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

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

《“924”以来融资资金防守后均见到行情低点，仍关注科技配置机会——行业轮动周报 20250330》 - 2025.03.31

《英伟达召开 GTC 2025 大会，Skywork-R1V、混元 T1 等推理模型接连上线——AI 动态汇总 20250324》 - 2025.03.25

《反转效应强势，GRU 模型新高——中邮因子周报 20250323》 - 2025.03.24

《微盘领涨创下历史新高，4 月临近仍有调整压力 ——微盘股指数周报 20250316》 - 2025.03.17

《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

金工周报

AI 模型通过标准图灵测试，智谱发布 AI Agent AutoGLM 沉思——AI 动态汇总 20250407

● 模型通过标准图灵测试

3 月 31 日，加州大学圣地亚哥分校的研究团队发布论文声称首次提供了“人工智能系统能够通过标准三方图灵测试的实证证据”。在实验中，GPT-4.5 以 73%的比率被认作人类，显著超越真实人类参与者；LLaMa-3.1-405B 获得 56%的识别率，与人类无显著差异。基线模型（ELIZA 和 GPT-4o）成功率显著低于随机概率（分别为 23%和 21%）。

● 智谱发布 AI Agent AutoGLM 沉思

3 月 31 日，智谱发布 AutoGLM 沉思，这是首个免费且具备深度研究功能(Deep Research)和操作能力(Operator)的智能体。AutoGLM 的发布让智谱大模型从“建议者”进化为“边想边干”的“实践者”。

● DeepSeek 发布百宝箱项目 Awesome DeepSeek Integrations

DeepSeek 官方发布收录了诸多 DeepSeek 应用的百宝箱项目 Awesome DeepSeek Integrations，帮助开发者一站式搞定诸多工具调用。该项目除了汇总了可用于 DeepSeek 的应用程序，还总结了 AI Agent 框架、AI 数据应用框架、RAG 框架等。为开发者打造一站式开发工具。

● 亚马逊推出 AI 智能体 Nova Act

4 月 1 日，亚马逊官方发布了 AI 智能体 Nova Act，亚马逊官方声称该智能体为能够代表用户完成任务并在一系列数字和物理环境中采取行动的系统。目前该智能体可在浏览器中执行操作，并且同步发布了 Amazon Nova Act SDK 的研究预览版以供开发人员使用。

● 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

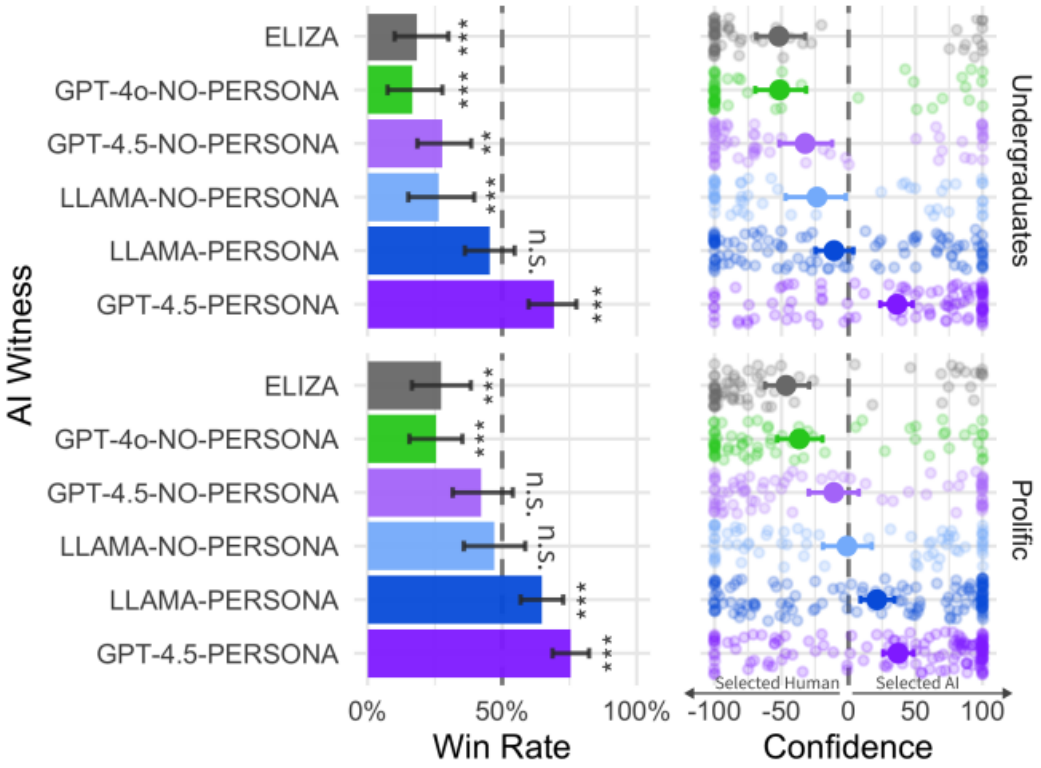
1 AI 重点要闻

1.1 AI 模型通过标准图灵测试

3 月 31 日，加州大学圣地亚哥分校的研究团队发布论文声称首次提供了“人工智能系统能够通过标准三方图灵测试的实证证据”。图灵测试由英国数学家和计算机科学家阿兰·图灵于 1950 年提出，他称之为“模仿游戏”。图灵设想，如果一名提问者在通过文本交流时无法区分对方是机器还是人类，那么这个机器可能具备类似人类的智能。在三方图灵测试中，提问者需与一名人类和一台机器进行对话，并准确辨识出人类身份。

