研究所
分析师：肖承志
SAC 登记编号：S1340524090001
Email：xiaochengzhi@cnpsec.com
研究助理：冯昱文
SAC 登记编号：S1340124100011
Email：fengyuwen@cnpsec.com

近期研究报告

《结合基本面和量价特征的 GRU 模型》 - 2025.06.05

《Claude 4 系列发布，谷歌上线编程智能体 Jules——AI 动态汇总 20250526》 - 2025.05.27

《谷歌发布智能体白皮书，Manus 全面开放注册——AI 动态汇总 20250519》 - 2025.05.20

《证监会修改《重组办法》，深化并购重组改革——微盘股指数周报 20250518》 - 2025.05.19

《通义千问发布 Qwen-3 模型，DeepSeek 发布数理证明大模型——AI 动态汇总 20250505》 - 2025.05.06

《基金 Q1 加仓有色汽车传媒，减仓电新食饮通信——公募基金 2025Q1 季报点评》 - 2025.04.30

《泛消费打开连板与涨幅高度，ETF 资金平铺机器人、人工智能与芯片——行业轮动周报 20250427》 - 2025.04.28

《国家队交易特征显著，短期指数仍交易补缺预期，TMT 类题材仍需等待——行业轮动周报 20250420》- 2025.04.21

《小市值持续，高低波风格交替——中邮因子周报 20250413》 - 2025.04.14

《4 月是否还会有“最后一跌”？——微盘股指数周报 20250406》- 2025.04.07

金工周报

谷歌更新 Gemini 2.5 Pro，阿里开源 Qwen3 新模型——AI 动态汇总 20250609【中邮金工】

- **谷歌更新 Gemini 2.5 Pro**

谷歌于 2025 年 6 月 5 日推出的 Gemini 2.5 Pro Preview 06-05 版本是其多模态大语言模型的最新迭代，标志着 AI 领域在编程、推理及多模态能力上的重大突破。此次更新基于 5 月 I/O 大会发布的 05-06 版本进一步优化，核心目标是通过技术升级巩固其在行业中的领先地位，并推动 AI 工具从实验室研究向实际生产力工具的转型。

- **阿里开源 Qwen3 新模型**

阿里巴巴于 2025 年 6 月 6 日开源的 Qwen3-Embedding 及 Reranker 系列模型，是基于 Qwen3 基础架构构建的文本表征与排序技术的最新突破，其设计理念和技术实现标志着国产大模型在语义理解与检索领域的重大进展。

- **英伟达推出 Fast-dLLM 框架**

英伟达于 2025 年 5 月底联合麻省理工学院、香港大学等机构推出的 Fast-dLLM 框架，是扩散式大语言模型（Diffusion-based LLM）推理加速领域的一项突破性技术。该框架通过创新的 分块 KV 缓存机制 和 置信度感知并行解码策略 ，在不需重新训练模型的前提下，实现了最高 27.6 倍的推理速度提升，同时将生成质量损失控制在 2%以内，标志着扩散模型从理论优势迈向实际应用的关键转折。

- **快手开源 Auto Think 大模型**

快手于 2025 年 6 月开源的 KwaiCoder-AutoThink-preview 模型，标志着大模型技术从"单向深度推理"向"动态思考调节"的范式转变。该模型通过创新的 双模思考机制 和 Step-SRPO 强化学习框架 ，实现了根据问题难度自动切换思考深度的突破性能力，被业界形象称为"DeepSeek-V3 与 R1 的技术合体"。

- **风险提示：**

以上内容基于历史数据完成，在政策、市场环境发生变化时存在失效的风险；历史信息不代表未来。

目录

图表目录

1 AI 重点要闻

1.1 谷歌更新 Gemini 2.5 Pro

谷歌于 2025 年 6 月 5 日推出的 Gemini 2.5 Pro Preview 06-05 版本是其多模态大语言模型的最新迭代，标志着 AI 领域在编程、推理及多模态能力上的重大突破。此次更新基于 5 月 I/O 大会发布的 05-06 版本进一步优化，核心目标是通过技术升级巩固其在行业中的领先地位，并推动 AI 工具从实验室研究向实际生产力工具的转型。

本次更新的首个亮点是技术架构与性能提升，新版本在底层架构上虽未公开细节，但实测表现显示其核心改进集中在三方面：

- **编程能力**：在 LMArena 编码排行榜以 1470 分（提升 24 分）和 WebDevArena 以 1443 分（提升 35 分）实现断层式领先，尤其在 Aider Polyglot 基准测试中以 82.2%通过率超越 Claude Opus 4 和 DeepSeek R1。其创新点在于通过单条提示生成完整交互式 Web 应用，例如将 YouTube 视频转化为带 UI 的学习应用，显著降低开发门槛。
- **推理与学术能力**：在 GPQA（科学问答）和 Humanity's Last Exam（HLE）等高难度测试中分别取得 86.4%和 21.6%的准确率，未依赖多数投票策略即实现 SOTA，凸显纯推理能力的提升。
- **多模态处理**：支持百万级 Token 上下文窗口，可解析 1 小时视频或 11 小时音频，并在 VideoMME 基准测试中以 84.8%得分实现视频到代码的端到端转换。

图表 1：Gemini 2.5 Pro TextArena 评分

图表 2：Gemini 2.5 Pro WebDevArena 评分

Text View →

Rank (UB)	Model	Score	Votes
1	gemini-2.5-pro-preview-06-05	1470	4,701
2	gemini-2.5-pro-preview-05-06	1446	10,386
2	o3-2025-04-16	1443	13,808
4	chatgpt-4o-latest-20250326	1431	18,302
4	gpt-4.5-preview-2025-02-27	1425	15,271
5	gemini-2.5-flash-preview-05-…	1419	9,970

资料来源：LMArena，中邮证券研究所

WebDev View →

Rank (UB)	Model	Score	Votes
1	Gemini-2.5-Pro-Preview-06-05	1443	1,872
1	Claude Opus 4 (20250514)	1412	2,466
2	Gemini-2.5-Pro-Preview-05-06	1408	3,858
2	Claude Sonnet 4 (20250514)	1389	2,078
5	Claude 3.7 Sonnet (20250219)	1357	7,481
6	Gemini-2.5-Flash-Preview-05-…	1312	2,626

