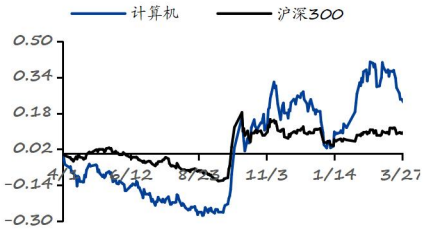
计算机

大模型硝烟再起，DeepSeek、通义千问、Google、OpenAI 先后迎来更新

强于大市（维持评级）

一年内行业相对大盘走势



团队成员

分析师： **钱劲宇(S0210524040006)**
QJY3773@hfzq.com.cn

相关报告

1、Deepseek 发布全新注意力机制 NSA——2025.02.23
2、美 AI 禁令或将升级——2025.01.13
3、把握 2025 年两大核心主线——2025 年计算机行业投资策略报告——2025.01.07

投资要点：

- **DeepSeek-V3 模型更新，各项能力全面进阶**

据 DeepSeek 官微，3 月 25 日，DeepSeek V3 模型已完成小版本升级，目前版本号 DeepSeek-V3-0324，DeepSeek-V3-0324 与之前的 DeepSeek-V3 使用同样的 base 模型，仅改进了后训练方法。私有化部署时只需要更新 checkpoint 和 tokenizer_config.json(tool calls 相关变动）。模型参数约 660B，开源版本上下文长度为 128K（网页端、App 和 API 提供 64K 上下文）。

- **通义千问 Qwen2.5-Omni-7B 正式开源，展现全模态优异性能**

据阿里云开发者官微，3 月 27 日，通义千问 Qwen2.5-Omni-7B 正式开源。作为通义系列模型中首个端到端全模态大模型，可同时处理文本、图像、音频和视频等多种输入，并实时生成文本与自然语音合成输出。Qwen2.5-Omni 以接近人类的多感官方式「立体」认知世界并与之实时交互，还能通过音视频识别情绪，在复杂任务中进行更智能、更自然的反馈与决策。目前，开发者和企业可免费下载商用 Qwen2.5-Omni，手机等终端智能硬件也可轻松部署运行。

- **谷歌发布“最先进复杂任务模型”Gemini 2.5 Pro，支持原生多模态**

据量子位，3 月 26 日，赶在 OpenAI 直播之前，谷歌发布 Gemini 2.5 Pro。谷歌介绍，相较于 Gemini 2.0 Flash Thinking 这个谷歌首个推理模型，Gemini 2.5 在基础模型和后训练技术上都有改进。不仅是在大模型竞技场上一举拿下高分，在各种推理、数学、科学、编程基准上，Gemini 2.5 Pro 都表现出色，属于是编程能跟 Claude 3.7 Sonnet 掰手腕，数学能跟 Grok 3 相媲美。

- **OpenAI 放出 GPT-4o 原生多模态图像生成功能**

据 InfoQ，3 月 26 日，OpenAI 发布了 GPT-4o image generation，图像生成技术模型。此初始版本仅专注于图像创建，并将在 ChatGPT Plus、Pro、Team 和 Free 订阅层中提供。值得注意的是价格，OpenAI 声称与 GPT-4 Turbo 相比，价格降低了 50%。更直观的对比是，GPT-4o 成本恰好是 10 倍 GPT-3.5；4o 是 5 美元 / 百万输入 token 和 15 美元 / 百万输出 token。3.5 是 0.50 美元 / 百万输入 token 和 1.50 美元 / 百万输出 token。价格下降尤其引人注目，因为 OpenAI 承诺也将向免费 ChatGPT 用户提供该模型——这是他们第一次直接向非付费客户提供“最佳”模型。

