



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

《小市值强势，动量风格依旧——中邮因子周报 20250309》 - 2025. 03. 10
《泛科技大幅回调，融资资金和 ETF 资金逆市流入行业轮动周报 20250302》 - 2025. 03. 03
《高波不再持续，多数风格切换——中邮因子周报 20250302》 - 2025. 03. 03
《3 月胜率最高的策略：多微盘空 1000——微盘股指数周报 20250302》 - 2025. 03. 02
《Deepseek 背景综述及在金融领域应用场景初探》 - 2025. 02. 26
《扩散指数有高位回调风险——微盘股指数周报 20250216》 - 2025. 02. 17
《基本面回撤，高波风格持续——中邮因子周报 20250209》 - 2025. 02. 10
《各资金持续流入机器人，短期注意回调风险，行业轮动开始超配成长——行业轮动周报 20250209》 - 2025. 02. 10
《全面牛市正在到来，微盘有望修复前高——微盘股指数周报 20250209》 - 2025. 02. 10
《基本面表现强势，风格切换加速——中邮因子周报 20250126》 - 2025. 01. 27
《节前融资资金大幅净流出，ETF 资金聚焦机器人以及红利——行业轮动周报 20250126》 - 2025. 01. 26
《信号全部翻多，微盘股高胜率月份即将到来——微盘股指数周报 20250126》 - 2025. 01. 26

金工周报

QwQ-32b 媲美 DeepSeek-R1-671b, Manus 引发智能体浪潮——AI 要闻周报 20250310

● 阿里推出 QwQ-32b 推理模型并开源，效果媲美 DeepSeek-R1

3 月 6 日阿里推出高性能推理模型 QwQ-32b，以 320 亿参数规模在多项任务中展现出与更大规模模型（如 6710 亿参数的 DeepSeek-R1）相媲美甚至超越的性能，成为当前 AI 领域的焦点之一。

● Manus 通用智能体发布引发 AI Agent 浪潮，多个智能体快速跟进

3 月 6 日凌晨，中国 Monica.im 团队推出通用人工智能 Manus，在 GAIA 测试三个难度测试中都达到了最先进（SOTA）表现。除此之外 Manus 在任务执行表现出了强大的功能，从演示实例中可以看到面对具体任务，Manus 可以控制电脑进行规划，自动挂载数据，调用工具执行任务。

● Openai 发布 GPT-4.5 研究预览版

OpenAI 于 2 月 28 日推出 GPT-4.5 AI 模型“研究预览版”，号称截至目前最好的对话大模型。GPT-4.5 提高了识别模式、建立联系和产生创造性见解的能力，知识库更广，更能理解用户意图，并且“情商”更高。

● 字节旗下 AI IDE Trae 发布国内版

字节早前在海外发布的“中国首个 AI 原生集成开发环境（AI IDE）”Trae 于 3 月 3 日官宣上线国内版。海外版本内置 GPT-4o, Claude-3.5-Sonnet 模型。与国内版本不同，国内版本内置豆包 1.5 Pro 大模型，同时支持使用满血版 DeepSeek R1、V3 模型。

● 风险提示：

本报告所有信息基于网络内容整理，不构成投资建议。

目录

| | | |
|-----|--|----|
| 1 | AI 重点要闻 | 4 |
| 1.1 | 字节旗下 AI IDE Trae 发布国内版 | 4 |
| 1.2 | Openai 发布 GPT-4.5 研究预览版 | 6 |
| 1.3 | 阿里推出 QwQ-32b 推理模型并开源，效果媲美 DeepSeek-R1 | 7 |
| 1.4 | Manus 通用智能体发布引发 AI Agent 浪潮，多个智能体快速跟进 | 9 |
| 2 | 企业动态 | 11 |
| 2.1 | AMD 推出 Radeon RX 9000 系新显卡 9070 XT 和 9070，聚焦推理和高性能计算 | 11 |
| 2.2 | 科大讯飞星火 X1 模型升级，对标世界一流模型 | 12 |
| 2.3 | 腾讯发布混元图生视频模型 | 13 |
| 3 | AI 行业洞察 | 14 |
| 3.1 | 杭州投资智谱，杭州城投、上城资本参与投资 | 14 |
| 3.2 | Lmarena.ai 公布大模型排名，Grok3 胜出，DeepSeek-R1 排名第六 | 15 |
| 4 | 技术前沿 | 17 |
| 4.1 | Karpathy 表达对大模型评估基准失效的担忧 | 17 |
| 4.2 | 2B 参数非 SFT 模型在视觉推理过程中重现 DeepSeek “顿悟时刻” | 18 |
| 4.3 | Dynasor-CoT：通过确定性探索实现更高效的思维链 | 19 |
| 5 | 风险提示 | 20 |

图表目录

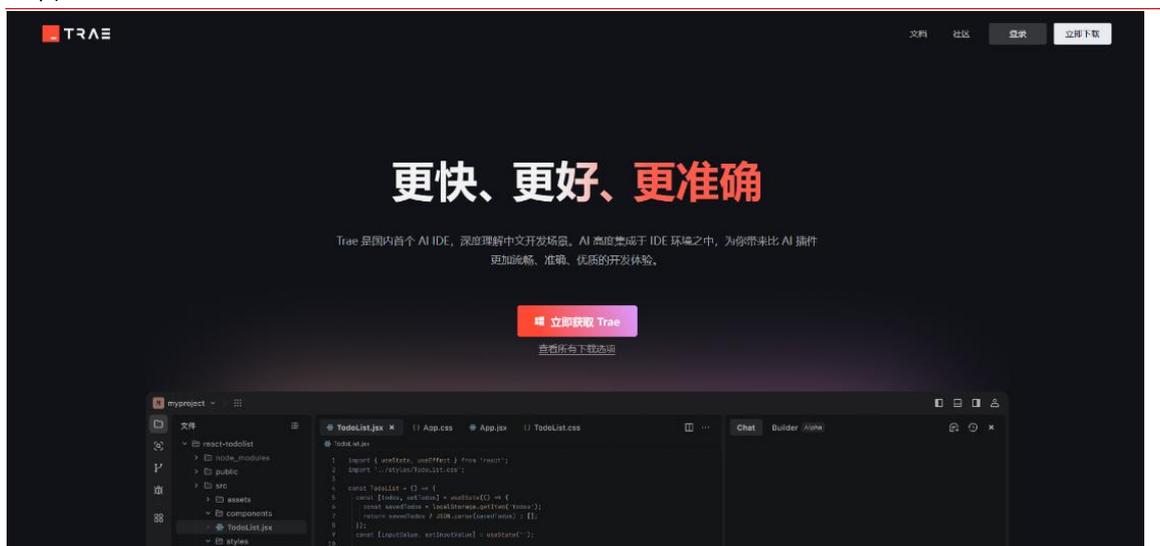
| | |
|--|----|
| 图表 1: Trae | 4 |
| 图表 2: MarsCode 将于 2025 年 4 月 1 日停止维护 | 5 |
| 图表 3: GPT-4.5 评测对比 | 6 |
| 图表 4: GPT-4.5 评测 | 7 |
| 图表 5: QwQ-32b 大模型评测 | 8 |
| 图表 6: Manus 测评 | 9 |
| 图表 7: OWL 架构 | 10 |
| 图表 8: OpenManus Stars 趋势 | 11 |
| 图表 9: RX 9070 系列参数 | 12 |
| 图表 10: 星火 X1 测评 | 13 |
| 图表 11: 腾讯混元大模型-I2V | 14 |
| 图表 12: GLM 大模型 | 15 |
| 图表 13: Imarena.ai 大模型排名综合榜 | 16 |
| 图表 14: Imarena.ai 大模型排名汇总 | 16 |
| 图表 15: Karpathy 质疑大模型评价基准 | 17 |
| 图表 16: VisualThinker-R1-Zero 项目 | 18 |
| 图表 17: 顿悟时刻 | 18 |
| 图表 18: DeepSeek 推理过长 | 19 |
| 图表 19: 有效降低 Token 使用 | 20 |

