计算机

证券研究报告
2024 年 03 月 09 日

大模型生态加速突破， 2024 年应用元年有望到来

海外大语言模型进入新一轮大模型技术突破期，国内也纷纷突破 GPT3.5 水平

国内外大语言模型进入新一轮突破期。在海外，OpenAI 推出 GPT-4V，多模态能力加强，Google 推出新一轮大语言模型 Gemini，其中 Ultra 模型在文本处理的基准测试优于 GPT4，在 2 月 15 日，Google 新发布了 1.5 版本相较于 1.0 的性能继续提升；Meta 近期公布正在架构算力研发 LLaMA3 并坚持开源；Anthropic 旗下的 Claude 模型也进化到 3 版本，整体性能略超 GPT-4，继续保持长文本性能和安全性特点突出，大模型 Mixtral 通过 MOE 结构较好的提质增效。同期在国内包括智谱、文心一言、科大讯飞和通义千问等的新版本都达到了赶超 GPT3.5 甚至部分能力接近 GPT4 的水平。

多模态生成大模型层出不穷，SORA 引领新一轮大模型创新浪潮

多模态生成大模型进入技术突破期，OpenAI 发布视频生成模型 SORA,采用 Diffusion Transformer 结构，使用时空 Latent patch 表示视频和图像，或成为模拟现实的基础；近期 Stability.ai 开源 Stable Video Diffusion 模型，Google 发布 VideoPoet，视频生成新技术不断涌现；文生图模型也逐步迭代，Midjourney 推出 V6 版本，图片生成能力更加优异；此外在数字人领域，微软推出 GAIA 大模型，阿里巴巴推出 Animate Anyone，我们认为这为垂类商业场景奠定了技术基础。

应用与算力齐头并进，海外安迪比尔定律持续演绎

除去大模型侧的创新，海外应用端在 GPTs 的带领下诞生了大量应用，2 个月内就有超过 300 万应用创建，OpenAI 还为此引入生态体系；在算力端，海外大厂的硬件投资依然持续，Meta 在 2024 年有望继续扩大 GPU 投资，微软、google 和 Amazon 都预计资本性支出在有望在新的一年继续扩大以支撑 AI 的投入。

考虑到国内外在模型能力和算力支出上的亮眼表现，我们在此推荐 AI 应用与算力板块的机会，建议关注：
应用：（1）办公软件：金山办公、福昕软件、彩讯股份（通信团队覆盖）
（2）多模态：万兴科技、美图公司（与海外团队联合覆盖）、虹软科技、光云科技
（3）B 端应用：用友网络、金蝶国际、致远互联、泛微网络、鼎捷软件、汉得信息
（4）金融、教育、医疗：同花顺、恒生电子、新致软件、科大讯飞、视源股份（与电子组联合覆盖）、润达医疗
基础设施：神州数码、烽火通信、拓维信息、高新发展、海光信息、星环科技、寒武纪、景嘉微（与电子团队联合覆盖）

风险提示：国内大模型效果提升不及预期、国产算力供应不及预期、国内应用场景落地不及预期

投资评级

行业评级	强于大市(维持评级)
上次评级	强于大市

作者

缪欣君 分析师
SAC 执业证书编号：S1110517080003
miaoxinjun@tfzq.com
刘鉴 联系人
liujianb@tfzq.com

行业走势图

资料来源：聚源数据

相关报告

1 《计算机-行业点评:广立微：持续领先的集成电路 EDA 软件和电性测试设备供应商》 2024-03-06
2 《计算机-行业专题研究:Text-to-Video 的 GPT-3 时刻已来：OpenAI 的 SORA 模型引领新技术突破》 2024-02-22
3 《计算机-行业点评:黄仁勋强调主权 AI，重申 AI 信创之华为链+四小龙机遇》 2024-02-19

内容目录

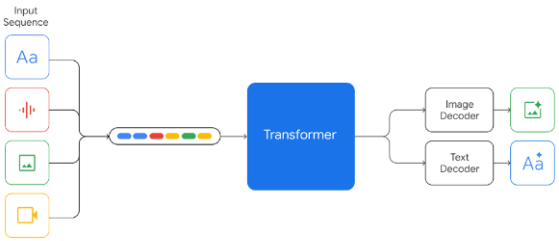
图表目录

1. 海外大模型形成一超多强格局，OpenAI 被加速追赶

1.1. Google Gemini：原生多模态且能力有望追平 GPT-4

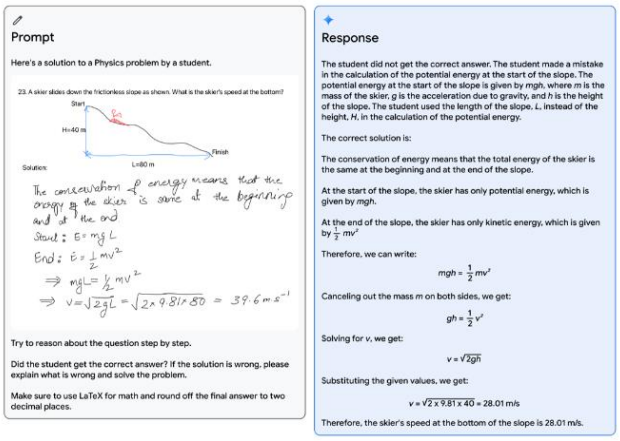
2023 年 12 月 7 日，Google 发布了新一款基于联合训练的原生多模态大模型 Gemini。谷歌所发布的 Gemini 基于文本、图片、语音和视频联合训练，形成了跨模态的强大泛化能力，并在多个测试中有优秀表现。在 Gemini 的模型报告中，Gemini 可以理解文档和手写笔迹，识别学生的推理步骤，并给出详细的解答，生成对应的 Latex 公式。

图 1：Gemini 采用原生多模态的模型结构



资料来源：《Gemini: A Family of Highly Capable Multimodal Models》（作者 Gemini Team, Google），天风证券研究所

图 2：Gemini 可用来修订学生作业



资料来源：《Gemini: A Family of Highly Capable Multimodal Models》（作者 Gemini Team, Google），天风证券研究所

Gemini 模型分为 3 种规模，适用于从大型数据中心到移动设备的各种场景。Ultra 是 Gemini 家族中最强大的模型，可以完成高度复杂的推理和多模态任务。Pro 在 Ultra 基础上进行了优化和平衡，仍然具有较强的推理性能和广泛的多模态能力。Nano 专为设备部署设计，Nano-1、Nano-2 参数量分别为 1.8B 和 3.25B，分别针对不同内存的设备。Nano 通过将模型参数转化为 4 位表示的形式，减小了模型大小，在提高部署效率的同时，维持了优良的性能。

图 3：Gemini 模型有三种等级的参数

Model size	Model description
Ultra	Our most capable model that delivers state-of-the-art performance across a wide range of highly complex tasks, including reasoning and multimodal tasks. It is efficiently serveable at scale on TPU accelerators due to the Gemini architecture.
Pro	A performance-optimized model in terms of cost as well as latency that delivers significant performance across a wide range of tasks. This model exhibits strong reasoning performance and broad multimodal capabilities.
Nano	Our most efficient model, designed to run on-device. We trained two versions of Nano, with 1.8B (Nano-1) and 3.25B (Nano-2) parameters, targeting low and high memory devices respectively. It is trained by distilling from larger Gemini models. It is 4-bit quantized for deployment and provides best-in-class performance.

资料来源：《Gemini: A Family of Highly Capable Multimodal Models》（作者 Gemini Team, Google），天风证券研究所

Gemini Ultra 在文本处理基准测试优于 GPT4，展现出一流的复杂推理能力。根据 Google 的测试，在综合知识（法律、生物和历史等）方面，Gemini Ultra 的 MMLU 指标达到 90.04%，超过人类专家的 89.8%，是首个超越人类的模型（GPT-4 的 MMLU 为 87.29%）。在数学领域，Gemini Ultra 的 GSM8K 指标为 94.4%，以往最高水平为 GPT-4 的 92.0%。在更加复杂的数学领域中，Gemini Ultra 的 MATH 指标（中学竞赛水平）得分为 53.2%，同样高于同期其他模型。在编程和复杂推理方面，编程模型 AlphaCode 2 基于 Gemini Pro，实现了在潜在程序空间大规模搜索。通过定制的过滤、聚类和排序机制，Gemini Pro 可以既生成候选代码，又产生奖励模型——用于选择出最合适的代码。与过去的 AlphaCode 相比，

AlphaCode 2 解决 Codeforces 分部中竞赛问题的数量，从原来的 25%，提高到了 43%，超过了 85%的参赛者。此外，在机器翻译、多语言能力、长文本处理和人类偏好评估方面，Gemini 同样展现了先进的性能。