- **风险提示**

市场需求不及预期，人工智能技术发展不及预期，政策发布不及预期，大模型商业落地不及预期的风险等。

正文目录

图表目录

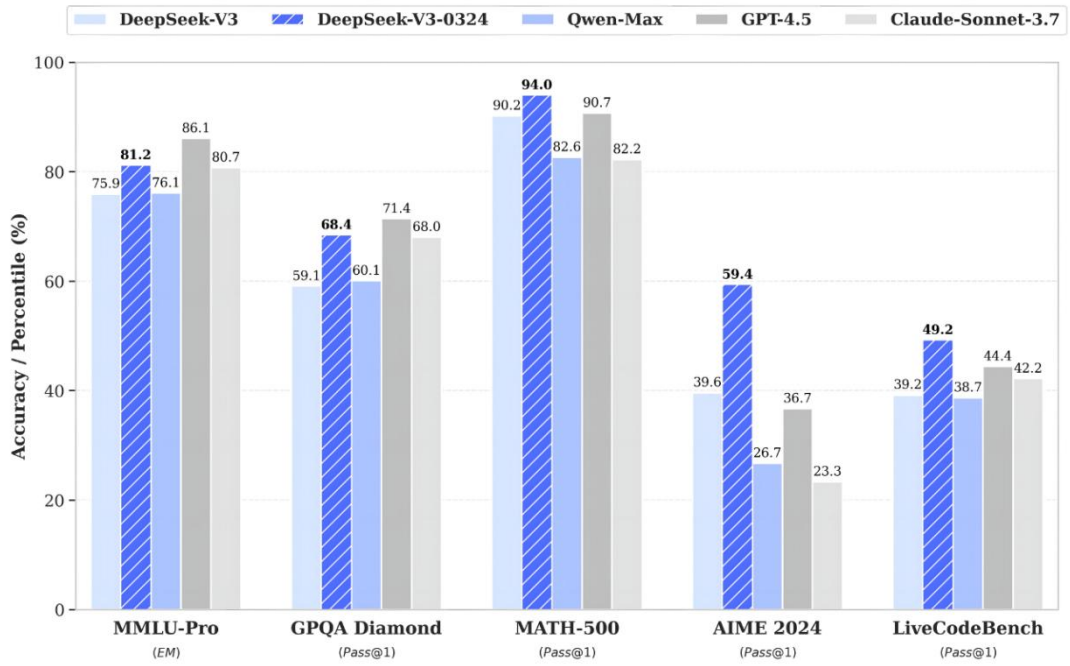
1 DeepSeek-V3 模型更新，各项能力全面进阶

据 DeepSeek 官微，3 月 25 日，DeepSeek V3 模型已完成小版本升级，目前版本号 DeepSeek-V3-0324，DeepSeek-V3-0324 与之前的 DeepSeek-V3 使用同样的 base 模型，仅改进了后训练方法。私有化部署时只需要更新 checkpoint 和 tokenizer_config.json（tool calls 相关变动）。模型参数约 660B，开源版本上下文长度为 128K（网页端、App 和 API 提供 64K 上下文）。

进阶能力包括：推理任务表现提高、前端开发能力增强、中文写作升级、中文搜索能力优化，此外，新版 V3 模型在工具调用、角色扮演、问答闲聊等方面也得到了一定幅度的能力提升。

推理任务方面，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。前端开发方面，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。中文写作方面，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。中文搜索方面，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。

