国产模型迭代持续超预期，美国制裁 22 家国内量子技术机构

行业评级	看好（维持）
国家/地区	中国
行业	计算机行业
报告发布日期	2024 年 05 月 12 日

核心观点

- **计算机板块上周下跌 3.2%，电力 IT 相关公司表现较好，而算力等行业表现相对较弱**。我们认为，投资者对行业基本面仍处于观察和确认期，而从产业的角度看，AI 距离产业化落地越来越近，而设备更新与超长期国债有望给工业、交通、市政等领域数字化带来需求拉动。
- **国内大模型持续超预期**：近期，国内众多模型进行了迭代与更新，包括商汤日日新 5.0、幻方 DeepSeek-V2、生数科技 Vidu、通义千问 2.5，这些模型在整体能力、部署方式、推理成本等各方面均有明显进展与优化，而 Kimi、秘塔搜索、360AI 搜索等应用在 4 月份访问量方面也有较明显环比增长。我们坚信，国内模型正在逐步跨过“可用”到“好用”的门槛，AI 应用的落地与普及前景乐观。
- **算力个股有望持续表现**。国内大模型从能力迭代到产品推广均呈现可喜进展，将带来算力需求提升；另一方面，国家数据局强调加快推进数字基建，上海、广东等地纷纷出台算力领域规划。而从 Q1 业绩来看，浪潮信息、海光信息、工业富联等公司均超预期。
- **设备更新将带动制造业以及交通、市政等领域数字化需求增长**。国务院常务会议周五审议通过《制造业数字化转型行动方案》，提出要加大对中小企业数字化转型的支持，与开展大规模设备更新行动、实施技术改造升级工程等有机结合。我们认为，设备更新将有效带动制造业、交通以及市政等领域数字化需求。
- **量子技术与低空经济有望反复活跃**。两会后，各地对低空经济等领域持续出台政策，量子技术近期也不断有新的进展与突破，相关产业目前均处发展早期，未来有着较大的空间，相关领域有望反复活跃。美国 5 月 9 日将中国 37 家企业列入实体清单，其中 22 个与量子技术相关，凸显量子技术的重要性以及在未来科技发展中的重要作用。

投资建议与投资标的

- **AI 应用领域**，重点关注：金山办公(688111，增持)、虹软科技(688088，未评级)、新致软件(688590，未评级)、彩讯股份(300634，买入)、星环科技-U(688031，未评级)、科大讯飞(002230，买入)等。
- **算力领域**，重点关注：海光信息(688041，买入)、浪潮信息(000977，未评级)、中科曙光(603019，买入)、寒武纪-U(688256，未评级)、润泽科技(300442，未评级)、华铁应急(603300，买入)、亚康股份(301085，未评级)等
- **制造业等领域数字化**，建议关注中控技术(688777，买入)、柏楚电子(688188，未评级)、宝信软件(600845，未评级)、瑞纳智能(301129，买入)、千方科技(002373，未评级)、通行宝(301339，未评级)等。
- **量子技术与低空经济领域**，建议关注国盾量子(688027，未评级)、神州信息(000555，买入)、吉大正元(003029，未评级)、信安世纪(688201，未评级)、莱斯信息(688631，未评级)、中科星图(688568，未评级)。

目 录

图表目录

一、本周行业观点

计算机板块上周下跌 3.2%，电力 IT 相关公司表现较好，而算力等行业表现相对较弱。我们认为，投资者对行业基本面仍处于观察和确认期，而从产业的角度看，AI 距离产业化落地越来越近，而设备更新与超长期国债有望给工业、交通、市政等领域数字化带来需求拉动。

国内大模型持续超预期：近期，国内众多模型进行了迭代与更新，包括商汤日日新 5.0、幻方 DeepSeek-V2、生数科技 Vidu、通义千问 2.5，这些模型在整体能力、部署方式、推理成本等各方面均有明显进展与优化，而 Kimi、秘塔搜索、360AI 搜索等应用在 4 月份访问量方面也有较明显环比增长。我们坚信，国内模型正在逐步跨过“可用”到“好用”的门槛，AI 应用的落地与普及前景乐观。

算力个股有望持续表现。国内大模型从能力迭代到产品推广均呈现可喜进展，将带来算力需求提升；另一方面，国家数据局强调加快推进数字基建，上海、广东等地纷纷出台算力领域规划。而从 Q1 业绩来看，浪潮信息、海光信息、工业富联等公司均超预期。

