东吴证券 SOOCHOW SECURITIES

电子行业深度报告

如何展望 GPT-5 带来的算力增长？从参数量、时间表、影响力三重视角——算力需求看点系列

2025 年 03 月 18 日

证券分析师 陈海进
执业证书：S0600525020001
chenhj@dwzq.com.cn
研究助理 李雅文
执业证书：S0600125020002
liyw@dwzq.com.cn

增持（维持）

投资要点

- **我们如何定义 GPT 下一代大模型？** 我们判断 OpenAI 对大模型的产品线与预期曾进行过调整。2024 年 7 月 OpenAI 首席技术官 Mira Murati 称，GPT-5 有望在 2025 年底或 2026 年初推出。但根据 2025 年 2 月 13 日 Altman 在社交平台上表明，GPT-5 几个月后面世。我们判断，GPT-5 的发布时间或提前，或由于 DeepSeek 近期的重磅更新和亮眼表现对 OpenAI 产品版图构成了威胁，进而希望加快产品迭代步伐。自 OpenAI 在 2015 年成立以来，通过多轮融资不断扩展其技术产品布局。从时间周期来看，平均 1 年半左右 OpenAI 会获得一次新的融资。从大的模型迭代节点来看，随着 ChatGPT 把热度推高，竞争对手持续推陈出新，市场对于 OpenAI 产品迭代速度的预期在加快。

- **GPT-5 预训练情况如何测算？** 我们基于普遍应用的算力供给需求公式，讨论公式中的核心参数变化趋势，以此给出我们的判断。已知 GPT-4 参数量为 1.8 万亿，在 2.5 万张 A100 上训练 90-100 天。已知 GPT-4.5 的计算量约为 GPT-4 模型的 10 倍，假设 Scaling Law 持续奏效，假设其具有 3w-5w 张 H100 的训练资源，其他条件与 GPT-4 基本保持一致，则可以推测，1）GPT-4.5 模型参数量约为 5.7 万亿，2）GPT-4.5 模型训练用时大约需要 148-247 天。关于 GPT-5：假设达到比 GPT-4 强 100 倍的运算能力，其他假设相同的情况下，可测算得到 1）GPT-5 模型参数量约为 18 万亿，2）模型训练用时大约需要 203-225 天。

- **GPT-5 推出对 AI 行业有何影响？** **影响#1** 虽然业内对于大模型发展的方向仍处于激烈讨论中，但头部大模型厂商的万卡集群建设未曾停歇。聚焦国内市场，从 GPT-4 能力的大模型发布时间表来看，普遍比 GPT-4 晚一年的时间。由此我们预计，GPT-5 若引发新一轮 AI 热潮，更多大集群的建设会提上日程。**影响#2** 24/12 月，ChatGPT 周活跃用户数已经超过 3 亿，目标是在未来一年内达到 10 亿用户。据“推理需求=2×参数量×token”的计算公式，在其他条件不变的前提下，2025 年推理市场空间有望达到 2024 年的三倍；若 GPT-5 带动参数量大幅提升（按 18 万亿计算），假设 26 年 ChatGPT 总体推理消耗的 tokens 为 25 年的 2 倍，按二八法则假设 26 年 tokens 消耗中仅有 20%为 GPT-5 的需求，则综合下来 26 年推理算力需求有望达到 25 年的 5.6 倍左右。

- **产业链相关公司：** 工业富联、沪电股份、胜宏科技、寒武纪、海光信息（与计算机联合覆盖）、龙芯中科、盛科通信（与通信联合覆盖）等。

- **风险提示：** AI 应用进展不及预期风险，Scaling Law 放缓或失效风险，GPU 技术升级不及预期的风险，万卡集群建设不及预期风险。

行业走势

相关研究

《GPGPU 与 ASIC 之争——算力芯片看点系列》

2025-03-12

内容目录

图表目录

1. 我们如何定义 OpenAI 下一代大模型？

1.1. “猎户座”（Orion）的进展预期如何变化？

关于“猎户座”（Orion）：这一表述最早出现在 2024 年 8 月 The Information 的一篇报道中：“OpenAI 开发的最新大模型“猎户座”或于 2025 年年初推出。”2024 年 9 月，OpenAI 日本公司 CEO 介绍了 GPT-Next 模型，被认为是“猎户座”，预期达到比 GPT-4 强 100 倍的运算能力，将成为大模型在语言处理和多模态功能上实现飞跃的重要里程碑。

