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

《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

《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

金工周报

Gemini 2.5 Pro 发布即屠榜，DeepSeek V3 完成模型更新——AI 动态汇总 20250331

- **谷歌发布 Gemini 2.5 Pro，发布即屠榜**

谷歌于 2025 年 3 月 25 日发布了 Gemini 2.5 Pro 模型，据介绍，Gemini 2.5 是思维模型（Thinking Models），能够在响应之前通过思考进行推理，从而提高性能和准确性。截至 2025 年 3 月 25 日，Gemini 2.5 Pro 已经登顶了 Lmarena 排行榜的第一位，而且创下了历史最大分数飞跃，截止 2025 年 3 月 30 日，Gemini 2.5 Pro 比第二名的 ChatGpt-4o-latest（2025-03-26）高出 35 分，更是比上月发布即屠榜的 Grok-3 分数高了接近 40 分。

- **DeepSeek V3 完成模型更新，各项能力全面进阶**

DeepSeek V3 模型已完成小版本升级，目前版本号 DeepSeek-V3-0324，本次模型更新提升主要围绕推理任务表现、前端开发能力、中文写作能力、中文搜索能力、工具使用能力展开。

- **ChatGPT-4o 更新，原生图像生成能力大幅提升**

3 月 25 日，山姆奥特曼亲自带队直播发布 ChatGPT-4o 更新，并现场利用 ChatGPT-4o 制作梗图，生成的吉博力风格图片引起网络上的模仿热潮。

- **昆仑万维发布全球首款音乐推理模型 Mureka O1**

继发布 Skywork 后，昆仑万维又发布一重量级大模型 Mureka O1，专门应用于音乐领域，发布后一举将同类模型 suno 拉下第一名的宝座。该模型基于 Mureka V6 基座，结合 CoT 技术，只需要一段提示词即可生成想要的音乐。Mureka O1 是全球首个将 CoT 用到音乐生成领域的模型，在众多音乐生成基准对比中表现超过同类大模型。

- **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

图表目录

1 AI 重点要闻

1.1 谷歌发布 Gemini 2.5 Pro，发布即屠榜

谷歌于 2025 年 3 月 25 日发布了 Gemini 2.5 Pro 模型，据介绍，Gemini 2.5 是思维模型（Thinking Models），能够在响应之前通过思考进行推理，从而提高性能和准确性。截至 2025 年 3 月 25 日，Gemini 2.5 Pro 已经登顶了 Lmarena 排行榜的第一位，而且创下了历史最大分数飞跃，截止 2025 年 3 月 30 日，Gemini 2.5 Pro 比第二名的 ChatGpt-4o-latest(2025-03-26)高出 35 分，更是比上月发布即屠榜的 Grok-3 分数高了接近 40 分。

图表1：Gemini 2.5 Pro 发布即屠榜

Rank* (UB)	Rank (StyleCtrl)	Model	Arena Score	95% CI	Votes	Organization	License
1	1	Gemini-2.5-Pro-Exp-03-25	1443	+11/-8	3474	Google	Proprietary
2	2	ChatGPT-4o-latest (2025-03-26)	1408	+11/-12	2676	OpenAI	Proprietary
2	4	Grok-3-Preview-02-24	1404	+6/-6	10397	xAI	Proprietary
2	2	GPT-4.5-Preview	1398	+6/-7	10907	OpenAI	Proprietary
5	7	Gemini-2.0-Flash-Thinking-Exp-01-21	1381	+4/-5	22987	Google	Proprietary
5	4	Gemini-2.0-Pro-Exp-02-05	1380	+5/-4	20289	Google	Proprietary
7	5	DeepSeek-R1	1360	+5/-4	13074	DeepSeek	MIT
7	12	Gemini-2.0-Flash-001	1355	+6/-4	18650	Google	Proprietary
7	4	o1-2024-12-17	1351	+5/-4	25363	OpenAI	Proprietary
10	12	Qwen2.5-Max	1340	+5/-5	17452	Alibaba	Proprietary
10	12	Gemma-3-27B-it	1339	+7/-5	7238	Google	Gemma
10	9	o1-preview	1335	+4/-3	33188	OpenAI	Proprietary