资料来源：LMArena，中邮证券研究所

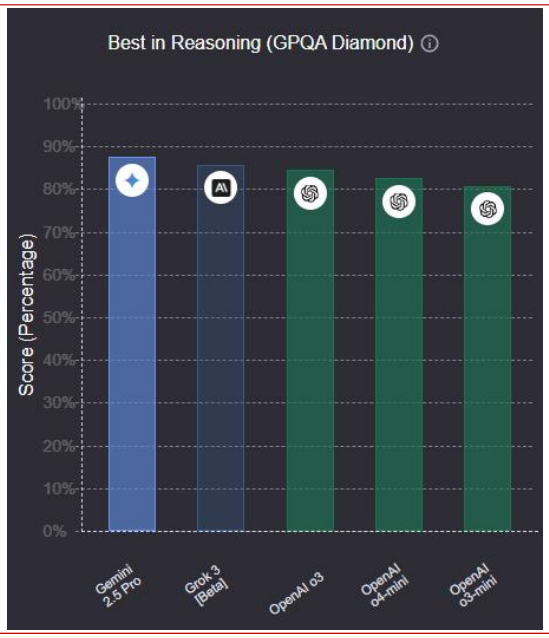
除此之外，本次更新还提升了功能创新与用户体验，谷歌首次引入“思考预算”（Thinking Budgets）功能，允许开发者通过调节 Token 消耗量平衡响应质量与成本，例如设定高预算时模型会进行更深入的逻辑推演。此外，针对用户反馈优化了非编码任务的响应风格，生成内容更具创意且结构清晰，例如自动格式化代码块和学术引用。在定价策略上维持输入 1.25 美元/百万 Token、输出 10 美元/百万 Token 的竞争力，仅为 Claude 4 的四分之一。

图表 3：Gemini 2.5 Pro 多基准评分

Benchmark	Gemini 2.5 Pro Preview 06-05 Thinking	OpenAI o3 High	OpenAI o4-mini High	Claude Opus 4 32k thinking	Grok 3 Beta Extended thinking	DeepSeek R1 05-28
Input price	$1.25	$10.00	$1.10	$15.00	$3.00	$0.55
Output price	$10.00	$40.00	$4.40	$75.00	$15.00	$2.19
Reasoning & knowledge Humanity's Last Exam (no tools)	21.6%	20.3%	14.3%	10.7%	—	14.0%*
Science GPQA diamond	86.4%	83.3%	81.4%	79.6% 83.3%	80.2% 84.6%	81.0%
Mathematics AIME 2025	88.0%	88.9%	92.7%	75.5% 90.0%	77.3% 93.3%	87.5%
Code generation LiveCodeBench	69.0%	72.0%	75.8%	51.1%	—	70.5%
Code editing Aider Polyglot	82.2%	79.6%	72.0%	72.0%	53.3%	71.6%
Agentic coding SWE-bench Verified	59.6% 67.2%	69.1%	68.1%	72.5% 79.4%	—	57.6%
Factuality SimpleQA	54.0%	48.6%	19.3%	—	43.6%	27.8%
Factuality FACTS Grounding	87.8%	69.6%	62.1%	77.7%	74.8%	—
Visual reasoning MMMU	82.0%	82.9%	81.6%	76.5%	76.0% 78.0%	—
Image understanding Vibe-Eval (Reka)	67.2%	—	—	—	—	—
Video understanding VideoMMMU	83.6%	—	—	—	—	—
Long context MRCR v2 (8-needle)	58.0% 16.4%	57.1%	36.3%	—	34.0%	—
Multilingual performance Global MMLU (Lite)	89.2%	—	—	—	—	—

资料来源：谷歌，中邮证券研究所

图表 4：Gemini 2.5 Pro GPQA 评分



资料来源：vellum，中邮证券研究所

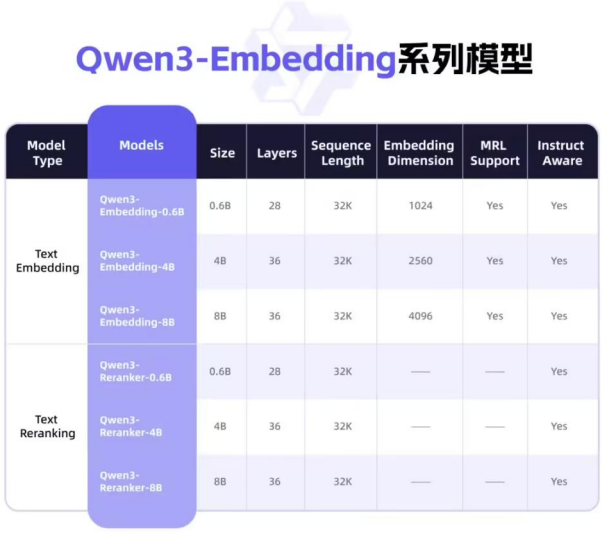
从行业影响与竞争格局　　来看，此次更新直接冲击了 OpenAI 的 o3、Anthropic 的 Claude 4 等竞品。例如在交通灯模拟编程测试中，Gemini 生成的 Python 代码在物理规律遵循和动画精细度上显著优于 GPT-4.5 和 Claude 3.7。谷歌 CEO 桑达尔·皮查伊强调，模型已通过 Replit 等平台集成，被开发者评价为“与高级工程师协作般的体验”。不过，部分测试显示其在 LiveCodeBench(75.8%)和 MMMU 视觉推理（82.9%）上仍略逊于 OpenAI o3，表明多模态细节处理尚有优化空间。

谷歌计划在未来几周内将 06-05 版本升级为稳定版（GA），并扩展上下文窗口至 200 万 Token。结合其已在 Vertex AI 和 Google AI Studio 开放的 API 访问，这一版本或将成为企业级 AI 应用的新基准，尤其在教育、自动化开发等领域。正如 DeepMind 首席执行官德米斯·哈萨比斯所言，Gemini 2.5 Pro 的迭代不仅关乎技术竞赛，更重新定义了“人类提出需求，AI 实现创意”的开发范式。

1.2 阿里开源 Qwen3 新模型

阿里巴巴于 2025 年 6 月 6 日开源的 Qwen3-Embedding 及 Reranker 系列模型，是基于 Qwen3 基础架构构建的文本表征与排序技术的最新突破，其设计理念和技术实现标志着国产大模型在语义理解与检索领域的重大进展。

图表 5：Qwen3-Embedding&Reranker 系列模型



Model Type	Models	Size	Layers	Sequence Length	Embedding Dimension	MRL Support	Instruct Aware
Text Embedding	Qwen3-Embedding-0.6B	0.6B	28	32K	1024	Yes	Yes
	Qwen3-Embedding-4B	4B	36	32K	2560	Yes	Yes
	Qwen3-Embedding-8B	8B	36	32K	4096	Yes	Yes
Text Reranking	Qwen3-Reranker-0.6B	0.6B	28	32K	—	—	Yes
	Qwen3-Reranker-4B	4B	36	32K	—	—	Yes
	Qwen3-Reranker-8B	8B	36	32K	—	—	Yes