图1：“猎户座”和“草莓”进展预期

新闻时间	进展预期	“草莓”Strawberry	“猎户座”Orion
2023年11月下旬	OpenAI新模型Q*进入公众视野：OpenAI几位研究人员给董事会写的警告信中首次提到内部名为Q*的下一代AI模型。	Q*是Strawberry项目的前身，能够回答数学问题，超越了其他商业化模型的能力。	
2024/7/13	根据路透社2024年5月份看到的一份 OpenAI内部文件副本，OpenAI正在开发Strawberry。	**Strawberry旨在增加OpenAI模型的推理能力**。推理能力是人工智能实现人类或超人级智能的关键。	
2024/8/7	OpenAI CEO在社交平台更新了一张关于草莓照片的动态。	引发公众对Strawberry项目的热议，为项目正式发布预热。	
2024/8/27	据The Information，OpenAI Strawberry项目计划最早于2024年秋季推出，“猎户座”Orion或于2025年初推出。	性能预告：①解决此前从未见过的数学问题；②解决涉及编程的问题，且不局限于回答技术性问题；③如果给予更多时间思考，还可以回答用户更“主观”的问题。	**OpenAI使用更大版本的Strawberry生成训练下一代旗舰模型“猎户座”Orion的数据。**Orion旨在帮助OpenAI获得对话式AI或大型语言模型的竞争力。
2024/9/3	KDDI峰会召开，OpenAI日本公司首席执行官长崎忠雄介绍了**GPT-Next模型。**		GPT Next将在使用与**GPT-4近似的计算资源情况下，预计有效计算容量将提升至100倍。**这一巨大跃迁不仅意味着简单的硬件升级，更涉及到算法架构的深层次变革和学习效率的极大提升。**GPT-Next模型被认为是“猎户座”。**
2024/9/10	OpenAI计划在未来两周内将Strawberry作为ChatGPT服务的一部分发布。	①与其他对话式AI最大的区别：在响应之前进行“思考”，思考阶段通常会持续 10 到 20 秒。 ②缺点：初始版本只能接收和生成文本，不具备多模态能力。	
2024/9/13	正式推出Strawberry模型的部分预览版，命名为**o1-preview**。还推出了更快、更小的o1 mini，使用与o1类似的框架进行训练，其成本比o1预览版便宜80%。	①性能：拥有进化的**推理能力**，在回答前进行缜密思考，生成**内部思维链**，在物理、生物、化学问题的基准测试中准确度超过了人类博士水平。 ②缺点：**只支持文本对话，不具备多模态能力。**	
2024/10/25	据The Verge，OpenAI计划2024年12月前推出下一代模型“猎户座”Orion。		①性能预告：具备处理文本、图像和视频等多模态数据的能力。 ②目标：**达到比GPT-4强100倍的能力。**
2024/12/6	**o1完整版**上线。同时推出了ChatGPT Pro，订阅费用200美元/月，可以无限制地访问模型。	性能升级：复杂问题，o1能够进行更深入、更全面的思考；简单问题，o1能够快速给出精准答案；**同时处理图像和文本信息**；o1 Pro为模型增添了更强大的思考能力。	
2024/12/21	OpenAI“连续12日圣诞发布”迎来大结局，OpenAI推出重磅收官新品，其迄今最强前沿推理模型的升级版——**o3**。OpenAI号称o3在一些条件下接近AGI。	o3有完整版和mini版，新功能是可将模型推理时间设置为低、中、高，模型思考时间越高，效果越好。mini版更精简，针对特定任务进行了微调，将在1月底推出，之后不久推出o3完整版。	
2025/2/1	**首个OpenAI免费推理模型o3-mini**发布,一共包含三个版本：low、medium和high。	整体和前一代o1-mini类似，也针对STEM（Science、Technology、Engineering、Mathematics）进行了优化，延续mini系列小而美的风格。仅o3-mini（medium），不但在数学编码上表现与o1系列相当，而且响应更快。	

数据来源：机器之心，量子位，网易新闻，每日经济新闻，券商中国，智东西，华尔街见闻，东吴证券研究所

关于 GPT-5: 市场曾预测 GPT-5 可能在 2023 年底或 2024 年夏季发布。2024 年 7

月左右传出 GPT-5 可能大幅推迟上线的消息，OpenAI 首席技术官 Mira Murati 称，GPT-5 有望在 2025 年底或 2026 年初推出，并表示 GPT-5 的性能将迎来重大飞跃，在特定任务中达到博士级智能水平。据中国经济网，GPT-5 内部代号为“Gobi”和“Arrakis”，是一个具有 52 万亿参数的多模态模型。

根据最新发布的 GPT-4.5（Orion）相关数据来看，我们判断 OpenAI 对大模型的产品线与预期曾进行过调整。2025 年 2 月 28 日，GPT-4.5（代号 Orion）发布，成为 GPT 系列最后一代非“思维链”模型，其计算量为上一代的 10 倍。而 2024 年 9 月，OpenAI 日本公司 CEO 介绍了 GPT-Next 模型，被认为是“猎户座”，预期达到比 GPT-4 强 100 倍的运算能力。**对比下来，二者口径有所调整。**我们认为，此前 OpenAI 或以“Orion”作为 GPT-5 大版本迭代产品，而据彭博社援引知情者消息，截至 2024 年 11 月，猎户座相比 GPT-4 的进步不及 GPT-4 超越 GPT-3.5 的表现（事实上，根据 25/3 月“Orion”发布后，实测反馈普遍也认为提升幅度不及预期）。同时，我们在 24Q3 陆续看到了关于“Strawberry”的消息传来，以及后续 o 系列大模型的问世。OpenAI 使用“Strawberry”生成训练下一代“Orion”大模型的数据，我们认为也是另辟蹊径之举。

而 GPT-5 的最新预期，我们认为也有所调整。2024 年 7 月 OpenAI 首席技术官 Mira Murati 称，GPT-5 有望在 2025 年底或 2026 年初推出。但根据 2025 年 2 月 13 日 Altman 在社交平台上表明，GPT-5 几个月后面世。**我们判断，GPT-5 的发布时间或提前，或由于 DeepSeek 近期的重磅更新和亮眼表现对 OpenAI 产品版图构成了威胁，进而希望加快产品迭代步伐。**