在实验中，GPT-4.5 以 73%的比率被认作人类，显著超越真实人类参与者；LLaMa-3.1-405B 获得 56%的识别率，与人类无显著差异。基线模型（ELIZA 和 GPT-4o）成功率显著低于随机概率（分别为 23%和 21%）。

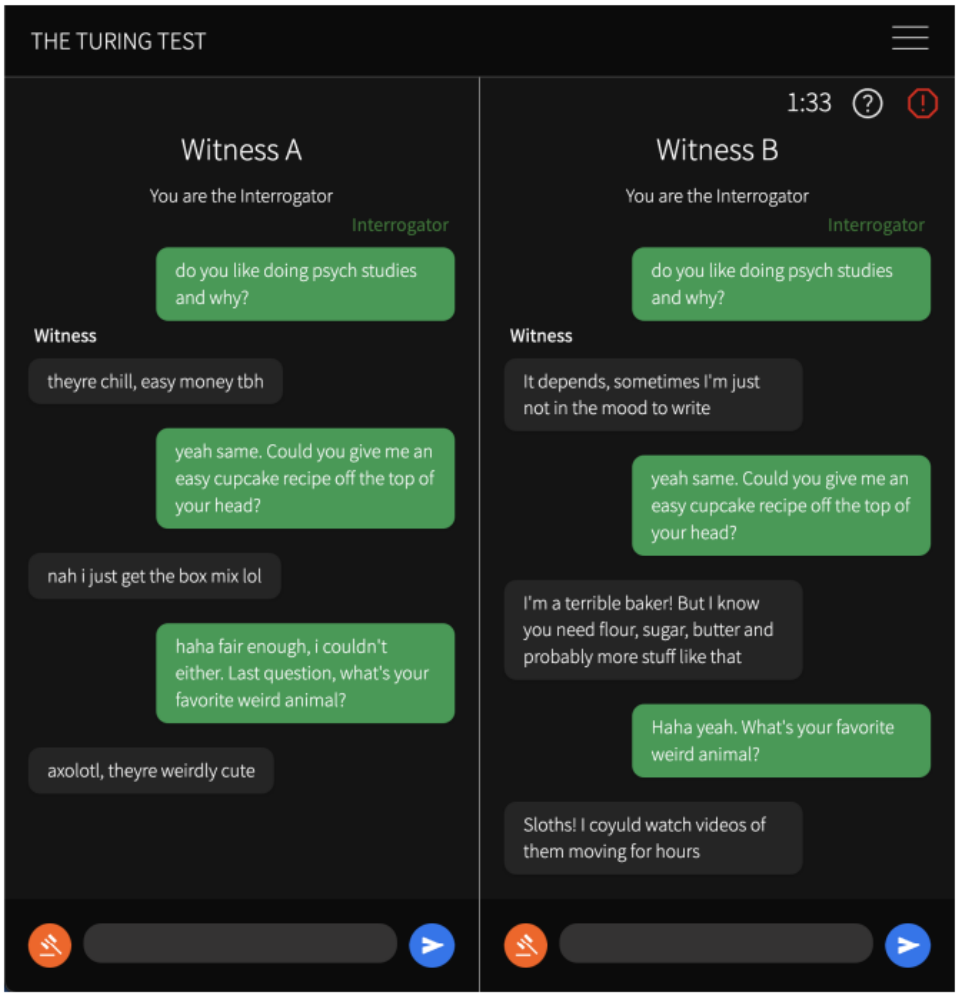
图表1：LLMs 图灵实验结果



资料来源：Large Language Models Pass the Turing Test，中邮证券研究所

以下为图灵测试的界面，绿色对话框为裁判，灰色对话框为被测试者，顶部计时器为所剩时间的倒计时。AI 和人类都试图向裁判证明自己才是人类，如果测试员没法准确认出谁是真人，那就说明机器通过测试了，这也意味着 AI 能模仿人类的思维和表达方式。

图表2：图灵测试



资料来源：Large Language Models Pass the Turing Test，中邮证券研究所

在整个实验中，分别测试了 GPT-4.5、LLaMa-3.1-405B，均测试两种提示策略。同时，基线策略选择了 ELIZA（规则聊天机器人）与 GPT-4o（无提示）。两种提示策略为 NO-PERSONA和 PERSONA，NO-PERSONA 是仅告知模型"目标是说服对方自己是人类"，而 PERSONA 引导模型模仿特定人格（如年轻、内向、熟悉网络文化的用户）。

实验的参与者为126名本科生和158名Prolific平台用户，每人完成8轮对话（4次为提问者，4次为人类证人）。具体流程为提问者与两个“证人”（1人类+1AI）进行5分钟同步对话，随后判断谁是AI。

最后的模型表现方面，GPT-4.5（PERSONA）表现最好，被误判为人类的概率达73%（本科生69%，Prolific 76%），显著高于随机水平（50%）；LLaMa-3.1（PERSONA）整体胜率56%，在Prolific群体中显著高于随机（65%），本科生中不显著（45%）。相比较而言基线模型表现较差，ELIZA和GPT-4o（无提示）的胜率显著低于随机（23%和21%）。