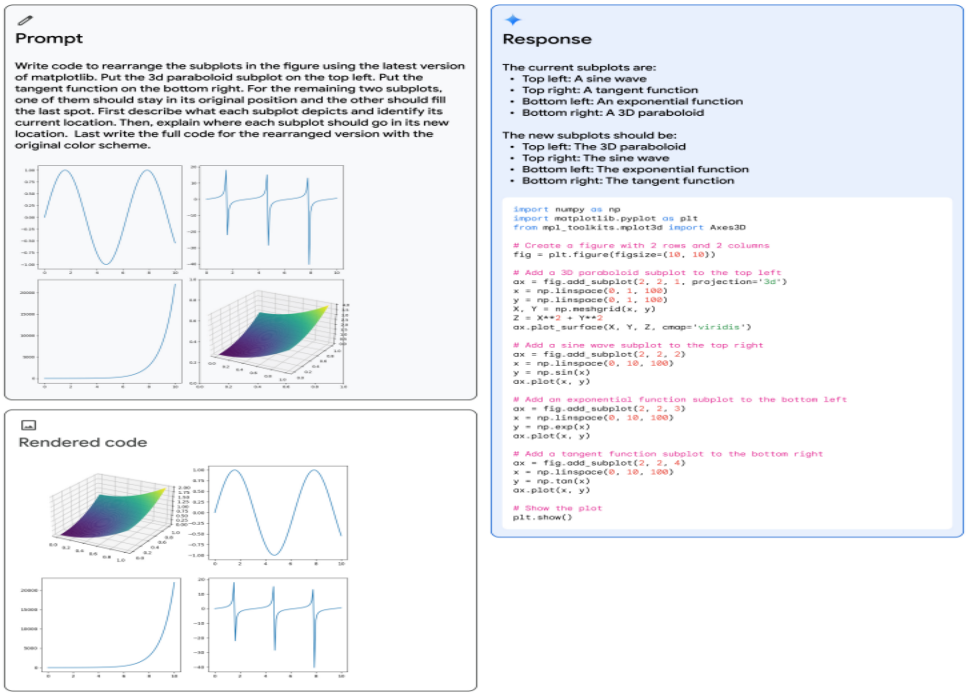
图 4：Gemini 基准测试量化结果

	Gemini Ultra	Gemini Pro	GPT-4	GPT-3.5	PaLM 2-L	Claude 2	Inflect-ion-2	Grok 1	LLAMA-2
MMLU Multiple-choice questions in 57 subjects (professional & academic) (Hendrycks et al., 2021a)	**90.04%** CoT@32* 83.7% 5-shot	79.13% CoT@8* 71.8% 5-shot	87.29% CoT@32 (via API**) 86.4% 5-shot (reported)	70% 5-shot	78.4% 5-shot	78.5% 5-shot CoT	79.6% 5-shot	73.0% 5-shot	68.0%***
GSM8K Grade-school math (Cobbe et al., 2021)	**94.4%** Maj1@32	86.5% Maj1@32	92.0% SFT & 5-shot CoT	57.1% 5-shot	80.0% 5-shot	88.0% 0-shot	81.4% 8-shot	62.9% 8-shot	56.8% 5-shot
MATH Math problems across 5 difficulty levels & 7 subdisciplines (Hendrycks et al., 2021b)	**53.2%** 4-shot	32.6% 4-shot	52.9% 4-shot (via API**) 50.3% (Zheng et al., 2023)	34.1% 4-shot (via API**)	34.4% 4-shot	—	34.8%	23.9% 4-shot	13.5% 4-shot
BIG-Bench-Hard Subset of hard BIG-bench tasks written as CoT problems (Srivastava et al., 2022)	**83.6%** 3-shot	75.0% 3-shot	83.1% 3-shot (via API**)	66.6% 3-shot (via API**)	77.7% 3-shot	—	—	—	51.2% 3-shot
HumanEval Python coding tasks (Chen et al., 2021)	**74.4%** 0-shot (IT)	67.7% 0-shot (IT)	67.0% 0-shot (reported)	48.1% 0-shot	—	70.0% 0-shot	44.5% 0-shot	63.2% 0-shot	29.9% 0-shot
Natural2Code Python code generation. (New held-out set with no leakage on web)	**74.9%** 0-shot	69.6% 0-shot	73.9% 0-shot (via API**)	62.3% 0-shot (via API**)	—	—	—	—	—
DROP Reading comprehension & arithmetic. (metric: F1-score) (Dua et al., 2019)	**82.4** Variable shots	74.1 Variable shots	80.9 3-shot (reported)	64.1 3-shot	82.0 Variable shots	—	—	—	—
HellaSwag (validation set) Common-sense multiple choice questions (Zellers et al., 2019)	87.8% 10-shot	84.7% 10-shot	**95.3%** 10-shot (reported)	85.5% 10-shot	86.8% 10-shot	—	89.0% 10-shot	—	80.0%***
WMT23 Machine translation (metric: BLEURT) (Tom et al., 2023)	**74.4** 1-shot (IT)	71.7 1-shot	73.8 1-shot (via API**)	—	72.7 1-shot	—	—	—	—

资料来源：《Gemini: A Family of Highly Capable Multimodal Models》（作者 Gemini Team, Google），天风证券研究所

作为原生多模态模型，Gemini 展现了较强的多模态能力。Gemini 可以在表格、图片、音频和视频中提取细节信息、空间布局和时间布局，并进行组合输出。例如，Gemini 可生成用于重新排列子画面的 matplotlib 代码，这表明 Gemini 结合了多种能力，如①识别子图；②逆推产生子图的代码；③从非直接的指示中，推理出子图新的排列顺序；④生成新的一组代码，重新生成、排列子图。

图 5：Gemini 模型结合了多种多模态能力



资料来源：《Gemini: A Family of Highly Capable Multimodal Models》（作者 Gemini Team, Google），天风证券研究所

Gemini 1.5 引入 MoE，最新的 1.5 pro 以更少计算量比肩 1.0 Ultra。Gemini 1.5 Pro 是 Gemini 家族的最新模型，于 2024 年 2 月 15 日发布，是一种高效的多模态混合专家模型，能够从数百万个上下文标记中回忆并推理出精细的信息，其中包括多个长文档和几小时的视频和音频。该模型在跨模态的长期上下文检索任务上实现了近乎完美的召回率，并在长文档问答、长视频问答和长上下文自动语音识别方面提高了现有最佳表现。该模型基于谷歌对 Transformer 和 MoE 的最新研究，与之前的版本相比，其性能在多个维度都有显著的改进，1.5 Pro 使用更少的计算实现了与 1.0 Ultra 相当的性能。

图 6：Gemini 1.5 pro 拟人（humaneval）能力对比

Capability	Benchmark	Gemini 1.0 Pro	Gemini 1.0 Ultra	Gemini 1.5 Pro
Math, Science & Reasoning	**Hellaswag** (Zellers et al., 2019)	84.7% 10-shot	87.8% 10-shot	**92.5% 10-shot**
	MMLU: Multiple-choice questions in 57 subjects (professional & academic). (Hendrycks et al., 2021a)	71.8% 5-shot	**83.7% 5-shot**	81.9% 5-shot
	GSM8K: Grade-school math problems. (Cobbe et al., 2021)	77.9% 11-shot	88.9% 11-shot	**91.7% 11-shot**
	MATH: Math problems ranging across 5 levels of difficulty and 7 sub-disciplines. (Hendrycks et al., 2021b)	32.6% 4-shot Minerva prompt	53.2% 4-shot Minerva prompt	58.5% 4-shot Minerva prompt **59.4% 7-shot**
	AMC 2022-23: 250 latest problems including 100 AMC 12, 100 AMC 10, and 50 AMC 8 problems.	22.8% 4-shot	30% 4-shot	**37.2% 4-shot**
	BigBench - Hard: A subset of harder tasks from Big Bench formatted as CoT problems. (Srivastava et al., 2022)	75.0% 3-shot	83.6% 3-shot	**84.0% 3-shot**
	DROP: Reading comprehension & arithmetic. (Metric: F1-Score). (Dua et al., 2019)	74.1% Variable shots	**82.4% Variable shots**	78.9% Variable shots
Coding	**HumanEval** chat preamble* (Metric: pass rate). (Chen et al., 2021)	67.7% 0-shot (PT)	**74.4% 0-shot (PT)**	71.9% 0-shot
	Natural2Code chat preamble* (Metric: pass rate).	69.6% 0-shot	74.9% 0-shot	**77.7% 0-shot**
Multilinguality	**WMT23:** sentence-level machine translation (Metric: BLEURT). (Tom et al., 2023)	71.73 (PT) 1-shot	74.41 (PT) 1-shot	**75.20 1-shot**
	MGSM: multilingual math reasoning. (Shi et al., 2023b)	63.45% 8-shot (PT)	78.95% 8-shot (PT)	**88.73% 8-shot**

资料来源：《Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context》（作者 Gemini Team, Google），天风证券研究所