1 AI 重点要闻

1.1 字节旗下 AI IDE Trae 发布国内版

字节早前在海外发布的“中国首个 AI 原生集成开发环境 (AI IDE)” Trae 于 3 月 3 日官宣上线国内版。海外版本内置 GPT-4o, Claude-3.5-Sonnet 模型。与海外版本不同,国内版本内置豆包 1.5 Pro 大模型,同时支持使用满血版 DeepSeek R1、V3 模型。

图表 1: Trae



资料来源: Trae, 中邮证券研究所

Trae 核心功能与亮点:

- **AI 驱动的智能编程**

Trae 内置多种 AI 大模型,国内版搭载豆包 1.5 Pro 和满血版 DeepSeek R1/V3 模型,支持自然语言生成代码、智能补全、代码理解与修复等功能。例如,用户输入“开发一个带用户登录功能的论坛”,Trae 可快速生成前后端框架代码,显著降低从零搭建项目的门槛。

- **多模态交互与 Builder 模式**

开发者可通过 Builder 模式用自然语言描述需求,由 AI 自动分解

任务、生成项目框架并持续优化代码。此外，Trae 支持上传图片（如设计草图或错误截图），通过多模态技术精准解析需求，进一步简化协作流程。

- **全场景开发支持**

涵盖代码编辑、版本控制、调试等传统 IDE 功能，同时支持实时协作与云端开发。其“全局上下文分析”能力可理解整个项目的代码逻辑，提供精准的代码建议和错误定位，适用于 Web 开发、工具应用、游戏定制等场景。

- **低门槛与跨平台兼容**

Trae 不仅服务于专业开发者，也面向非技术用户（如产品经理或爱好者），支持直接导入开源项目并快速运行调试。例如，用户可对英文界面的开源游戏 HexGL 进行中文翻译和功能扩展，无需复杂配置。

目前国内版支持 Windows 10/11 和 macOS, Linux 版本正在开发中。

Trae 是继 MarsCode IDE、Coze IDE、掘金 AI 后字节开发的又一 AI 产品，与 Cursor、Windsurf 相似，目标为 AI 辅助编程（L2）之上的 AI 结对编程（L3）。字节在发布 Trae 的同时，Marscode 团队也宣布运营将于 2025 年 4 月 1 日结束，之后不再维护，由此也可看出字节在该赛道将全力发展 Trae。

图表 2: MarsCode 将于 2025 年 4 月 1 日停止维护



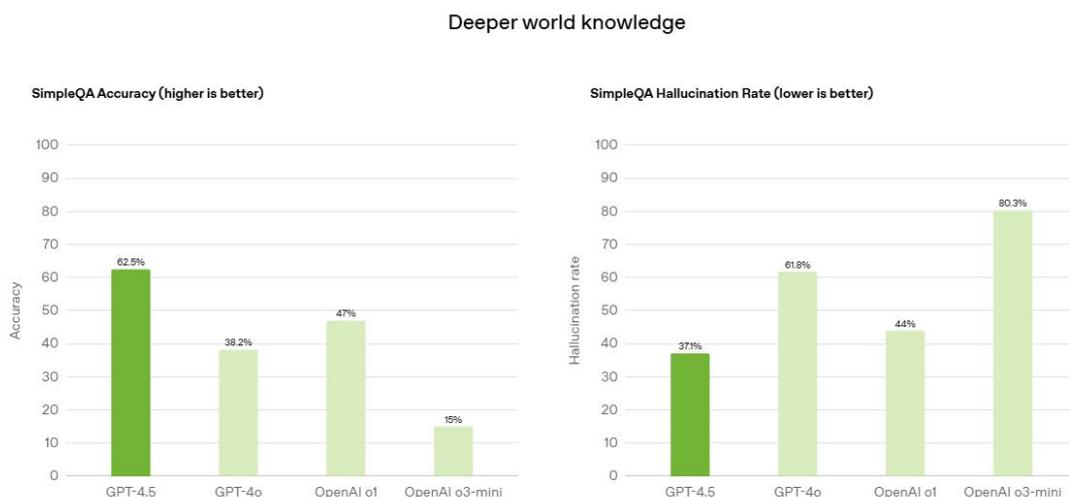
资料来源：MarsCode，中邮证券研究所

1.2 Openai 发布 GPT-4.5 研究预览版

OpenAI 于 2 月 28 日推出 GPT-4.5 AI 模型“研究预览版”，号称截至目前最好的对话大模型。GPT-4.5 提高了识别模式、建立联系和产生创造性见解的能力，知识库更广，更能理解用户意图，并且“情商”更高。

首先 GPT-4.5 扩展了无监督学习和推理，OpenAI 认为无监督学习提高了世界模型的准确性和直觉性（GPT-3.5、GPT-4 和 GPT-4.5 的范式），同时扩展推理教会模型在做出反应之前思考并产生思路链，从而使它们能够解决复杂的 STEM 或逻辑问题（o1、o3-mini 的范式）。GPT-4.5 在 Microsoft Azure AI 超级计算机上进行了训练。结果是模型拥有更广泛的知识和对世界的更深入理解，从而减少了幻觉并在广泛的主题上提高了可靠性。