资料来源：通义千问，中邮证券研究所

图表 6：排序模型评测

Model	Param	MTEB-R	CMTEB-R	MMTEB-R	MLDR	MTEB-Code	FollowIR
Qwen3-Embedding-0.6B	0.6B	61.82	71.02	64.64	50.26	75.41	5.09
Jina-multilingual-reranker-v2-base	0.3B	58.22	63.37	63.73	39.66	58.98	-0.68
gte-multilingual-reranker-base	0.3B	59.51	74.08	59.44	66.33	54.18	-1.64
BGE-reranker-v2-m3	0.6B	57.03	72.16	58.36	59.51	41.38	-0.01
Qwen3-Reranker-0.6B	0.6B	65.80	71.31	66.36	67.28	73.42	5.41
Qwen3-Reranker-4B	4B	69.76	75.94	72.74	69.97	81.20	14.84
Qwen3-Reranker-8B	8B	69.02	77.45	72.94	70.19	81.22	8.05

资料来源：通义千问，中邮证券研究所

以下从技术架构、训练范式、性能表现及行业影响三个维度展开深度解析：

- 技术架构创新

Qwen3-Embedding 采用双塔结构设计，通过提取最后一层[EOS]标记的隐藏状态作为语义向量，支持动态调整输出维度（768/1024/4096 等），这种多分辨率向量（MRL）机制使得模型能根据硬件条件灵活平衡精度与效率。其创新性体现在指令感知设计上——将任务指令与查询文本拼接为统一输入上下文，而文档保持独立处理，这种架构使同一模型能适配检索、语义相似度计算、分类等多场景需求，同时保留 32K 长文本处理能力。Reranker 则采用单塔交互结构，将查询-文档对拼接后通过二元分类模板（输出"Yes/No"概率）计算相关性得分，结合 RoPE 位置编码与双块注意力机制（Dual Chunk Attention），有效解决长文档语义连贯性问题。

- 训练范式突破

Embedding 模型采用三阶段训练流程：首阶段利用 Qwen3-32B 合成的 1.5 亿多语种弱监督文本对进行对比学习预训练，通过动态生成的 Prompt 体系突破传统数据采集局限；第二阶段融合 1200 万弱监督对与 700 万人工标注数据进行监督微调，采用改进的 InfoNCE 损失函数优化；最终通过球面线性插值（slerp）

融合多个检查点提升泛化性。消融实验表明，弱监督预训练和模型融合分别贡献约 15%和 8%的性能提升。Reranker 则直接采用高质量标注数据监督训练，通过指令微调可使特定领域排序准确率提升 3%-5%。

- 性能表现与行业影响

在 MTEB 多语言基准测试中，Qwen3-Embedding-8B 以 70.58 分超越 Google Gemini-Embedding 等商业模型，其代码检索任务 nDCG@10 达 80.68，中文检索得分 77.45，均刷新开源记录。值得注意的是，0.6B 轻量版在边缘设备部署时仅需 1.5GB 显存，性能却接近 7B 参数的 gte-Qwen2。Reranker-8B 在 mMARCO 跨语言检索中 MRR@10 达 0.42，对 100 文档排序延迟控制在 80ms 内（A100），使企业知识库问答准确率提升 40%。

图表 7：Qwen-3-Embedding 模型 MTEB 评测跑分

	Qwen3-Embedding-8B	Qwen3-Embedding-4B	Qwen3-Embedding-0.6B	Gemini Embedding	Cohere-embed-multilingual-v3.0	text-embedding-3-large	multilingual-e5-large-instruct	gte-Qwen2-7B-instruct
MMTEB Mean-Task	70.58	69.45	64.33	68.37	61.12	58.93	63.22	62.51
MTEB (en v2) Mean-Task	75.22	74.60	70.70	73.30	66.01	66.43	65.53	70.72
MTEB-Code	80.68	80.06	75.41	74.66	51.94	58.95	65.00	56.41

资料来源：通义千问，中邮证券研究所

实际应用中，二者构成端到端检索链路：Embedding 负责初步召回，Reranker 进行精细排序。例如在法律文书检索场景，32K 上下文窗口可直接处理完整卷宗，避免传统切片导致的信息损失。开发者可通过 Hugging Face 或 ModelScope 获取模型，阿里云 API 支持 5 行代码集成。这种开源策略正在推动文本检索从关键词匹配迈向"语义理解+动态交互"的新范式，为多模态应用奠定基础。

1.3 英伟达推出 Fast-dLLM 框架

英伟达于 2025 年 5 月底联合麻省理工学院、香港大学等机构推出的 Fast-dLLM 框架，是扩散式大语言模型（Diffusion-based LLM）推理加速领域的一项突破性技术。该框架通过创新的 分块 KV 缓存机制 和 置信度

感知并行解码策略，在不需重新训练模型的前提下，实现了最高 27.6 倍的推理速度提升，同时将生成质量损失控制在 2%以内，标志着扩散模型从理论优势迈向实际应用的关键转折。

传统扩散模型虽采用双向注意力机制支持多标记并行生成，但缺乏高效的键值（KV）缓存设计，导致每次推理需重复计算全部注意力状态，且并行解码易破坏标记间依赖关系。Fast-dLLM 通过双重技术路径解决这一核心矛盾：

图表 8：分块 KV 缓存（Block-Wise KV Cache）设计

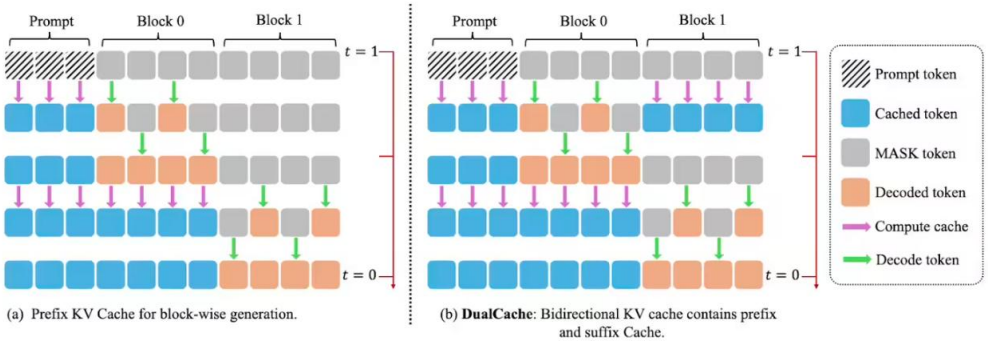


Figure 2 | **Illustration of our Key-Value Cache for Block-Wise Decoding.** (a) During prefix-only caching, the KV cache is computed once for the prompt and reused across multiple decoding steps within each block. The cache is updated after completing a block to maintain consistency, with negligible overhead. (b) DualCache extends this approach by caching both prefix and masked suffix tokens, further accelerating decoding. The high similarity of KV activations across steps allows effective reuse with minimal approximation error.