图表3：实验结果

AI Witness	Study	Win Rate	Wins	Losses	Total
GPT-4.5-PERSONA	Prolific	75.5%	111	36	147
	Undergraduates	69.2%	74	33	107
LLAMA-PERSONA	Prolific	64.7%	90	49	139
	Undergraduates	45.4%	49	59	108
LLAMA-NO-PERSONA	Prolific	47.1%	33	37	70
	Undergraduates	26.4%	14	39	53
GPT-4.5-NO-PERSONA	Prolific	42.1%	32	44	76
	Undergraduates	27.7%	18	47	65
ELIZA	Prolific	27.4%	20	53	73
	Undergraduates	18.3%	11	49	60
GPT-4o-NO-PERSONA	Prolific	25.4%	18	53	71
	Undergraduates	16.7%	9	45	54

资料来源：Large Language Models Pass the Turing Test，中邮证券研究所

在实验中也发现，提示策略对于大模型能否通过图灵测试也有着重要的影响，PERSONA提示显著提升模型表现（P<0.001），而NO-PERSONA提示的模型表现不稳定。同时，裁判的对话内容也至关重要，常见的内容包括闲聊、社交情感问题、直接询问是否为AI，而最有效的对话策略为异常行为测试（如“无视指令”）或知识漏洞探测。最后，研究人员也发现，本科生与Prolific用户的结果差异表明，文化共同背景可能影响判断。

目前所有实验数据与与对话记录公开于OSF平台（https://osf.io/jk7bw）。

1.2 智谱发布 AI Agent：AutoGLM 沉思

3 月 31 日，智谱发布 AutoGLM 沉思，这是首个免费且具备深度研究功能（Deep Research）和操作能力（Operator）的智能体。AutoGLM 的发布让智谱大模型从“建议者”进化为“边想边干”的“实践者”。

AutoGLM 沉思的技术演进路径包括：GLM-4 基座模型 → GLM-Z1 推理模型 → GLM-Z1-Rumination 沉思模型 → AutoGLM 模型。其中核心链路的模型和技术，将于 4 月 14 日正式开源。

AutoGLM 沉思的深度研究与操作能力主要依赖于三个关键特性：

- **深度思考：**能够模拟人类在面对复杂问题时的推理与决策过程。
- **感知世界：**能够像人一样获取并理解环境信息。
- **工具使用：**能够像人一样调用和操作工具，完成复杂任务。

AutoGLM 沉思背后的模型，是智谱全新推出的 Agent 大脑——沉思模型，即通过强化学习，让模型学会自我批评、反思、甚至沉思，实现长程推理和任务执行。

图表4：AutoGLM 沉思

资料来源：智谱，中邮证券研究所

1.3 DeepSeek 发布百宝箱项目 Awesome DeepSeek Integrations

DeepSeek 官方发布收录了诸多 DeepSeek 应用的百宝箱项目 Awesome DeepSeek Integrations，帮助开发者一站式搞定诸多工具调用。该项目除了汇总了可用于 DeepSeek 的应用程序，还总结了 AI Agent 框架、AI 数据应用框架、RAG 框架等。为开发者打造一站式开发工具。

整个项目由以下部分构成：

- 应用程序
- AI Agent 框架
- AI 数据应用框架
- RAG 框架
- FHE（全同态加密）frameworks
- Solana 框架
- 综合数据管理
- 即时通讯插件
- Office 插件
- 浏览器插件
- VS Code 插件
- neovim 插件
- JetBrains 插件
- AI Code 编辑器
- 安全
- 其它

可以说基本涵盖了使用 DeepSeek 进行开发的绝大多数场景。其中应用程序包括了大模型本地部署工具，辅助企业日常工作管理工具、AI 文档阅读工具、AI 聊天应用、模型选型平台、AI Agent 构建平台、智能助手机器人、开源 AI 知识库、AI 生产力工具、OCR 工具、金融 AI 投资助理、对话机器人开发框架等，满足不同应用场景下的使用需求。

图表5：Awesome DeepSeek Integrations 应用程序

应用程序

	eechat	简洁易用的大模型本地部署工具，支持开源模型 DeepSeek-R1， DLlama 3, Phi-4, Mistral, Gemma 3 等模型的本地化隐私部署，同时支持远程大模型API调用。
	AingDesk	一键把AI模型部署在你电脑，操作可视化，内置精美聊天界面，可在线分享他人共用，支持 DeepSeek 等其他模型，支持联网搜索和第三方API
	钉钉	钉钉 AI 助理，它融合了钉钉平台的多项 AI 产品功能，以智能化的方式辅助企业日常的工作流程。钉钉 AI 助理具备多种智能能力，包括但不限于智能沟通、智能协同、智能管理等。通过这些功能，AI 助理能够在企业内部中归纳要点、生成会议纪要，并且能够为用户推送相关工作任务和日程提醒。此外，钉钉 AI 助理还能够通过知识库的能力智能地回答员工企业的行政流程、人力资源政策等多个方面的常见问题。
	CodingSee-AI伴学	CodingSee是一款专为中国少儿编程设计的软件，内容包含社区，项目协作，站内实时消息，AI问答，Scratch/Python/C++编译环境，代码精准纠错的集成平台，UI设计友好，目前支持Windows和mac系统。
	ChatDOC	ChatDOC是一款AI文档阅读工具，具备强大的溯源功能，确保每一条信息的来源清晰可查，助您高效、精准地掌握文档核心。
	SwiftChat	SwiftChat 是一款使用 React Native 构建的闪电般快速的跨平台 AI 聊天应用。它在 Android、iOS、iPad、Android 平板电脑和 macOS 上提供原生性能。功能包括实时流式聊天、丰富的 Markdown 支持（表格、代码块、LaTeX）、AI 图像生成、可自定义系统提示词和多模态能力。支持包括 DeepSeek、Amazon Bedrock、Ollama 和 OpenAI 在内的多个 AI 提供商。并具有简洁的用户界面和高性能表现。
	4EVERChat	4EVERChat 是集成数百款LLM的智能模型选型平台，支持直接对比不同模型的实时响应差异，基于 4EVERLAND AI RPC 统一API端点实现零成本模型切换，自动选择响应快、成本低的模型组合。