图表 1：新版 V3 模型的百科知识、数学、代码任务表现均有提升



来源：DeepSeek，华福证券研究所

2 通义千问 Qwen2.5-Omni-7B 正式开源，展现全模态优异性能

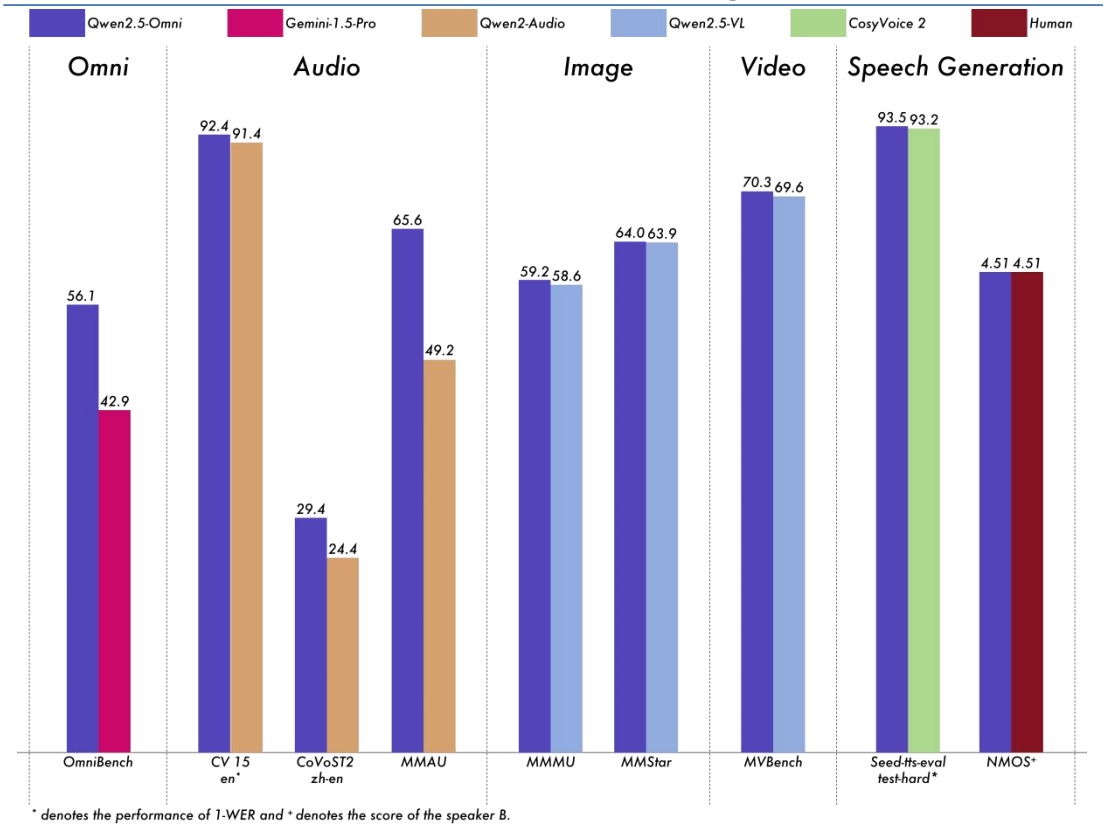
据阿里云开发者官微，3 月 27 日，通义千问 Qwen2.5-Omni-7B 正式开源。作为通义系列模型中首个端到端全模态大模型，可同时处理文本、图像、音频和视频等

多种输入，并实时生成文本与自然语音合成输出。Qwen2.5-Omni 以接近人类的多感官方式「立体」认知世界并与之实时交互，还能通过音视频识别情绪，在复杂任务中进行更智能、更自然的反馈与决策。目前，开发者和企业可免费下载商用 Qwen2.5-Omni，手机等终端智能硬件也可轻松部署运行。

Qwen2.5-Omni 采用了通义团队全新首创的 Thinker-Talker 双核架构、Position Embedding （位置嵌入）融合音视频技术、位置编码算法 TMRoPE（Time-aligned Multimodal RoPE）。双核架构 Thinker-Talker 让 Qwen2.5-Omni 拥有了人类的“大脑”和“发声器”，形成了端到端的统一模型架构，实现了实时语义理解与语音生成的高效协同。具体而言，Qwen2.5-Omni 支持文本、图像、音频和视频等多种输入形式，可同时感知所有模态输入，并以流式处理方式实时生成文本与自然语音响应。