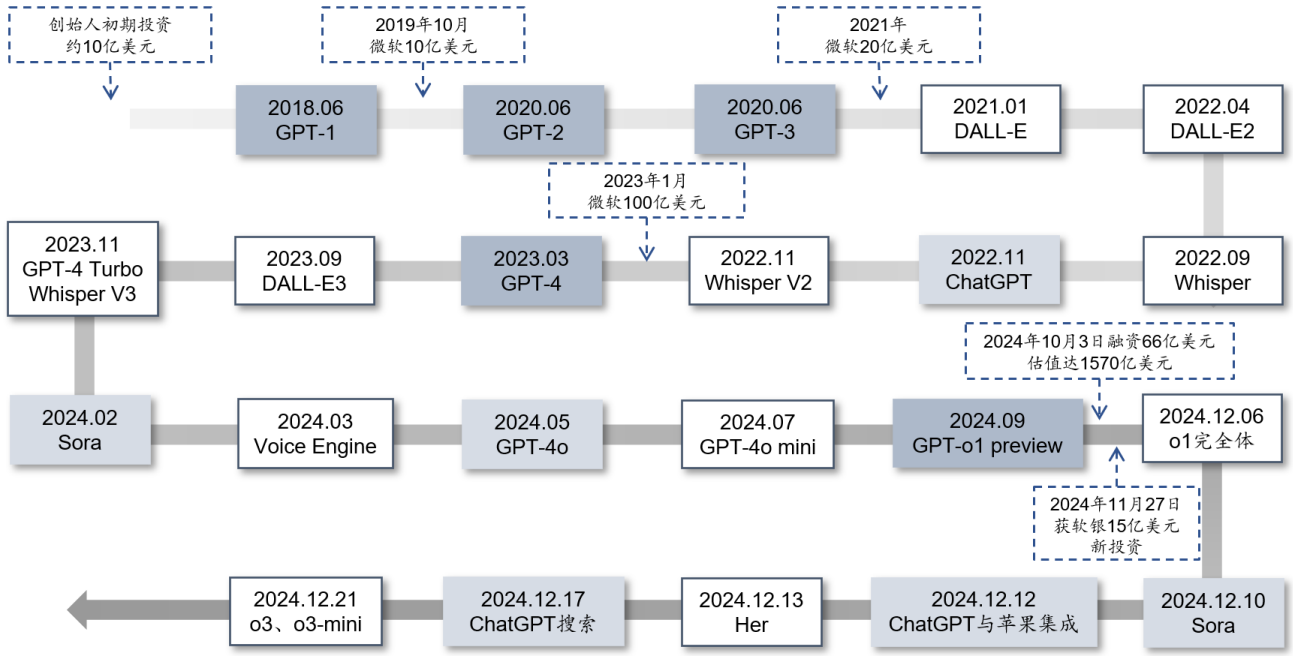
1.2. GPT-5 发布时间如何判断？——从融资视角

自 OpenAI 在 2015 年成立以来，通过多轮融资不断扩展其技术产品布局。最初，OpenAI 通过创始人初期投资启动其研究工作，专注于人工智能的基础研究，于 2018 年 6 月推出 GPT-1。随后，OpenAI 与微软达成合作，分别在 2019 年、2021 年获得了总计 30 亿美元的投资。2022 年 11 月见证了 ChatGPT 的问世。2023 年 1 月，微软再次投入大规模资金支持，金额高达 100 亿美元。其后，GPT 系列不断推陈出新，从 2023 年的 GPT-4、GPT-4 Turbo，到 2024 年的 GPT-4o 和 GPT-4o mini，多样化的产品先后发布，以满足不同客户的需求。2024 年 10 月 3 日，OpenAI 宣布获得 66 亿美元融资，这是该公司迄今最大的风投交易。OpenAI 的投后估值冲破 1570 亿美元，短短 9 个月时间公司估值接近翻倍，创下硅谷历史最高纪录。

从时间周期来看，平均 1 年半左右 OpenAI 会获得一次新的融资。以当前最后一次融资时间 2024Q4 计算，**下一次融资时间或为 2026 年初。**从大的模型迭代节点来看，虽然大版本迭代时间约为 2-3 年，但随着 ChatGPT 把热度推高，竞争对手持续推陈出新，市场对于 OpenAI 产品迭代速度的预期在加快。自 GPT-4 发布以来，约 1 年半的时间后 GPT-o1 preview 问世。由此，**我们推断下一次大的版本更新不会超过一年半的时间，或**