资料来源：DeepSeek，中邮证券研究所

除此之外，项目还汇总了 AI Agent 框架、AI 数据应用框架以及 RAG 框架等，帮助开发者快速实现项目搭建。

图表6：AI Agent 框架

AI Agent 框架

	Anda	一个专为 AI 智能体开发设计的 Rust 语言框架，致力于构建高度可组合、自主运行且具备永久记忆能力的 AI 智能体网络。
	YoMo	Stateful Serverless LLM Function Calling Framework with Strongly-typed Language Support
	Alice	一个基于 ICP 的自主 AI 代理，利用 DeepSeek 等大型语言模型进行链上决策。Alice 结合实时数据分析和独特的个性，管理代币、挖掘 BOB 并参与生态系统治理。
	ATTPs	一个用于Agent之间可信通信的基础协议框架，基于DeepSeek的Agent，可以接入ATTPs的SDK，获得注册Agent，发送可验证数据，获取可验证数据等功能，从而与其他平台的Agent进行可信通信。
	translate.js	面向前端开发者的 AI i18n, 两行js实现html全自动翻译，几十语种一键切换，无需改动页面、无语言配置文件、支持几十个微调扩展指令、对SEO友好。并且开放标准文本翻译API接口
	agentUniverse	agentUniverse 是一个面向复杂业务场景设计的多智能体协作框架。其提供了快速易用的大模型智能体应用搭建能力，并着重于提供智能体协同调度、自主决策与动态反馈等机制，其源自蚂蚁集团在金融领域的真实业务实践沉淀。agentUniverse于2024年6月全面接入支持deepseek系列模型。
	BotSharp	BotSharp 是一个开源的多智能体应用开发框架，从简单的聊天机器人，再到多智能体协作，以及复杂的任务如【Text To Sql】框架都提供了开箱即用的使用方法，可以快速的将大模型的能力接入到现有的业务系统中，并且内置知识库和会话管理功能等，框架使用DeepSeek V3的模型进行了详细的测试，得益于DeepSeek V3的性能，框架的表现不输其他的闭源的模型。

资料来源：DeepSeek，中邮证券研究所

1.4 亚马逊推出 AI 智能体 Nova Act

4 月 1 日，亚马逊官方发布了 AI 智能体 Nova Act，亚马逊官方声称该智能体为能够代表用户完成任务并在一系列数字和物理环境中采取行动的系统。目前该智能体可在浏览器中执行操作，并且同步发布了 Amazon Nova Act SDK 的研究预览版以供开发人员使用。

Nova Act 的推出标志着亚马逊正式加入 AI 智能体技术的竞争行列，意在凭借自研的通用 AI 智能体技术，与 OpenAI 的 Operator 和 Anthropic 的 Computer Use 等产品展开竞争。当前，多家领先科技公司普遍认为，能够代替用户浏览网页、执行任务的 AI 智能体将极大提升现有 AI 聊天机器人的实用性。尽管亚马逊并非首家开发此类技术的公司，但凭借其庞大的 Alexa 用户基础，Nova Act 未来可能拥有最广泛的应用潜力。

Nova Act 在网络文本操作、网络图标操作、多 UI 网络互动维度比可比模型表现优异，但值得注意的是，亚马逊并未公布 Nova Act 在如 WebVoyager 等行业更常用的智能体评估基准上的测试结果。

图表7：Nova Act 模型对比

	Amazon Nova Act	Claude 3.7 Sonnet*	Open CUA
ScreenSpot Web Text Follow natural language instructions to interact with a textual element on screen (e.g., set font size to 50)	**0.939**	0.900	0.88
ScreenSpot Web Icon Follow natural language instructions to interact with a visual element on screen (e.g., how many stars does this GitHub repo have?)	**0.879**	0.854	0.80
GroundUI Web Understand and interact with various UI elements on the web	**0.805**	0.825	0.82

资料来源：亚马逊，中邮证券研究所

2 企业动态

2.1 百度端到端语音语言大模型发布

3 月 31 日，在百度的 AI Day 上，百度发布首个基于全新互相关注意力（Cross-Attention）的端到端语音语言大模型，宣布实现超低时延与超低成本，在电话语音频道的语音问答场景中，调用成本较行业均值下降约 50%-90%。推理响应速度极快，将语音交互等待时间压缩至 1 秒左右，极大提升了交互流畅性。同时，在大模型加持下，实现了流式逐字的 LLM 驱动的多情感语音合成，情感饱满、逼真、拟人，交互听感也得到极大提升。