图 7：Gemini 1.5 pro 多模态能力对比

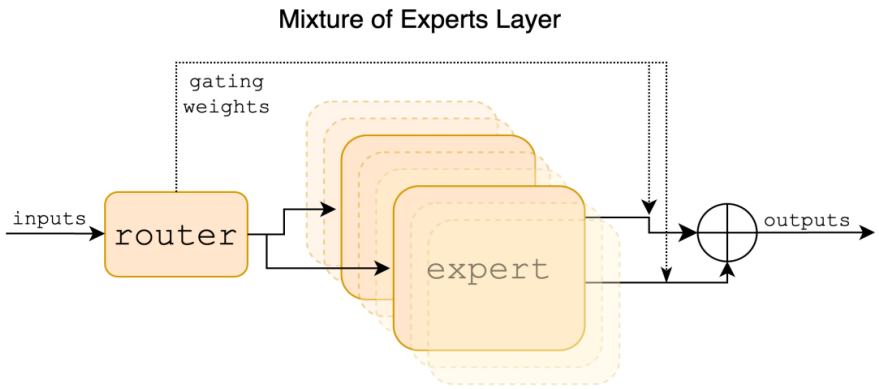
Capability	Benchmark	Gemini 1.0 Pro	Gemini 1.0 Ultra	Gemini 1.5 Pro
Image Understanding	**MMMU (val)** Multi-discipline college-level problems 0-shot (Yue et al., 2023)	47.9%	**59.4%**	58.5%
	Ai2D (test) Science diagrams 0-shot (Kembhavi et al., 2016)	73.9%	79.5%	**80.3%**
	MathViSTA (testmini) Mathematical reasoning 0-shot (Lu et al., 2023)	45.2%	**53.0%**	52.1%
	ChartQA (test) Chart understanding 0-shot (Masry et al., 2022)	74.1%	80.8%	**81.3%**
	VQAv2 (test-dev) Natural image understanding 0-shot (Goyal et al., 2017)	71.2%	**77.8%**	73.2%
	TextVQA (val) Text reading on natural images 0-shot (Singh et al., 2019)	74.6%	**82.3%**	73.5%
	DocVQA (test) Document understanding 0-shot (Mathew et al., 2021)	88.1%	**90.9%**	86.5%
	InfographicVQA (test) Infographic understanding 0-shot (Mathew et al., 2022)	75.2%	**80.3%**	72.7%
Video understanding	**VATEX (test)** English video captioning 4-shot (Wang et al., 2019)	57.4	62.7	**63.0**
	VATEX ZH (val) Chinese video captioning 4-shot (Wang et al., 2019)	39.7	50.8	**54.9**
	YouCook2 (val) English video captioning 4-shot (Zhou et al., 2018)	123.2	**135.4**	134.2
	ActivityNet-QA (test) Video question answering 0-shot (Yu et al., 2019)	49.8%	52.2%	**56.7%**
	EgoSchema (test) Video question answering 0-shot (Mangalam et al., 2023)	55.7%	61.5%	**63.2%**

资料来源：《Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context》（作者 Gemini Team, Google），天风证券研究所

1.2. 大模型 Mixtral 通过专家混合结构提质增效

MoE 方法每次只取用部分参数，同处理规模下推理较快。为了提升模型质量，研究人员不断增大参数规模，大模型的训练难度和推理成本也随之增大。为了实现大模型的高效训练和推理，人们提出了多种方法，包括 mamba 架构、URIAL 方法和 MoE 方法。MoE 在面对多个领域的复杂问题时，先分析任务，将其分发给多个领域的专家，再汇总结论。由于 MoE 结构处理单个 token 的时候只取用部分参数，在保持同等处理规模的情况下，实现了较快的推理速度。

图 8：混合专家层结构



资料来源：《Mixtral of Experts》（作者 Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux 等），天风证券研究所

Mixtral 采用了多个 70 亿参数量的 MoE 组合，结果跑分优于多个主流模型。2023 年 12 月 11 日，Mistral 发布 Mixtral 8x7B，该模型包括了 8 个“专家”模块，分别来自 ArXiv、Github、PhilPapers、StackExchange、DM Mathematics、Gutenberg、PubMed Abstracts 和 Wikipedia (en)。Mixtral 8x7B 的参数量仅为 470 亿。在综合基准测试中，Mixtral 优于或等于 LLaMA 2 70B 和 GPT-3.5。特别地，在数学、代码生成和多语言基准测试中，Mixtral 显著优于 LLaMA 2 70B。Mixtral 针对指令微调的模型——Mixtral 8x7B-Instruct 在人类基准测试中超过了 GPT-3.5Turbo、Claude-2.1、Gemini Pro 和 LLaMA2 70B 聊天模型。Mixtral 8x7B 在多语言和长文本领域也有较好表现。

图 9：Mixtral 8*7B 标化测试结果

	LLaMA 2 70B	GPT-3.5	Mixtral 8x7B
MMLU (MCQ in 57 subjects)	69.9%	70.0%	**70.6%**
HellaSwag (10-shot)	**87.1%**	85.5%	86.7%
ARC Challenge (25-shot)	85.1%	85.2%	**85.8%**
WinoGrande (5-shot)	**83.2%**	81.6%	81.2%
MBPP (pass@1)	49.8%	52.2%	**60.7%**
GSM-8K (5-shot)	53.6%	57.1%	**58.4%**
MT Bench (for Instruct Models)	6.86	**8.32**	8.30

资料来源：《Mixtral of Experts》（作者 Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux 等），天风证券研究所

图 10：LMSys 排行榜（2023 年 12 月 22 日）

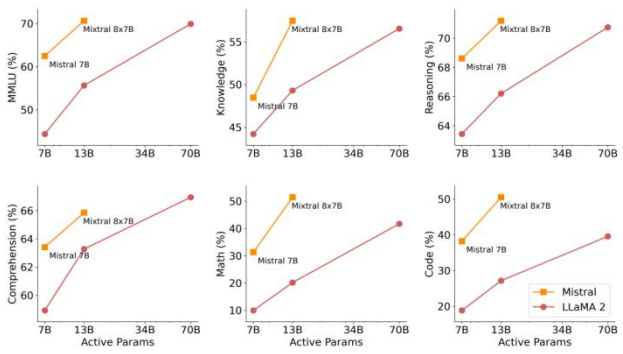
Model	Arena Elo rating	MT-bench (score)	License
GPT-4-Turbo	1243	9.32	Proprietary
GPT-4-0314	1192	8.96	Proprietary
GPT-4-0613	1158	9.18	Proprietary
Claude-1	1149	7.9	Proprietary
Claude-2.0	1131	8.06	Proprietary
Mixtral-8x7b-Instruct-v0.1	1121	8.3	Apache 2.0
Claude-2.1	1117	8.18	Proprietary
GPT-3.5-Turbo-0613	1117	8.39	Proprietary
Gemini Pro	1111		Proprietary
Claude-Instant-1	1110	7.85	Proprietary
Tulu-2-DPO-70B	1110	7.89	AI2 ImpACT Low-risk
Yi-34B-Chat	1110		Yi License
GPT-3.5-Turbo-0314	1105	7.94	Proprietary
Llama-2-70b-chat	1077	6.86	Llama 2 Community

资料来源：《Mixtral of Experts》（作者 Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux 等），天风证券研究所

Mixtral 8x7B 模型的推理计算与存储成本显著缩小，SMoEs 更适用于并行运算。Mixtral 8x7B 仅使用 130 亿活动参数，多项跑分高于 700 亿活动参数的 LLaMA 2 70B。在不考虑内存成本和硬件利用率的情况下，活动参数越多，推理计算成本越大。因此 Mixtral 8x7B 在保持性能的同时，有效降低了推理成本。此外，在与存储成本相关的稀疏参数量方面，Mixtral 8x7B 仅有 470 亿参数量，显著小于 LLaMA 2 70B 的参数量。设备利用率方面，在单时间步中，SMoEs 层的路由机制引入了额外的运算量，用于在单个设备上加载数个“专家”模块。因此，使用 SMoEs 层更适用于并行运算，批量地处理 token，可以提高设备利用率。

因此我们认为，在供给侧，Mixtral 8x7B 所采用的方法以其小型化的特点，为边缘、终端部署大模型有力赋能；在需求侧，Mixtral SMoEs 会对硬件的并行运算能力提出新的要求。

图 11：Mixtral 8x7B、LLaMA2 测试结果对比



资料来源：《Mixtral of Experts》（作者 Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux 等），天风证券研究所