对应 2025 年底。

图2：OpenAI 融资与产品 Roadmap 时间表



数据来源：钛媒体，华尔街见闻，第一财经，财联社，EC Innovations，机器之心，壹沓科技，每日经济新闻，BFT 白芙堂 B 站官方，东吴证券研究所

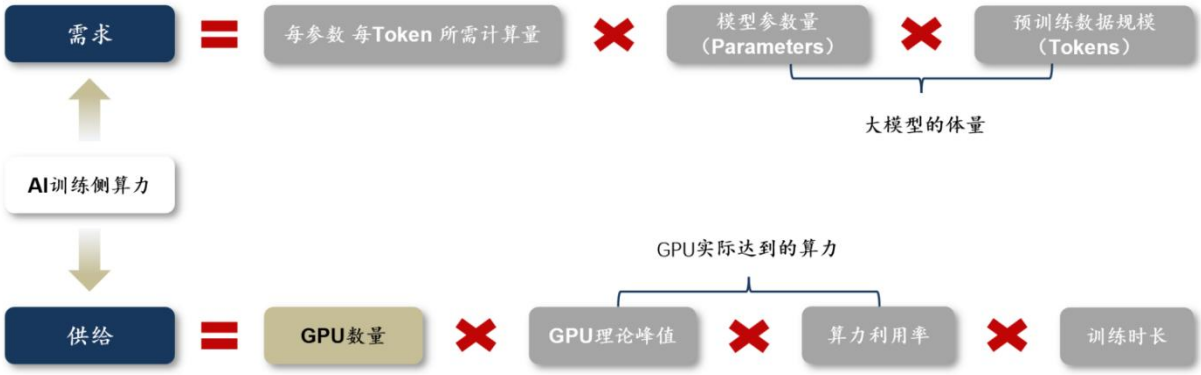
为什么 GPT-4.5（Orion）发布时间为 25Q1？OpenAI 为解决开发“猎户座”时遇到的挑战——高质量训练数据的供应减少，而积极探索合成数据生成。GPT-o1 拥有进化的推理能力，在“回答”前能够进行缜密思考，生成内部思维链，因此其重要的应用之一是为“猎户座”生成高质量的训练数据，o1 生成的高质量训练数据可以帮助“猎户座”减少生成的错误数量（也称为“幻觉”）。我们已知 GPT-o1 预览版已于 9 月推出，并可以为“猎户座”提供训练数据。假设 9 月相关训练数据已经到位，已知 GPT-4 曾采用 2.5 万张 A100 训练 100 天得到，假设由于训练难度等原因“猎户座”训练时间长于 GPT-4，则**“猎户座”最早发布时间为 25Q1**。

GPT-5 的发布时间为何一再推迟？我们认为可以从融资时间轴窥探一二，在时间相对充裕（从“猎户座 Orion”训练完成到下一轮融资节点还有大半年时间）的情况下，我们认为 OpenAI 会安排比“Orion”更为“重量级”的产品（如内部代号为“Gobi”和“Arrakis”的具有 52 万亿参数的多模态模型）命名为 GPT-5，而将 GPT-4.5（Orion）作为过渡性质的产品。

2. GPT-5 预训练情况如何测算？

我们基于普遍应用的算力供给需求公式，讨论公式中的核心参数变化趋势，以此给出我们的判断。

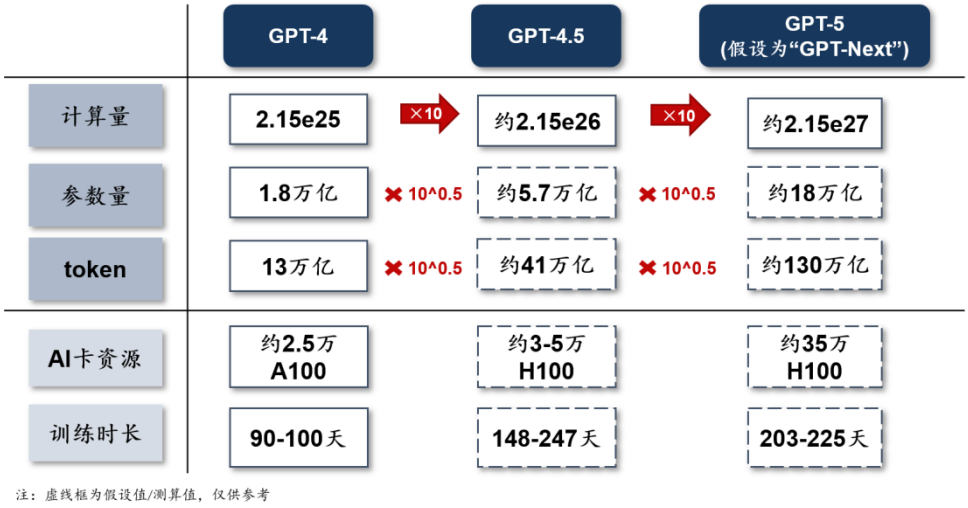
图3：文本大模型 AI 训练侧算力供给需求公式



数据来源：NVIDIA&Stanford University&Microsoft Research《Efficient Large-Scale Language Model Training on GPU Clusters Using Megatron-LM》，新智元，CIBA 新经济，东吴证券研究所

已知 GPT-4 参数量为 1.8 万亿，在 2.5 万张 A100 上训练 90-100 天。市场此前普遍预测“GPT-5”（也即实际命名为 GPT-4.5 的大模型，预期变化的推演详见本文 1.1）可能需要 3w-5w 张 H100 训练，由此测算 AI 卡资源约为 GPT-4 的 1.2-2 倍。在 FP16 精度下，H100 算力为 989TFLOPS，A100 为 312TFLOPS，H100 算力约为 A100 的 3.2 倍。已知 GPT-4.5 的计算量约为 GPT-4 模型的 10 倍，假设 Scaling Law 持续奏效，假设 GPT-4、GPT-4.5 与预测中的 GPT-5 均在 FP16 精度下完成训练，且算力利用率、训练迭代次数大致相同，按 GPT-4 训练用时为均值 95 天计算，**由上述公式推测，1）GPT-4.5 模型参数量约为 5.7 万亿，2）GPT-4.5 模型训练用时大约需要 148-247 天。**

