浅谈 DeepSeek 的创新性，对比科技巨头和 OpenAI

华泰研究

2025 年 1 月 31 日 | 美国　　动态点评

互联网　　增持（维持）

研究员　何翩翩
SAC No. S0570523020002　purdyho@htsc.com
SFC No. ASI353　+(852) 3658 6000

联系人　易楚妍
SAC No. S0570124070123　yichuyan@htsc.com
+(86) 21 2897 2228

低成本敲响美国科技界“警钟”，DS 引发科技股抛售，掀起 AI 主导权之争

DeepSeek（DS）的低开发成本引发全球投资者对美国科技巨头高成本投资的质疑，以及对算力与相关产业的担忧，1 月 27 日纳指出现拥挤性抛售。英伟达、ARM 和博通股价下跌 17.0/10.2/17.4%。知名科技投资人 Marc Andreessen 认为 DS 是一项重大突破，并引发各界对美国在 AI 领域主导地位的质疑。而特朗普更将 DS 描述为对美国科技产业敲响了“警钟”。从科技界看，Meta 首席科学家 Yann LeCun 认为 DS 受益于开源生态，并认为开源模型正在赶超闭源模型，Meta 的 Llama 同属开源。Scale AI CEO Alexandr Wang 表示 DS 的模型“惊天动地”，性能高且大致与美国最好的模型相当，并认为中美之间的 AI 竞赛加剧。OpenAI CEO Sam Altman 称赞 R1 的高性价比，但表示其将推出的改进模型或再次引领行业发展。不过，据彭博报道，微软和 OpenAI 正在调查 DS 是否以未经授权方式“蒸馏”了 OpenAI 的数据输出作为其训练依据。英伟达也于 1 月 31 日宣布 DS-R1 模型现已在 NVIDIA NIM 微服务预览版上提供。该微服务在单个 HGX H200 系统上每秒最多可提供 3872 个 Tokens。

对比科技巨头和 OpenAI 的模型，DS 是否真正创新？

我们认为 DS 的 R1 模型主要是在现有技术路线的基础上创新，并进行深度优化和改进。这是学术研究惯用的研究方法，也反映了 AI 发展中开源的趋势。Meta CEO 扎克伯格在 24Q4 财报里表示，DS 的崛起“只会加强我们对 Meta AI 战略和投资的信心”。他强调须建立美国 AI 技术标准，尤其在面对开源模型带来的全球竞争。他也承认 DS 采用的几种创新算法可更有效、更经济地训练模型，Meta 目前正评估并考虑将一些算法整合到自己的模型中。DS 于 2024 年推出 V1-V3 模型迭代，主要基于 MOE（专家混合模型，Mixture of Experts）和 MLA（多头潜在注意力，Multi-head Latent Attention）算法，以解决 AI 计算的两大瓶颈：内存与算力。

DeepSeek 的 MoE 算法能如何降低算力需求？

MOE 是神经网络之父 Geoff Hinton 于 90 年代提出，目前相关算法已被 Mistral、谷歌、腾讯、OpenAI 等广泛应用。MOE 将大型模型划分为多个专门处理特定任务或数据的小型子模型。而每个子模型仅在其特定知识相关时才被激活。尽管 DS 的 V3 总共有 6710 亿个参数，但实际上一次只使用 370 亿个参数。而 R1 是以 V3 基础，通过强化学习实现高效推理。传统 MOE 算法（如谷歌 GShard）通过激活不同专家来处理任务，但较难确保专家获得的知识不重叠。而 DS 的 MOE 特点在于使用了：1）细粒度专家分割（Fine-grained Expert Segmentation）将专家划分为更小单元以更灵活激活专家组合；2）Shared Experts Isolation 将部分专家设定为共享专家，以捕捉整合上下文的共同知识，降低其他专家中的参数冗余。相比之下，其他公司虽也有使用类似算法，但主要集中于提升单一专家能力。

DeepSeek 的 MLA 算法如何降低所需内存？

MLA 是对于谷歌 GQA（分组查询注意力，Grouped Query Attention）方法的深度改进，通过数学变换优化计算和内存使用，关键在于合并和简化矩阵运算。谷歌算法的基本原理为共享 KV 矩阵，从而减少计算量和内存使用。而 Meta 算法在于探索稀疏 KV 缓存路线，以减少 KV 缓存的体积和计算复杂度。然而，MLA 的特点在于不直接存储 KV 矩阵，而是仅存储经过合并吸收后的低秩压缩向量，从而减少内存占用。R1 也采用的 MTP（多标记预测，Multi-Token Prediction）可同时预测多个 Token，减少 KV 缓存的访问次数，提高复杂任务的整体性能。