图 12：Mixtral 8x7B 与 LLaMA2、GPT-3.5 对比

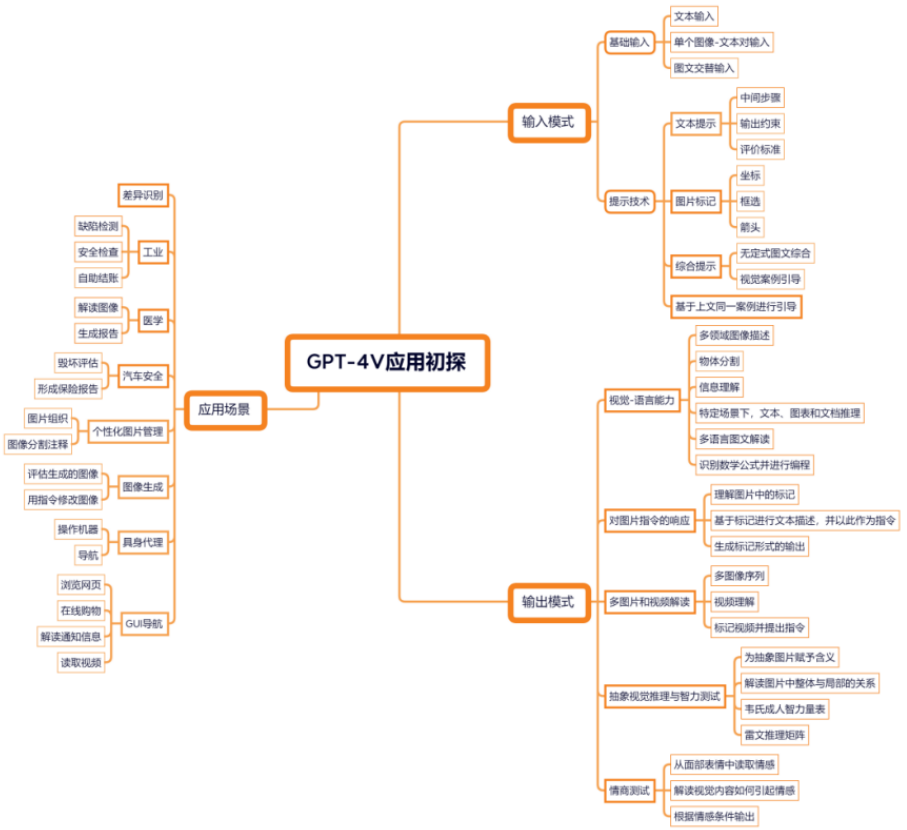
	LLaMA 2 70B	GPT-3.5	Mixtral 8x7B
MMLU (MCQ in 57 subjects)	69.9%	70.0%	**70.6%**
HellaSwag (10-shot)	**87.1%**	85.5%	86.7%
ARC Challenge (25-shot)	85.1%	85.2%	**85.8%**
WinoGrande (5-shot)	**83.2%**	81.6%	81.2%
MBPP (pass@1)	49.8%	52.2%	**60.7%**
GSM-8K (5-shot)	53.6%	57.1%	**58.4%**
MT Bench (for Instruct Models)	6.86	**8.32**	8.30

资料来源：《Mixtral of Experts》（作者 Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux 等），天风证券研究所

1.3. OpenAI 推出 GPT-4V 并持续保持领先

OpenAI 在 2023 年 9 月发布的 GPT-4V 在处理交织的多模态互动方面体现了通用性和强大的处理能力。在输入模式方面，GPT-4V 具备图片标记互动、识别无定式图文输入和接受案例引导的能力。在输出模式方面，强大的多模态处理能力使得 GPT-4V 可以完成事件划分、视频解读和情感解读任务。基于丰富的功能，GPT-4V 衍生出了医学图像解读、具身代理和 GUI 导航等场景应用。还有许多潜在的功能等待使用者发掘。

图 13：GPT-4V 输入、输出模式和应用场景



资料来源：《The Dawn of LMMs: Preliminary Explorations with GPT-4V(ision)》（作者 Zhengyuan Yang, Linjie Li, Kevin Lin 等），天风证券研究所

在工业具身代理、GUI 导航和多物体识别等领域，GPT-4V 测试表现优于 Gemini Pro。在基础的图像识别任务中，GPT-4V 和 Gemini Pro 均表现良好，在复杂的公式和表格信息处理方面存在差异。在图像推理和情绪理解方面，2 个模型都展现了理解多种情绪的能力。在 IQ 测试和多物体识别中，GPT-4V 略强，但 Gemini Pro 在单物体识别方面表现更好，且单图像+文本识别能力更强。在工业应用领域，尤其是包含了具身智能代理和 GUI 导航方面，GPT-4V 更具优势。GPT-4V 和 Gemini Pro 均为能力优秀的多模态大模型，GPT-4V 在某些领域略优于 Gemini Pro。

表 1：GPT-4V 与 Gemini 对比

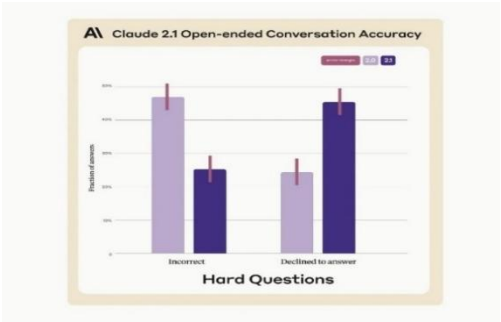
	GPT-4V	Gemini
基本图像识别	较好	较好
IQ 测试	略强	
多物体识别	略强	
单物体识别		在某些方面表现更好
图像-文本理解		更强
带有单图像的文本推理	同水平	同水平
具身代理	更强	
GUI 导航	更强	

资料来源：《Gemini vs GPT-4V : A Preliminary Comparison and Combination of Vision-Language Models Through Qualitative Cases》（作者 Zhangyang Qi, Ye Fang 等），天风证券研究所

1.4. Claude3 震撼发布，能力略超 GPT-4

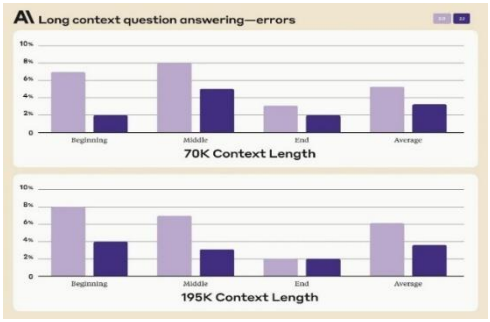
Claude 2.1 显著减少幻觉率，处理更长文本可靠性更高。2023 年 11 月 21 日，Anthropic 推出 Claude 2.1。在开放式对话和复杂 QA 方面，Claude2.1 幻觉率（Hallucination Rates）减少近 50%，为企业提供了更高的可靠性。与同期模型相比，Claude2.1 在“诚实度（Honesty）”测试中，可以输出不确定性结果（例如，“我不确定玻利维亚人口第五大城市是什么”），而非输出一个错误结果（例如，“玻利维亚人口第五大城市是蒙特罗”）。在文本处理方面，Claude2.1 将 200k 长文本处理能力产品化，为业界首创。在处理可靠性要求高的文本（法律报告、财务报告和技术规范）时，Claude2.1 错误答案减少了 30%，错误引用率减少为原来的 25%-33%。

图 14：Claude2.1 开放式 Q&A 精度提升



资料来源：Anthropic 官网，天风证券研究所

图 15：Claude2.1 减少长文本错误率



资料来源：Anthropic 官网，天风证券研究所

Claude3 正式发布效果追平甚至超过 GPT4，成本梯度明显。Claude3 在 3 月 4 日发布，目前有 3 个型号 Haiku, Sonnet 和 Opus，其中 Opus 能力最强但成本最高，Haiku 能力最弱但最有性价比，Opus 和 Sonnet 现在可以在 Claude.ai 上使用，效果上包括以下几点特征：（1）在 MMLU、GPQA、基础数学 GSM8K 等测试机上，Opus 在表现出接近人类的理解力和流畅性，效果基本追平或者超过 GPT-4；（2）响应速度快，Haiku 最快，Sonnet 比上一代 2.1 块 2 倍，Opus 与上一代速度相似但水平更高；（3）视觉能力强，在视觉理解上可以理解照片、图表、图形等，效果追平 GPT-4V 和 Gemini1.0；（4）拒绝率下降，拒绝回答的概率相较于上一代下降一倍；（5）幻觉下降，准确度更高，开放问答的错误率更低；（6）长文本更长，Claude3 支持正常 200k 长文本，客户有需求可以接受最多 100

万 token 的长文本，同时‘大海捞针’的召回能力更强。价格上，Opus 的价格目前每百万 tokens 输入 15 美元，输出 75 美元，Sonnet 为百万 tokens 输入 3 美元，输出 15 美元，Haiku 为百万 tokens 输入 0.25 美元，输出 1.25 美元。我们认为 Claude3 发布接近基本超越了 GPT4。