资料来源：英伟达，中邮证券研究所

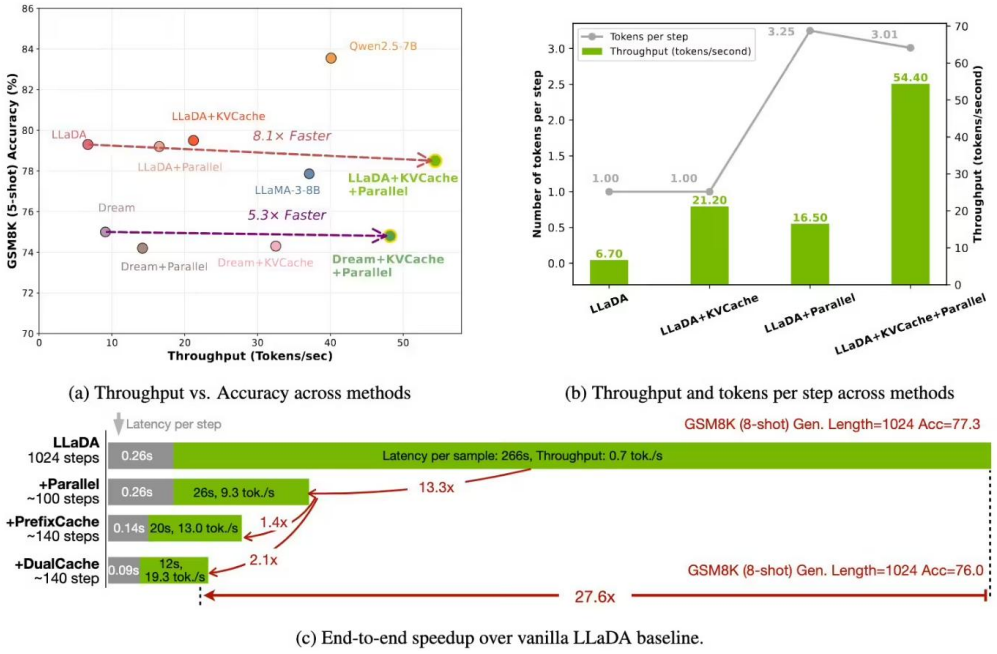
1. **分块 KV 缓存（Block-Wise KV Cache）**：将序列划分为多个块，预先计算并缓存前缀（Prompt）和后缀（Masked Tokens）的注意力激活值（KV Cache）。其创新性体现在 DualCache 设计——同时缓存双向上下文，利用相邻推理步骤间高达 0.98 的余弦相似度（实验验证）实现 90%以上的激活值复用，单步计算量降低 94%。例如在 LLaDA 模型中，1024 标记的长文本生成任务端到端延迟从 266 秒压缩至 12 秒。

2. **置信度感知并行解码**：通过动态阈值筛选高置信度标记（如概率≥0.9）进行并行解码，低置信度标记留待后续步骤处理。数学上证明当(n+1)ϵ≤1（n 为并行标记数，ϵ为置信度偏差）时，该策略与顺序解码结果一致，既保持吞吐量又避免生成"high house"等逻辑冲突。算法伪代码显示，该机制通过交替执行缓存更新（步骤 13）与置信度验证（步骤 7-8）实现高效协同。

在多项标准测试中，Fast-dLLM 展现出显著优势。最值得关注的突破在于速度突破：GSM8K 数学推理任务（1024 标记）实现 27.6 倍加速，吞吐量从 0.7 token/秒提升至 19.3 token/秒；代码生成任务 HumanEval 加速 3.7 倍，MBPP 任务加速 9.2 倍。这种加速效应随序列长度增加而放大，例如在 8-shot 设置下，生成长度从 256 增至 1024 标记时，加速比从 9.4 倍跃升至 27.6 倍。

除此之外，该架构在质量保持方面表现优异。GSM8K-5shot 准确率仅下降 0.8%（78.5%→77.7%），MATH-4shot 保持 33.2%准确率，部分任务如 HumanEval 甚至因缓存优化提升 1.2%。对比实验显示，单独使用 KV 缓存或并行解码仅能实现 3-5 倍加速，而二者结合产生指数级增益。

图表 9：Fast-dLLM 速度提升 27.6 倍



(a) Throughput vs. Accuracy across methods

(b) Throughput and tokens per step across methods

(c) End-to-end speedup over vanilla LLaDA baseline.

资料来源：英伟达，中邮证券研究所

Fast-dLLM 的零训练成本特性使其能即插即用地集成至现有系统。例如在 Replit 等开发平台中，开发者可直接调用该框架将 LLaDA、Dream 等扩散模型的代码生成速度提升 8 倍以上。其技术意义在于 1）**打破自回归模型垄断**：通过解决扩散模型的核心效率瓶颈，使其在实时交互、长文本生成等场景具备与 GPT 类模型竞争的实力。2）**优化计算资源利用**：相比依赖硬

件升级的加速方案，Fast-dLLM 通过算法创新实现"软加速"，为边缘设备部署提供可能。例如轻量版在 A100 显卡上对 100 文档排序延迟仅 80ms。

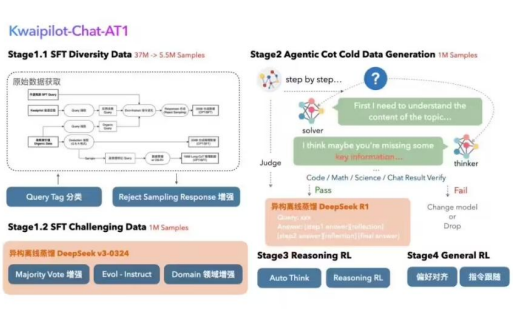
未来，该团队计划扩展至更大规模模型（如千亿参数级）和视频生成等多模态任务。开发者可通过 GitHub 仓库获取代码，或阅读 arXiv 论文深入理解其数学证明。这项研究不仅推进了非自回归生成模型的工程落地，更重新定义了"速度-质量"权衡的优化范式。