图4：GPT-5 预训练情况测算

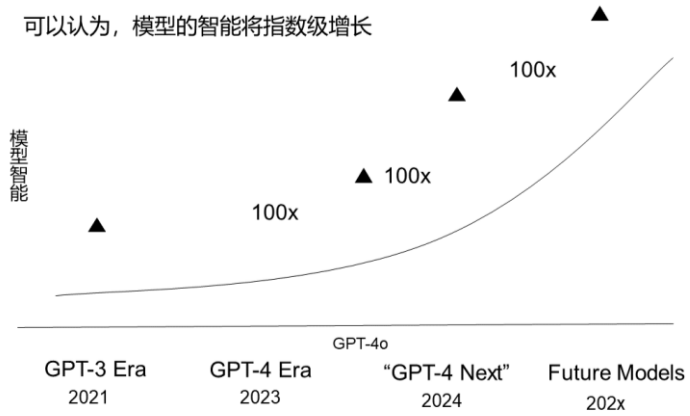


数据来源：海外独角兽，机器之心，英伟达，爱集微，新智元，东吴证券研究所

关于 GPT-5:

1）参数量预期: 当前市场预期 GPT-Next 有望达到比 GPT-4 强 100 倍的运算能力，若假设 Scaling Law 持续奏效，则 100 倍可拆解为参数量 **10 倍**、预训练数据规模 10 倍。假设 GPT-Next 代表 GPT-5 可能会达到的能力，则可以此作为 GPT-5 的体量预期。

图5: OpenAI 日本公司首席执行官长崎忠雄介绍 GPT-Next 模型



数据来源：快科技，东吴证券研究所

2）训练时长预期: 在对计算量有了基本的预期假设后，我们通过假设 AI 集群卡数进一步判断 GPT-5 可能需要的训练时长。2023-24H1，各厂商陆续建成的 5 万卡以下集群，其中比较有代表性的是 Meta 于 24/03 月宣布的两个 24k GPU 集群（共 49152 个 H100），此前提到 24 年底的目标有大幅增长，预计建成 35 万卡 H100 集群。24H2 以来市场最为关注的是 xAI 建设的 10 万卡 H100 集群，2025 年目标或将扩展至 100 万卡。在当前时点，我们假设同样作为头部厂商的 OpenAI 也已具有约 35 万卡 H100 集群的计算资源，倒推可测算得到 GPT-5 模型训练用时大约需要 203-225 天。

图6: 主流科技公司公开宣布的万卡集群情况

厂商	算力建设			
	时间	GPU类型	GPU数量（万张）	备注
Microsoft	2020年	-	1.0	
Google	2023-05	H100	2.6	
	-	TPU v5p	0.9	
Meta	2022年	A100	1.6	
	2024年初	H100	2.5	共2个集群
	2024年底目标	H100	35.0	
AWS	2023-07	H100	2.0	
	正在部署	Trainium2	40.0	
xAI	2024年	H100	10.0	
	计划	-	100.0	

数据来源：消费日报网，机械之心，通信产业网，半导体行业观察，格隆汇 APP，东吴证券研究所

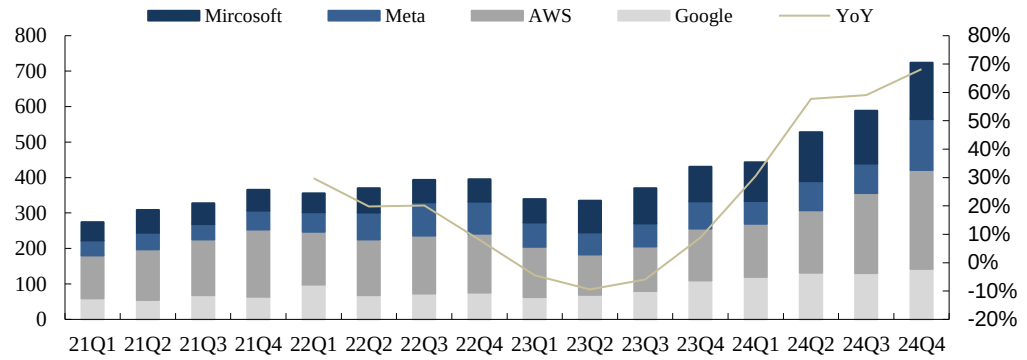
注：本图为非完全统计

3. GPT-5 推出对 AI 行业有何影响？

3.1. 影响#1 大模型“标杆”再上新台阶，引发 AI 行业新一轮竞赛

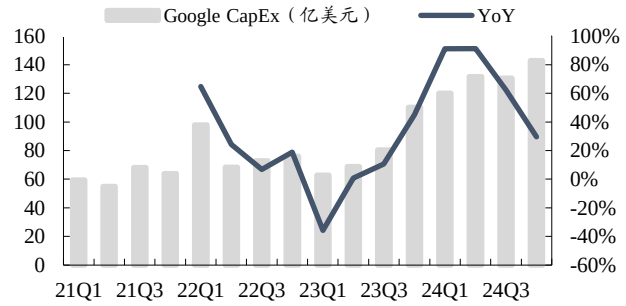
关于互联网大厂 CapEx 投入的担忧始终存在。23Q2 以来，四大云厂商在 CapEx 增速上画出完美的上行弧线，但 24Q2-Q3 开始出现增速持平或放缓的情况。英伟达表示，云服务厂商（CSP）占了数据中心业务的近一半营收，消费互联网公司和企业大约占了另一半。因此，英伟达 GPU 卡的出货量与云服务厂商对于 AI 大模型的态度紧密相关。**而 24Q4 四大云厂商合计增速进一步提升，释放出积极信号。**