相较于动辄数千亿参数的闭源大模型，Qwen2.5-Omni 以 7B 的小尺寸让全模态大模型在产业上的广泛应用成为可能。即便在手机上，也能轻松部署和应用 Qwen2.5-Omni 模型。当前，Qwen2.5-Omni 已在魔搭社区和 Hugging Face 同步开源，用户也可在 Qwen Chat 上直接体验。

图表 2：在权威的多模态融合任务 OmniBench 等测评中 Qwen2.5-Omni 刷新业界纪录



来源：阿里云开发者，华福证券研究所

3 谷歌发布“最先进复杂任务模型”Gemini 2.5 Pro，支持原生多模态

据量子位，3 月 26 日，赶在 OpenAI 直播之前，谷歌发布 Gemini 2.5 Pro。谷歌

介绍，相较于 Gemini 2.0 Flash Thinking 这个谷歌首个推理模型，Gemini 2.5 在基础模型和后训练技术上都有改进。不仅是在大模型竞技场上一举拿下高分，在各种推理、数学、科学、编程基准上，Gemini 2.5 Pro 都表现出色，属于是编程能跟 Claude 3.7 Sonnet 掰手腕，数学能跟 Grok 3 相媲美。

Gemini 2.5 Pro 的上下文窗口是 1M tokens，并且支持原生多模态：可以理解庞大数据集并处理来自不同信息源的复杂问题，包括文本、音频、图像、视频，甚至是整个代码库。目前，Gemini 2.5 Pro 已经面向 Gemini Advanced 付费用户开放，开发人员也可以在 Google AI Studio 中试用。谷歌表示，未来几周内还将在 Vertex AI 上推出该模型。

图表 3：Gemini 2.5 Pro 测试结果

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	**91.5%**	36.3%	48.8%	—	—	—
	1M	83.1%	—	—			
Multilingual performance Global MMLU (Lite)		89.8%	—	—	—	—	—

来源：量子位，华福证券研究所

4 OpenAI 放出 GPT-4o 原生多模态图像生成功能

据 InfoQ，3 月 26 日，OpenAI 发布了 GPT-4o image generation，图像生成技术模型。此初始版本仅专注于图像创建，并将在 ChatGPT Plus、Pro、Team 和 Free 订阅层中提供。值得注意的是价格，OpenAI 声称与 GPT-4 Turbo 相比，价格降低了 50%。更直观的对比是， GPT-4o 成本恰好是 10 倍 GPT-3.5；4o 是 5 美元 / 百万输入 token 和 15 美元 / 百万输出 token。3.5 是 0.50 美元 / 百万输入 token 和 1.50 美元 / 百万输出 token。价格下降尤其引人注目，因为 OpenAI 承诺也将向免费 ChatGPT 用户提供该模型——这是他们第一次直接向非付费客户提供“最佳”模型。

OpenAI 研究负责人 Gabriel Goh 在接受媒体采访时表示：“该模型比以前的模型有了很大的改进”，并补充说，团队使用了 GPT-4o“全模态”——一种可以生成任何类型数据（如文本、图像、音频和视频）的模型——作为该功能的基础。

据 OpenAI 官方说明，GPT-4o 在多个方面相较于过去的模型进行了改进：1）更好的文本集成：与过去那些难以生成清晰、恰当位置文字的 AI 模型不同，GPT-4o 现在可以准确地将文字嵌入图像中；2）增强的上下文理解：GPT-4o 通过利用聊天历史，允许用户在互动中不断细化图像；3）改进的多对象绑定：过去的模型在正确定位场景中的多个不同物体时存在困难，而 GPT-4o 现在可以一次处理多达 10 至 20 个物体；4）多样化风格适应：该模型可以生成或将图像转化为多种风格，支持从手绘草图到高清写实风格的转换。