风险提示：大模型技术研发进展不及市场预期，贸易科技摩擦风险。

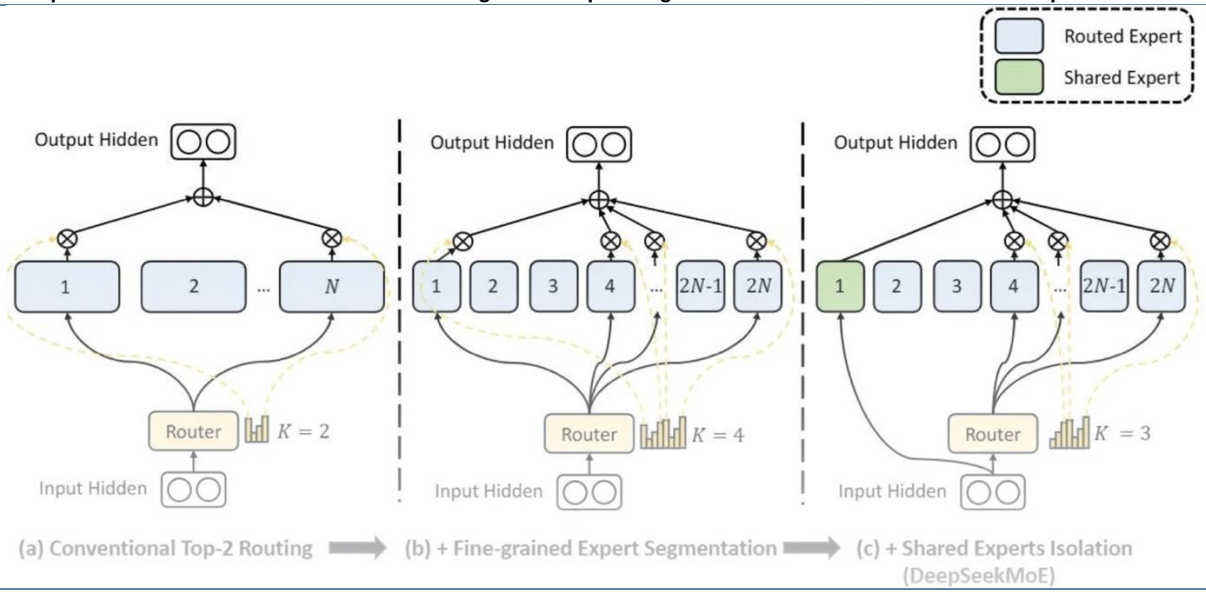
DS 的 MOE 特点详解

1）细粒度专家分割（Fine-grained Expert Segmentation）将专家划分为更小单元以更灵活激活专家组合，实现更精细的任务分解和专业化。DS 采用 256 个专家，而 Mistral、腾讯和 OpenAI 分别采用 8/16/16 个，因此每次任务激活的相应参数减少。2）Shared Experts Isolation 将部分专家设定为共享专家，以捕捉整合上下文的共同知识，降低其他专家中的参数冗余。相比之下，其他公司虽也有使用类似算法，但主要集中于提升单一专家能力，例如 Mistral 每次只激活两个专家，以增强该专家对特定问题的理解、腾讯则探索根据每个专家的能力，动态分配 Token，赋予每个专家不同表达能力。R1 是以 V3 模型为基础，通过强化学习实现高效推理。不同于传统方法依赖 SFT（监督微调，Supervised Fine-tuning）与 PPO（近段策略优化，Proximal Policy Optimization，由 OpenAI 提出）以单个输出作为交互单位，DS 的核心在于 GRPO（群组相对策略优化，Group Relative Policy Optimization）对多个输出作为交互整体进行评分，优势包括：1）不依赖独立的价值函数模型，从而简化训练过程并降低内存消耗；2）以群体输出的平均奖励作为基准，更契合奖励模型的训练需求，从而减少对复杂价值模型的依赖，提高强化学习训练效率。DS 还开发了一种独特的负载均衡策略（Load Bearing Strategy）通过使用更动态的调整，而不是用可导致性能下降的传统基于惩罚的方法，来确保没有任何一个专家工作过载或欠载。DS 还使用了一种称为推理时计算缩放（Inference Time Compute Scaling）的技术，允许模型根据手头的任务向上或向下调整其计算工作量，而不是始终以全功率运行。