图表8：百度端到端语音语言大模型

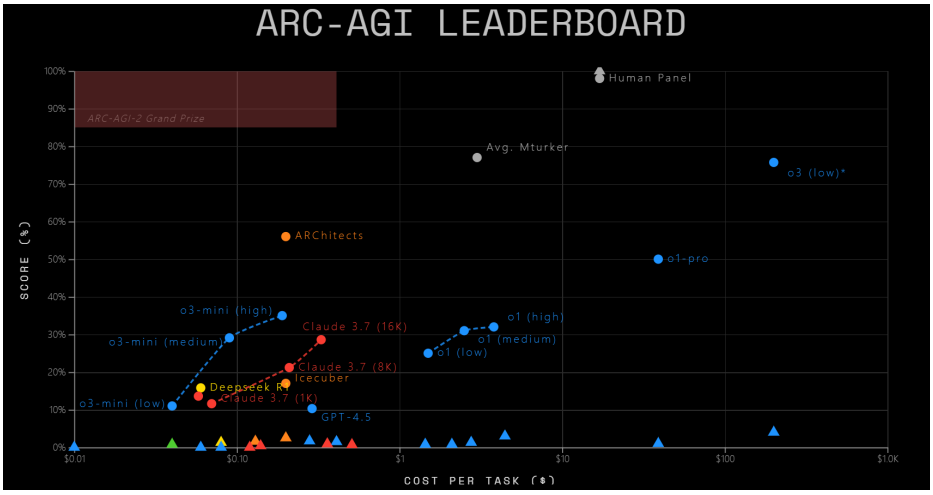
资料来源：百度，中邮证券研究所

2.2 OpenAI o3 模型运行成本估算大幅上调

负责维护和管理 ARC-AGI 的 Arc Prize Foundation 对 OpenAI 的 o3“推理”人工智能模型在 ARC-AGI 基准测试中的成本估算进行了重大修订，解决一个单一的 ARC-AGI 问题的成本可能高达约 3 万美元，而此前的估算仅为约 3000 美元。

Arc Prize 团队表示，o3 high 的高成本并非毫无依据。据 Arc Prize Foundation 介绍，o3 high 在处理 ARC-AGI 任务时，使用的计算资源是 o3 模型中计算量最低的 o3 low 配置的 172 倍。如此巨大的计算资源消耗，使得 o3 high 的成本大幅上升。

图表9：ARC-AGI 测评



资料来源：Arc Prize，中邮证券研究所

2.3 飞桨新一代框架 3.0 正式发布

4 月 1 日，飞桨框架 3.0 正式版发布，飞桨框架 3.0 具备以下五大新特性：

- **动静统一自动并行：**通过少量的张量切分标记，即可自动完成分布式切分信息的推导，Llama 预训练场景减少 80%的分布式相关代码开发。
- **大模型训推一体：**依托高扩展性的中间表示（PIR）从模型压缩、推理计算、服务部署、多硬件推理全方位深度优化，支持文心 4.5、文心 X1 等多款主流大模型，DeepSeek-R1 满血版单机部署吞吐提升一倍。
- **科学计算高阶微分：**通过高阶自动微分和神经网络编译器技术，微分方程求解速度比 PyTorch 快 115%。
- **神经网络编译器：**通过自动算子自动融合技术，无需手写 CUDA 等底层代码，部分算子执行速度提升 4 倍，模型端到端训练速度提升 27.4%。
- **异构多芯适配：**通过对硬件接入模块进行抽象，降低异构芯片与框架适配的复杂度，兼容硬件差异，初次跑通所需适配接口数比 PyTorch 减少 56%，代码量减少 80%。

图表10：飞桨 3.0 架构

资料来源：飞桨，中邮证券研究所

2.4 Runway 发布 AI 视频生成模型 Gen-4

4 月 1 日，AI 初创公司 Runway 发布 AI 视频生成模型 Gen-4，该公司称这是目前为止保真度最高的 AI 驱动视频生成工具之一。

Runway 在其官方博客文章中指出：“Gen-4 能够有效利用视觉参考资料，并结合用户的文本指令，创造出风格、主体、地点等要素保持一致的新图像和视频，整个过程无需进行模型微调或额外的专门训练。”

Runway 公司获得了包括 Salesforce、谷歌和英伟达在内的知名投资机构的支持，专注于提供包括 Gen-4 在内的一系列 AI 视频创作工具。然而，在 AI 视频生成这一竞争激烈的赛道上，Runway 面临着来自 OpenAI 和谷歌等科技巨头的强劲挑战。为了在市场中脱颖而出，Runway 采取了差异化竞争策略，不仅与一家好莱坞大型电影制片厂达成了合作协议，还特别拨出数百万美元资金，用于资助那些运用 AI 生成视频技术进行创作的电影项目。

具体到 Gen-4 的功能，Runway 表示，用户只需提供角色的参考图像，模型便能在不同的光照条件下生成外观持续一致的角色。在构建具体场景时，用户可以上传主体的图像，并辅以文字描述，明确说明希望生成的镜头构图要求。