图表 3：GPT-4.5 评测对比

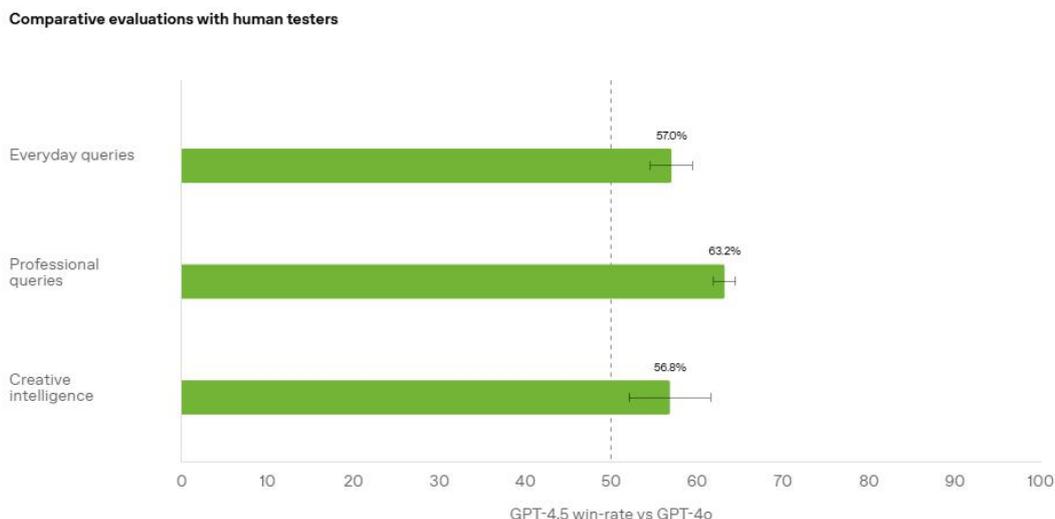


资料来源：OpenAI，中邮证券研究所

为了解决让大模型更好更深入理解人类需求和意图的问题，GPT-4.5 使用了可扩展技术方法，能够利用从小型模型衍生的数据训练更大、更强大的模型。这些技术显著提升了 GPT-4.5 的可操控性、对细微差别的理解能力以及自然对话流畅度。通过将对世界的深度理解与改进的协作能力相结合，该模型能够在温暖

且直觉化的对话中自然融合观点，更加贴合人类协作的节奏。GPT-4.5 能更精准捕捉人类的意图，以更强的“情商”解读隐含的线索或未明说的期望。

图表 4：GPT-4.5 评测



资料来源：OpenAI，中邮证券研究所

1.3 阿里推出 QwQ-32b 推理模型并开源，效果媲美 DeepSeek-R1

3月6日阿里推出高性能推理模型 QwQ-32b，以 320 亿参数规模在多项任务中展现出与更大规模模型（如 6710 亿参数的 DeepSeek-R1）相媲美甚至超越的性能，成为当前 AI 领域的焦点之一。以下是其核心特点及技术亮点：

- **参数效率与性能突破：**

参数规模与性能对比：QwQ-32B 仅拥有 320 亿参数，约为 DeepSeek-R1 的 1/20，但通过强化学习（RL）优化，其在数学推理（如 AIME24 评测集准确率 89.7%）、代码生成（LiveCodeBench 一次通过率 82%）及通用能力（如 Meta 的 LiveBench、Google 的 IFEval）等任务中表现与 DeepSeek-R1 相当，甚至部分指标实现超越。

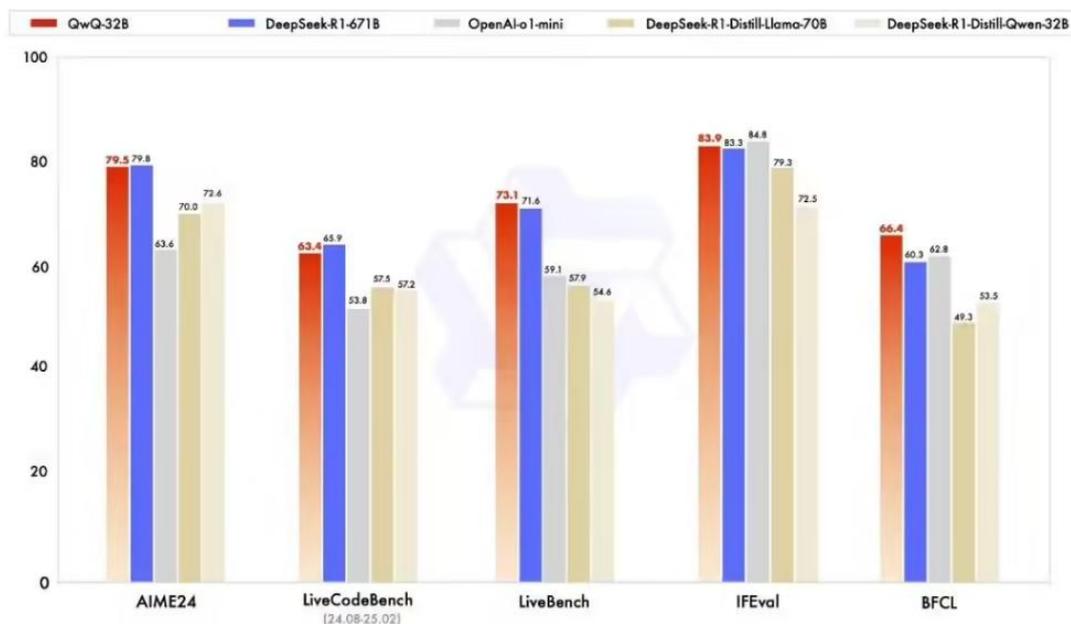
多阶段强化学习训练：模型采用分阶段 RL 训练策略。第一阶段：针对数学和编程任务，通过答案正确性校验和代码执行反馈优化模型，确保逻辑严谨性。第二阶段：引入通用奖励模型和规则验证器，提升指令遵循、工具调用等通用能力，同时避免数学与编程性能下降。