资料来源：Lmarena，中邮证券研究所

本次 Gemini 2.5 Pro 模型主要有以下亮点：

- **推理和代码能力大幅提升**

  在常见的推理，科学、数学、代码生成、视觉推理、图片识别长上下文以及多语言表现领域，Gemini 2.5 Pro 均有不俗的表现。

图表2：Gemini 2.5 Pro跑分

Benchmark		Gemini 2.5 Pro Experimental (03-25)	OpenAI o3-mini High	OpenAI GPT-4.5	Claude 3.7 Sonnet 64k Extended Thinking	Grok 3 Beta Extended Thinking	DeepSeek R1
Reasoning & knowledge Humanity's Last Exam (no tools)		**18.8%**	14.0%*	6.4%	8.9%	—	8.6%*
Science GPQA diamond	single attempt (pass@1)	**84.0%**	79.7%	71.4%	78.2%	80.2%	71.5%
	multiple attempts	—	—	—	**84.8%**	84.6%	—
Mathematics AIME 2025	single attempt (pass@1)	**86.7%**	86.5%	—	49.5%	77.3%	70.0%
	multiple attempts	—	—	—	—	**93.3%**	—
Mathematics AIME 2024	single attempt (pass@1)	**92.0%**	87.3%	36.7%	61.3%	83.9%	79.8%
	multiple attempts	—	—	—	80.0%	**93.3%**	—
Code generation LiveCodeBench v5	single attempt (pass@1)	70.4%	**74.1%**	—	—	70.6%	64.3%
	multiple attempts	—	—	—	—	**79.4%**	—
Code editing Aider Polyglot		**74.0% / 68.6%** whole / diff	60.4% diff	44.9% diff	64.9% diff	—	56.9% diff
Agentic coding SWE-bench verified		63.8%	49.3%	38.0%	**70.3%**	—	49.2%
Factuality SimpleQA		52.9%	13.8%	**62.5%**	—	43.6%	30.1%
Visual reasoning MMMU	single attempt (pass@1)	**81.7%**	no MM support	74.4%	75.0%	76.0%	no MM support
	multiple attempts	—	no MM support	—	—	78.0%	no MM support
Image understanding Vibe-Eval (Reka)		69.4%	no MM support	—	—	—	no MM support
Long context MRCR	128k (average)	**94.5%**	61.4%	64.0%	—	—	—
	1M (pointwise)	83.1%	—	—			
Multilingual performance Global MMLU (Lite)		89.8%	—	—	—	—	—

Methodology

Gemini results: All Gemini 2.5 Pro scores are pass @1 (no majority voting or parallel test time compute unless indicated otherwise). They are all run with the AI Studio API for the model-id gemini-2.5-pro-exp-03-25 with default sampling settings. To reduce variance, we average over multiple trials for smaller benchmarks. Vibe-Eval results are reported using Gemini as a judge.

Non-Gemini results: All the results for non-Gemini models are sourced from providers' self reported numbers. All SWE-bench Verified numbers follow official provider reports, using different scaffolding and infrastructure. Google's scaffolding includes drawing multiple trajectories and re-scoring them using model's own judgement.

Thinking vs not-thinking: For Claude 3.7 Sonnet: GPQA, AIME 2024, MMMU come with 64k extended thinking, Aider with 32k, and HLE with 16k. Remaining results come from the non thinking model due to result availability. For Grok-3 all results come with extended reasoning except for SimpleQA (based on xAI reports).

Single attempt vs multiple attempts: When two numbers are reported for the same eval higher number uses majority voting with n=64 for Grok models and internal scoring with parallel test time compute for Anthropic models.

Result sources: Where provider numbers are not available we report numbers from leaderboards reporting results on these benchmarks: Humanity's Last Exam results are sourced from https://agi.safe.ai/ and https://scale.com/leaderboard/humanitys_last_exam, AIME 2025 numbers are sourced from https://matharena.ai/. LiveCodeBench results are from https://livecodebench.github.io/leaderboard.html (10/1/2024 - 2/1/2025 in the UI). Aider Polyglot numbers come from https://aider.chat/docs/leaderboards/. For MRCR we include 128k results as a cumulative score to ensure they can be comparable with previous results and a pointwise value for 1M context window to show the capability of the model at full length.

* indicates evaluated on text problems only (without images)

资料来源：Google，中邮证券研究所

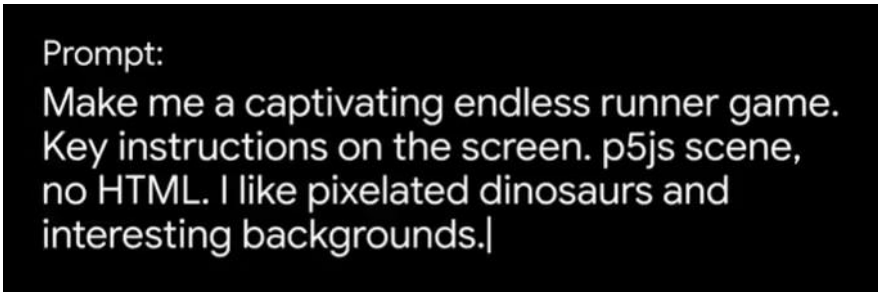
除此之外，在各类需要高级推理能力的基准测试中，它都达到了SOTA水平。

无需使用测试阶段会增加计算成本的技术（如多数投票法），2.5 Pro就能在GPQA和AIME 2025等数学和科学基准评测中表现卓越。而且，在不使用任何外部工具的条件下，它就在挑战人类知识和推理能力的极限前沿“人类最后的考试”中取得了18.8%的准确率，达到业界领先。

- **编程能力大幅提升**

相比较于Gemini 2.0，Gemini 2.5 Pro在编程方面有了长足的飞跃，2.5 Pro擅长创建视觉上引人注目的Web应用程序和代理代码应用程序，以及代码转换和编辑。在代理代码评估的行业标准SWE-Bench Verified上，Gemini 2.5 Pro使用自定义代理设置得分为63.8%。谷歌团队通过demo展示了Gemini 2.5 Pro如何运用强大推理，仅通过一行提示词，就能生成可执行代码，来创建完整的动画和游戏。

图表3：提示词demo



资料来源：Google，中邮证券研究所

- **原生多模态和超长上下文**

Gemini 2.5继承并发扬了Gemini模型的优势——原生多模态能力和超长上下文长度。自己发布之初，2.5 Pro就支持100万token的上下文窗口，性能显著超越了前代模型。这能让它理解海量数据集，并处理来自多种信息源的复杂问题，包括文本、音频、图像、视频，甚至完整的代码仓库。