设备更新将带动制造业以及交通、市政等领域数字化需求增长。国务院常务会议周五审议通过《制造业数字化转型行动方案》，提出要加大对中小企业数字化转型的支持，与开展大规模设备更新行动、实施技术改造升级工程等有机结合。我们认为，设备更新将有效带动制造业、交通以及市政等领域数字化需求。

量子技术与低空经济有望反复活跃。两会后，各地对低空经济等领域持续出台政策，量子技术近期也不断有新的进展与突破，相关产业目前均处发展早期，未来有着较大的空间，相关领域有望反复活跃。美国 5 月 9 日将中国 37 家企业列入实体清单，其中 22 个与量子技术相关，凸显量子技术的重要性以及在未来科技发展中的重要作用。

二、本周行业专题：国产模型近期更新不断，能力提升值得重视

日日新 5.0：全面对标 GPT-4 Turbo

4 月 24 日，商汤在技术交流会上正式发布日日新 5.0 基座大模型，全面对标 GPT-4-Turbo。自去年 4 月首次发布以来，商汤“日日新 SenseNova”大模型体系已正式推出五个大版本迭代。基于超过 10TB tokens 训练、覆盖大量合成数据，全新的日日新 5.0 采用 MoE 架构，总参数 6000 亿，推理时上下文窗口可以有效到 200K 左右。本次更新主要聚集增强了知识、数学、推理及代码能力，全面对标 GPT-4 Turbo，主流客观评测上达到或超越 GPT-4 Turbo。

图 1：日日新 5.0 全面对标 GPT-4 Turbo

Category	Benchmark	GPT-4 Turbo (1106)	GPT-3.5 Turbo	Gain	Llama2-70B Instruct	Llama2-70B Chat	Gain	SenseChat V5
综合考试	MMLU (5-Shot)	83.61	69.07	21.05%	80.52	59.95	34.36%	84.78
	CMMLU (5-shot)	71.94	53.89	33.49%	70.11	43.07	62.78%	78.92
	CEval (test)	69.67	52.49	32.73%	66.91	32.21	107.73%	77.82
语言 & 知识	TriviaQA (1-shot)	73.08	63.81	14.53%	89.80	82.45	8.91%	89.36
	NaturalQuestions (1-shot)	27.89	26.98	3.37%	40.06	31.83	25.86%	48.67
	RACE-High (0-shot)	89.25	80.33	11.10%	89.39	81.62	9.52%	92.68
常识推理	WinoGrande (5-shot)	80.66	58.80	37.18%	69.69	53.12	31.19%	93.61
	HellaSwag (10-shot)	92.72	79.53	16.58%	87.72	74.18	18.25%	97.52
	BigBench-Hard (3-shot)	82.74	70.11	18.01%	80.45	54.94	46.43%	82.98
数学&科学	GSM8K (4-shot, w/o code)	80.52	78.17	3.01%	90.22	60.65	48.76%	92.49
	MATH (0-shot, w/o code)	61.90	28.00	121.07%	47.08	8.34	464.51%	58.30
	GPQA (diamond, w/o code)	40.40	29.60	36.49%	38.89	26.26	48.10%	42.93
代码	HumanEval (0-shot)	74.39	73.17	1.67%	72.56	37.20	95.05%	78.05
	MBPP (3-shot)	78.60	60.20	30.56%	71.60	40.47	76.92%	75.10

*表中所有的数字是基于 OpenCompass 公开测试代码针对模型或 API 实测获得

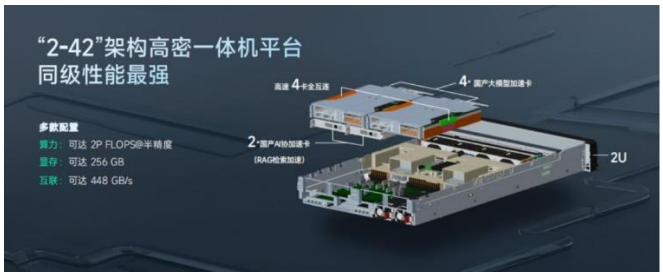
多模态能力比肩 GPT-4V。日日新 5.0 的多模态能力和谷歌 Gemini 类似，采用了原生的多模态技术，而业界包括很多 GPT-4V 在内，很多模型的多模态能力还是基于在语言模型上嫁接其他模态的能力而达到的。日日新 5.0 的多模态图文感知能力达到了全球领先水平，在应用产品层面也实现了更卓越的多模态能力，支持高清长图的解析和理解以及文生图交互式生成，还可以实现复杂的跨文档知识抽取及总结问答展示，还具备丰富的多模态交互能力。