- **创新与功能集成**

动态推理与智能体能力：模型集成 Agent 相关功能，可结合工具使用进行批判性思考，并根据环境反馈动态调整推理策略，适用于复杂交互场景。

高效训练机制：采用“分步验证”和实时反馈机制，例如数学问题逐行验证、代码生成动态优化，显著提升输出准确性。

图表 5: QwQ-32b 大模型评测



资料来源：通义千问，中邮证券研究所

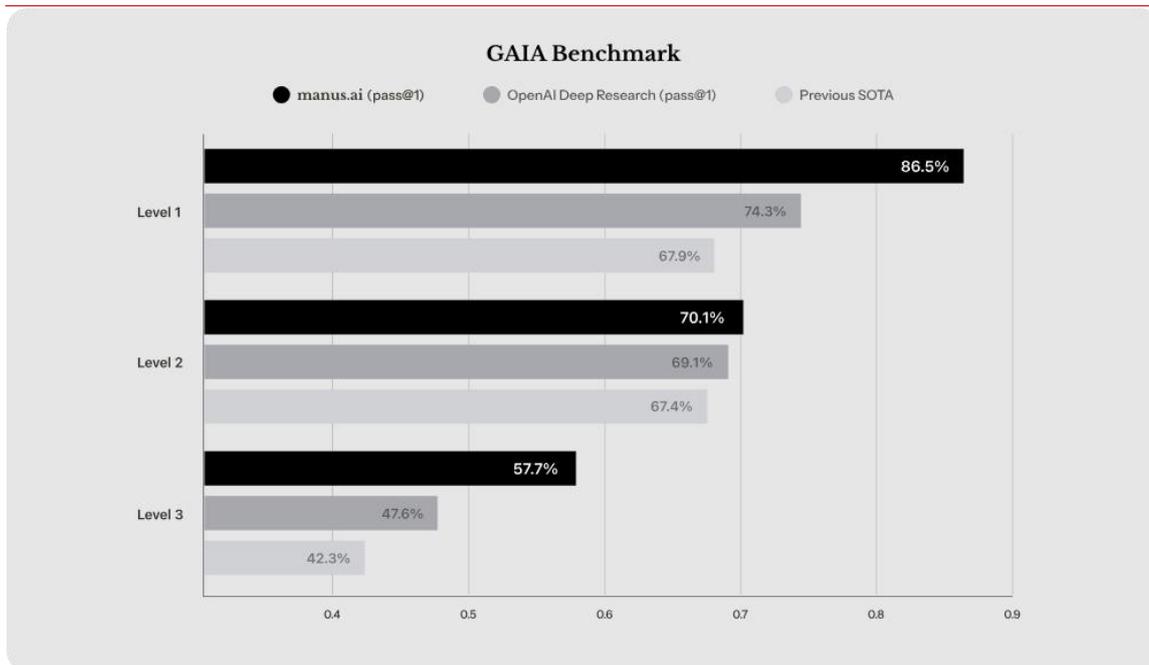
在考察数学能力的基准 AIME24、考察大模型代码能力的基准 LiveCodeBench、“最难大模型评测榜” LiveBench、测试中，QwQ-32b 表现和 DeepSeek-R1 (671b) 相当且远超 GPT-o1-mini；在考察指令遵循能力的评测基准 IFEval 打分上，QwQ-32b 与 DeepSeek-R1 (671b) 以及 GPT-o1-mini 表现相当；在考察模型函数调用与使用的测试基准 BFCL 上，QwQ-32b 甚至超过了 DeepSeek-R1 (671b) 等众多对标模型，体现出超强实力。

QwQ-32B 凭借“小参数、高性能”的设计理念和开源策略，不仅降低了 AI 应用门槛，还为教育、科研、企业数字化等领域提供了高效工具，成为推动 AI 普惠化的重要里程碑。其技术路径验证了“强化学习+强大基座模型”在实现高效推理中的潜力，未来或将成为行业新标杆。

1.4 Manus 通用智能体发布引发 AI Agent 浪潮，多个智能体快速跟进

3月6日凌晨，中国 Monica.im 团队推出通用人工智能 Manus，在 GAIA 测试三个难度测试中都达到了最先进（SOTA）表现。除此之外 Manus 在任务执行表现出了强大的功能，从演示实例中可以看到面对具体任务，Manus 可以控制电脑进行规划，自动挂载数据，调用工具执行任务。

图表 6: Manus 测评



资料来源：Manus，中邮证券研究所

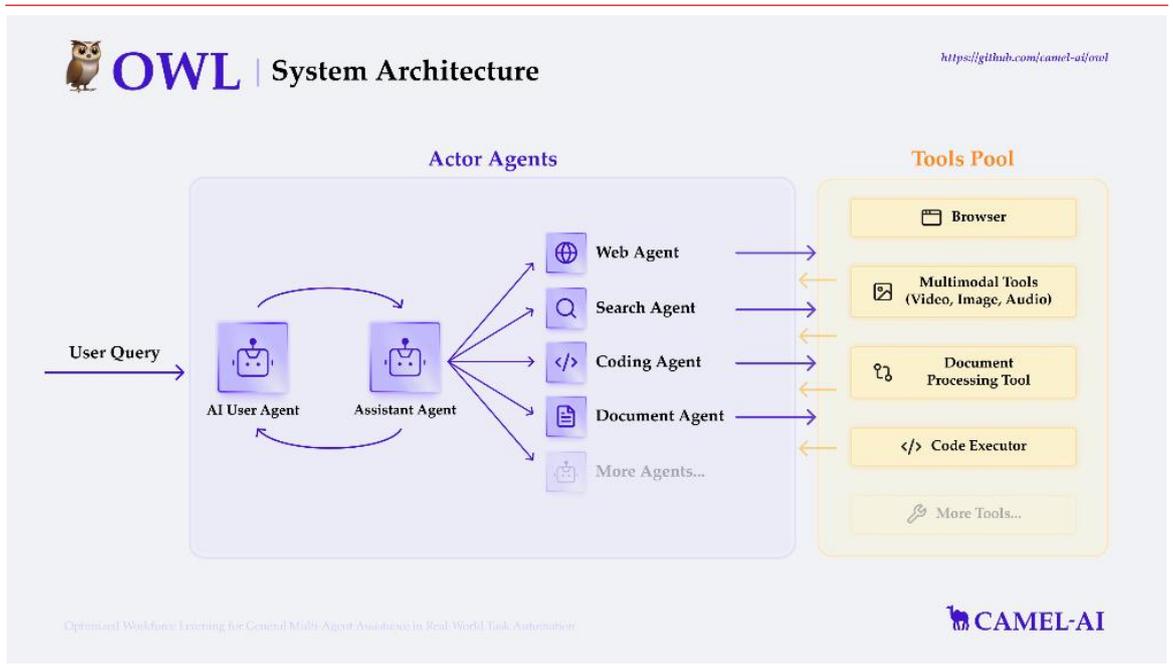
在 Manus 发布后，号称复现 Manus 的开源项目 OWL 也快速上线，并在文章中分析了 Manus 的工作流程：