图表4：Gemini 2.5 Pro模型指标

Model deployment status	Experimental
Supported data types for input	Text, Image, Video, Audio
Supported data types for output	Text
Supported # tokens for input	1M
Supported # tokens for output	64k
Knowledge cutoff	January 2025
Tool use	Function calling Structured output Search as a tool Code execution
Best for	Reasoning Coding Complex prompts
Availability	Google AI Studio Gemini API Gemini App

资料来源：Google，中邮证券研究所

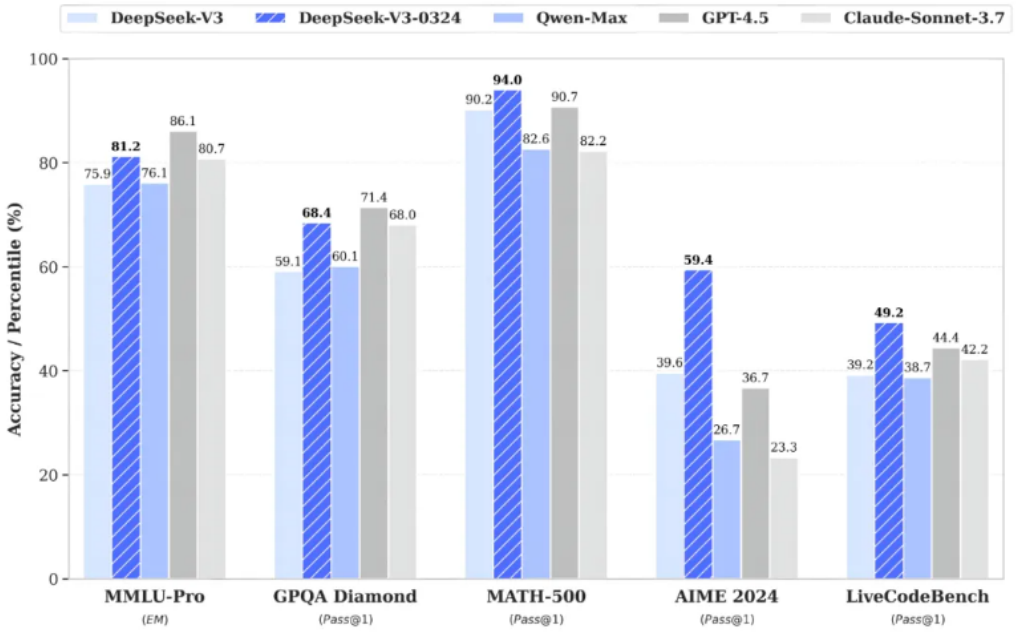
1.2 DeepSeek V3完成模型更新，各项能力全面进阶

DeepSeek V3模型已完成小版本升级，目前版本号DeepSeek-V3-0324，本次模型更新提升主要围绕以下方面：

- **推理任务表现提高**

  新版V3模型借鉴DeepSeek-R1模型训练过程中所使用的强化学习技术，大幅提高了在推理类任务上的表现水平，在数学、代码类相关评测集上取得了超过GPT-4.5的得分成绩。