DS 的 MLA 特点详解

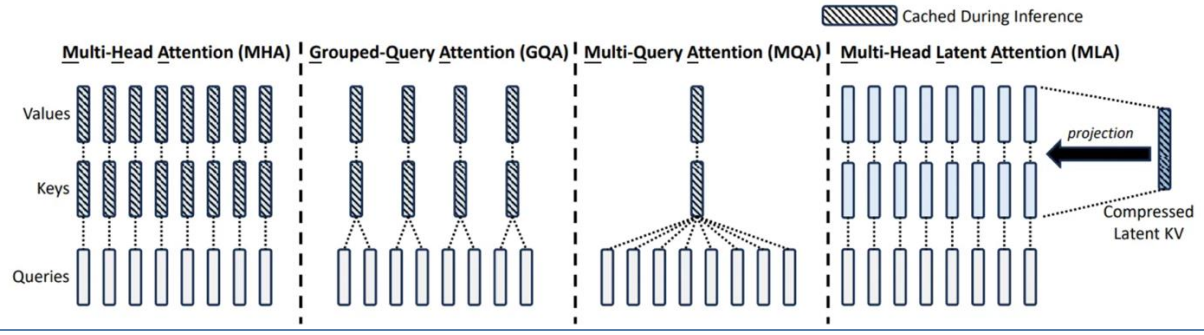
传统 Transformer 中使用的 MHA（多头注意力机制，Multi Head Attention），随着前置序列变长，需存储完整的 Key-Value（KV）矩阵也将变长，内存占用大增。谷歌于 2019 年、2023 年分别提出 MQA（多查询注意力，Multi Query Attention）与 GQA，基本原理均为共享 KV 矩阵，从而减少计算量和内存使用。然而，MLA 的特点在于不直接存储 KV 矩阵，而是仅存储经过合并吸收后的低秩压缩向量，从而减少内存占用。除共享机制外，Meta 联合 CMU 探索稀疏 KV 缓存路线，核心思想在于仅储存关键 KV 矩阵，以减少 KV 缓存的体积和计算复杂度。R1 模型采用的 MTP（多标记预测，Multi-Token Prediction）可同时预测多个 Token，减少 KV 缓存的访问次数，提高复杂任务的整体性能。主流大模型采用 Decoder-Based 结构，在训练和推理时均以 Token-by-Token 方式生成序列。每次生成新标记都需频繁访问 KV 缓存，并执行多层前向计算。而 MTP 通过同时预测多个 Token，可提高信号密度，减少上下文漂移和重复内存读取与计算步骤，从而提升数学、代码生成和文本摘要等任务的效率。相较 Meta 于 2024 年 4 月提出的 MTP 方法，R1 引入因果链（Causal Chain）的连接关系，并在 Embedding 层增加残差链接，从而提高信息流的传递效率。

图表1：DeepSeek MoE 架构采用细粒度专家分割（Fine-grained Expert Segmentation）以及共享专家（Shared Experts Isolation）方式