1. 启动一个 Ubuntu 容器（Agent 远程工位就位）
2. 知识召回（把之前学过的内容捡起来用）
3. 连接数据源（数据库、网盘、云存储全覆盖）
4. 把数据挂载到 Ubuntu（Agent 的搬砖时刻）
5. 自动生成 todo.md（规划任务+写待办清单）
6. Ubuntu 工具链+外接工具组合拳，执行全流程任务

除此之外，OWL 团队也分析了智能体使用过程中需要用到的工具并梳理：

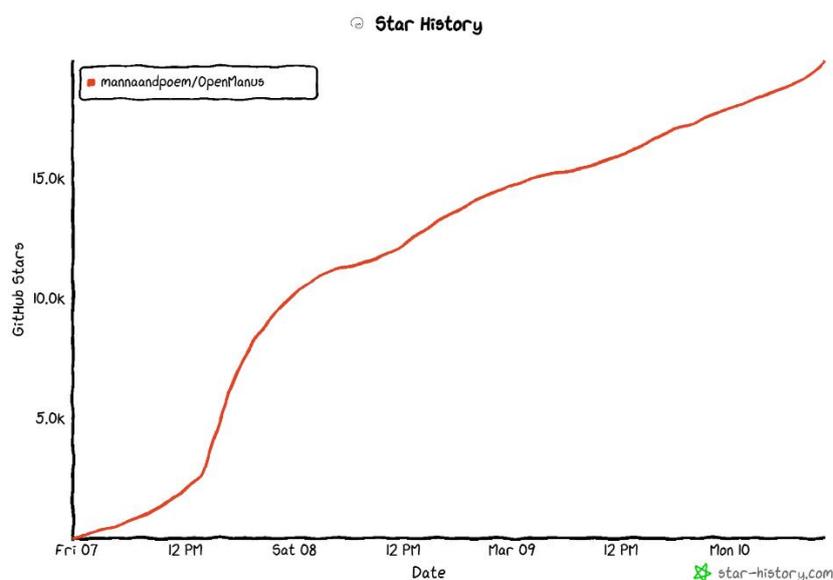
- 终端命令执行（运维、部署全搞定）
- 文件解析：PDF 转 Markdown、网页爬取
- 自动创建、编辑文件，生成 todo.md
- 浏览器操作：滚动、点击、输入全会
- 在线搜索+实时信息检索，一秒找到关键资料
- 自动生成报告、代码、文档，直接交付成果
- 一键把服务部署到公网，输出报告直接上线

图表 7：OWL 架构



资料来源：CAMEL-AI，中邮证券研究所

除此之外，也有团队推出 OpenManus 开源项目，快速复现了 Manus 智能体，在 Github 上已收获 19.8k Stars。

图表 8: OpenManus Stars 趋势


资料来源: OpenManus, 中邮证券研究所

2 企业动态

2.1 AMD 推出 Radeon RX 9000 系新显卡 9070 XT 和 9070，聚焦推理和高性能计算

RX 9070 XT 和 RX 9070 基于 AMD 最新的 RDNA 4 架构，采用台积电 4nm 工艺，晶体管数量达 539 亿，密度为 151 MTr/mm²，显著提升了能效比和性能上限。两款显卡均配备 16GB GDDR6 显存（256-bit 位宽，640 GB/s 带宽），支持 PCIe 5.0 接口，并搭载 HDMI 2.1b 和 DisplayPort 2.1a 接口，兼容 8K 144Hz 和 4K 240Hz 显示器。

核心参数：

图表 9：RX 9070 系列参数

| 型号 | RX 9070 XT | RX 9070 |
|---------|-------------|-------------|
| 计算单元 | 64 CU | 56 CU |
| 流处理器 | 4096 | 3584 |
| 加速频率 | 2970 MHz | 2520 MHz |
| 基准频率 | 2400 MHz | 2070 MHz |
| FP32算力 | 48.7 TFLOPs | 36.1 TFLOPs |
| TBP功耗 | 304W | 220W |
| 显存带宽 | 640 GB/s | 640 GB/s |
| 光线追踪加速器 | 64个 | 56个 |

资料来源：AMD，中邮证券研究所

RX 9070 系显卡配备第二代 AI 加速器：支持 FP8 数据类型和结构化稀疏性优化，INT8 吞吐量相比 RDNA 3 提升 8 倍，可高效运行生成式 AI 应用（如 Stable Diffusion）和创意 AI 任务（如视频降噪）；同时配备第三代光线追踪加速器：每计算单元的光追吞吐量是前代的两倍，显著提升 4K 光追游戏中的光影效果与帧率。

市场定价方面 RX 9070 国行起售价 4499 元，RX 9070 XT 为 4999 元，主打“以 1440p 价格提供 4K 性能”的高性价比策略，目标直指中高端市场。

2.2 科大讯飞星火 X1 模型升级，对标世界一流模型

星火 X1 模型数学能力全面对标国际顶尖模型，星火 X1 在参数量仅为 70B（远低于 DeepSeek R1 的 671B）的情况下，通过国产算力训练实现了数学能力与 DeepSeek R1 及 OpenAI o1 的全面对标。其升级后在中小学作业批改、竞赛级难题解答（如奥数、高考题）等任务中表现显著提升，尤其在中文数学任务中全面领先。