图表5：DeepSeek-V3-0324 评测对比

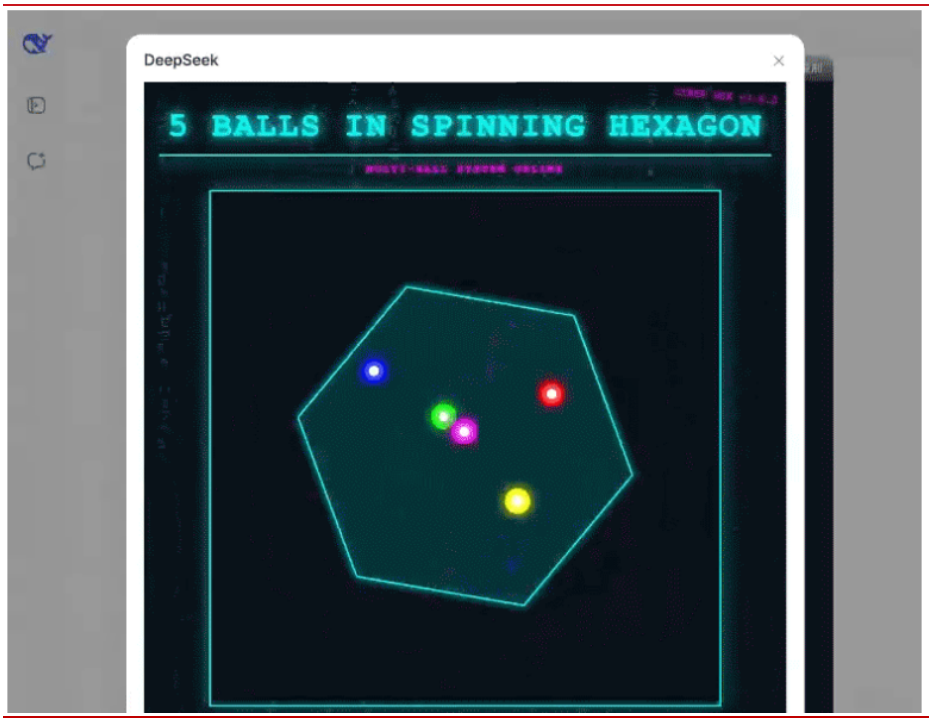


资料来源：DeepSeek，中邮证券研究所

- **前端开发能力增强**

  在 HTML 等代码前端任务上，新版 V3 模型生成的代码可用性更高，视觉效果也更加美观、富有设计感。

图表6：DeepSeek-V3-0324 前端开发案例



资料来源：DeepSeek，中邮证券研究所

- **中文写作升级**

  在中文写作任务方面，新版 V3 模型基于 R1 的写作水平进行了进一步优化，同时特别提升了中长篇文本创作的内容质量。

- **中文搜索能力优化**

  新版 V3 模型可以在联网搜索场景下，对于报告生成类指令输出内容更为详实准确、排版更加清晰美观的结果。

- **函数调用能力增强**

  提高函数调用的准确性，修复之前 V3 版本中的问题。

1.3 ChatGPT-4o 更新，原生图像生成能力大幅提升

3 月 25 日，山姆奥特曼亲自带队直播发布 ChatGPT-4o 更新，并现场利用 ChatGPT-4o 制作梗图，生成的吉博力风格图片引起网络上的模仿热潮。

图表7：ChatGPT-4o 原生图展示

资料来源：OpenAI，中邮证券研究所

GPT-4 的图像生成功能擅长准确渲染文本，精确遵循提示词，并利用 GPT-4o 固有的知识库和聊天上下文——包括转换上传的图像或将其用作视觉灵感。这些能力让用户可以更加容易地创建想象中的画面，帮助通过视觉更有效地沟通，并将图像生成发展成为一种具有精确性和强大功能的实用工具。毕竟，只有当图像配上指代共享语言和经验的符号时，才能传达精确的含义。

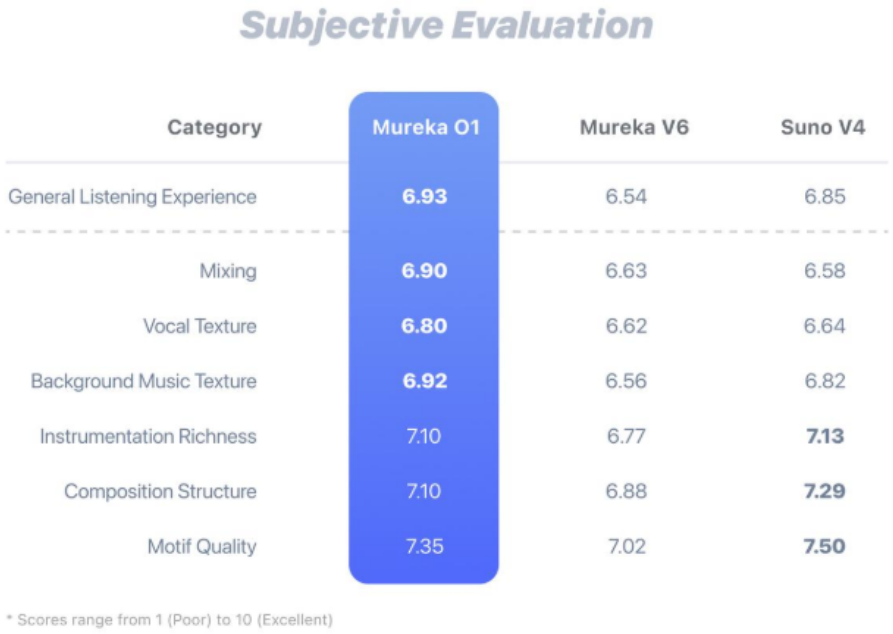
与之前相比，这次针对原生图的改进主要有以下特点：**1)能力大幅增强：**通过线上图像和文本的联合分布训练，模型不仅能学会图像如何与语言相关联，还能知道它们之间的相互关系。结合积极的后训练优化，最终的模型展现出惊人的视觉表现力，能够生成实用、一致且具有上下文感知能力的图像。**2)文本渲染优化：**GPT-4o 将精确的符号与图像融合的能力使图像生成成为视觉交流的有力工具。**3)多轮交互生成：**GPT-4o 能够基于聊天上下文中的图像和文本进行构建，确保始终保持一致性。**4)指令遵循优化：**GPT-4o 的图像生成功能不仅可以遵循详细的提示词，而且还十分注重细节。相比起其他只能处理 5-8 个物体的模型，GPT-4o 可以轻松搞定多达 10-20 个不同的物体。**5)根据上下文进行学习：**GPT-4o 能够分析并学习用户上传的图像，将图像细节无缝整合到上下文中，用于辅助图像生成。**6)图片风格多样：**通过对反映多种多样图像风格的图像进行训练，模型能够以令人信服的方式创建或转换图像。

1.4 昆仑万维发布全球首款音乐推理模型 Mureka O1

继发布 Skywork 后，昆仑万维又发布一重量级大模型 Mureka O1，专门应用于音乐领域，发布后一举将同类模型 suno 拉下第一名的宝座。

该模型基于 Mureka V6 基座，结合 CoT 技术，只需要一段提示词即可生成想要的音乐。Mureka O1 是全球首个将 CoT 用到音乐生成领域的模型，在众多音乐生成基准对比中表现超过同类大模型。