图 2：日日新 5.0 文生图对比业内主流文生图模型

率先完成“云端边”全栈布局，端侧模型位列行业首位，边侧推出企业级应用一体机。24 年是端侧大模型应用的元年，为了满足移动终端用户对大模型技术的应用需求，商汤也推出参数量 1.8B 的端侧大模型，同等尺度性能最优，推理速度也实现业内最快，可在中端平台实现 18.3 字/s 的平均生成速度，旗舰平台更是达到了 78.3 字/s。对于边侧场景如金融、代码、医疗、政务等重点行业日益增长的 AI 应用需求，商汤推出了企业级大模型一体机，可同时支持企业级千亿模型加速和知识检索硬件加速，实现本地化部署，即买即用，降低企业应用大模型的门槛。相比行业同类产品，推理成本节约 80%，检索大大加速，CPU 工作负载 50%。

图 3：日日新端侧大模型推理速度业内最快

图 4：商汤推出企业级应用一体机



Vidu：中国首个高动态性视频大模型

4月27日，生数科技联合清华大学发布中国首个长时长、高一致性、高动态性的视频大模型Vidu。Vidu是自Sora发布之后，全球率先取得重大突破的视频大模型。该模型采用团队自创架构U-ViT，是全球首个Diffusion与Transformer融合的架构。Vidu快速的技术突破源自于团队在贝叶斯机器学习和多模态大模型的长期积累，完全为自主研发。在中关村论坛上，生数科技正式推出“Vidu大模型合作伙伴计划”，诚邀产业链上下游企业，共建合作生态，进一步拓展多模态通用能力的边界。

图5：生数科技联合清华大学正式发布视频大模型Vidu

Vidu在视频生成领域展示出了卓越的性能，性能全面对标国际顶尖水平。Vidu支持根据文本描述一键生成长达16秒且分辨率高达1080P的高清视频内容。它不仅能够模拟真实物理世界，创作出细节丰富且符合实际物理规律的场景，还具备丰富的创意，能够生成一些深邃、非现实的虚构画面，如画作中的船只正迎着海浪驶向镜头。此外，Vidu还具备了生成复杂动态镜头的能力，它能够在一段画面里实现远景、近景、中景、特写等不同镜头的转换，并灵活添加长镜头、追焦、转场等效果，极大地丰富了视频的表现力。由于Vidu是基于U-ViT的架构，所以不会涉及任何中间插帧和拼接等步骤，使得视频生成后在观感上非常流畅，从头到尾没有插帧痕迹。根据生数科

技首席科学家朱军在论坛现场展示与介绍，Vidu 还可以生成特有的中国元素视频例如熊猫等中国特有的象征性元素，为视频添加了独特的文化韵味。

图 6：生成视频片段：画作中的船只正迎着海浪驶向镜头

图 7：生成视频片段：具有中国象征性元素的熊猫

DeepSeek-V2：幻方开源第二代 MoE 模型

幻方量化旗下 Deepseek 正式发布二代 MoE 模型。5 月 6 日，DeepSeek-V2 正式发布，这是一款拥有更多参数，更强能力且超低成本的大模型。从公布的性能指标上来看，DeepSeek-V2 拥有着一流的中文综合能力，优于其他开源大模型，且与 GPT-4-Turbo 等闭源大模型处于同一梯队。在英文综合能力上，DeepSeek-V2 也有较好表现，处于国产大模型的第一梯队，同时超越了同属 MoE 模型的 Mixtral 8x22B。DeepSeek-V2 在知识、数学、推理、编程方面也有着较优的综合能力，性能总体表现出色。