图表 10：星火 X1 测评


资料来源：科大讯飞，中邮证券研究所

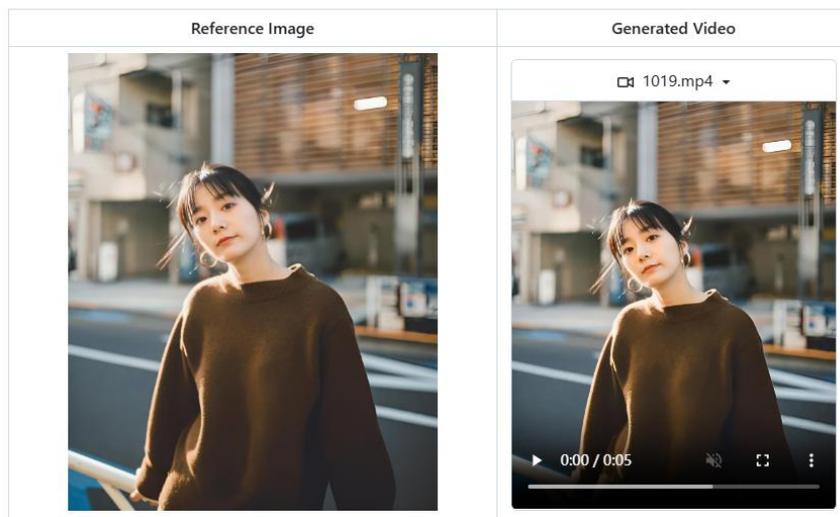
除此之外，星火 X1 还针对具体场景开发模型。基于星火 X1 升级的星火医疗大模型 X1，在诊断推荐、健康咨询、检查报告解读等场景中，效果超越 GPT-4o 和 DeepSeek R1。其通过亿级医学知识库和诊疗数据训练，结合循证思维链技术，显著降低“医疗幻觉”问题，并提升了复杂推理的可解释性。例如，在用药建议中能结合患者既往病史提供个性化方案。

2.3 腾讯发布混元图生视频模型

3月6日，腾讯混元发布图生视频模型“HunyuanVideo-I2V”并开源，可以做到根据图片生成5秒短视频，并可以配背景音。

图表 11：腾讯混元大模型-I2V

Frist Frame Consistency Demo



资料来源：腾讯混元，中邮证券研究所

该模型是混元文生视频模型的延续，模型参数维持 13B，项目现已开源，项目地址：<https://github.com/Tencent/HunyuanVideo-I2V>。

3 AI 行业洞察

3.1 杭州投资智谱，杭州城投、上城资本参与投资

智谱 AI 于 2025 年 3 月 3 日宣布完成超 10 亿元人民币的战略融资，投资方包括杭州城投产业基金和上城资本等地方国资平台。此次融资是继 2024 年 12 月完成 30 亿元融资后的又一次重要资本动作，也是 2025 年大模型“六小虎”中的首次公开融资。

新资金将重点投向 GLM 大模型技术体系的迭代升级，包括基座模型、推理模型、多模态模型及智能体 (Agent) 的研发，并计划在 2025 年全面开源这些模型，推动 AI 技术的普惠化。智谱是国内最早开源大模型的企业之一，其 GLM 架构已兼容 40 余种国产 GPU 芯片，且在幻觉率控制上表现领先。

图表 12: GLM 大模型

| 模型名称 | 参数规模 | 类型 | 说明 |
|------------------|------|--------|----------------------------------|
| ChatGLM3-6B | 6亿 | 大语言模型 | 最新一代模型，支持工具调用、代码执行和Agent任务等复杂场景。 |
| ChatGLM3-6B-32K | 6亿 | 大语言模型 | 支持32K上下文的版本，适用于更长文本的处理。 |
| ChatGLM2-6B | 6亿 | 大语言模型 | 第二代版本，支持中英双语。 |
| ChatGLM3-6B-Base | 6亿 | 大语言模型 | 基础模型，采用更多样的训练数据和合理的训练策略。 |
| GLM-4-9B-Chat | 9亿 | 大语言模型 | GLM-4的对齐版本，专注于人类偏好的优化。 |
| AgentLM-7B | 7亿 | 智能体模型 | 基于AgentTuning方法进行微调，具备强大的智能体能力。 |
| AgentLM-13B | 13亿 | 智能体模型 | 同样基于AgentTuning方法，适用于复杂任务。 |
| AgentLM-70B | 70亿 | 智能体模型 | 更大规模的智能体模型，适合高复杂度应用场景。 |
| CogVideoX | - | 视频生成模型 | 视频生成模型，国内首个通过API来使用的文生视频和图生视频服务。 |
| GLM-4-Voice | - | 情感语音模型 | 首个开源的端到端多模态模型 |

资料来源：量子位，中邮证券研究所

3.2 Lmarena. ai 公布大模型排名，Grok3 胜出，DeepSeek-R1 排名第六

3月4日，Lmarena.ai 在 X 上公布了最新的大模型天梯榜，2月24日发布的 Grok3 在总体评价中以一票优势登顶。截至2025年3月10日，GPT-4.5 在发布后排名快速上升，与 Grok3 并列总体评价第一名，来自中国的 DeepSeek-R1 和 Qwen2.5-Max 分别位列第六位和第九位。在数学能力中，DeepSeek-R1 并列第一名。在编程能力排名中，GPT-4.5 和 Grok3 并列第一名，DeepSeek-R1 并列第三名。

图表 13: Imarena.ai 大模型排名综合榜

| Category | | Apply filter | | Overall Questions | | | |
|------------|------------------|--|--|---------------------|--------------------------|--------------|-------------|
| Overall | | <input type="checkbox"/> Style Control | <input type="checkbox"/> Show Deprecated | #models: 211 (100%) | #votes: 2,736,442 (100%) | | |
| Rank* (UB) | Rank (StyleCtrl) | Model | Arena Score | 95% CI | Votes | Organization | License |
| 1 | 2 | Grok-3-Preview-02-24 | 1412 | +8/-10 | 3364 | xAI | Proprietary |
| 1 | 1 | GPT-4.5-Preview | 1411 | +11/-11 | 3242 | OpenAI | Proprietary |
| 3 | 5 | Gemini-2.0-Flash-Thinking-Exp-01-21 | 1384 | +6/-5 | 17487 | Google | Proprietary |
| 3 | 3 | Gemini-2.0-Pro-Exp-02-05 | 1380 | +5/-6 | 15466 | Google | Proprietary |
| 3 | 2 | ChatGPT-4o-latest...(2025-01-29) | 1377 | +5/-4 | 17221 | OpenAI | Proprietary |
| 6 | 3 | DeepSeek-R1 | 1363 | +8/-6 | 8580 | DeepSeek | MIT |
| 6 | 10 | Gemini-2.0-Flash-001 | 1357 | +6/-5 | 13257 | Google | Proprietary |
| 7 | 3 | o1-2024-12-17 | 1352 | +4/-6 | 19785 | OpenAI | Proprietary |
| 9 | 10 | Qwen2.5-Max | 1336 | +7/-5 | 11930 | Alibaba | Proprietary |
| 9 | 7 | o1-preview | 1335 | +4/-3 | 33167 | OpenAI | Proprietary |
| 9 | 10 | o3-mini-high | 1329 | +8/-6 | 9102 | OpenAI | Proprietary |
| 11 | 13 | DeepSeek-V3 | 1318 | +5/-4 | 22007 | DeepSeek | DeepSeek |