图表8：Mureka 模型对比



资料来源：Mureke，中邮证券研究所

2 企业动态

2.1 GPT-4o 再次升级，新版本已面向全部付费用户开放

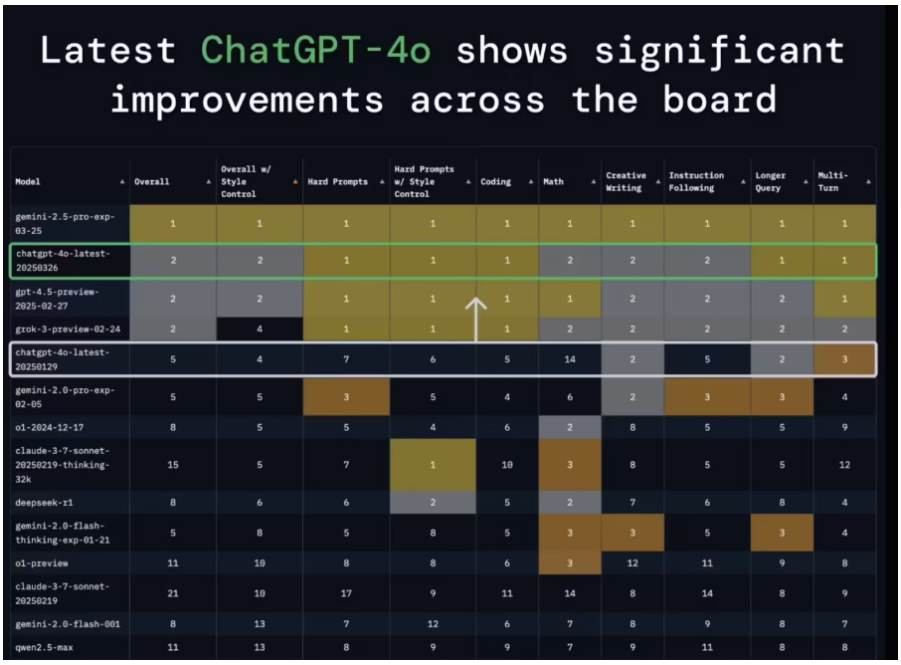
3 月 28 日，OpenAI 宣布 GPT-4o 带来了一些功能上的更新，并确认升级版 GPT-4o 已面向所有付费用户开放，而免费用户还要再等几周。

GPT-4o 本次更新聚焦四大核心部分：

- **多指令解析优化：**显著提升对包含多重需求的复杂指令理解能力；
- **技术问题处理增强：**强化复杂技术及编程问题的解析与解决方案生成；
- **逻辑推理与创造力提升：**增强创新性思维与跨领域知识融合能力；
- **交互界面精简：**减少表情符号使用频率，优化专业场景对话体验。

除此之外，在 AI 基准测试平台 Lmarena 上，最新的 ChatGPT GPT-4o（2025-03-26）模型已经提升到了第二名的位置，甚至超过了其上个月推出的 GPT-4.5。与 2025-01-29 那次测试成绩相比，此次更新后成绩足足提高了 30 分。

图表9：GPT-4o 更新后排名上升



Model	Overall	Overall w/ Style Control	Hard Prompts	Hard Prompts w/ Style Control	Coding	Math	Creative Writing	Instruction Following	Longer Query	Multi-Turn
gemini-2.5-pro-exp-03-25	1	1	1	1	1	1	1	1	1	1
chatgpt-4o-latest-20250326	2	2	1	1	1	2	2	2	1	1
gpt-4.5-preview-2025-02-27	2	2	1	1	1	1	2	2	2	1
grok-3-preview-02-24	2	4	1	1	1	2	2	2	2	2
chatgpt-4o-latest-20250129	5	4	7	6	5	14	2	5	2	3
gemini-2.0-pro-exp-02-05	5	5	3	5	4	6	2	3	3	4
o1-2024-12-17	8	5	5	4	6	2	8	5	5	9
claude-3-7-sonnet-20250219-thinking-32k	15	5	7	1	10	3	8	5	5	12
deepseek-r1	8	6	6	2	5	2	7	6	8	4
gemini-2.0-flash-thinking-exp-01-21	5	8	5	8	5	3	3	5	3	4
o1-preview	11	10	8	8	6	3	12	11	9	8
claude-3-7-sonnet-20250219	21	10	17	9	11	14	8	14	8	9
gemini-2.0-flash-001	8	13	7	12	6	7	8	9	8	7
qwen2.5-max	11	13	8	9	9	7	9	11	8	8

资料来源：Lmarena，中邮证券研究所

2.2 蚂蚁集团采用国产芯片训练 AI：性能匹配 H800，成本显著降低

据悉，蚂蚁集团使用中国制造的芯片开发 AI 模型训练技术，这将使成本降低 20%。蚂蚁集团使用了包括阿里巴巴集团控股有限公司和华为技术有限公司在内的国内芯片，采用混合专家模型（MoE）机器学习方法，训练结果与英伟达公司 H800 芯片匹敌。