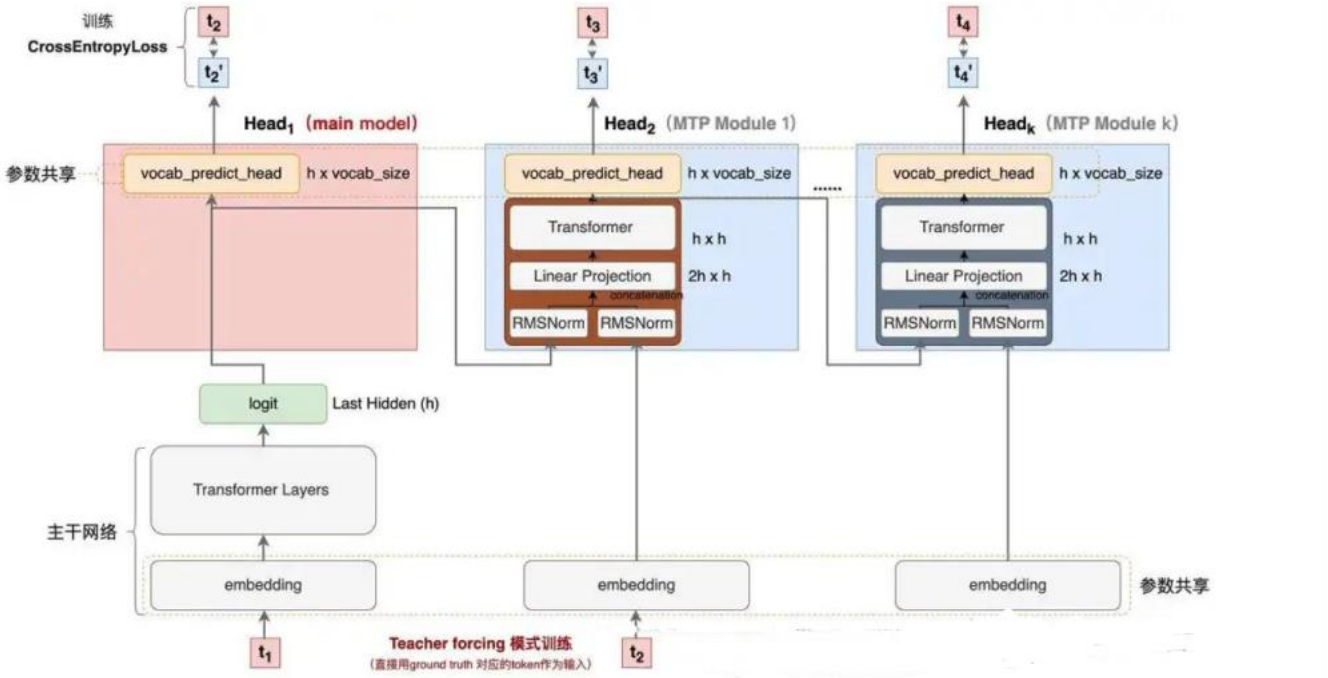
资料来源：DeepSeek 2024 年 1 月论文《DeepSeekMoE：Towards Ultimate Expert Specialization in Mixture-of-Experts Language Models》、Medium、华泰研究

图表2：MHA（Google）、GQA（Google）、MQA（Google）、MLA（DeepSeek）在 KV 缓存处理中的作用示意图



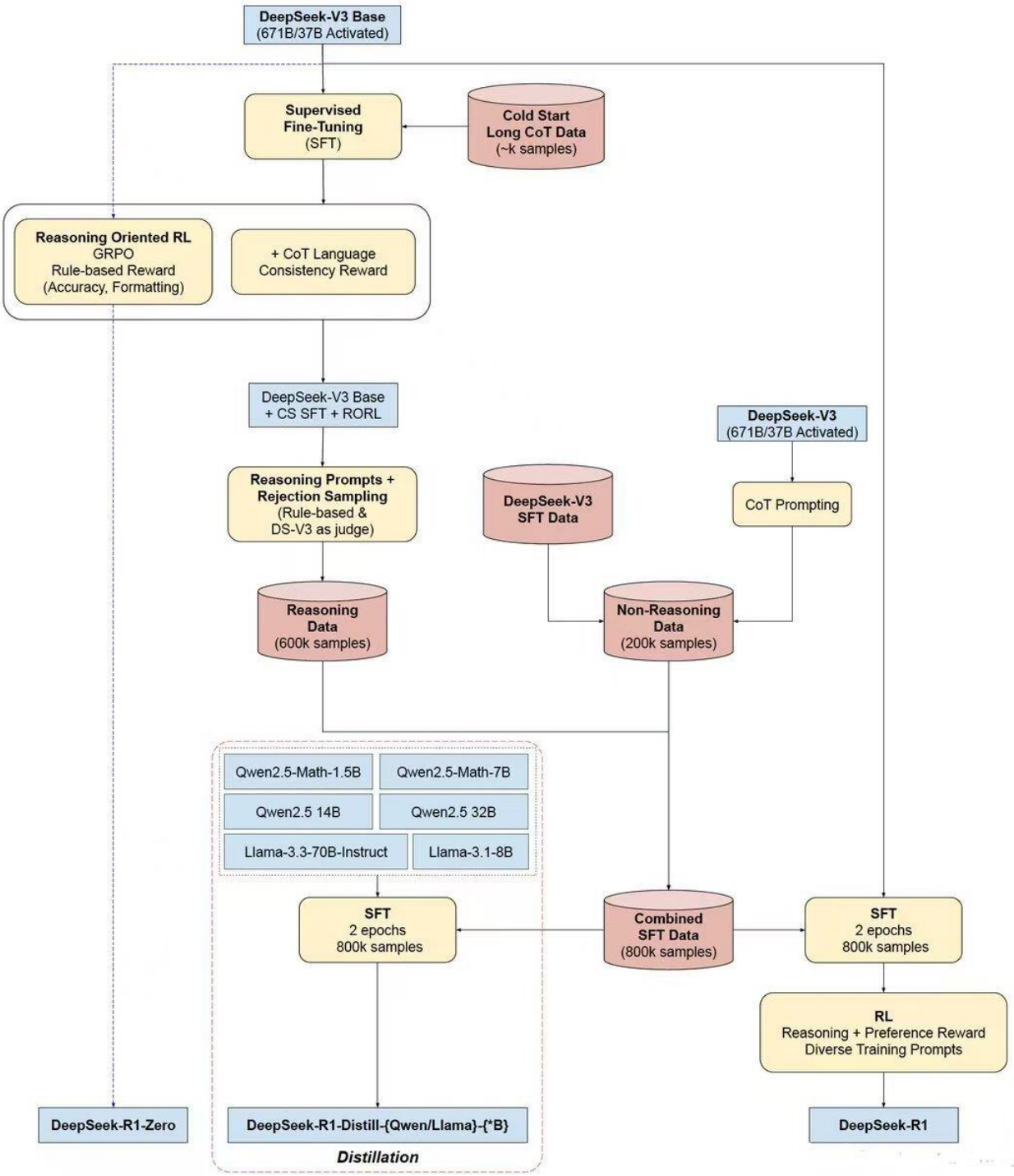
资料来源：DeepSeek 2024 年 5 月论文《DeepSeek-V2：A Strong, Economical, and Efficient Mixture-of-Experts Language Model》、Medium、华泰研究