资料来源: Imarena.ai, 中邮证券研究所

图表 14: Imarena.ai 大模型排名汇总

| Model | Overall | Overall w/ Style Control | Hard Prompts | Hard Prompts w/ Style Control | Coding | Math | Creative Writing | Instruction Following | Longer Query | Multi-Turn |
|-------------------------------------|---------|--------------------------|--------------|-------------------------------|--------|------|------------------|-----------------------|--------------|------------|
| grok-3-preview-02-24 | 1 | 2 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 2 |
| gpt-4.5-preview-2025-02-27 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| gemini-2.0-flash-thinking-exp-01-21 | 3 | 5 | 2 | 2 | 3 | 1 | 2 | 2 | 2 | 2 |
| gemini-2.0-pro-exp-02-05 | 3 | 3 | 1 | 2 | 3 | 1 | 2 | 3 | 2 | 2 |
| chatgpt-4o-latest-20250129 | 3 | 2 | 5 | 3 | 3 | 11 | 1 | 3 | 1 | 2 |
| deepseek-r1 | 6 | 3 | 3 | 1 | 3 | 1 | 3 | 3 | 6 | 2 |
| gemini-2.0-flash-001 | 6 | 10 | 4 | 9 | 3 | 1 | 6 | 7 | 6 | 4 |
| o1-2024-12-17 | 7 | 3 | 3 | 2 | 3 | 1 | 6 | 3 | 2 | 7 |
| qwen2.5-max | 9 | 10 | 5 | 5 | 7 | 5 | 8 | 8 | 7 | 7 |
| o1-preview | 9 | 7 | 5 | 4 | 4 | 1 | 9 | 8 | 7 | 7 |
| o3-mini-high | 9 | 10 | 3 | 2 | 3 | 1 | 16 | 8 | 4 | 9 |
| deepseek-v3 | 11 | 13 | 14 | 16 | 15 | 13 | 7 | 11 | 7 | 7 |
| glm-4-plus-0111 | 12 | 16 | 15 | 19 | 18 | 17 | 9 | 14 | 14 | 13 |

资料来源: Imarena.ai, 中邮证券研究所

4 技术前沿

4.1 Karpathy 表达对大模型评估基准失效的担忧

Karpathy 近日在 X 上发帖称其认为当前存在一场评估危机。以致于现在真的不知道应该关注哪些指标了。并且提到几个评估基准并评价道：

MMLU (大规模多任务语言理解)：曾是一个优秀且实用的基准，但它的黄金期早已过去。

SWE-Bench Verified (真实、实用、已验证的问题集)：我非常喜欢且认为很有价值，但它本身覆盖范围太窄。

Chatbot Arena：受到了过多关注，导致各大 LLM 实验室开始严重过拟合它——通过 API 请求的提示挖掘、对私有评估集的狂轰滥炸，甚至更糟糕的是直接将排名结果作为训练监督信号。我认为它目前还算可用，但缺乏“更好的替代品”，且其信号价值似乎正在下降。

针对这些问题，Karpathy 提出的解决方法是：目前涌现了许多私有评估集，将它们组合使用或许是一条有潜力的路径。与此同时，在缺乏全面优秀评估体系的情况下，我尝试转而依赖“氛围检查”（直觉判断），但现在担心它们会误导人，因为存在太多确认偏误的可能性，样本量也过小，总之并不理想。

图表 15: Karpathy 质疑大模型评价基准



资料来源：Andrej Karpathy，中邮证券研究所

4.2 2B 参数非 SFT 模型在视觉推理过程中重现 DeepSeek“顿悟时刻”

这是一款名叫 VisualThinker-R1-Zero 的项目，思路是很多团队在复现 DeepSeek-R1 的研究过程中发现，很难复现出 DeepSeek-R1 所表现出的回答长度增加和思考模式，也就是 DeepSeek-R1 模型的自我进化和“顿悟时刻”所表现出的涌现现象。但来自 UCLA 的团队成功在多模态推理中复现了“顿悟时刻”，并且只用了参数量为 2B 的未经监督微调的模型。

图表 16: VisualThinker-R1-Zero 项目

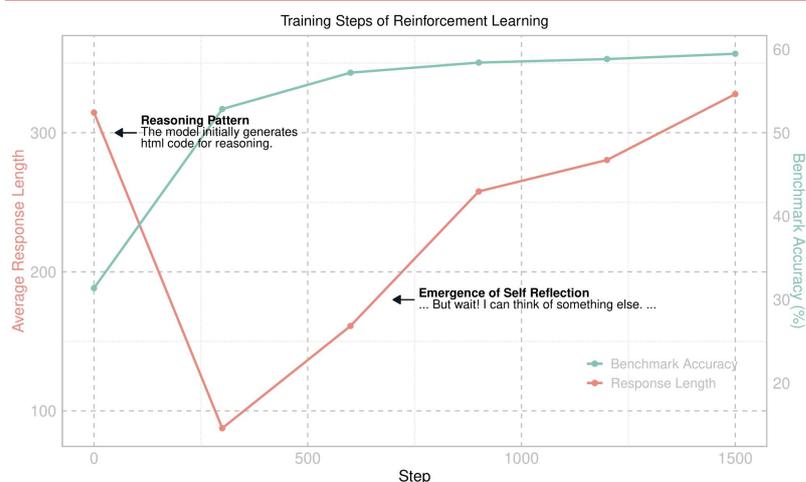
R1-Zero's "Aha Moment" in Visual Reasoning on a 2B Non-SFT Model

Hengguang Zhou*, Xirui Li*, Ruochen Wang†, Minhao Cheng, Tianyi Zhou and Cho-Jui Hsieh

资料来源: VisualThinker-R1-Zero, 中邮证券研究所

该项目是全球首次在未经 SFT 的 Qwen2-VL-2B 基础模型上，通过纯强化学习 (RL) 实现了多模态推理的“顿悟时刻”，并在 CVBench 视觉基准测试中达到 59.47% 准确率，较基础模型提升 30%，超越指令微调模型约 5%。

图表 17: 顿悟时刻



资料来源: VisualThinker-R1-Zero, 中邮证券研究所

训练策略：

直接对未微调的基础模型应用 RL，而非传统先 SFT 后 RL 的路径。与此同时采用了和 DeepSeek-R1 相同的 GRPO 算法，结合基于规则的奖励函数，奖励分别为准确性奖励和格式奖励。

