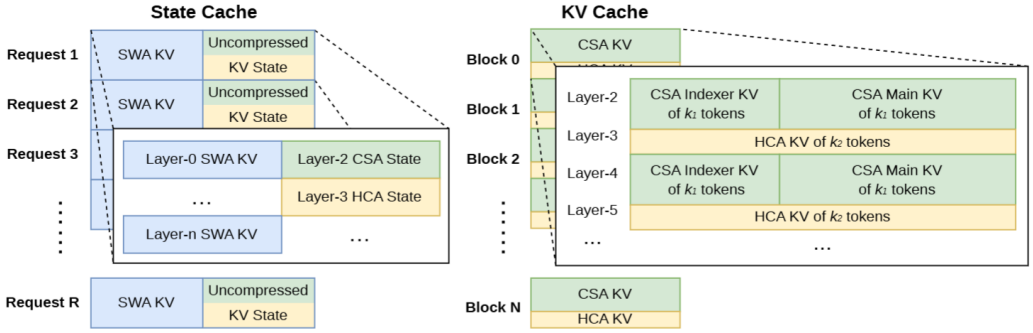
提供批量无关、可确定性内核库，保障训练与推理全程位级可复现；

针对 MoE 专家权重与索引器 QK 计算路径，落地 FP4 量化感知训练，降低内存占用与计算开销；

训练框架层面：拓展自动微分模块，支持张量级检查点，实现精细化重计算控制；搭配适配 Muon 优化器的混合 ZeRO 策略、基于重计算与融合内核的低成本 mHC 部署方案、面向压缩注意力的双阶段上下文并行策略，全方位提升训练效率；

推理框架层面：设计异构 KV 缓存架构，结合磁盘存储策略，实现共享前缀的高效复用。

图表3：V4 的 KV 缓存布局由两个主要部分构成：用于 CSA/HCA 的经典 KV 缓存，以及用于 SWA 和 CSA/HCA 中尚未准备好压缩的 token 的状态缓存



来源：deepseek 技术报告，国金证券研究所

我们认为，作为开源模型，DeepSeek V4 的各项技术创新可以被其他国产模型学习，有利于国内整体大模型水平的提高。

国产算力协同适配——Day 0 双首发，助力产业自主可控

3.1 寒武纪

DeepSeek-V4 发布当日，寒武纪已基于 vLLM 推理框架完成 Day 0 适配，代码开源到 GitHub 社区。这一成果得益于寒武纪长期积累的自研 NeuWare 软件生态与芯片设计技术，也是寒武纪对芯片与算法联合创新持续投入的延续。此前，寒武纪已对 DeepSeek 系列模型开展深入的软硬件协同性能优化，达成业界领先的算力利用率水平。

极致性能优化，释放 DeepSeek-V4 推理潜能

针对 DeepSeek-V4 的新结构，寒武纪通过自研高性能融合算子库 Torch-MLU-Ops，对 Compressor、mHC 等模块进行专项加速；利用 BangC 高性能编程语言，编写稀疏/压缩 Attention、GroupGemm 等热点算子的极致优化 Kernel，充分释放硬件底层性能。

在推理框架优化层面，寒武纪在 vLLM 中全面支持 TP/PP/SP/DP/EP 5D 混合并行、通信计算并行、低精度量化以及 PD 分离部署等优化技术，通过策略优化，在满足延时约束下达到最佳的词元吞吐能力，显著提升端到端推理效率。

硬件特性同样被深度挖掘：利用 MLU 访存与排序加速能力，有效加速稀疏 Attention、Indexer 等结构；高互联带宽与低通信延时，将 Prefill 和 Decode 两种不同工作负载场景下的通信占比降至最低，最大化分布式推理的利用率。