图 8：DeepSeek-V2 综合能力

Model	是否开源	中文综合 AlignBench	英文综合 MT-Bench	知识 MMLU	基础算数 GSM8K	数学解题 MATH	逻辑推理 BBH	编程 HumanEval
DeepSeek-V2	✓	7.91	8.97	77.8	92.2	53.9	79.7	81.1
GPT-4-Turbo-1106	✗	8.01	9.32	84.6	93.0	64.1	-	82.2
GPT-4-0613	✗	7.53	8.96	86.4	92.0	52.9	83.1	84.1
GPT-3.5	✗	6.08	8.21	70.0	57.1	34.1	66.6	48.1
Gemini 1.5 Pro	✗	7.33	8.93	81.9	91.7	58.5	84.0	71.9
Claude 3 Opus	✗	7.62	9.00	86.8	95.0	61.0	86.8	84.9
Claude 3 Sonnet	✗	6.70	8.47	79.0	92.3	40.5	82.9	73.0
Claude 3 Haiku	✗	6.42	8.39	75.2	88.9	40.9	73.7	75.9
abab-6.5（MiniMax）	✗	7.97	8.82	79.5	91.7	51.4	82.0	78.0
abab-6.5s（MiniMax）	✗	7.34	8.69	74.6	87.3	42.0	76.8	68.3
ERNIE-4.0（文心一言）	✗	7.89	7.69	-	91.3	52.2	-	72.0
GLM-4（智谱清言）	✗	7.88	8.60	81.5	87.6	47.9	82.3	72.0
Moonshot-v1（月之暗面）	✗	7.22	8.59	-	89.5	44.2	-	82.9
Baichuan 3（百川）	✗	-	8.70	81.7	88.2	49.2	84.5	70.1
Qwen1.5 72B（通义千问）	✓	7.19	8.61	76.2	81.9	40.6	65.9	68.9
LLaMA 3 70B	✓	7.42	8.95	80.3	93.2	48.5	80.1	76.2
Mixtral 8x22B	✓	6.49	8.66	77.8	87.9	49.8	78.4	75.0

图 9：DeepSeek-V2 API 价格

Model	API价格/百万Tokens 输入（元）	输出（元）
DeepSeek-V2	1	2
GPT-4-Turbo-1106	72	217
GPT-4-0613	217	434
GPT-3.5	11	14
Gemini 1.5 Pro	51	152
Claude 3 Opus	109	543
Claude 3 Sonnet	22	109
Claude 3 Haiku	2	9
abab-6.5（MiniMax）	30	30
abab-6.5s（MiniMax）	10	10
ERNIE-4.0（文心一言）	120	120
GLM-4（智谱清言）	100	100
Moonshot-v1（月之暗面）	24	24
Qwen1.5 72B（通义千问）	20	20
LLaMA 3 70B	27	82
Mixtral 8x22B	14	43

DeepSeek-V2 全新的模型结构带来超高性价比。DeepSeek-V2 并没有沿用主流的 Dense 与 Sprase 结构，而是提出全新的 MLA（多头隐式注意力）架构，可以大幅降低计算量与推理显存，并自主研发了 Sprase 结构，进一步降低计算量。官方估计，DeepSeek-V2 以 236B 总参数、21B 的激活参数，大致能够达到 70B-110B Dense 模型的能力，消耗的显存只有同级别 Dense 模型的 1/5~1/100，这意味着每 token 成本的大幅降低。DeepSeek-V2 API 的定价为：每百万 tokens 输入 1 元、输出 2 元（支持 32K 上下文），在拥有高性能的条件下，付费价格却几乎低于市面上所有其他大模型，拥有极高性价比。

图 10：主流模型的中文能力-API 价格图示

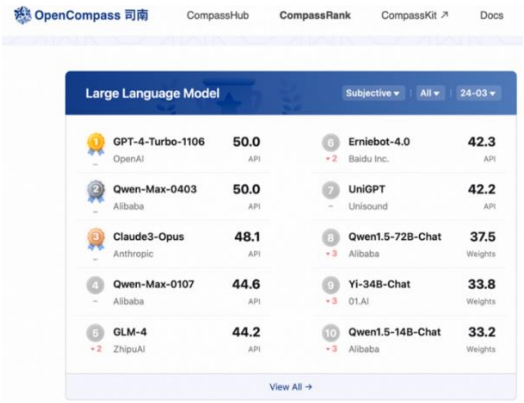
DeepSeek 表示，采用 8xH800 GPU 的单节点峰值吞吐量可达到每秒输出超 5 万个 token。如果仅按输出 token 的 API 的报价计算，每个节点每小时的收费 50.4 美元。假设利用率完全充分，按照一个 8xH800 节点的成本为每小时 15 美元来计算，DeepSeek 每台服务器每小时的收益可达 35.4 美元，甚至能实现 70%以上的毛利率。即使是在这样的低定价下，DeepSeek 仍然保持足够的盈利空间。