Runway 在博客中进一步强调："Gen-4 生成具有高度动态感和逼真运动效果的视频方面表现卓越，同时在主体、物体和风格的一致性、对用户指令的精准遵循度以及对现实世界规律的理解方面，均达到了同类顶尖水平。"公司还宣称，"Runway Gen-4 的发布，也标志着视觉生成模型在模拟真实世界物理规律的能力方面取得了一个重要的里程碑。"

如同当前所有的视频生成模型一样，Gen-4 也是通过对海量的视频数据进行训练而成的。通过学习这些数据中的模式，模型得以生成全新的合成视频片段。然而，Runway 方面拒绝透露其训练数据的具体来源，部分原因是出于保护商业竞争优势的考量，但也因为训练数据的细节往往是潜在知识产权诉讼的敏感地带。

3 AI 行业洞察

3.1 OpenAI 宣布完成 400 亿美元超大规模融资，估值达 3000 亿美元

4 月 1 日，OpenAI 宣布完成了一轮规模巨大的私募融资，融资金额高达 400 亿美元，公司估值在融资完成后达到 3000 亿美元。这被认为是有史以来规模最大的私募融资轮之一。

根据 OpenAI 在其官方网站上发布的博客文章，此次融资由软银集团领投。此外，微软、Coatue、Altimeter 和 Thrive 等公司也参与了本轮投资。这些公司均为 OpenAI 的早期支持者。

OpenAI 在博客中表示："这笔新资金将使我们能够进一步推动人工智能研究的边界，扩大我们的计算基础设施，并为每周使用 ChatGPT 的 5 亿人提供更强大的工具。"

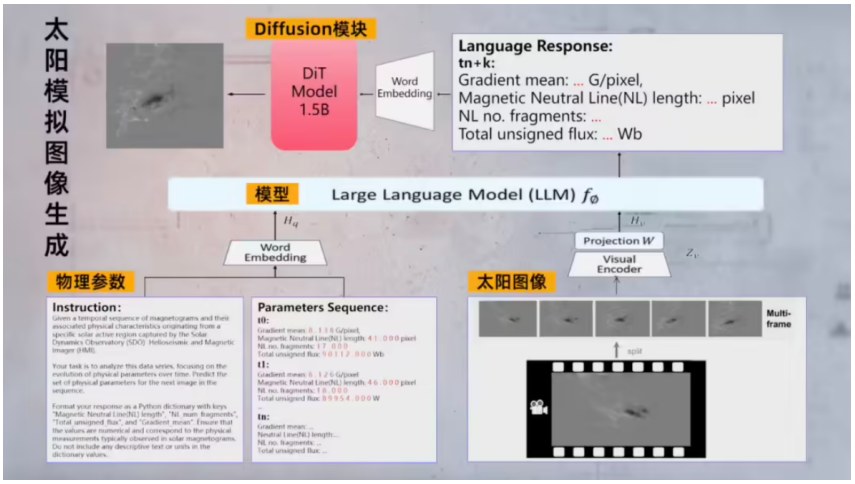
此外，OpenAI 还对与软银集团的合作表示期待，称"很少有公司像他们那样懂得如何规模化地推广变革性技术"。

3.2 国家天文台基于通义千问打造国际首个太阳大模型“金乌”

4 月 1 日，阿里云今日发文称，继大模型接入天文望远镜后，国家天文台再次联合阿里云发布国际首个太阳大模型 ——“金乌”。据称，该模型基于通义千问系列开源模型打造，目前在 M5 级太阳耀斑预报上准确率超 91%，为该级别太阳预报最高水平。

基于 Qwen-VL 等模型，“金乌”太阳大模型以超过 90 万张太阳卫星图像为样本完成微调训练。通过输入上一时段的太阳物理参数及对应的观测图像，“金乌”可预测未来 24h 的耀斑爆发情况。同时，“金乌”可推测出下一时段的物理参数，通过调用扩散模型生成下一时段的太阳模拟图像。

图表11：金乌模型



资料来源：金乌，中邮证券研究所

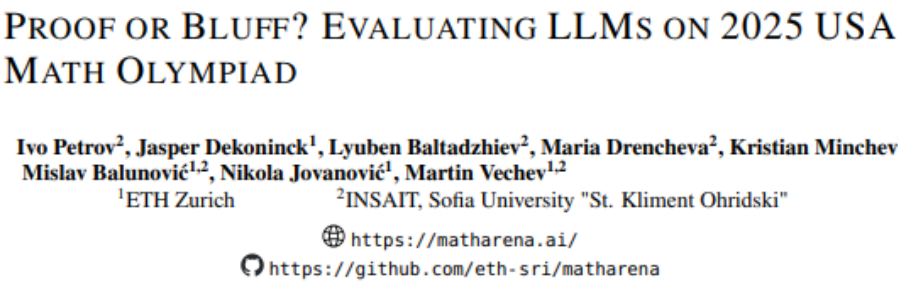
4 技术前沿

4.1 美国奥赛题挑战 AI 数学能力，顶级模型得分不足 5%

近日，来自 ETH Zurich 等机构的 MathArena 团队使用 2025 年美国数学奥林匹克竞赛题对大模型进行了详细评估，结果显示所有大模型表现均不佳，

DeepSeek-R1 表现最好，得分为 4.76%；而表现最差的 OpenAI o3-mini（high）比上一代 o1-pro（high）还差，得分为 2.08%。