图 16：Claude3 模型的部分测试结果超过 GPT-4

	Claude 3 Opus	Claude 3 Sonnet	Claude 3 Haiku	GPT-4	GPT-3.5	Gemini 1.0 Ultra	Gemini 1.0 Pro
Undergraduate level knowledge *MMLU*	86.8% 5 shot	79.0% 5-shot	75.2% 5-shot	86.4% 5-shot	70.0% 5-shot	83.7% 5-shot	71.8% 5-shot
Graduate level reasoning *GPQA, Diamond*	50.4% 0-shot CoT	40.4% 0-shot CoT	33.3% 0-shot CoT	35.7% 0-shot CoT	28.1% 0-shot CoT	–	–
Grade school math *GSM8K*	95.0% 0-shot CoT	92.3% 0-shot CoT	88.9% 0-shot CoT	92.0% 5-shot CoT	57.1% 5-shot	94.4% Maj1@32	86.5% Maj1@32
Math problem-solving *MATH*	60.1% 0-shot CoT	43.1% 0-shot CoT	38.9% 0-shot CoT	52.9% 4-shot	34.1% 4-shot	53.2% 4-shot	32.6% 4-shot
Multilingual math *MGSM*	90.7% 0-shot	83.5% 0-shot	75.1% 0-shot	74.5% 8-shot	–	79.0% 8-shot	63.5% 8-shot
Code *HumanEval*	84.9% 0-shot	73.0% 0-shot	75.9% 0-shot	67.0% 0-shot	48.1% 0-shot	74.4% 0-shot	67.7% 0-shot
Reasoning over text *DROP, F1 score*	83.1 3-shot	78.9 3-shot	78.4 3-shot	80.9 3-shot	64.1 3-shot	82.4 Variable shots	74.1 Variable shots
Mixed evaluations *BIG-Bench-Hard*	86.8% 3-shot CoT	82.9% 3-shot CoT	73.7% 3-shot CoT	83.1% 3-shot CoT	66.6% 3-shot CoT	83.6% 3-shot CoT	75.0% 3-shot CoT
Knowledge Q&A *ARC-Challenge*	96.4% 25-shot	93.2% 25-shot	89.2% 25-shot	96.3% 25-shot	85.2% 25-shot	–	–
Common Knowledge *HellaSwag*	95.4% 10-shot	89.0% 10-shot	85.9% 10-shot	95.3% 10-shot	85.5% 10-shot	87.8% 10-shot	84.7% 10-shot

资料来源：Anthropic 官网，天风证券研究所

1.5. Meta 持续打造开源生态，加购算力研发 LLaMA3

2023 年 7 月 19 日发布的 LLaMA 2 最多有 700 亿参数量，标化测试水平略高于 PaLM。2023 年 7 月 19 日，Meta 发布免费可商用模型 LLaMA 2，最大参数量为 700 亿，可处理内容长度为第一代的 2 倍。与 GPT-3.5、GPT-4、PaLM 和 PaLM-2-L 相比，LLaMA 2 70B 的 MMLU 最低，GSM8K 较低；TriviaQA 和 Natural Questions 水平介于 PaLM 和 PaLM-2-L 之间；HumanEval 和 BIG-Bench Hard 水平与 PaLM 相近，低于其他模型。

图 17：LLaMA1 与 LLaMA2 模型家族的参数和性能等

	Training Data	Params	Context Length	GQA	Tokens	LR
LLaMA 1	*See Touvron et al. (2023)*	7B	2k	✗	1.0T	3.0×10^{-4}
		13B	2k	✗	1.0T	3.0×10^{-4}
		33B	2k	✗	1.4T	1.5×10^{-4}
		65B	2k	✗	1.4T	1.5×10^{-4}
LLaMA 2	*A new mix of publicly available online data*	7B	4k	✗	2.0T	3.0×10^{-4}
		13B	4k	✗	2.0T	3.0×10^{-4}
		34B	4k	✓	2.0T	1.5×10^{-4}
		70B	4k	✓	2.0T	1.5×10^{-4}

资料来源：《LLaMA 2: Open Foundation and Fine-Tuned Chat Models》（作者 HugoTouvron，Louis Martin, Kevin Stone†），天风证券研究所

图 18：闭源模型基准测试对比结果

Benchmark (shots)	GPT-3.5	GPT-4	PaLM	PaLM-2-L	LLaMA 2
MMLU (5-shot)	70.0	**86.4**	69.3	78.3	68.9
TriviaQA (1-shot)	–	–	81.4	**86.1**	85.0
Natural Questions (1-shot)	–	–	29.3	**37.5**	33.0
GSM8K (8-shot)	57.1	**92.0**	56.5	80.7	56.8
HumanEval (0-shot)	48.1	**67.0**	26.2	–	29.9
BIG-Bench Hard (3-shot)	–	–	52.3	**65.7**	51.2

资料来源：《LLaMA 2: Open Foundation and Fine-Tuned Chat Models》（作者 HugoTouvron，Louis Martin, Kevin Stone†），天风证券研究所

LLaMA2 投入 330 万 A100 计算小时进行训练，LLaMA3 或累计投入 60 万台 H100，Meta

正加速研发。在 LLaMA2 训练过程中，投入了 330 万 GPU 小时（A100-80GB），用于 2 万亿 token 数据的预训练。微调数据包括公开可用的指令数据集，以及超过 100 万人工标注案例的实例。Meta 指出，预训练和微调数据集均不包括用户数据。

图 19：LLaMA2 软硬件投入

Hardware and Software (Section 2.2)	
Training Factors	We used custom training libraries, Meta's Research Super Cluster, and production clusters for pretraining. Fine-tuning, annotation, and evaluation were also performed on third-party cloud compute.
Carbon Footprint	Pretraining utilized a cumulative 3.3M GPU hours of computation on hardware of type A100-80GB (TDP of 350-400W). Estimated total emissions were 539 tCO_2eq, 100% of which were offset by Meta's sustainability program.
Training Data (Sections 2.1 and 3)	
Overview	LLAMA 2 was pretrained on 2 trillion tokens of data from publicly available sources. The fine-tuning data includes publicly available instruction datasets, as well as over one million new human-annotated examples. Neither the pretraining nor the fine-tuning datasets include Meta user data.
Data Freshness	The pretraining data has a cutoff of September 2022, but some tuning data is more recent, up to July 2023.

资料来源：《LLaMA 2: Open Foundation and Fine-Tuned Chat Models》（作者 HugoTouvron，Louis Martin, Kevin Stone †），天风证券研究所

LLaMA3 有望比肩 GPT-4，最高 1400 亿参数，研究人员计划放松安全限制。2024 年 1 月 18 日扎克伯格透露，Meta 正在训练下一代大模型 LLaMA3，未来一年，将投入 35 万台 H100 GPU 用于构建计算基础设施。近期 The Information 报道，Meta 计划在今年 7 月份发布 Llama 3 大模型，旨在与 OpenAI 的 GPT-4 模型相竞争。参数量方面，LLaMA3 的最大参数量可能超过 1400 亿，对标的 GPT-4 模型参数规模约为 1.8 万亿。Meta 在开发 LLaMA 2 时引入了安全机制，以避免回答可能引发争议的问题，这种过于谨慎的设计，在公司高层和模型研究人员中引发了关于其“过度安全”的担忧。鉴于此，研究人员计划在 LLaMA3 中放宽这些限制，以促进模型与用户之间的更多互动，提供更丰富的背景信息，而不仅仅是回避争议话题。

2. 国内大语言模型能力突破，逐步达到甚至超过 GPT3.5 水平

2.1. 智谱推出 GLM4，能力超过 CodeGeex2-6B

GLM-4 整体性能超过 GPT3.5，部分能力比肩 GPT-4，128k 文本能力测试超越 Claude2.1。2024 年 01 月 16 日，智谱 AI 推出 GLM-4。在基础能力方面，GLM-4 在各项测评中达到了 GPT-4 90%以上水平。长文本方面，基于 LongBench（128k）的总结、信息抽取、复杂推理和代码场景中，GLM-4 测试结果超过 Claude2.1；在“大海捞针”（Needle in the Heystack，将一段信息放在一段长文本中的任意位置，检测大模型的回答准确率）测试中，GLM-4 实现在 128k 文本长度内几乎 100%的精度召回。此外，智谱 AI 在发布 GLM-4 的同时，发布了“定制化的个人 GLM 大模型”GLMs 和 GLM Store，对标 OpenAI 的 GPTs。

图 20：GLM-4 基础能力和中文对齐能力

资料来源：GLM 大模型公众号，天风证券研究所

图 21：智谱 GLM Store

资料来源：新智元公众号，天风证券研究所

2.2. 百度推出文心一言 4.0

基于文心一言 4.0，百度打造应用商店和 4 层架构平台，服务企业。在模型领域，2023 年 10 月 17 日百度发布文心大模型 4.0，个人和企业客户可通过百度智能云千帆大模型平台接入使用，百度可延伸提供企业级一站式客户服务，打通芯片+平台+模型+应用的 4 层架构，实现应用落地。在应用层面，百度 AI 原生应用商店引入 5 大领域应用，涵盖智能办公、营销服务、行业智能、生产提效和分析决策领域，助力企业业务提效和创新发展。