1.4 快手开源 Auto Think 大模型

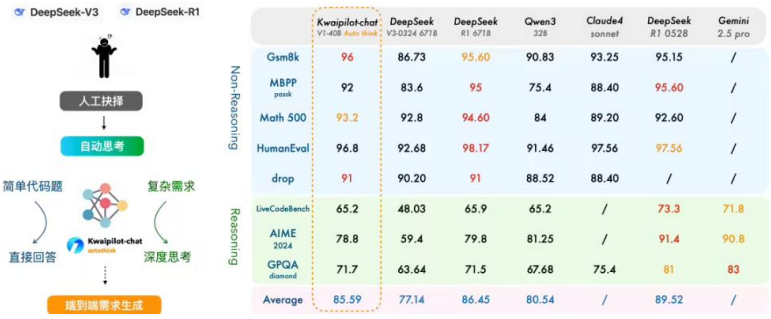
快手于 2025 年 6 月开源的 KwaiCoder-AutoThink-preview 模型，标志着大模型技术从"单向深度推理"向"动态思考调节"的范式转变。该模型通过创新的 双模思考机制 和 Step-SRPO 强化学习框架 ，实现了根据问题难度自动切换思考深度的突破性能力，被业界形象称为"DeepSeek-V3 与 R1 的技术合体"。

AutoThink 的核心突破在于解决了大模型普遍存在的"过度思考"问题——即对简单问题（如"1+1 等于几"）进行不必要的复杂推理。其技术实现基于三阶段训练范式：首先通过 Ellipsis Prompt 技术 （在输入中添加省略号）引导模型分化出快慢两种思维路径，建立基础的模式切换能力；随后采用 异构离线蒸馏 方法，分别用 DeepSeek-V3 和 R1 作为教师模型优化快思考（直接响应）与慢思考（深度推理）模式；最终通过创新的 Step-SRPO 算法进行强化学习，该算法在传统 GRPO 基础上引入过程监督，对不同 token 根据未来收益动态调整优势函数计算，使模型判断思考必要性的准确率提升 37%。这种训练方式使得模型在 GSM8K 数学基准测试中达到 96 分，比传统方法节省 40%计算资源。

图表 10：Kwaipilot-chat 架构



图表 11：Kwaipilot-chat 评分



	Kwaipilot-chat V1-40B Auto Work	DeepSeek V3-0324 671B	DeepSeek R1 671B	Qwen3 32B	Claude4 sonnet	DeepSeek R1 0528	Gemini 2.5 pro
Gsm8k	96	86.73	95.60	90.83	93.25	95.15	/
MBPP pass@k	92	83.6	95	75.4	88.40	95.60	/
Math 500	93.2	92.8	94.60	84	89.20	92.60	/
HumanEval	96.8	92.68	98.17	91.46	97.56	97.56	/
drop	91	90.20	91	88.52	88.40	/	/
LiveCodeBench	65.2	48.03	65.9	65.2	/	73.3	71.8
AIME 2024	78.8	59.4	79.8	81.25	/	91.4	90.8
GPQA diamond	71.7	63.64	71.5	67.68	75.4	81	83
Average	85.59	77.14	86.45	80.54	/	89.52	/

资料来源：快手，中邮证券研究所　　资料来源：快手，中邮证券研究所

在 8 项核心评测中，AutoThink 展现出显著的效率-精度平衡优势：非推理任务如 MBPP 代码生成（95.6 分）和 HumanEval 函数补全（96.8 分）采用快思考模式实现 3-5 倍响应加速；推理任务如 AIME 数学竞赛题（78.8 分）和 GPQA 专业问答（71.7 分）通过慢思考模式提升 20 分以上准确率。特别值得注意的是，其　　动态上下文窗口技术　　能在 16K-32K 长度间自适应调整，相比固定窗口模型减少 17%的冗余计算。不过模型体积达 80GB，需至少两张 A100 显卡部署，在成本敏感场景可能面临挑战。

该模型已展现出跨领域应用潜力：在智能客服场景，通过自动区分简单咨询与复杂问题，使平均响应延迟从 12 秒降至 3 秒；在教育领域能根据学生水平动态调节讲解深度，如对小学数学题直接输出关键步骤，对奥赛题则生成完整推导过程；在视频创作中实现"镜头级思考调节"，简单运镜快速生成，复杂转场精细推演。其开源策略（包括 Hugging Face 模型库和即将发布的技术报告）正推动行业突破"参数内卷"，转向更本质的　　思考经济学　　优化——正如团队负责人所言："让模型学会不做啥，比让它能做啥更难能可贵"。

2 企业动态

2.1 Manus 推出文生视频功能

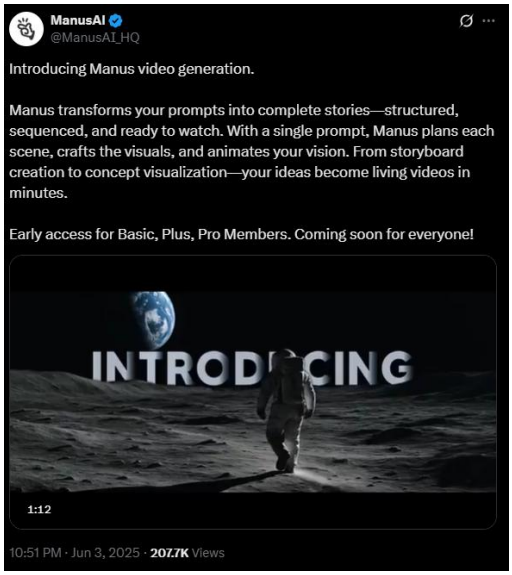
Manus 于 2025 年 6 月 4 日推出的"文生视频"功能，标志着中国 AI 初创公司在多模态生成领域取得重大突破。该技术通过深度重构视频创作范式，将传统"创意构思-脚本编写-拍摄剪辑"的线性流程压缩为"文本输入-智能生成"的即时转化。

Manus 采用"跨模态生成模型+叙事逻辑引擎"的双层架构设计，底层通过分析海量视频数据建立文本到视觉元素的映射关系，上层则运用基于强化学习的镜头调度算法确保叙事连贯性。具体实现中，系统会先解析文本中的时空线索（如

"黄昏时分的森林"）生成关键帧，再通过动态扩散模型补全中间帧，最后采用双流注意力机制协调场景切换节奏。例如在《山海经》神话生物演示中，模型能准确捕捉"人鱼鳞片反光"等细节描述，生成每秒 24 帧的 4K 视频。与 OpenAI Sora 相比，其独特优势在于支持多场景自动拼接——用户输入"制作包含三个场景的产品广告"等复合指令时，AI 会先分解子任务生成独立片段，再通过过渡特效智能合成完整视频。