2.3 阿里通义千问发布新一代端到端多模态旗舰模型 Qwen2.5-Omni，现已开源

3 月 27 日，阿里云发布通义千问 Qwen 模型家族中新一代端到端多模态旗舰模型 Qwen2.5-Omni，现已在 Github 等平台开源。

该模型专为全方位多模态感知设计，能够无缝处理文本、图像、音频和视频等多种输入形式，并通过实时流式响应同时生成文本与自然语音合成输出。主要亮点如下：

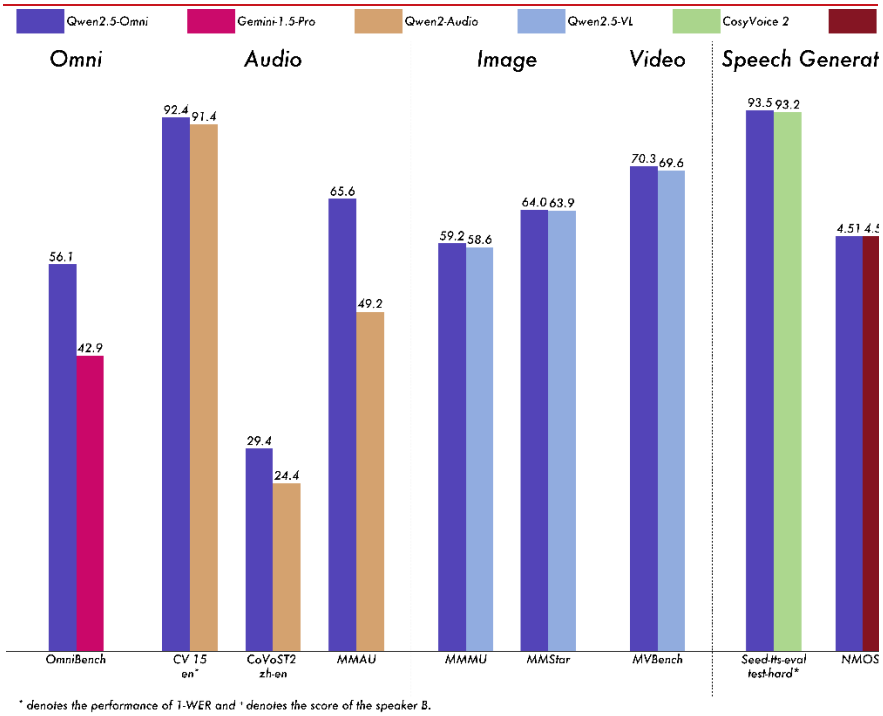
- **全能创新架构：**Qwen 团队提出了一种全新的 Thinker-Talker 架构，这是一种端到端的多模态模型，旨在支持文本、图像、音频、视频的

跨模态理解，同时以流式方式生成文本和自然语音响应。Qwen 提出了一种新的位置编码技术，称为 TMRoPE（Time-aligned Multimodal RoPE），通过时间轴对齐实现视频与音频输入的精准同步。

- **实时音视频交互：**架构旨在支持完全实时交互，支持分块输入和即时输出。
- **自然流畅的语音生成：**在语音生成的自然性和稳定性方面超越了许多现有的流式和非流式替代方案。
- **全模态性能优势：**在同等规模的单模态模型进行基准测试时，表现出卓越的性能。Qwen2.5-Omni 在音频能力上优于类似大小的 Qwen2-Audio，并与 Qwen2.5-VL-7B 保持同等水平。
- **卓越的端到端语音指令跟随能力：**Qwen2.5-Omni 在端到端语音指令跟随方面表现出与文本输入处理相媲美的效果，在 MMLU 通用知识理解和 GSM8K 数学推理等基准测试中表现优异。

在多种测评中，Qwen2.5-Omni 达到了 SOTA 的表现。

图表10：Qwen2.5-Omni 评测



资料来源：通义千问，中邮证券研究所

2.4 百度发布国内首个对话式应用开发平台秒哒

3月24日，百度宣布国内首个“对话式”应用开发平台秒哒正式全量上线，用户仅需通过自然语言描述需求，即可自动生成完整功能代码。据官方介绍，秒哒采用“无代码编程+多智能体协作+多工具调用”的技术组合，用户仅需通过自然语言描述需求，即可自动生成完整功能代码，实现“3分钟生成+1小时迭代”的极致开发体验；“智能体协作矩阵”内置十余个垂直领域智能体，用户可根据任务需求动态调整策略和行为，灵活组建不同技能的虚拟开发团队；此外，平台还集成了多种第三方工具和服务，能够实现与各种数据源和工具的无缝对接，构建从需求到部署的全链路支持。

图表11：秒哒介绍

资料来源：秒哒，中邮证券研究所

秒哒主要功能主要包括：**1)0代码编程：**用户可以通过图形化界面和自然语言来开发软件，无需编写代码。**2)多智能体协作：**软件内置多个智能体，能协同工作，处理复杂的任务和流程。**3)规模化工具调用：**用户可以便捷地调用各种工具和API，实现功能扩展和集成。**4)直观操作：**提供易于理解的界面和操作方式，非技术人员也能轻松上手。**5)创意实现：**用户可以用“秒哒"将自己的想法快速转化为实际的软件应用。**6)自动化流程：**通过智能体的自动化处理，简化和加速软件开发和部署过程。**7)模块化构建：**支持模块化设计，用户可以像搭积木一样构建软件，提高开发效率。