图 22：百度“芯片+平台+模型+应用“4 层架构

资料来源：百度智能云千帆大模型官网，天风证券研究所

图 23：千帆 AI 原生应用商店覆盖 B 端 5 大领域

资料来源：百度智能云千帆大模型官网，天风证券研究所

自研预训练框架 ERNIE 赋能文心大模型，2023 年 10 月 17 日更新至 ERNIE 4.0。百度在 2019 年发布 ERNIE 1.0 版本。实证结果显示，ERNIE 3.0 在 54 项中文 NLP 任务上优于同期最先进的模型，其英文版在 SuperGLUE 基准测试中排名第一（2021 年 7 月 3 日），超过人类表现+0.8%（90.6%对 89.8%）。2023 年 10 月 17 日 ERNIE 更新至 4.0 版本，功能对标 GPT-4，在理解、生成、推理和记忆方面能力有所加强，进一步赋能文心大模型。

图 24：ERNIE 赋能文心产业级知识增强大模型

资料来源：百度智能云千帆大模型官网，天风证券研究所

图 25：ERNIE 效果对比

Model	BoolQ	CB	COPA	MultiRC	ReCoRD	RTE	WiC	WSC	Score
Human Baseline	89.0	95.8/98.9	100	81.8/51.9	91.7/91.3	93.6	80.0	100	89.8
T5+Menna	**91.4**	95.8/97.6	98.0	88.3/63.0	94.2/93.5	93.0	**77.9**	96.6	90.4
DeBERTa	90.4	95.7/97.6	**98.4**	88.2/**63.7**	94.5/94.1	**93.2**	77.5	95.9	90.3
ERNIE 3.0	91.0	**98.6/99.2**	97.4	**88.6**/63.2	**94.7/94.2**	92.6	77.4	**97.3**	**90.6**

资料来源：《ERNIE 3.0: LARGE-SCALE KNOWLEDGE ENHANCED PRE-TRAINING FOR LANGUAGE UNDERSTANDING AND GENERATION》（作者 Yu Sun, Shuohuan Wang 等），天风证券研究所

2.3. 讯飞正式发布星火大模型 3.5，能力比肩 GPT-4

基于全国产化算力平台“飞星一号“，星火大模型 V3.5 实现七大能力提升。2023 年 10 月 24 日，讯飞联合昇腾生态共同发布“飞星一号”大模型算力平台，并启动对标 GPT-4 的更大参数规模的星火大模型训练，自主生态平台为世界提供了新的选择。星火 V3.5 于 2024 年 1 月 30 日发布，在文本生成、语言理解和知识问答等七大能力有所提升。在语言理解、数学能力方面超越 GPT-4 Turbo，在代码能力和多模态能力方面逼近 GPT-4 Turbo、GPT-4V。此外，讯飞计划在 2024 年上半年发布 4.0 版本。

图 26：用于训练新一代星火大模型的“飞星一号”平台

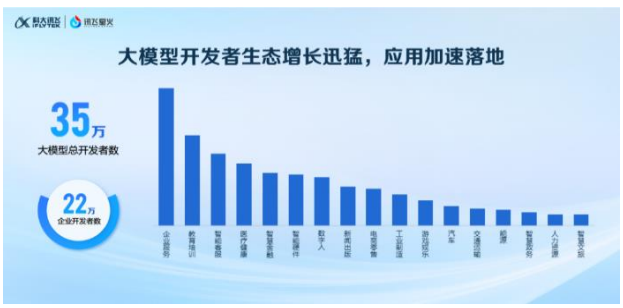
资料来源：科大讯飞公众号，天风证券研究所

图 27：讯飞星火 V3.5 七大能力提升

资料来源：讯飞开放平台公众号，天风证券研究所

讯飞 AI 应用生态增长迅猛，发布星火开源-13B 进一步赋能开发者。自星火大模型发布以来，大模型开发者总数超 35 万，其中企业开发者 22 万，遍布企业服务、教育培训、智能客服、医疗健康等领域。此外，科大讯飞还首次开源了深度适配国产算力，拥有 130 亿参数的 iFlytekSpark-13B 模型（星火开源-13B）。不仅场景应用效果领先，而且还对学术和企业研究完全免费。具体来说，此次开源不仅包括基础模型 iFlytekSpark-13B-base、精调模型 iFlytekSpark-13B-chat，还有微调工具 iFlytekSpark-13B-Lora，以及人设定制工具 iFlytekSpark-13B-Charater。基于这些全栈自主创新的套件，企业和机构可以方便地训练自己的大模型。

图 28：讯飞大模型总开发者总数超 35 万



资料来源：讯飞开放平台公众号，天风证券研究所

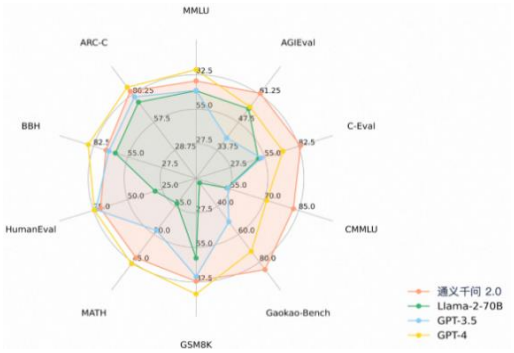
图 29：星火开源-13B 上线

资料来源：新智元公众号，天风证券研究所

2.4. 通义千问推出 2.0 版本，能力赶超 GPT 3.5

通义千问能力评测超越 GPT-3.5，8 大行业应用组团上线。2023 年 10 月 31 日，阿里云正式发布千亿级参数大模型通义千问 2.0。在 10 项权威测评中，通义千问 2.0 综合性能超过 GPT-3.5，正在加速追赶 GPT-4。在 MMLU、C-Eval、4.GSM8K、HumanEval、MATH 等 10 个主流 Benchmark 测评集上，通义千问 2.0 的得分整体超越 Meta 的 LLaMA-2-70B，相比 OpenAI 的 Chat-3.5 是九胜一负，相比 GPT-4 则是四胜六负，与 GPT-4 的差距进一步缩小。与此同时，基于通义大模型训练的 8 大行业模型组团上线，8 大行业模型面向当下最受欢迎的多个垂直场景，使用领域数据进行专门训练。用户可以在官网直接体验模型功能，开发者可以通过网页嵌入、API/SDK 调用等方式，将模型能力集成到自己的大模型应用和服务中。

图 30：通义千问 2.0 主流评测结果



资料来源：阿里云开发者社区，天风证券研究所

图 31：通义大模型训练的 8 大行业模型

资料来源：通义千问官网，天风证券研究所

2.5. Minimax 在国内推出 MoE 模型 abab6

MiniMax 在国内上线 MoE 模型，部分效果超过 Mistral 商用版。2024 年 1 月 16 日，国内首个 MoE 大模型 abab6 上线。与上一代模型相比，其参数量还增大了一个数量级。从测评数据来看，在指令遵从、中文综合能力和英文综合能力上，abab6 大幅超过了 GPT-3.5；和 Claude 2.1 相比，abab6 也在指令遵从、中文综合能力和英文综合能力上略胜一筹；相较于 Mistral 的商用版本 Mistral-Medium，abab6 在指令遵从和中文综合能力上都优于 Mistral-Medium，在英文综合能力上与 Mistral-Medium 旗鼓相当。

图 32：abab6 测评数据

	abab6	abab5.5	Claude	Mistral	GPT-3.5	GPT-4
IFEval	0.67	0.49	0.57	0.56	0.55	0.75
MT-Bench	8.61	6.63	8.18	8.61	8.39	9.32
AlignBench	7.41	5.50	6.62	6.42	6.08	8.01

资料来源：MiniMax 稀宇科技公众号，天风证券研究所

2.6. 百川智能上线 Baichuan3，中文、医疗能力表现优秀

百川智能发布 Baichuan 3，中文评测超越 GPT-4。2024 年 1 月 29 日，百川智能发布超千亿参数的大语言模型 Baichuan 3。在 CMMLU、GAOKAO 等多个中文评测榜单上，Baichuan 3 超越 GPT-4。同时，在多个英文评测中，Baichuan 3 表现出色，达到接近 GPT-4 的水平。

Baichuan 3 的数学和代码能力介于 GPT-3.5 与 GPT-4 之间。此外，在 MT-Bench、IFEval 等对齐榜单的评测中，Baichuan 3 超越了 GPT-3.5、Claude 等大模型，处于行业领先水平。Baichuan 3 在多个权威医疗评测任务中表现优异，不仅 MCMLE、MedExam、CMExam 等中文医疗任务的评测成绩超过 GPT-4，USMLE、MedMCQA 等英文医疗任务的评测成绩也逼近了 GPT-4 的水准。

图 33：Baichuan 3 中英文、数学和代码评测

	数学		代码			
	GSM8K	MATH	HumanEval	MBPP	Crux-I	Crux-O
Baichuan 3	88.17	49.20	**70.12**	**68.20**	57.88	58.38
GPT-4	**92.00**	**52.90**	67.00	63.60	**69.8**	**68.7**
GPT-3.5	57.10	13.96	52.44	61.40	49.0	49.4
Baichuan 3/GPT-4	95.8%	93.0%	104.7%	107.2%	82.9%	85.0%
Baichuan 3/GPT-3.5	154.4%	352.4%	133.7%	111.1%	118.1%	118.2%