Manus 构建了梯度分明的三级订阅体系：Basic 版（19 美元/月）支持 1080P 基础生成，Plus 版（39 美元/月）开放 4K 分辨率和风格定制，Pro 版（199 美元/月）则提供 API 接口和 30 分钟长视频生成能力。这种定价策略较 Sora 专业版（200 美元/月）形成差异化竞争，尤其针对中小企业批量生产需求——Pro 会员单次调用可并行生成 20 条视频，使电商商品视频的制作成本降低 92%。实际测试显示，从输入"制作太空探险动画"到获得 60 秒成片仅需 3 分 28 秒，且支持通过语义修改（如"把火箭改成蓝色"）实时调整输出，这种交互式创作模式使教育机构课件制作效率提升 8 倍。

图表 12：ManusAI 宣布推出文生视频功能



资料来源：Manus，中邮证券研究所

尽管未完全公开技术细节，业界推测其核心模型可能基于阿里通义千问架构改进，这在 GAIA 基准测试中已展现优于 GPT-4 的叙事连贯性得分(87.3 vs 79.1)。

不过早期版本仍存在物理规律模拟缺陷，例如生成的"飞鸟"翅膀摆动频率不符合空气动力学，这暴露出当前文生视频技术的共性瓶颈。值得关注的是，Manus 采用"过程监督"训练范式，通过人工标注的 200 万条镜头衔接规则数据，使生成视频的镜头语言专业度达到业余摄影师水平。

该功能的推出直接冲击了 Runway、Pika 等专业视频工具市场，其母公司 Butterfly Effect 近期获得 Benchmark 领投的 B 轮融资，估值已达 18 亿美元。随着计划中的全用户开放，Manus 或将重塑从短视频营销到影视预可视化的全产业链，正如其 CTO 所言："我们不是在替代创作者，而是将他们的创意释放速度提升 1000 倍"。未来迭代方向包括引入 NeRF 技术实现 3D 场景生成，以及通过 LoRA 微调支持企业专属风格训练，这些进展将持续推动 AI 视频从工具向协作者的角色进化。

2.2 英伟达推出 ProRL 方法

英伟达于 2025 年 6 月提出的 ProRL（Prolonged Reinforcement Learning）方法，是强化学习领域针对大语言模型推理能力优化的突破性框架。该方法通过系统性重构训练范式，首次实证了强化学习不仅能优化已有解决方案的采样效率，更能解锁基础模型无法触及的全新推理策略。

ProRL 的核心在于打破传统强化学习的训练时长限制，将训练周期延长至 2000 步以上，并引入三大关键技术组件：首先采用改进的 GRPO（Group Relative Policy Optimization）算法替代传统 PPO，通过消除独立价值模型需求简化训练结构；其次设计动态 KL 散度控制机制，在保持策略稳定性的同时防止熵塌缩——具体通过周期性重置参考策略（每 500 步将在线策略快照更新为参考策略）避免模型陷入局部最优；最后结合高温采样（温度系数 1.2）与 DAPO（Decoupled Advantage Policy Optimization）的动态采样技术，使模型在 13.6 万跨领域任务样本（涵盖数学、编程、STEM 等五大类）中持续探索新解空间。这种训练范式使得 15 亿参数的 Nemotron-Research-Reasoning-Qwen-1.5B 模型在 GPQA Diamond 科学推理任务中准确率提升 25.9%，逻辑谜题解决能力提升 54.8%，部分任务 pass@1 准确率从基础模型的 0%跃升至 100%。

图表 13：ProRL 方法效果

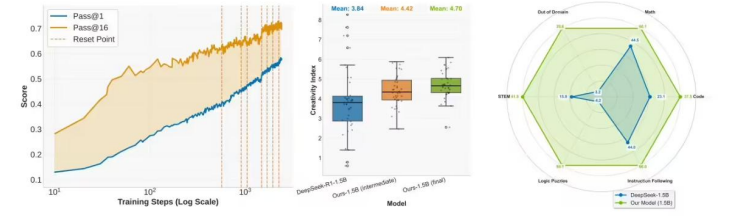
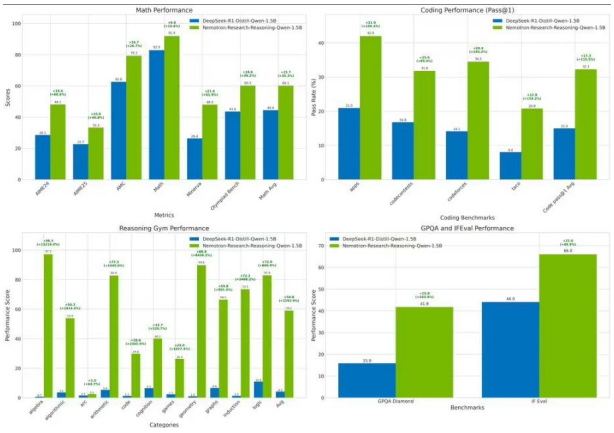


Figure 1: Benefits of prolonged reinforcement learning (ProRL). **Left**: Pass@1 and Pass@16 scales with ProRL training. **Middle**: ProRL leads to more novel solutions reflected by higher Creativity Index [12]. **Right**: Our model greatly surpass base model across diverse tasks.

资料来源：英伟达，中邮证券研究所

图表 14：ProRL 方法测评



资料来源：英伟达，中邮证券研究所

实验数据揭示了强化学习改进程度与基础模型初始能力间的负相关规律：在 DeepSeek-R1-1.5B 表现最弱的逻辑推理任务上，ProRL 带来最大幅度提升（54.8%），而在基础模型已较强的数学领域提升幅度为 14.7%。研究团队提出的

创造力指数　　量化了模型输出的新颖性，证明 ProRL 生成的推理路径与预训练数据重叠度降低 37%，表明其真正形成了新的认知模式。特别值得注意的是，该方法在分布外评估（如 acre、boxnet 等未见任务）中展现出近乎完美的泛化能力，验证了其内化抽象推理规则而非记忆特定解的本质特征。

ProRL 的成功实践挑战了"强化学习仅优化采样效率"的传统认知，为小参数模型实现强推理能力提供了可行路径。其开源的训练框架（基于 verl 改进）已支持在 NVIDIA H100 集群上复现，1.6 万 GPU 小时的投入虽显昂贵，但相比 7B 参数模型的推理成本仍具性价比优势。该方法正在医疗诊断、科学发现等领域验证迁移效果，未来或将与英伟达新一代 Rubin 芯片的推理优化特性（如 1024GB 显存支持的长上下文处理）形成协同效应。正如论文作者所述："当训练时长突破临界点，强化学习就能从量变引发质变，这种非线性进步规律正在重塑我们对 AI 能力边界的所有假设"。