图表3：DeepSeek R1 MTP（Multi-Token Prediction）示意图



注：此图仅为 DeepSeek R1 模型 MTP 方法的示意图，实际操作步骤可能与示意图存在差异。
资料来源：DataFunTalk 公众号、华泰研究；

图表4：DeepSeek R1-Training Pipeline 示意图



注：此图仅为 DeepSeek R1 模型训练流程的示意图，实际训练步骤可能与示意图存在差异。

资料来源：DeepSeek 2025 年 1 月论文《DeepSeek-R1：Incentivizing Reasoning Capability in LLMs via Reinforcement Learning》、CSDN、Reddit、华泰研究

风险提示

大模型技术研发进展不及市场预期：大模型研发具有较高的不确定性，可能因技术突破受阻、算法优化困难、计算资源不足等因素导致进展缓慢不及市场预期。

贸易科技摩擦风险：若中美贸易与科技摩擦风险加剧，或将导致 DeepSeek 数据使用遭受审查，对于公司产品迭代造成潜在负面影响。

本研报中涉及到未上市公司或未覆盖个股内容，均系对其客观公开信息的整理，并不代表本研究团队对该公司、该股票的推荐或覆盖。

免责声明

分析师声明

本人，何翩翩，兹证明本报告所表达的观点准确地反映了分析师对标的证券或发行人的个人意见；彼以往、现在或未来并无就其研究报告所提供的具体建议或所表达的意见直接或间接收取任何报酬。

一般声明及披露

本报告由华泰证券股份有限公司（已具备中国证监会批准的证券投资咨询业务资格，以下简称“本公司”）制作。本报告所载资料是仅供接收人的严格保密资料。本报告仅供本公司及其客户和其关联机构使用。本公司不因接收人收到本报告而视其为客户。

本报告基于本公司认为可靠的、已公开的信息编制，但本公司及其关联机构(以下统称为“华泰”)对该等信息的准确性及完整性不作任何保证。

本报告所载的意见、评估及预测仅反映报告发布当日的观点和判断。在不同时期，华泰可能会发出与本报告所载意见、评估及预测不一致的研究报告。同时，本报告所指的证券或投资标的的价格、价值及投资收入可能会波动。以往表现并不能指引未来，未来回报并不能得到保证，并存在损失本金的可能。华泰不保证本报告所含信息保持在最新状态。华泰对本报告所含信息可在不发出通知的情形下做出修改，投资者应当自行关注相应的更新或修改。