资料来源：百川大模型公众号，天风证券研究所

图 34：Baichuan 3 对齐测试和医疗评测结果

	中文			英文		
	MCMLE	MedExam	CMExam	USMLE	MedMCQA	PubMedQA
Baichuan3	**85.76**	**84.64**	**86.9**	**80.2**	**72.08**	**76**
GPT-3.5 (5 shot)	52.92	54.33	47.25	53.57	56.39	71.60
GPT-4 (5 shot)	74.58	75.17	67.31	81.38	72.4	75.2
GPT-4 (Med.prompt)	-	-	-	90.20	79.1	82.0
Baichuan 3/GPT-4	1.15	1.13	1.29	0.99	1.0	1.0

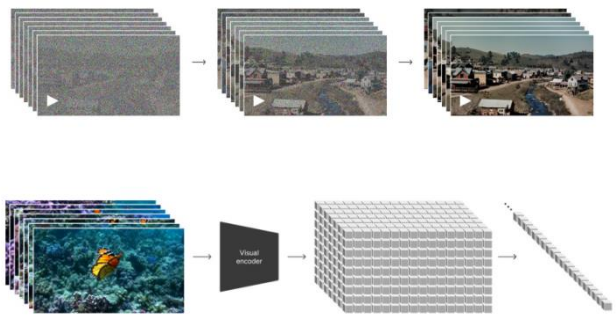
资料来源：百川大模型公众号，天风证券研究所

3. 多模态生成新技术不断突破，正处于技术突破的关键期

3.1. SORA 模型横空出世，视频生成模型的 GPT3 时刻来临

Sora 采用 Diffusion Transformer 结构，使用时空 Latent patch 表示视频和图像，或成为模拟现实的基础。Sora 建立在 DALL·E 和 GPT 的基础上，它采用扩散模型，以类似静态噪声的视频为起点，通过多个步骤去除噪声来逐渐产生视频。此外，Sora 引入了 Transformer 结构，OpenAI 团队用 patch 作为基本单位，把视频和图像表示为 patch 的组合（类似于 GPT 中的 token）。patch 的表示方法扩大了 OpenAI 的数据集，因此 DM+Transformer 的训练可以引入不同持续时间、分辨率和纵横比的数据。OpenAI 认为，Sora 是 AI 理解和模拟真实世界的基础，是 AGI 的重要里程碑。

图 35：Sora 采用 DM+Transformer 结构



资料来源：OpenAI 官网，天风证券研究所

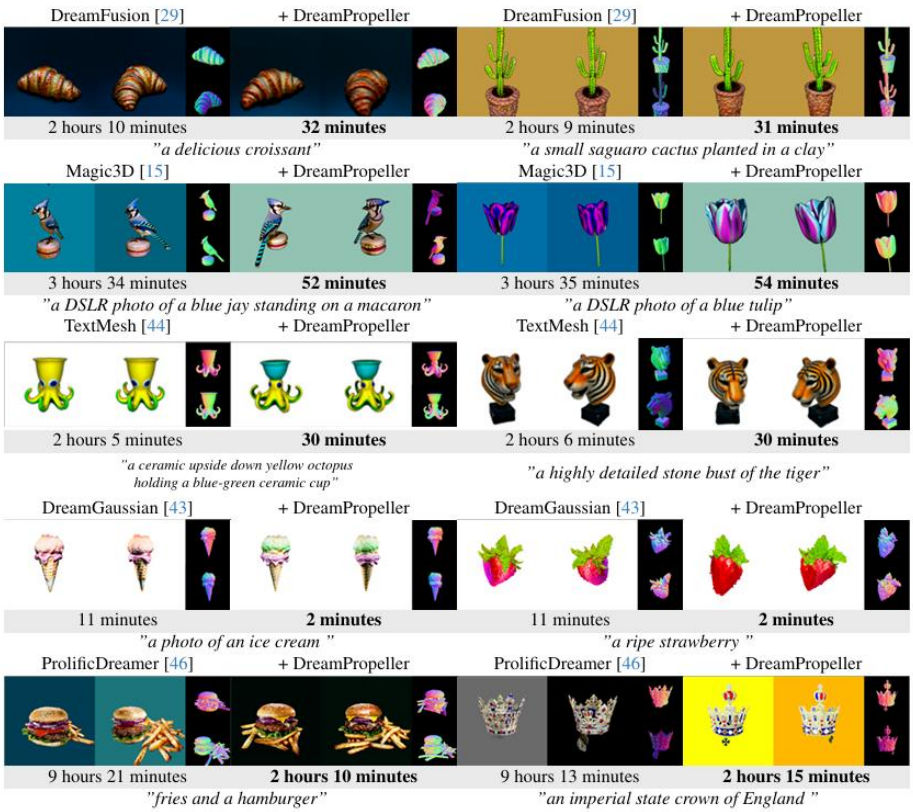
图 36：Sora 视频生成效果

资料来源：OpenAI 官网，天风证券研究所

3.2. pika 推出 1.0，模型效果快速提升

Pika labs 发布 1.0 产品，DreamPropeller 助力视频生成保质提速。2023 年 11 月 29 日，位于美国的初创企业 Pika Labs，对外正式发布了其全新的视频生成与编辑软件——Pika 1.0。该软件具备视频处理能力，可生成并编辑 3D 动画、动漫、卡通以及电影等多种形式的视频内容。值得一提的是，Pika 1.0 的使用门槛极低，用户仅需输入一句话，即可生成多种风格的视频。同时，用户还可以通过简单的描述，对视频中的形象和风格进行个性化调整。Pika 提出的 DreamPropeller 方法，以并行计算换取速度，将该方法用于 DreamGaussian 和 ProlificDreamer 后，在保证生成质量的同时，实现了超过 4 倍的加速。

图 37：DreamPropeller 方法提升视频生成速度效果展示



资料来源：《DreamPropeller: Supercharge Text-to-3D Generation with Parallel Sampling》（作者 Linqi Zhou, Andy Shih 等），天风证券研究所

3.3. Stability.ai 发布并开源 Stable Video Diffusion 模型

2023 年 11 月份上线的 Stable Video Diffusion 模型可用于视频、图片生成，性能优于部分同期模型。2023 年 11 月 21 日，Stability.AI 基于研究目的发布 Stable Video Diffusion（SVD）模型和微调版本 SVD-XT。功能方面，SVD 可用于文本-视频生成、图片-视频生成和多视图合成。性能方面，SVD 的文生视频 FVD 测试得分（类似图像的 FID 指标，越小则越接近真实案例）优于 Make-A-Video 和 MagciVideo 等多个模型；此外，在图生视频方面，SVD 比 Pika 和 GEN-2 更受欢迎。

图 38：SVD 文本-视频生成、图片-视频生成和多视图合成案例

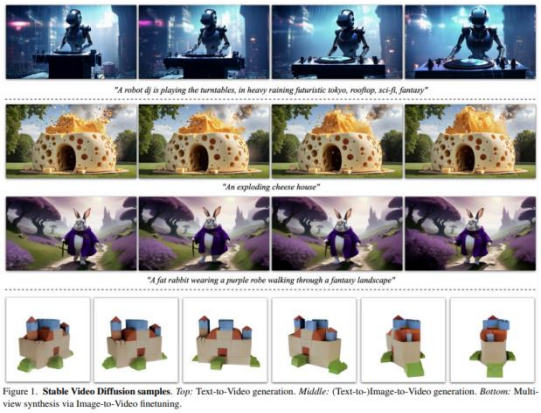


Figure 1. **Stable Video Diffusion samples**. *Top:* Text-to-Video generation. *Middle:* (Text-to-)Image-to-Video generation. *Bottom:* Multi-view synthesis via Image-to-Video finetuning.

资料来源：《Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets》（作者 Andreas Blattmann, Tim Dockhorn, Sumith Kulal 等），天风证券研究所

图 39：SVD 定量比较效果

Table 2. UCF-101 zero-shot text-to-video generation. Comparing our base model to baselines (numbers from literature).

Method	FVD (↓)
CogVideo (ZH) [43]	751.34
CogVideo (EN) [43]	701.59
Make-A-Video [82]	367.23
Video LDM [9]	550.61
MagicVideo [115]	655.00
PYOCO [29]	355.20
SVD (*ours*)	**242.02**

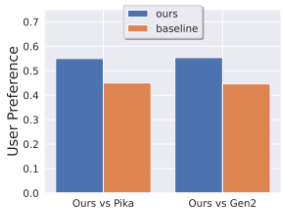


Figure 6. Our 25 frame Image-to-Video model is preferred by human voters over GEN-2 [74] and PikaLabs [54].