3 AI 行业洞察

3.1 TAO 方法微调 Llama 模型，FinanceBench 跑分超 GPT-4o

3 月 26 日，Databricks 发布新型大语言模型微调方法 TAO（Test-time Adaptive Optimization），通过无标注数据和强化学习技术，在显著降低企业成本的同时提升模型性能。

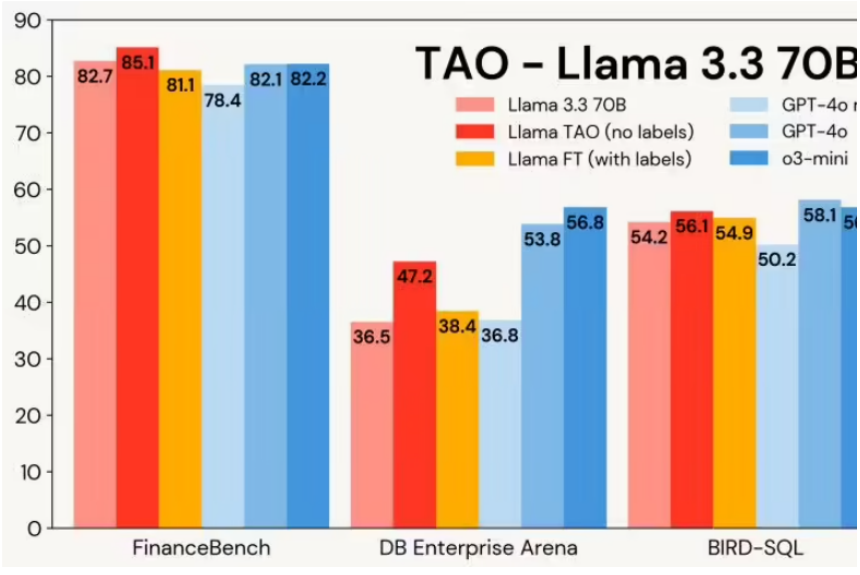
图表12：TAO

	Prompting	Fine-Tuning	Classic Test-Time Compute	Test-time Adap Optimization (T
Data Required	None	Input-Output Examples Typically thousands of high-quality outputs	None	Input Examples O Existing LLM usage data annotations needed
Quality Achievable	Medium Limited by complexity & size of prompts supported	High But limited by quality of human outputs	High Particularly so for logical reasoning tasks	High Shown on enterprise ta this blog
Inference Cost	Low	Low	High Long reasoning chains, thinking tokens, etc.	Low Same as for the original

资料来源：Databricks，中邮证券研究所

测试显示，在金融文档问答和 SQL 生成任务中，通过 TAO 微调后的 Llama 3.3 70B 模型，表现甚至超越传统标注微调方法，逼近 OpenAI 顶级闭源模型。

图表13：TAO 测评



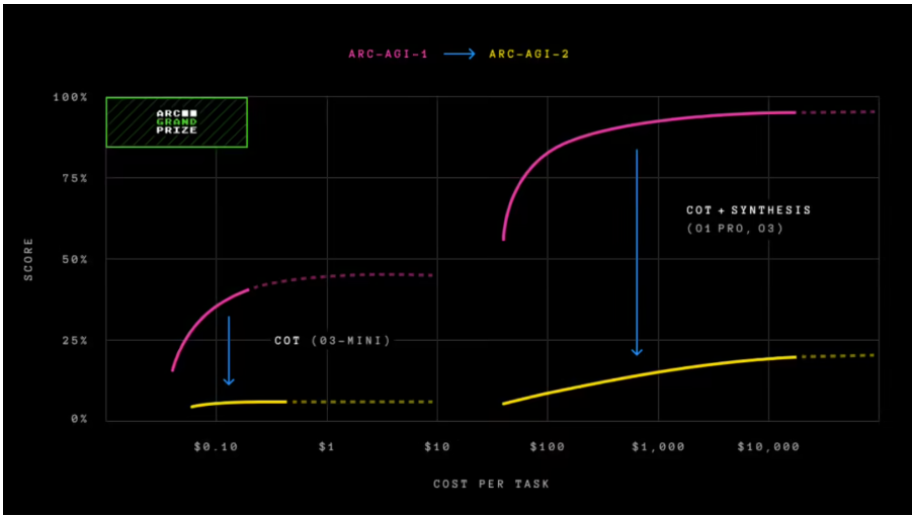
资料来源：Databricks，中邮证券研究所

3.2 ARC-AGI-2 测试登场：AI 模型得分惨淡

Arc Prize 基金会宣布推出一个名为 ARC-AGI-2 的全新测试，旨在衡量领先人工智能模型的通用智能水平。这项测试的难度极高，截至目前，大多数 AI 模型都在该测试中表现不佳。

根据 Arc Prize 排行榜的数据显示，以推理能力著称的 AI 模型，如 OpenAI 的 o1-pro 和 DeepSeek 的 R1，在 ARC-AGI-2 测试中的得分仅为 1%至 1.3%。而包括 GPT-4.5、Claude3.7 Sonnet 和 Gemini 2.0 Flash 等强大的非推理型模型，得分也仅在 1%左右。

图表14：ARC-AGI-1 vs. ARC-AGI-2



资料来源：Arc Prize，中邮证券研究所

4 技术前沿

4.1 昆仑万维首创 MusiCoT 框架

近日，昆仑万维团队发表论文" Analyzable Chain-of-Musical-Thought Prompting for High-Fidelity Music Generation"，提出了 MusiCoT，这是一种专为音乐生成设计的创新性思维链提示技术，旨在解决自回归（AR）模型在生成高保真音乐时的局限性，提升生成音乐的连贯性和创造性。