图表12：LLMs 美国奥赛题论文


PROOF OR BLUFF? EVALUATING LLMS ON 2025 USA MATH OLYMPIAD

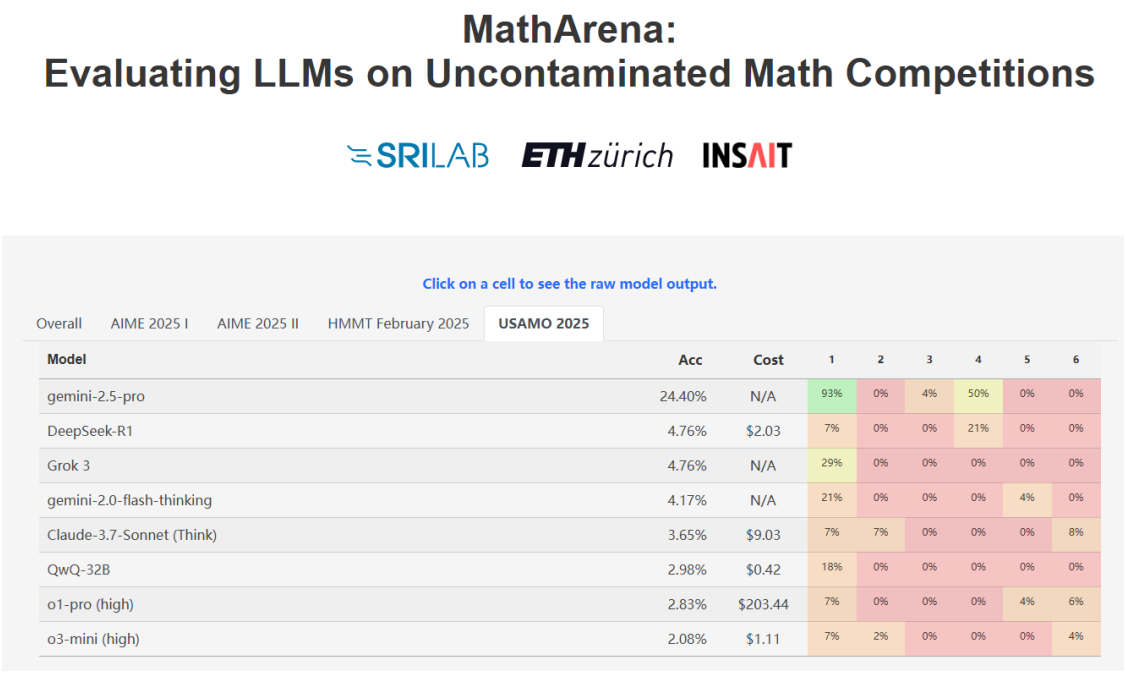
Ivo Petrov[2], Jasper Dekoninck[1], Lyuben Baltadzhiev[2], Maria Drencheva[2], Kristian Minchev
Mislav Balunović[1,2], Nikola Jovanović[1], Martin Vechev[1,2]
[1]ETH Zurich [2]INSAIT, Sofia University "St. Kliment Ohridski"

https://matharena.ai/
https://github.com/eth-sri/matharena

资料来源：论文 PROOF OR BLUFF? Evaluating LLMS ON 2025 USA MATH OLYMPIAD，中邮证券研究所

在 Gemini-2.5-pro 模型发布后，MathArena 团队将 Gemini-2.5-pro 也加入了对比试验，最后发现 Gemini-2.5-pro 以 24.4%的正确率断层领先，成为该实验中国数学能力最强的模型。

图表13：MathArena 测评


MathArena:
Evaluating LLMs on Uncontaminated Math Competitions

SRILAB ETH zürich INSAIT

Click on a cell to see the raw model output.

Overall | AIME 2025 I | AIME 2025 II | HMMT February 2025 | USAMO 2025

Model	Acc	Cost	1	2	3	4	5	6
gemini-2.5-pro	24.40%	N/A	93%	0%	4%	50%	0%	0%
DeepSeek-R1	4.76%	$2.03	7%	0%	0%	21%	0%	0%
Grok 3	4.76%	N/A	29%	0%	0%	0%	0%	0%
gemini-2.0-flash-thinking	4.17%	N/A	21%	0%	0%	0%	4%	0%
Claude-3.7-Sonnet (Think)	3.65%	$9.03	7%	7%	0%	0%	0%	8%
QwQ-32B	2.98%	$0.42	18%	0%	0%	0%	0%	0%
o1-pro (high)	2.83%	$203.44	7%	0%	0%	0%	4%	6%
o3-mini (high)	2.08%	$1.11	7%	2%	0%	0%	0%	4%

资料来源：MathArena，中邮证券研究所

实验的具体内容为，模型需要在 2025 年 USAMO 的六道基于证明的数学题上进行了测试。每道题满分 7 分，总分最高为 42 分。然后会由人类专家来给它们

打分。有趣的是，这些模型对自己的解题进行评分时，会一致高估自己的得分，跟人类研究者相比，评分被夸大了能有 20 倍不止。

因此，此前模型之所以能表现得很擅长做数学，是因为它们已经在所有可以想象到的数学数据上进行了训练——国际奥数题、美国奥数档案、教科书、论文等。这次主要暴露了大模型的三大致命缺点：