3 AI 行业洞察

3.1 Karpathy 教你如何正确使用 ChatGPT

在 ChatGPT 日益普及的今天，许多用户每月支付 20 美元 Plus 会员费甚至 200 美元 Pro 费用却未能充分发挥其价值。OpenAI 联合创始人、AI 领域权威 Andrej Karpathy 近期公开分享的专业级使用策略，揭示了高效利用 ChatGPT 的核心方法论，这套体系通过科学分类使用场景与深度挖掘记忆系统，可使专业用户的综合效用提升 2 倍以上。

Karpathy 直指 OpenAI 模型命名体系的混乱现状（如 GPT-4o、o3、o4-mini 等混杂），通过大量实践验证构建了四象限选择框架：日常简单问题（如膳食纤维查询）选用响应迅捷的 GPT-4o 模型，这类场景占比约 40%；关键复杂问题（如税务规划）启用推理能力最强的 o3 模型，同样占据 40%使用量；编程开发场景专属 GPT-4.1 模型优化代码生成，约占 10%；深度研究任务则启动 Deep Research 功能，该系统基于 o3 模型但增加了自动联网检索与多源信息整合能力。值得注意的是，Karpathy 特别警示应规避 o4-mini 等过渡性模型，认为其当前缺乏实用价值。

ChatGPT 区别于其他 AI 的核心优势在于其分层记忆架构。初创工程师 Eric Hayes 通过逆向工程揭示，该系统包含主动记忆（Saved Memory）与智能历史（Chat History）两大模块。前者通过 bio tool 技术实现用户显式偏好的存储与矛盾检测（如声明"我不是软件工程师"时会要求澄清而非直接覆盖）；后者则包含三层结构：当前会话的短期记忆、支持两周内精准引用的跨对话记忆，以及通过每周批处理生成的用户画像标签（如"偏好技术细节"）。数据表明，这种用户洞察引擎贡献了 ChatGPT 智能感知提升的 80%以上，使其能预判未言明的需求。

Karpathy 在其技术演讲中强调，基础语言模型本质仅是文档补全工具，ChatGPT 的能力跃迁来自四阶段训练法则：千卡 GPU 集群的预训练消耗 99%算力成本；监督微调注入高质量问答数据塑造助手雏形；奖励建模建立人类偏好评估体系；最终通过 RLHF 技术优化回答质量。记忆系统则采用键值检索机制，将记忆信息动态注入对话上下文，而用户洞察系统通过聚类优化算法将海量对话压缩

为高价值特征向量。这种设计使得当用户要求分析"激光雷达技术突破"时，系统能自动关联历史讨论中的成本控制需求，实现真正的智能协作。

遵循该策略的科技分析师工作流显示：启动 Deep Research 功能分析行业动态时，系统会自动关联上周讨论的焦点；使用 o3 模型起草技术报告可自动生成含竞争对手分析的数据表格；切换 GPT-4.1 验证代码能精准识别传感器兼容性问题。这种用法不仅使 200 美元 Pro 会员费产生超额回报，更揭示了 AI 应用的深层规律——最优工具组合的本质，在于精确匹配任务复杂度与模型能力谱系。正如 Karpathy 开源 minbpe 项目所昭示的，对技术本质的透彻理解，才是最大化 AI 价值的终极密钥。

4 技术前沿

4.1 DeepMind：智能体需要世界模型

论文《General Agents Need World Models》由 Jonathan Richens 等学者共同完成，从理论层面探讨了世界模型（world models）在通用智能体（general agents）中的必要性，并提出了形式化证明和算法框架。研究核心围绕一个根本性问题展开： 是否所有能够执行多步目标导向任务的智能体都必须学习环境的世界模型？ 通过严格的数学推导和实验验证，论文不仅给出了肯定答案，还揭示了世界模型精度与智能体能力之间的量化关系，对 AI 安全性、可解释性及能力边界等关键领域具有深远影响。

图表 15：General agents need world models

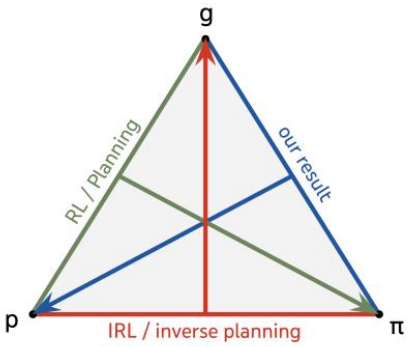
General agents need world models

Jonathan Richens [1] David Abel [1] Alexis Bellot [1] Tom Everitt [1]

资料来源：论文 General agents need world models，中邮证券研究所

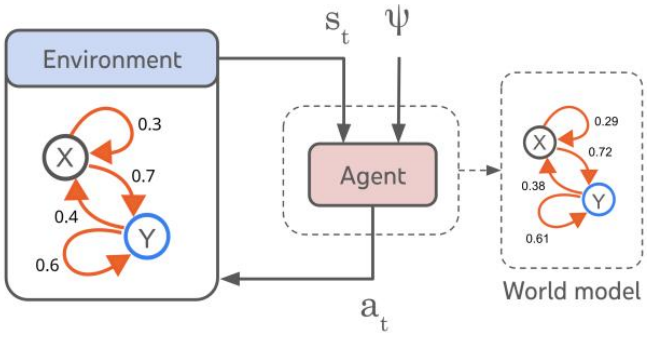
研究基于控制马尔可夫过程（cMP）环境假设，将智能体定义为　　目标条件策略　　（goal-conditioned policy），其能力通过“目标深度”（goal depth）和“遗憾边界”（regret bound）量化。目标深度指智能体需完成的子目标序列长度，而遗憾边界要求智能体达成目标的概率接近最优策略的（1-δ）倍。论文的核心定理（Theorem 1）证明：　　任何满足遗憾边界的智能体，其策略本身编码了一个误差有界的世界模型　　，且模型精度随目标深度增加而提升（误差上界为√(2p(1-p)/(n(1-δ)))），其中 p 为状态转移概率，n 为目标深度）。这一结论颠覆了 Brooks“世界即自身最佳模型”的经典观点，表明模型无关（model-free）学习无法绕过世界建模的底层需求。