图表15：MusiCoT 论文

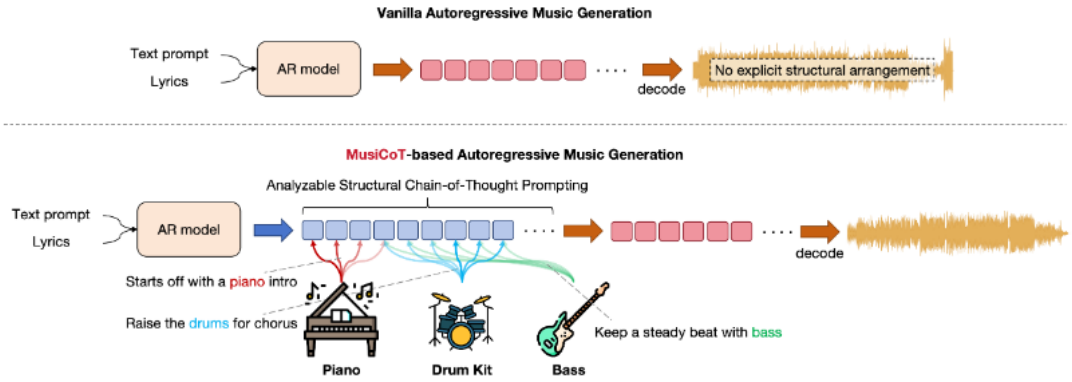
Analyzable Chain-of-Musical-Thought Prompting for High-Fidelity Music Generation

Max W. Y. Lam*†, Yijin Xing*, Weiya You, Jingcheng Wu, Zongyu Yin, Fuqiang Jiang, Hangyu Liu, Feng Liu, Xingda Li, Wei-Tsung Lu, Hanyu Chen, Tong Feng, Tianwei Zhao, Chien-Hung Liu, Xuchen Song†, Yang Li, Yahui Zhou
Skywork AI
{maxwy.lam, xuchen.song}@kunlun-inc.com

资料来源：论文 Analyzable Chain-of-Musical-Thought Prompting for High-Fidelity Music Generation，中邮证券研究所

MusiCoT 跳脱了传统模型的局限，先通过全局视角预生成音乐结构，在精细化填充音频细节，基于 CLAP 模型构建，无需人工标注即具备高扩展性，大幅提升生成音乐可解释性与艺术感染力。除此之外，MusiCoT 通过使 AR 模型的创作过程与音乐思维保持一致，提升了高保真音乐生成的质量，具有结构可分析性和音乐参考支持等优势，为生成式人工智能在音乐领域的发展开辟了新方向。

图表16：MusiCoT 架构



资料来源：论文 Analyzable Chain-of-Musical-Thought Prompting for High-Fidelity Music Generation，中邮证券研究所

4.2 TAO：使用测试时间计算来训练没有标记数据的高效 LLM

Databricks 发表论文“TAO: Using test-time compute to train efficient LLMs without labeled data”，提出了测试时自适应优化（TAO）方法，利用测

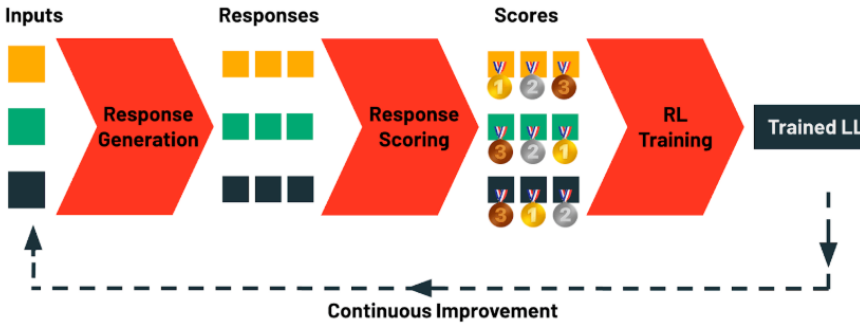
试时计算和强化学习，在无需标记数据的情况下优化大语言模型，提升其在特定任务上的性能，降低成本。

TAO 是一种新的模型调优方法，通过测试时计算让模型探索任务的合理响应，再用强化学习基于这些响应评估更新 LLM，仅需 LLM 的使用数据（输入示例），就能提升模型性能。

在实验设计方面，整个实验包括响应生成、响应评分、强化学习训练和持续改进四个阶段。响应生成收集任务的输入提示并生成多样的候选响应；响应评分对生成的响应进行质量评估；强化学习训练根据高评分响应更新 LLM；持续改进利用用户与 LLM 交互产生的输入数据，不断优化模型。

通过实验发现，在文档问答、SQL 生成等专业企业任务上，TAO 超越了使用数千个标记示例的传统微调方法。以 Llama 8B 和 70B 模型为例，经 TAO 优化后，在 FinanceBench、DB Enterprise Arena 和 BIRD-SQL 等基准测试中，性能显著提升，质量可与昂贵的专有模型（如 GPT-4o 和 o3-mini）相媲美。

图表17：TAO 架构



资料来源：Databricks，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。

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