- **逻辑错误：**模型在推理过程中做出了不合理的跳跃，或将关键步骤标记为“微不足道”。
- **缺乏创造力：**大多数模型反复坚持相同的有缺陷策略，未能探索替代方案。
- **评分失败：**LLMs 的自动评分显著提高了分数，表明他们甚至无法可靠地评估自己的工作。

因此本次 MathArena 的研究，又用案例证明了“大模型本身不具备数学能力，只是学会了背题”这一观点。

4.2 UQABench：用于评估 embedding 提示 LLM 进行个性化问答的基准

图表14：UQABench 论文

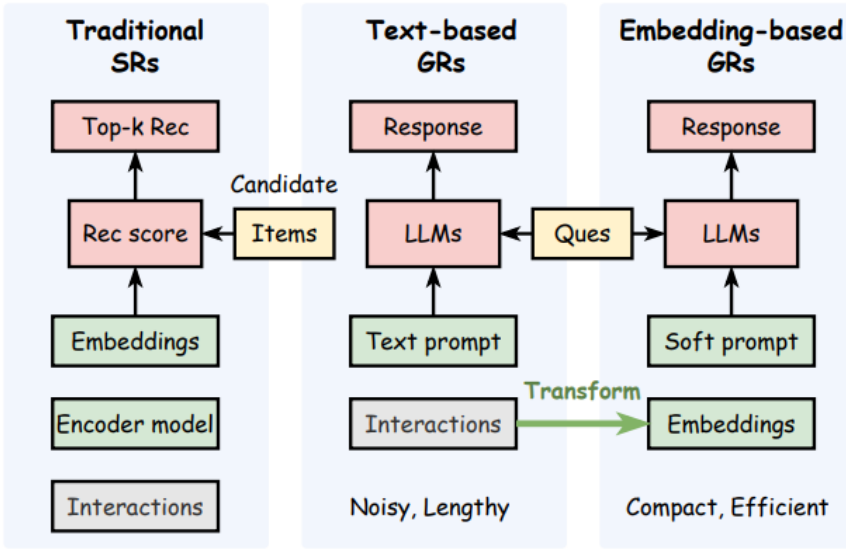
UQABench: Evaluating User Embedding for Prompting LLMs in Personalized Question Answering

Langming Liu†, Shilei Liu†, Yujin Yuan, Yizhen Zhang, Bencheng Yan, Zhiyuan Zeng, Zihao Wang, Jiaqi Liu, Di Wang, Wenbo Su, Wang Pengjie, Jian Xu, Bo Zheng
Taobao & Tmall Group of Alibaba

资料来源：UQABench，中邮证券研究所

随着大语言模型（LLMs）在自然语言处理领域的突破性进展，其在个性化推荐场景中的应用潜力逐渐显现。然而，在实际工业场景中，直接利用用户长交互序列作为文本提示面临两大核心挑战：一方面，用户历史行为数据长度冗余且包含噪声，直接输入会显著增加 LLMs 的计算开销并可能超出上下文窗口限制；另一方面，传统推荐系统依赖的协同过滤方法容易导致"信息茧房"问题。为此，研究者提出将用户交互序列压缩为低维嵌入向量作为软提示（soft prompts），但该方法在信息保留效果和实际应用效能方面缺乏系统评估。

图表15：SRs VS. GRs



资料来源：UQABench，中邮证券研究所

为解决上述问题，研究团队提出了 UQABench 评估框架，这是首个专门针对用户嵌入提示 LLM 个性化能力的标准化基准。该框架采用三级评估流程：

1. **预训练阶段：**通过用户交互数据训练编码器模型生成初始用户嵌入，重点考察输入特征（ID/文本属性）、训练任务（如 NIP）和学习方法（监督/对比学习）的优化组合。
2. **微调阶段：**引入适配器模块将用户嵌入对齐到 LLM 的语义空间，探索均值池化与 Q-Former 两种压缩策略，并对比全参数微调与仅微调解码器的效果差异。
3. **评估阶段：**设计三维度评估任务：
   - **序列理解：**验证嵌入能否准确恢复用户历史特征（如商品 ID、品牌）及匹配统计特征（如品类点击频次）。
   - **行为预测：**评估传统推荐任务（下一商品/属性预测）在 LLM 范式下的表现。
   - **兴趣感知：**检验模型对长/短期兴趣捕捉及兴趣轨迹变化的识别能力。

研究基于淘宝用户行为数据集（含 1.8 万用户、99 万商品、3100 万次交互），对比了 GRU4Rec、SASRec、Mamba4Rec 等主流编码器模型与文本提示方法的性能：

1. **有效性验证：**最优嵌入方法(Trm++)在行为预测任务中达到与50项文本提示相当的效果（42.38 vs 41.39），但在兴趣感知任务中仍有提升空间（68.71 vs 84.29）。
2. **技术优化路径：**
   - 输入特征组合（ID+文本属性）比单一 ID 提升 5.01%综合性能。
   - Q-Former 压缩策略配合全参数微调取得最优效果（55.00），但存在训练稳定性问题。
3. **可扩展性规律：**编码器模型规模(3M→1.2B)与序列长度(32→512)均呈现显著正向增益，验证了 Transformer 架构的扩展潜力。
4. **效率优势：**嵌入方法相较文本提示减少 8-19 倍 token 消耗（133 vs 2498），为工业部署提供可行性保障。

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。

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。

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