3.2 华为昇腾

同日，昇腾官方宣布通过与 deepseek 双方芯模技术紧密协同，实现昇腾超节点全系列产品支持 DeepSeek V4 系列模型。昇腾 950 通过融合 kernel 和多流并行技术降低 Attention 计算和访存开销，大幅提升推理性能，结合多种量化算法，实现了高吞吐、低时延的 DeepSeek V4 模型推理部署。昇腾 A3 超节点系列产品也全面适配，同时为便于用户快速微调，提供了基于昇腾 A3 集群的**训练**参考实现。，

华为公布了适配 DeepSeek-V4 的具体性能指标，具备重要参考价值：

基于 DeepSeek V4-Pro 模型，在 8K 输入场景，昇腾 950 超节点可实现 TPOT 约 20ms 时单卡 Decode 吞吐 4700TPS。DeepSeek V4模型，8K 长序列输入场景下可实现 TPOT 约 10ms 时单卡 Decode 吞吐 1600TPS。

极低时延的实现源于昇腾 950 代际底层架构的三大升级：

原生精度加速： 全面支持 FP8 、MXFP8、MXFP4 等数据格式，在保证模型精度的同时，可实现内存占用降低 50%+，计算能力翻倍。

稀疏访存优化： 针对 MoE 模型的离散访存特征，通过大幅提升硬件级稀疏访存能力，有效解决了专家路由过程中的带宽瓶颈。

增强 Vector 与 Cube 间的数据通路： 创新的存储架构设计，实现了向量单元（Vector）与矩阵单元（Cube）的 Memory 通路，极大地降低了端到端推理时延。

我们认为，Day0 适配说明 DeepSeek 与国产芯片厂商在技术预研、软硬协同和测试流程上已形成深入合作机制，有利于国内大模型软硬件产业的协同发展。

投资建议

相关标的：

国内算力：寒武纪、东阳光、海光信息、利通电子、协创数据、浪潮信息、华勤技术、网宿科技、芯原股份、华丰科技、亿田智能、豫能控股、星环科技、首都在线、神州数码、百度集团、中芯国际、华虹半导体、中科曙光、润泽科技、大位科技、润建股份、奥飞数据、云赛智联、瑞晟智能、科华数据、潍柴重机、金山云、欧陆通、杰创智能。

海外算力/存储：胜宏科技、中际旭创、东山精密、欧科亿、天孚通信、天岳先进、新易盛、工业富联、兆易创新、大普微、源杰科技、景旺电子、英维克、唯科科技、领益智造等；Lumentum、闪迪、博通、marvell、铠侠、美光、SK 海力士、中微公司、北方华创、拓荆科技、长川科技。

CPU：海光信息、中科曙光、澜起科技、禾盛新材、中国长城、龙芯中科、兴森科技、深南电路、宏和科技、广合科技。

AI 应用：1）大模型&自定义 Agent：智谱、Minimax、腾讯控股、阿里巴巴、科大讯飞。2）星环科技、德才股份、美年健康、真爱美家、中控技术、金蝶国际、迪普科技、云知声、多点数智、聚水潭、迈富时、卓博集团、范式智能、汇量科技等 AI INFRA&高景气&高壁垒。其他：空天时代、具身智能等。

风险提示

行业竞争加剧的风险： 在信创等政策持续加码支持计算机行业发展的背景下，众多新兴玩家参与到市场竞争之中， 若市场竞争进一步加剧，竞争优势偏弱的企业或面临出清，某些中低端品类的毛利率或受 到一定程度影响。
技术研发进度不及预期的风险： 计算机行业技术开发需投入大量资源，如果相关厂商新品研发进程不及预期，表观层面将 呈现出投入产出在较长时期的滞后特征。
特定行业下游资本开支周期性波动的风险： 部分计算机公司系顺周期行业，下游资本开支波动与行业周期性相关性较强，或在个别年 份对于上游软件厂商的营收表现产生扰动。