图表 16：研究结果与 RL，IRL 之间关系



资料来源：论文 General agents need world models，中邮证券研究所

图表 17：智能体环境系统



资料来源：论文 General agents need world models，中邮证券研究所

研究通过构造　　复合目标　　（composite goals）的二分决策问题，将世界模型提取转化为策略行为的逆向推理。例如，设计目标 $\psi_a(k,n)$ 要求智能体在状态 s 执行动作 a 后转移至 s' 不超过 k 次，而 ψ_b(k,n) 要求超过 k 次。通过分析智能体在 $\psi_a \vee \psi\_b$ 下的动作选择（如优先执行 a 或 b），可推断转移概率 $P_{s,s'}(a)$ 的置信区间。论文提出的　　Algorithm 1　　通过线性时序逻辑（LTL）生成此类目标，并证明其提取的世界模型误差随 n 增大呈 $O(1/\sqrt{n})$ 衰减。实验验证中，即便智能体违反部分假设（如某些目标 δ=1），算法仍能有效恢复环境动态（Figure 3 显示平均误差（ e） 随 n 增长而降低）。

研究揭示了世界模型在智能体能力演进中的 双重角色 ：一方面，它是实现长期目标规划的必要条件（Theorem 2 证明短视智能体无需学习转移概率）；另一方面，其精度直接制约智能体的泛化能力。例如，在安全关键领域，高精度世界模型可支持风险预测和干预验证。论文还探讨了与机械可解释性（mechanistic interpretability）的关联：策略中隐含的世界模型可通过干预实验解码，但本文方法无需访问网络激活，仅依赖策略输入输出。此外，研究对比了与逆向强化学习（IRL）和因果世界模型的差异，指出 任务泛化比领域泛化所需的环境知识更少 ，因为后者需识别变量间因果结构。

理论框架目前限于完全观测的马尔可夫环境，部分观测和时序依赖的扩展尚未解决。实验部分在稀疏转移的随机环境中验证，但复杂现实场景的适应性待进一步研究。作者建议未来探索 通用任务集 设计，以更高效地诱导世界模型学习，并开发基于该框架的安全验证工具。值得注意的是，论文提出的“策略即世界模型载体”观点，为理解大语言模型的涌现能力（如零样本推理）提供了新视角——这些能力可能源于训练过程中隐含的世界建模。

5 风险提示

以上内容基于历史数据完成，在政策、市场环境发生变化时存在失效的风险；历史信息不代表未来。

中邮证券投资评级说明

投资评级标准	类型	评级	说明
报告中投资建议的评级标准：报告发布日后的6个月内的相对市场表现，即报告发布日后的6个月内的公司股价（或行业指数、可转债价格）的涨跌幅相对同期相关证券市场基准指数的涨跌幅。 市场基准指数的选取：A股市场以沪深300指数为基准；新三板市场以三板成指为基准；可转债市场以中信标普可转债指数为基准；香港市场以恒生指数为基准；美国市场以标普500或纳斯达克综合指数为基准。	股票评级	买入	预期个股相对同期基准指数涨幅在20%以上
		增持	预期个股相对同期基准指数涨幅在10%与20%之间
		中性	预期个股相对同期基准指数涨幅在-10%与10%之间
		回避	预期个股相对同期基准指数涨幅在-10%以下
	行业评级	强于大市	预期行业相对同期基准指数涨幅在10%以上
		中性	预期行业相对同期基准指数涨幅在-10%与10%之间
		弱于大市	预期行业相对同期基准指数涨幅在-10%以下
	可转债评级	推荐	预期可转债相对同期基准指数涨幅在10%以上
		谨慎推荐	预期可转债相对同期基准指数涨幅在5%与10%之间
		中性	预期可转债相对同期基准指数涨幅在-5%与5%之间
		回避	预期可转债相对同期基准指数涨幅在-5%以下

分析师声明

撰写此报告的分析师（一人或多人）承诺本机构、本人以及财产利害关系人与所评价或推荐的证券无利害关系。

本报告所采用的数据均来自我们认为可靠的目前已公开的信息，并通过独立判断并得出结论，力求独立、客观、公平，报告结论不受本公司其他部门和人员以及证券发行人、上市公司、基金公司、证券资产管理公司、特定客户等利益相关方的干涉和影响，特此声明。

免责声明

中邮证券有限责任公司（以下简称“中邮证券”）具备经中国证监会批准的开展证券投资咨询业务的资格。

本报告信息均来源于公开资料或者我们认为可靠的资料，我们力求但不保证这些信息的准确性和完整性。报告内容仅供参考，报告中的信息或所表达观点不构成所涉证券买卖的出价或询价，中邮证券不对因使用本报告的内容而导致的损失承担任何责任。客户不应以本报告取代其独立判断或仅根据本报告做出决策。

本报告所载的意见、评估及预测仅为本报告出具日的观点和判断。该等意见、评估及预测无需通知即可随时更改。过往的表现亦不应作为日后表现的预示和担保。在不同时期，中邮证券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。

中邮证券及其所属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，也可能为这些公司提供或者计划提供投资银行、财务顾问或者其他金融产品等相关服务。

《证券期货投资者适当性管理办法》于2017年7月1日起正式实施，本报告仅供中邮证券签约客户使用，若您非中邮证券签约客户，为控制投资风险，请取消接收、订阅或使用本报告中的任何信息。本公司不会因接收人收到、阅读或关注本报告中的内容而视其为签约客户。

中邮证券对于本申明具有最终解释权。

公司简介

中邮证券有限责任公司，2002 年 9 月经中国证券监督管理委员会批准设立，是中国邮政集团有限公司绝对控股的证券类金融子公司。

公司经营范围包括:证券经纪，证券自营，证券投资咨询，证券资产管理，融资融券，证券投资基金销售，证券承销与保荐，代理销售金融产品，与证券交易、证券投资活动有关的财务顾问等。

公司目前已经在北京、陕西、深圳、山东、江苏、四川、江西、湖北、湖南、福建、辽宁、吉林、黑龙江、广东、浙江、贵州、新疆、河南、山西、上海、云南、内蒙古、重庆、天津、河北等地设有分支机构，全国多家分支机构正在建设中。

中邮证券紧紧依托中国邮政集团有限公司雄厚的实力，坚持诚信经营，践行普惠服务，为社会大众提供全方位专业化的证券投、融资服务，帮助客户实现价值增长，努力成为客户认同、社会尊重、股东满意、员工自豪的优秀企业。

中邮证券研究所

北京
邮箱：yanjiusuo@cnpsec.com
地址：北京市东城区前门街道珠市口东大街 17 号
邮编：100050

上海
邮箱：yanjiusuo@cnpsec.com
地址：上海市虹口区东大名路 1080 号邮储银行大厦 3 楼
邮编：200000

深圳
邮箱：yanjiusuo@cnpsec.com
地址：深圳市福田区滨河大道 9023 号国通大厦二楼
邮编：518048