图7：北美四大云厂商各季度 CapEx 投入情况及增速（单位：亿美元）



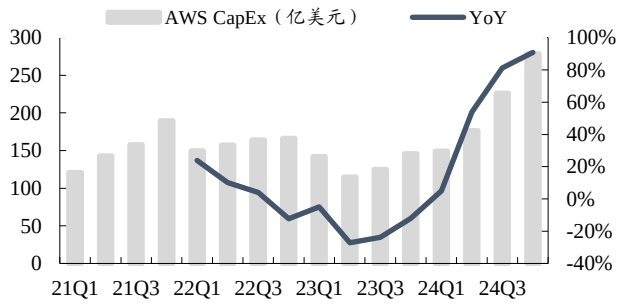
数据来源：各公司公告，Bloomberg，东吴证券研究所

图8：Google 各季度 CapEx 投入情况



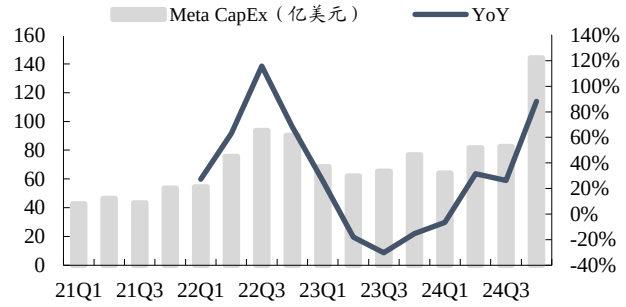
数据来源：公司公告，Bloomberg，东吴证券研究所

图9：AWS 各季度 CapEx 投入情况



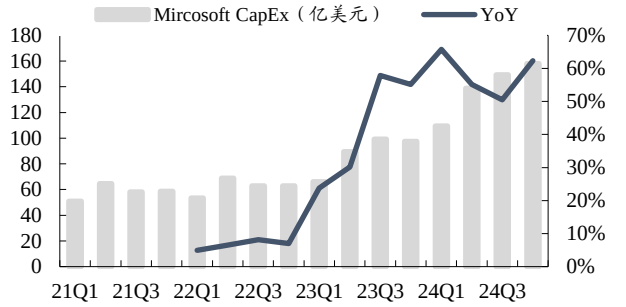
数据来源：公司公告，Bloomberg，东吴证券研究所

图10：Meta 各季度 CapEx 投入情况



数据来源：公司公告，Bloomberg，东吴证券研究所

图11：Microsoft 各季度 CapEx 投入情况



数据来源：公司公告，Bloomberg，东吴证券研究所

我们预计北美四大云厂商均具备 10 万卡集群能力，但 AI 初创公司、乃至国内云厂商或许对 GPT-5 仍持观望态度。而 OpenAI 的“标杆”作用正在于此。聚焦国内市场，从 GPT-4 能力的大模型发布时间表来看，普遍比 GPT-4 晚一年的时间。由此我们预计，**GPT-5 若引发新一轮 AI 热潮，更多大集群的建设会提上日程。**

图12：海外主流 AI 大模型训练侧算力供给需求情况

	单位	GPT	GPT2	GPT3	GPT4	GPT-4o	Gopher	PaLM	PaLM 2	Gemini1.0	LLaMA	LLaMA 2	LLaMA 3	LLaMA 3.1	LLaMA 3.2	LLaMA 3.3
基本信息																
发布机构		OpenAI	OpenAI	OpenAI	OpenAI	OpenAI	DeepMind	Google	Google	Google	Meta	Meta	Meta	Meta	Meta	Meta
发布时间		2018-06	2019	2020-05	2023-03	2024-05	2021-12	2022-04	2023-05	2023-12	2023-02	2023-07	2024-04	2024-07	2024-09	2024-12
AI训练																
大模型算力需求																
参数量	亿	1	15	1746	18000	约2000	2800	5400	3400		70-650	70-700	80-700	80-4050	10-900	700
预训练数据规模（token）	万亿			0.3	13.0		0.3	0.8	3.6		1-1.4	2.0	15+	15+		15+
GPU算力供给																
GPU产品				V100	A100		TPU v3	TPU v4		TPU v4,TPU v5e	A100-80GB	A100-80GB	H100-80GB	H100-80GB	H100-80GB	H100-80GB
GPU数量	片			10000	25000		4096	6144								
理论峰值FP16 TC	TFlops			125	312						312	312	1000	1000	1000	1000
算力利用率				21%	34%		33%	46%								

数据来源：OpenAI 论文，Google 论文，Microsoft 论文，Meta 论文，Github Llama 开源 Model Card，新京报，大数据文摘，新智元，英伟达，谷歌研究院，腾讯科技，机器之心，中关村在线，AIGC 开放社区，Llama 中文社区，河北省科学技术厅，东吴证券研究所