本公司不是 FINRA 的注册会员，其研究分析师亦没有注册为 FINRA 的研究分析师/不具有 FINRA 分析师的注册资格。

华泰力求报告内容客观、公正，但本报告所载的观点、结论和建议仅供参考，不构成购买或出售所述证券的要约或招揽。该等观点、建议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，在任何时候均不构成对客户私人投资建议。投资者应当充分考虑自身特定状况，并完整理解和使用本报告内容，不应视本报告为做出投资决策的唯一因素。对依据或者使用本报告所造成的一切后果，华泰及作者均不承担任何法律责任。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。

除非另行说明，本报告中所引用的关于业绩的数据代表过往表现，过往的业绩表现不应作为日后回报的预示。华泰不承诺也不保证任何预示的回报会得以实现，分析中所做的预测可能是基于相应的假设，任何假设的变化可能会显著影响所预测的回报。

华泰及作者在自身所知情的范围内，与本报告所指的证券或投资标的不存在法律禁止的利害关系。在法律许可的情况下，华泰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，为该公司提供投资银行、财务顾问或者金融产品等相关服务或向该公司招揽业务。

华泰的销售人员、交易人员或其他专业人士可能会依据不同假设和标准、采用不同的分析方法而口头或书面发表与本报告意见及建议不一致的市场评论和/或交易观点。华泰没有将此意见及建议向报告所有接收者进行更新的义务。华泰的资产管理部门、自营部门以及其他投资业务部门可能独立做出与本报告中的意见或建议不一致的投资决策。投资者应当考虑到华泰及/或其相关人员可能存在影响本报告观点客观性的潜在利益冲突。投资者请勿将本报告视为投资或其他决定的唯一信赖依据。有关该方面的具体披露请参照本报告尾部。

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、发布的机构或人员，也并非意图发送、发布给因可得到、使用本报告的行为而使华泰违反或受制于当地法律或监管规则的机构或人员。

中国香港

本报告由华泰证券股份有限公司制作,在香港由华泰金融控股（香港）有限公司向符合《证券及期货条例》及其附属法律规定的机构投资者和专业投资者的客户进行分发。华泰金融控股（香港）有限公司受香港证券及期货事务监察委员会监管，是华泰国际金融控股有限公司的全资子公司，后者为华泰证券股份有限公司的全资子公司。在香港获得本报告的人员若有任何有关本报告的问题,请与华泰金融控股（香港）有限公司联系。

香港-重要监管披露

- 华泰金融控股（香港）有限公司的雇员或其关联人士没有担任本报告中提及的公司或发行人的高级人员。
- 有关重要的披露信息，请参华泰金融控股（香港）有限公司的网页 https://www.htsc.com.hk/stock_disclosure

其他信息请参见下方 **“美国-重要监管披露”**。

美国

在美国本报告由华泰证券（美国）有限公司向符合美国监管规定的机构投资者进行发表与分发。华泰证券（美国）有限公司是美国注册经纪商和美国金融业监管局（FINRA）的注册会员。对于其在美国分发的研究报告，华泰证券（美国）有限公司根据《1934 年证券交易法》（修订版）第 15a-6 条规定以及美国证券交易委员会人员解释，对本研究报告内容负责。华泰证券（美国）有限公司联营公司的分析师不具有美国金融监管（FINRA）分析师的注册资格，可能不属于华泰证券（美国）有限公司的关联人员，因此可能不受 FINRA 关于分析师与标的公司沟通、公开露面和所持交易证券的限制。华泰证券（美国）有限公司是华泰国际金融控股有限公司的全资子公司，后者为华泰证券股份有限公司的全资子公司。任何直接从华泰证券（美国）有限公司收到此报告并希望就本报告所述任何证券进行交易的人士，应通过华泰证券（美国）有限公司进行交易。

美国-重要监管披露

- 分析师何翩翩本人及相关人士并不担任本报告所提及的标的证券或发行人的高级人员、董事或顾问。分析师及相关人士与本报告所提及的标的证券或发行人并无任何相关财务利益。本披露中所提及的“相关人士”包括 FINRA 定义下分析师的家庭成员。分析师根据华泰证券的整体收入和盈利能力获得薪酬，包括源自公司投资银行业务的收入。
- 华泰证券股份有限公司、其子公司和/或其联营公司，及/或不时会以自身或代理形式向客户出售及购买华泰证券研究所覆盖公司的证券/衍生工具，包括股票及债券（包括衍生品）华泰证券研究所覆盖公司的证券/衍生工具，包括股票及债券（包括衍生品）。
- 华泰证券股份有限公司、其子公司和/或其联营公司，及/或其高级管理层、董事和雇员可能会持有本报告中所提到的任何证券（或任何相关投资）头寸，并可能不时进行增持或减持该证券（或投资）。因此，投资者应该意识到可能存在利益冲突。