通义千问 2.5：行业落地不断推进

5 月 9 日，阿里云正式发布通义千问 2.5，且在中文处理能力上行业领先。相比较于前代产品通义千问 2.1 版本，通义千问 2.5 新模型在多个方面都取得了长足进步，能够更高效、处理信息，例如理解能力提升了 9%，逻辑推理能力提升 16%，指令遵循提升 19%，代码能力提升 10%。同时对比国外巨头 GPT-4，通义千问 2.5 在中文语境下，文本理解、文本生成、知识问答及生活建议、闲聊及对话、安全风险等多项能力赶超 GPT-4。通义千问 2.5 在权威基准 OpenCompass 测评中，得分追平 GPT-4-Turbo，表现出色。

图 11：阿里云峰会发布通义千问 2.5，能力全面升级

图 12：通义千问 2.5 得分追平 GPT-4-Turbo



阿里云持续推进大模型的应用落地。在阿里云 AI 峰会现场，公司宣布小米旗下 AI 助手“小爱同学”已与阿里云通义大模型达成合作，强化多方面的多模态 AI 生成能力如图片生成，且会在各类设备落地，这将提升小米用户的智能化体验。通过合作，“小爱同学”将能够不断学习和优化自身的算法和模型，以提供更加智能化、个性化的服务。阿里云数据显示，通义大模型目前已服务超 9 万企业客户，并在各行各业实现应用落地，例如为完美世界提供云+AI 助力游戏开发工具全面优化，提升创作效率与玩家体验。

图 13：通义大模型成为最受中国企业欢迎的大模型

国产模型对 GPT-4 的追赶不停，目前已能和 GPT-4 Turbo 比拼。自 GPT-4 在 2023 年 3 月发布以来，国产大模型的对 OpenAI 的追赶一直没有停歇。从一开始仅能和 GPT-3.5 比较，到了 23 年下半年，GPT-4 就开始出现在各大国产模型厂商的对比名单中，在部分指标上能够达到甚至超越 GPT-4。进入 2024 年后，智谱 GLM-4 和讯飞星火 V3.5 等国产厂商都在发布会上表示，自家模型综合能力上已经逼近 GPT-4 水平。而 4 月以来，日日新 5.0、通义千问 2.5 等模型的发布，则将对比的目标换成了能力更强的 GPT-4 Turbo。

国产模型能力值得重视，有望带动 AI 应用生态成长。随着国产大模型能力的飞速发展，目前头部厂商的模型在各项性能测试中基本已经接近 GPT-4 的水平，甚至在某些特定领域的表现超过了 GPT-4。国产模型的可用性也逐步提升，正在从“可用”迈向“好用”的阶段。同时从成本端来

看，随着训练和推理过程的不断优化以及模型本身的效率提升，目前国产模型的成本已经降低到一个较低的水平，逐步有了使用性价比。国产模型的能力与性价比提升有望推动国内 AI 应用的生态不断成长，24 年以来，我们已经看到国内 2C 端的 AI 应用如 Kimi、秘塔 AI 搜索等多次引起了较高热度，2B 端的金山办公也正式推出了面向企业的商业化版本 WPS 365。我们认为，随着国产大模型的能力边界持续拓展，国内 AI 应用有望在 24 年陆续迎来突破，现阶段国产模型的能力值得我们重视。

投资建议与投资标的

- AI 应用领域，重点关注：金山办公(688111，增持)、虹软科技(688088，未评级)、新致软件(688590，未评级)、彩讯股份(300634，买入)、星环科技-U(688031，未评级)、科大讯飞(002230，买入)等。
- 算力领域，重点关注：海光信息(688041，买入)、浪潮信息(000977，未评级)、中科曙光(603019，买入)、寒武纪-U(688256，未评级)、润泽科技(300442，未评级)、华铁应急(603300，买入)、亚康股份(301085，未评级)等
- 制造业等领域数字化，建议关注中控技术(688777，买入)、柏楚电子(688188，未评级)、宝信软件(600845，未评级)、瑞纳智能(301129，买入)、千方科技(002373，未评级)、通行宝(301339，未评级)等。
- 量子技术与低空经济领域，建议关注国盾量子(688027，未评级)、神州信息(000555，买入)、吉大正元(003029，未评级)、信安世纪(688201，未评级)、莱斯信息(688631，未评级)、中科星图(688568，未评级)。