。

行业投资评级的说明：

买入：预期未来 3—6 个月内该行业上涨幅度超过大盘在 15%以上；

增持：预期未来 3—6 个月内该行业上涨幅度超过大盘在 5%—15%；

中性：预期未来 3—6 个月内该行业变动幅度相对大盘在 -5%—5%；

减持：预期未来 3—6 个月内该行业下跌幅度超过大盘在 5%以上。

特别声明：

国金证券股份有限公司经中国证券监督管理委员会批准，已具备证券投资咨询业务资格。

本报告的产生基于国金证券及其研究人员认为可信的公开资料或实地调研资料，但国金证券及其研究人员对这些信息的准确性和完整性不作任何保证。本报告反映撰写研究人员的不同设想、见解及分析方法，故本报告所载观点可能与其他类似研究报告的观点及市场实际情况不一致，国金证券不对使用本报告所包含的材料产生的任何直接或间接损失或与此有关的其他任何损失承担任何责任。且本报告中的资料、意见、预测均反映报告初次公开发布时的判断，在不作事先通知的情况下，可能会随时调整，亦可因使用不同假设和标准、采用不同观点和分析方法而与国金证券其它业务部门、单位或附属机构在制作类似的其他材料时所给出的意见不同或者相反。

本报告仅为参考之用，在任何地区均不应被视为买卖任何证券、金融工具的要约或要约邀请。本报告提及的任何证券或金融工具均可能含有重大的风险，可能不易变卖以及不适合所有投资者。本报告所提及的证券或金融工具的价格、价值及收益可能会受汇率影响而波动。过往的业绩并不能代表未来的表现。

客户应当考虑到国金证券存在可能影响本报告客观性的利益冲突，而不应视本报告为作出投资决策的唯一因素。证券研究报告是用于服务具备专业知识的投资者和投资顾问的专业产品，使用时必须经专业人士进行解读。国金证券建议获取报告人员应考虑本报告的任何意见或建议是否符合其特定状况，以及（若有必要）咨询独立投资顾问。报告本身、报告中的信息或所表达意见也不构成投资、法律、会计或税务的最终操作建议，国金证券不就报告中的内容对最终操作建议做出任何担保，在任何时候均不构成对任何人的个人推荐。

在法律允许的情况下，国金证券的关联机构可能会持有报告中涉及的公司所发行的证券并进行交易，并可能为这些公司正在提供或争取提供多种金融服务。

本报告并非意图发送、发布给在当地法律或监管规则下不允许向其发送、发布该研究报告的人员。国金证券并不因收件人收到本报告而视其为国金证券的客户。本报告对于收件人而言属高度机密，只有符合条件的收件人才能使用。根据《证券期货投资者适当性管理办法》，本报告仅供国金证券股份有限公司客户中风险评级高于 C3 级（含 C3 级）的投资者使用；本报告所包含的观点及建议并未考虑个别客户的特殊状况、目标或需要，不应被视为对特定客户关于特定证券或金融工具的建议或策略。对于本报告中提及的任何证券或金融工具，本报告的收件人须保持自身的独立判断。使用国金证券研究报告进行投资，遭受任何损失，国金证券不承担相关法律责任。

若国金证券以外的任何机构或个人发送本报告，则由该机构或个人为此发送行为承担全部责任。本报告不构成国金证券向发送本报告机构或个人的收件人提供投资建议，国金证券不为此承担任何责任。

上海
电话：021-80234211
邮箱：researchsh@gjzq.com.cn
邮编：201204
地址：上海浦东新区芳甸路 1088 号紫竹国际大厦 5 楼

北京
电话：010-85950438
邮箱：researchbj@gjzq.com.cn
邮编：100005
地址：北京市东城区建内大街 26 号新闻大厦 8 层南侧

深圳
电话：0755-86695353
邮箱：researchsz@gjzq.com.cn
邮编：518000
地址：深圳市福田区金田路 2028 号皇岗商务中心 18 楼 1806

【小程序】
国金证券研究服务

【公众号】
国金证券研究