新加坡

华泰证券（新加坡）有限公司持有新加坡金融管理局颁发的资本市场服务许可证，可从事资本市场产品交易，包括证券、集体投资计划中的单位、交易所交易的衍生品合约和场外衍生品合约，并且是《财务顾问法》规定的豁免财务顾问，就投资产品向他人提供建议，包括发布或公布研究分析或研究报告。华泰证券（新加坡）有限公司可能会根据《财务顾问条例》第 32C 条的规定分发其在华泰内的外国附属公司各自制作的信息/研究。本报告仅供认可投资者、专家投资者或机构投资者使用，华泰证券（新加坡）有限公司不对本报告内容承担法律责任。如果您是非预期接收者，请您立即通知并直接将本报告返回给华泰证券（新加坡）有限公司。本报告的新加坡接收者应联系您的华泰证券（新加坡）有限公司关系经理或客户主管，了解来自或与所分发的信息相关的事宜。

评级说明

投资评级基于分析师对报告发布日后 6 至 12 个月内行业或公司回报潜力（含此期间的股息回报）相对基准表现的预期（A 股市场基准为沪深 300 指数，香港市场基准为恒生指数，美国市场基准为标普 500 指数，台湾市场基准为台湾加权指数，日本市场基准为日经 225 指数，新加坡市场基准为海峡时报指数，韩国市场基准为韩国有价证券指数，英国市场基准为富时 100 指数），具体如下：

行业评级

增持：预计行业股票指数超越基准
中性：预计行业股票指数基本与基准持平
减持：预计行业股票指数明显弱于基准

公司评级

买入：预计股价超越基准 15%以上
增持：预计股价超越基准 5%~15%
持有：预计股价相对基准波动在-15%~5%之间
卖出：预计股价弱于基准 15%以上
暂停评级：已暂停评级、目标价及预测，以遵守适用法规及/或公司政策
无评级：股票不在常规研究覆盖范围内。投资者不应期待华泰提供该等证券及/或公司相关的持续或补充信息

法律实体披露

中国：华泰证券股份有限公司具有中国证监会核准的“证券投资咨询”业务资格，经营许可证编号为：91320000704041011J

香港：华泰金融控股（香港）有限公司具有香港证监会核准的“就证券提供意见”业务资格，经营许可证编号为：AOK809

美国：华泰证券（美国）有限公司为美国金融业监管局（FINRA）成员，具有在美国开展经纪交易商业务的资格，经营业务许可编号为：CRD#:298809/SEC#:8-70231

新加坡：华泰证券（新加坡）有限公司具有新加坡金融管理局颁发的资本市场服务许可证，并且是豁免财务顾问。公司注册号：202233398E

华泰证券股份有限公司

南京
南京市建邺区江东中路 228 号华泰证券广场 1 号楼/邮政编码：210019
电话：86 25 83389999/传真：86 25 83387521
电子邮件：ht-rd@htsc.com

北京
北京市西城区太平桥大街丰盛胡同 28 号太平洋保险大厦 A 座 18 层/邮政编码：100032
电话：86 10 63211166/传真：86 10 63211275
电子邮件：ht-rd@htsc.com

深圳
深圳市福田区益田路 5999 号基金大厦 10 楼/邮政编码：518017
电话：86 755 82493932/传真：86 755 82492062
电子邮件：ht-rd@htsc.com

上海
上海市浦东新区东方路 18 号保利广场 E 栋 23 楼/邮政编码：200120
电话：86 21 28972098/传真：86 21 28972068
电子邮件：ht-rd@htsc.com

华泰金融控股（香港）有限公司
香港中环皇后大道中 99 号中环中心 53 楼
电话：+852-3658-6000/传真：+852-2567-6123
电子邮件：research@htsc.com
http://www.htsc.com.hk

华泰证券（美国）有限公司
美国纽约公园大道 280 号 21 楼东（纽约 10017）
电话：+212-763-8160/传真：+917-725-9702
电子邮件: Huatai@htsc-us.com
http://www.htsc-us.com

华泰证券（新加坡）有限公司
滨海湾金融中心 1 号大厦, #08-02, 新加坡 018981
电话：+65 68603600
传真：+65 65091183