作为 ChatGPT 中的默认图像生成工具，4o 图像生成功能从即日起开始向 Plus、Pro、Team 及 Free 用户全面开放。Enterprise 及 Edu 访问权限将后续开放。Sora 也可享受到此次功能升级。对于希望继续使用 DALL-E 的用户来说，则可通过专门的 DALL-E GPT 访问这项新功能。开发人员很快就能通过 API 使用 GPT-4o 生成图像功能，访问权限将在未来几周内开放。OpenAI 表示，整个图像创建与自定义过程，就像与 GPT-4o 聊天一样简单——只需描述你的需求，包含画面比例、使用十六进制代码的精确色彩或透明背景等细节即可。由于此模型能够生成涉及更多细节的图像，因此渲染时间可能更长，最多可能达到 1 分钟。

5 风险提示

市场需求不及预期，人工智能技术发展不及预期，政策发布不及预期，大模型商业落地不及预期的风险等。

分析师声明

本人具有中国证券业协会授予的证券投资咨询执业资格并注册为证券分析师，以勤勉的职业态度，独立、客观地出具本报告。本报告清晰准确地反映了本人的研究观点。本人不曾因，不因，也将不会因本报告中的具体推荐意见或观点而直接或间接收到任何形式的补偿。

一般声明

华福证券有限责任公司（以下简称“本公司”）具有中国证监会许可的证券投资咨询业务资格。本报告仅供本公司的客户使用。本公司不会因接收人收到本报告而视其为客户。在任何情况下，本公司不对任何人因使用本报告中的任何内容所引致的任何损失负任何责任。

本报告的信息均来源于本公司认为可信的公开资料，该等公开资料的准确性及完整性由其发布者负责，本公司及其研究人员对该等信息不作任何保证。本报告中的资料、意见及预测仅反映本公司于发布本报告当日的判断，之后可能会随情况的变化而调整。在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的报告。本公司不保证本报告所含信息及资料保持在最新状态，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。

在任何情况下，本报告所载的信息或所做出的任何建议、意见及推测并不构成所述证券买卖的出价或询价，也不构成对所述金融产品、产品发行或管理人作出任何形式的保证。在任何情况下，本公司仅承诺以勤勉的职业态度，独立、客观地出具本报告以供投资者参考，但不就本报告中的任何内容对任何投资做出任何形式的承诺或担保。投资者应自行决策，自担投资风险。

特别声明

投资者应注意，在法律许可的情况下，本公司及其本公司的关联机构可能会持有本报告中涉及的公司所发行的证券并进行交易，也可能为这些公司正在提供或争取提供投资银行、财务顾问和金融产品等各种金融服务。投资者请勿将本报告视为投资或其他决定的唯一参考依据。

投资评级声明

类别	评级	评级说明
公司评级	买入	未来 6 个月内，个股相对市场基准指数指数涨幅在 20%以上
	持有	未来 6 个月内，个股相对市场基准指数指数涨幅介于 10%与 20%之间
	中性	未来 6 个月内，个股相对市场基准指数涨幅介于-10%与 10%之间
	回避	未来 6 个月内，个股相对市场基准指数涨幅介于-20%与-10%之间
	卖出	未来 6 个月内，个股相对市场基准指数涨幅在-20%以下
行业评级	强于大市	未来 6 个月内，行业整体回报高于市场基准指数 5%以上
	跟随大市	未来 6 个月内，行业整体回报介于市场基准指数-5%与 5%之间
	弱于大市	未来 6 个月内，行业整体回报低于市场基准指数-5%以下

备注：评级标准为报告发布日后的 6~12 个月内公司股价（或行业指数）相对同期基准指数的相对市场表现。其中 A 股市场以沪深 300 指数为基准；香港市场以恒生指数为基准，美股市场以标普 500 指数或纳斯达克综合指数为基准（另有说明的除外）

联系方式

华福证券研究所　上海
公司地址：上海市浦东新区浦明路 1436 号陆家嘴滨江中心 MT 座 20 层
邮编：200120
邮箱：hfyjs@hfzq.com.cn