实验设计：

数据集选用了 SAT 训练集的 12,000 个空间推理问题，对比设置为冻结视觉编码器或大语言模型参数，验证 RL 对推理能力的真实影响。

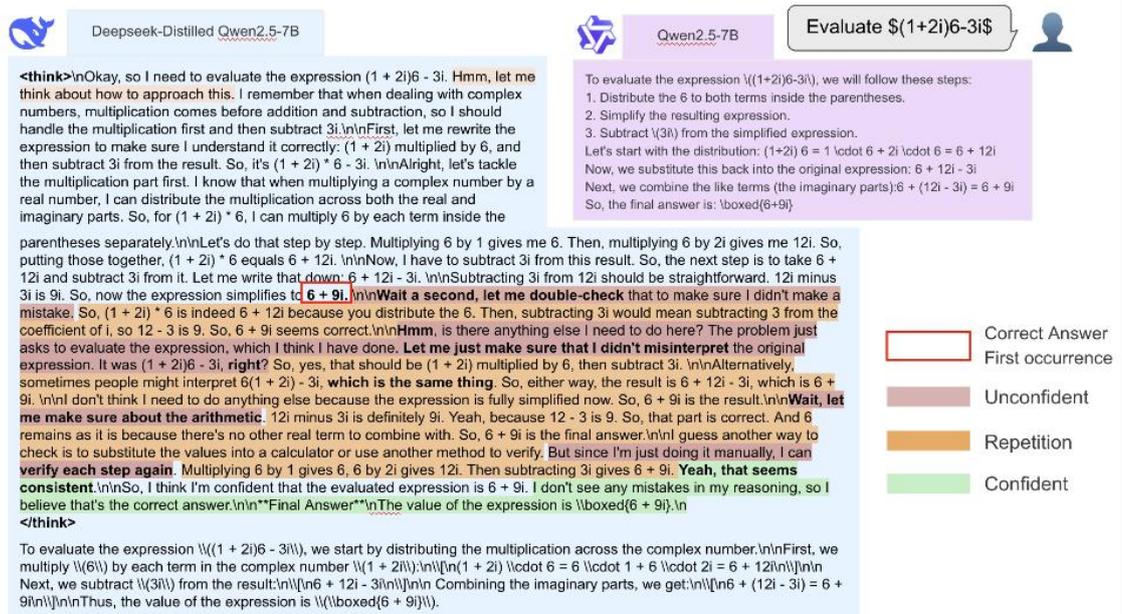
项目现已开源，项目链接：

<https://github.com/turningpoint-ai/VisualThinker-R1-Zero>

4.3 Dynasor-CoT：通过确定性探索实现更高效的思维链

近日，清华和 UCSD 的研究人员提出了 Dynasor-CoT，最终结果证明该方法可以有效降低 Token 使用。研究人员在之前的研究中发现，大模型其实在输出前就已经获得了问题的答案，但会在反复确认的环节中消耗大量 Token，经测算 DeepSeek 模型浪费在这些验算环节的资源高达 70%。

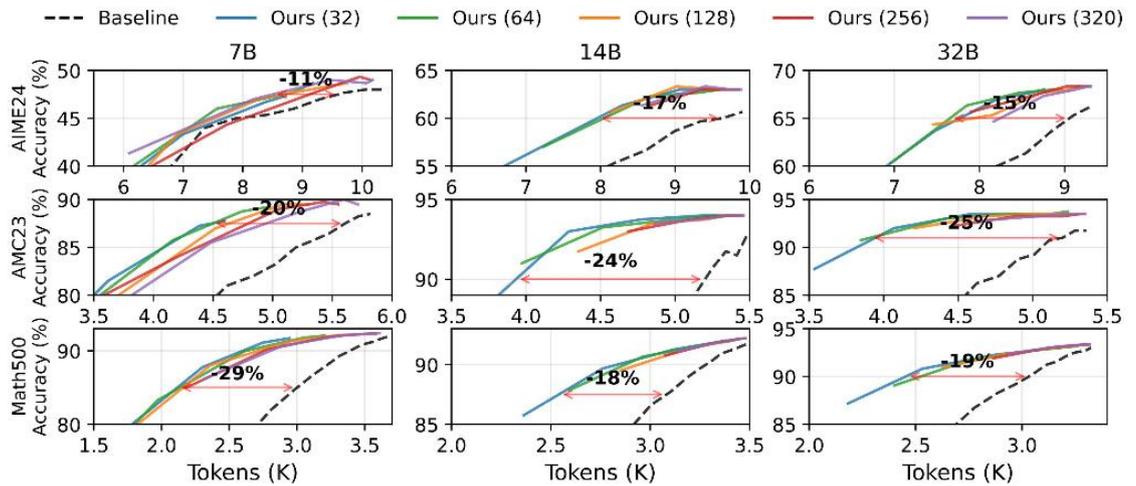
图 18：DeepSeek 推理过长



资料来源：Dynasor-CoT，中邮证券研究所

为了解决大模型自我怀疑的问题，团队推出了 Dynasor-CoT 方法用于 CoT 上。该技术通过思维扫描的方法，可以动态确定思维链应该终止的点，以达到主动终止推理的效果，通过实验发现该方法可以有效节省计算资源。

图表 19：有效降低 Token 使用



资料来源：Dynasor-CoT，中邮证券研究所

5 风险提示

本报告所有信息基于网络内容整理，不构成投资建议。

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。

中邮证券可发出其它与本报告所载信息不一致或有不同结论的报告。报告所载资料、意见及推测仅反映研究人员于发出本报告当日的判断，可随时更改且不予通告。

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客户中的专业投资者使用，若您非中邮证券客户中的专业投资者，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专业投资者。

本报告版权归中邮证券所有，未经书面许可，任何机构或个人不得存在对本报告以任何形式进行翻版、修改、节选、复制、发布，或对本报告进行改编、汇编等侵犯知识产权的行为，亦不得存在其他有损中邮证券商业性权益的任何情形。如经中邮证券授权后引用发布，需注明出处为中邮证券研究所，且不得对本报告进行有悖原意的引用、删节或修改。

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，注册资本50.6亿元人民币。中邮证券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。此外，公司还具有：证券经纪人业务资格；企业债券主承销资格；沪港通；深港通；利率互换；投资管理人受托管理保险资金；全国银行间同业拆借；作为主办券商在全国中小企业股份转让系统从事经纪、做市、推荐业务资格等业务资格。

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