注 1：由于各公司对于大模型的训练数据披露口径不一，以上为本文非完全统计

注 2：GPT4 算力利用率在 32-36%区间，本文取中值粗略计算

注 3：英伟达 V100 理论峰值为官网所示“深度学习 | NVLink 版本”性能

图13：国内主流 AI 大模型训练侧算力供给需求情况

	单位	Hunyuan	Hunyuan-Pro	Hunyuan-Large	Qwen-72B	Qwen2	Qwen2.5	DeepSeek-MoE	DeepSeek v2	DeepSeek v3	GLM-130B	ChatGLM-6B	ChatGLM2-6B	Baichuan	Baichuan2	Baichuan3
基本信息																
发布机构		腾讯	腾讯	腾讯	阿里	阿里	阿里	幻方	幻方	幻方	智谱清言	智谱清言	智谱清言	百川智能	百川智能	百川智能
发布时间		2023-09	2024-04	2024-11	2023-11	2024-06	2024-09	2024-01	2024-05	2024-12	2022-08	2023-03	2023-06	2023-07	2023-09	2024-01
AI训练																
大模型算力需求																
参数量	亿	超千亿	万亿	3890	720	5-720	5-720	20-1446	2360	6710	1300	60	60	70-130	70-130	超千亿
预训练数据规模（token）	万亿	2.0		7.0	3.0	4.5-12	18.0	0.1-2.45	8.1	14.8	0.4+	1	1.4	1.2-1.4	2.6	
GPU算力供给																
GPU产品								A100,H800	H800	H800	A100-40GB				A800	
GPU数量	片									2048	768				1024	1000+
理论峰值FP16 TC	TFlops								990	990					180	
算力利用率																

数据来源：腾讯混元论文&公众号，通义千问官网&论文&公众号&GitHub 网页，DeepSeek 论文&公众号，智谱论文&公众号&Github 网页，百川大模型论文&公众号&Github 网页，腾讯云，AI 新榜，市界，华尔街见闻，东吴证券研究所

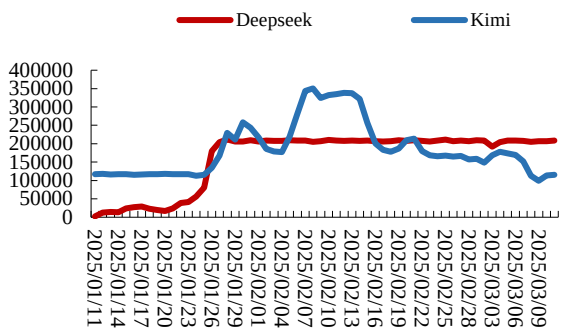
注 1：由于各公司对于大模型的训练数据披露口径不一，以上为本文非完全统计

注 2：Hunyuan、Hunyuan-Pro、Baichuan3 参数量披露口径较为模糊，分别为超千亿参数/万亿参数/超千亿参数，在本图中不涉及左侧第二列单位

3.2. 影响#2 更智能的 AI 带来更优质的体验，AI 推理需求提升

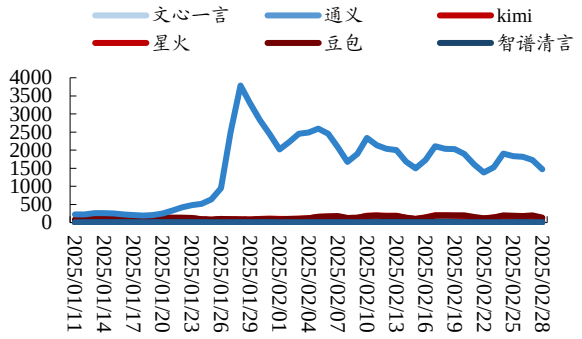
DeepSeek App 全球上线后用户量飙升，AI 竞争加剧，或带动推理算力需求进一步增长。从 2025 年 1 月 11 日至 2025 年 1 月 31 日，DeepSeek 全平台（Web+App）日活跃用户从 124 万涨到峰值的 4541 万，25 年 1 月 Web 月活量达 7068 万。据称 DeepSeek-V3 在预训练阶段每处理 1T token 仅需 180K H800 GPU 小时，即在配备 2048 个 H800 GPU 的集群上仅需 3.7 天。因此，整个预训练阶段在不到两个月内完成，总计使用了 2664K GPU 小时。

图14：Deepseek、Kimi 下载量（IOS+安卓，单位：次）



数据来源：点点数据，东吴证券研究所

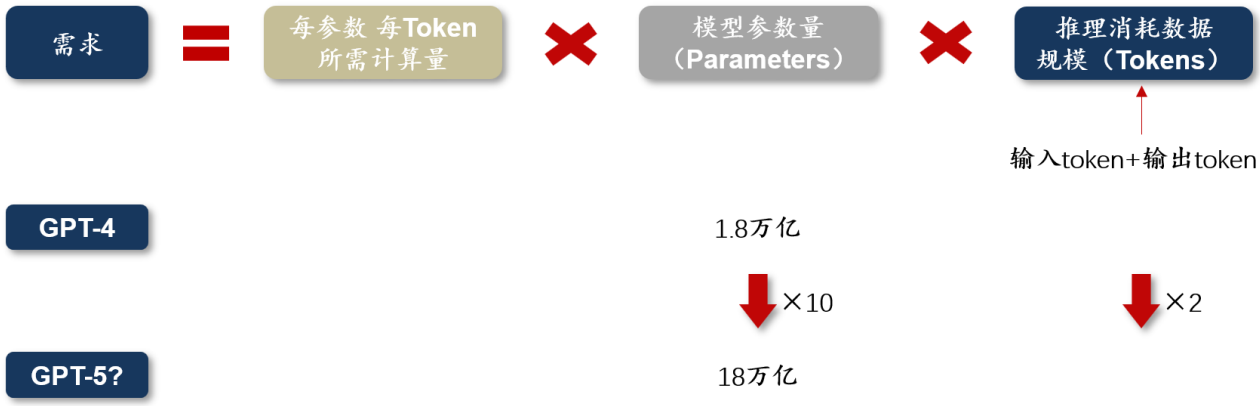
图15：国产 AI 大模型日度访问量（单位：万次）



数据来源：SimilarWeb，东吴证券研究所

GPT-5 若顺利发布，有望带动 2026 年推理需求数倍提升。以 OpenAI 为例，24/08 月 ChatGPT 周活跃用户数突破 2 亿；24/12 月，Altman 也在出席纽约活动时表明，该数量已经超过 3 亿，每一天用户们都向 ChatGPT 发送超过 10 亿条信息。OpenAI 计划通过推出可以帮助用户执行网络信息收集和购物等任务的 AI 智能体以及 ChatGPT 与苹果设备的集成，实现进一步扩张，其目标是在未来一年内达到 10 亿用户——**若仅按用户量推断 25 年推理算力需求，25 年推理算力需求有望达到 24 年的 3 倍以上。**据“推理需求=2 × 参数量 × token”的计算公式，新推出模型的参数量若仍处于 GPT-4 水平，则增长幅度不会太大，推理需求的增长速度以用户消耗的 token 规模为依据；若 GPT-5 带动参数量大幅提升（按本文第二章测算结果，GPT-5 参数量或为 18 万亿），假设 26 年 ChatGPT 总体推理消耗的 tokens 为 25 年的 2 倍，按二八法则假设 26 年 tokens 消耗中仅有 20%为 GPT-5 的需求，**则综合下来 26 年推理算力需求有望达到 25 年的 5.6 倍左右**。

图16：文本大模型 AI 推理算力需求测算



数据来源：OpenAI《Scaling Laws for Neural Language Models》，思瀚产业研究院，极市平台，东吴证券研究所
注：GPT-5 参数量为东吴测算结果，不代表产品实际发布情况

4. 风险提示

AI 应用进展不及预期风险。算力的长期需求是建立在 AI 应用逐步发展之上，在初期大模型训练带来大量算力需求之外，AI 应用带来的推理需求是长期维度上市场空间增长的前提。如果 AI 应用进展不及预期，将对算力各环节需求产生影响。

Scaling Law 放缓或失效风险。以 OpenAI 为代表的大模型厂商大多数沿 Scaling Law 进行迭代升级，当前该法则或存在放缓或失效风险，或对于大模型的发展以及算力的需求产生影响。

GPU 技术升级不及预期的风险。大模型的训练及推理效果除了与 AI 技术发展本身有关，也受到 GPU 等硬件设施的影响。若 GPU 技术升级受到阻碍，或将影响大模型迭代进程。

东吴证券 SOOCHOW SECURITIES

免责声明

东吴证券股份有限公司经中国证券监督管理委员会批准，已具备证券投资咨询业务资格。

本研究报告仅供东吴证券股份有限公司（以下简称“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，本报告中的信息或所表述的意见并不构成对任何人的投资建议，本公司及作者不对任何人因使用本报告中的内容所导致的任何后果负任何责任。任何形式的分享证券投资收益或者分担证券投资损失的书面或口头承诺均为无效。

在法律许可的情况下，东吴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并进行交易，还可能为这些公司提供投资银行服务或其他服务。

市场有风险，投资需谨慎。本报告是基于本公司分析师认为可靠且已公开的信息，本公司力求但不保证这些信息的准确性和完整性，也不保证文中观点或陈述不会发生任何变更，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。

东吴证券投资评级标准

投资评级基于分析师对报告发布日后 6 至 12 个月内行业或公司回报潜力相对基准表现的预期（A 股市场基准为沪深 300 指数，香港市场基准为恒生指数，美国市场基准为标普 500 指数，新三板基准指数为三板成指（针对协议转让标的）或三板做市指数（针对做市转让标的），北交所基准指数为北证 50 指数），具体如下：

公司投资评级：

买入：预期未来 6 个月个股涨跌幅相对基准在 15%以上；

增持：预期未来 6 个月个股涨跌幅相对基准介于 5%与 15%之间；

中性：预期未来 6 个月个股涨跌幅相对基准介于-5%与 5%之间；

减持：预期未来 6 个月个股涨跌幅相对基准介于-15%与-5%之间；

卖出：预期未来 6 个月个股涨跌幅相对基准在-15%以下。

行业投资评级：

增持： 预期未来 6 个月内，行业指数相对强于基准 5%以上；

中性： 预期未来 6 个月内，行业指数相对基准-5%与 5%；

减持： 预期未来 6 个月内，行业指数相对弱于基准 5%以上。

我们在此提醒您，不同证券研究机构采用不同的评级术语及评级标准。我们采用的是相对评级体系，表示投资的相对比重建议。投资者买入或者卖出证券的决定应当充分考虑自身特定状况，如具体投资目的、财务状况以及特定需求等，并完整理解和使用本报告内容，不应视本报告为做出投资决策的唯一因素。

东吴证券研究所
苏州工业园区星阳街 5 号
邮政编码：215021
传真：（0512）62938527
公司网址： http://www.dwzq.com.cn