资料来源：《Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets》（作者 Andreas Blattmann, Tim Dockhorn, Sumith Kulal 等），天风证券研究所

3.4. Google 发布 VideoPoet，基于 LLM 的技术路径表现出亮眼的视频生成能力

VideoPoet 能力覆盖各种视频任务，效果亮眼。2023 年 12 月 19 日，谷歌发布视频生成大模型 VideoPoet，能够执行各种视频生成任务，包括文本到视频、图像到视频、视频风格化、视频修复和扩展，以及视频转音频。测试结果方面，VideoPoet 在零样本文本到视频基准测试上（MSR-VTT 和 UCF-101）实现了先进的性能。

图 40：VideoPoet 能力概览

资料来源：《VideoPoet: A Large Language Model for Zero-Shot Video Generation》（作者 Dan Kondratyuk, Lijun Yu,等），天风证券研究所

图 41：VideoPoet 测评效果

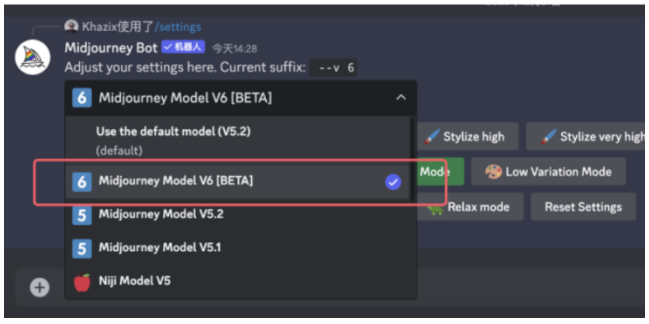
Model	MSR-VTT		UCF-101	
	CLIPSIM	FVD	FVD	IS
CogVideo (EN) [36]	0.2631	1294	702	25.27
MagicVideo [85]	-	998	655	-
Video LDM [5]	0.2929	-	551	33.45
ModelScopeT2V [67]	0.2930	550	-	-
InternVid [70]	0.2951	-	617	21.04
VideoFactory [69]	0.3005	-	410	-
Make-A-Video [56]	0.3049	-	367	33.00
Show-1 [81]	0.3072	538	394	35.42
VideoPoet (Pretrain)	0.3049	**213**	**355**	**38.44**
VideoPoet (Task adapt)	**0.3123**	-	-	-

资料来源：《VideoPoet: A Large Language Model for Zero-Shot Video Generation》（作者 Dan Kondratyuk, Lijun Yu,等），天风证券研究所

3.5. Midjourney 推出 V6，大版本迭代带来更优异的图片生成能力

Midjourney v6 优化 Prompt 模式，生成图片相较于前一代版本更加准确自然。2023 年 12 月 21 日，Midjourney v6 发布 BETA 版。与 v5 相比，v6 的主要变化，就是图像质量更好、语义理解更强、能嵌入英文单词、更容纳更多 token 了。从生成效果来看，v6 的效果更自然，已经达到了电影级别的质量。从光影效果来看，v6 更丰富、真实，有光追效果。

图 42：Midjourney V6 BETA 发布



资料来源：新智元公众号，天风证券研究所

图 43：Midjourney V6 用户分享

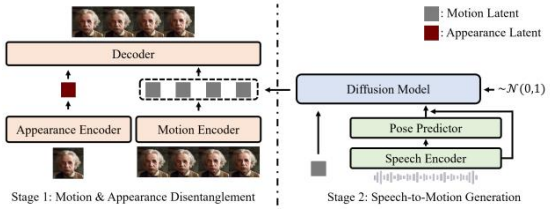
资料来源：新智元公众号，天风证券研究所

3.6. 微软推出针对数字人的大模型 GAIA

GAIA（Generative AI for Avatar）基于语音和单张人像图片，生成说话视频，比 Real Video 等模型效果更自然。2023 年 11 月 26 日，微软发表论文《GAIA: ZERO-SHOT TALKING AVATAR GENERATION》。将语音和人像图片输入 GAIA 可以得到对应的说话视频。原理上，GAIA 由一个变分自编码器（橙色模块）和一个扩散模型（蓝色和绿色模块）组成。效果上，与 Real Video、MakeItTalk、Audio2Head 和 SadTalker 进行定性比较，GAIA 实现

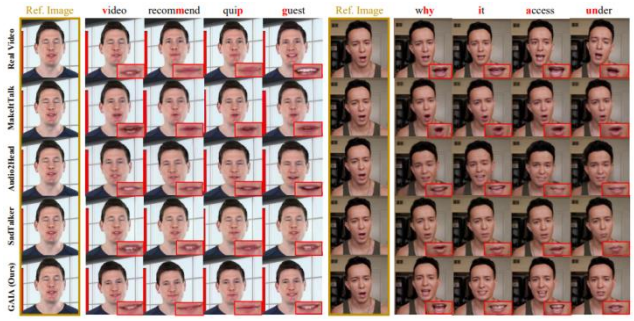
了更高的自然度、唇同步质量、视觉质量和运动多样性。

图 44：GAIA 原理示意图



资料来源:《GAIA: ZERO-SHOT TALKING AVATAR GENERATION》作者 Tianyu He, Junliang Guo, Runyi Yu 等），天风证券研究所

图 45：GAIA 效果定性比较

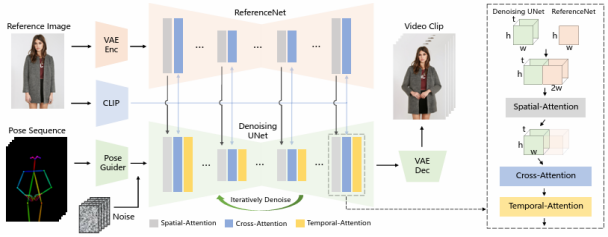


资料来源:《GAIA: ZERO-SHOT TALKING AVATAR GENERATION》作者 Tianyu He, Junliang Guo, Runyi Yu 等），天风证券研究所

3.7. 阿里巴巴推出 Animate Anyone，让图片动的更自然

输入图片与姿势序列，Animate Anyone 模型引导图片产生预置运动。方法上，Animate Anyone 引入 Pose Guider 对特定舞蹈动作的姿势序列进行编码，混入噪声后加入去噪 UNet。参考图片给定后，被 VAE 编码，再通过 ReferenceNet 提取详细特征，用于空间注意力；同时用 CLIP 提取参考图像的语义特征，用于跨时空注意力。时间注意力用于去噪 UNet 在时域上的运行。对去噪 UNet 的输出进行 VAE 解码，产生视频（切面）。从生成效果上看，给定参考图像，Animate Anyone 可以实现连贯可控的角色动画效果，生成的视频结果清晰、稳定，且可以保持与参考图像的外观细节一致。

图 46：Animate Anyone 模型结构



资料来源：《Animate Anyone: Consistent and Controllable Image-to-Video Synthesis for Character Animation》（作者 Li Hu, Xin Gao 等），天风证券研究所

图 47：Animate Anyone 模型效果

资料来源：《Animate Anyone: Consistent and Controllable Image-to-Video Synthesis for Character Animation》（作者 Li Hu, Xin Gao 等），天风证券研究所

在时装和舞蹈合成方面，Animate Anyone 优于同期模型。时装视频合成方面，Animate Anyone 的 SSIM（结构相似性指数）和 PSNR（峰值信噪比）指标高于 MRAA（Motion Representations for Articulated Animation）等模型，其 LPIPS（图像感知相似度）和 FVD 指标低于 MRAA 等模型。在舞蹈合成方面，Animate Anyone 的 SSIM、PSNR、LPIPS 和 FVD 参数优于 FOMM（First Order Motion Model for Image Animation）等模型。

图 48：Animate Anyone 时装视频合成测试结果

	SSIM ↑	PSNR ↑	LPIPS ↓	FVD ↓
MRAA[37]	0.749	-	0.212	253.6
TPSMM[60]	0.746	-	0.213	247.5
BDMM[54]	0.918	24.07	0.048	148.3
DreamPose[19]	0.885	-	0.068	238.7
DreamPose*	0.879	34.75	0.111	279.6
Ours	**0.931**	**38.49**	**0.044**	**81.6**

资料来源：《Animate Anyone: Consistent and Controllable Image-to-Video Synthesis for Character Animation》（作者 Li Hu, Xin Gao 等），天风证券研究所

图 49：Animate Anyone 舞蹈视频合成测试结果

	SSIM ↑	PSNR ↑	LPIPS ↓	FVD ↓
FOMM[35]	0.648	29.01	0.335	405.2
MRAA[37]	0.672	29.39	0.296	284.8
TPSMM[60]	0.673	29.18	0.299	306.1
Disco[45]	0.668	29.03	0.292	292.8
Ours	**0.718**	**29.56**	**0.285**	**171.9**

资料来源：《Animate Anyone: Consistent and Controllable Image-to-Video Synthesis for Character Animation》（作者 Li Hu, Xin Gao 等），天风证券研究所

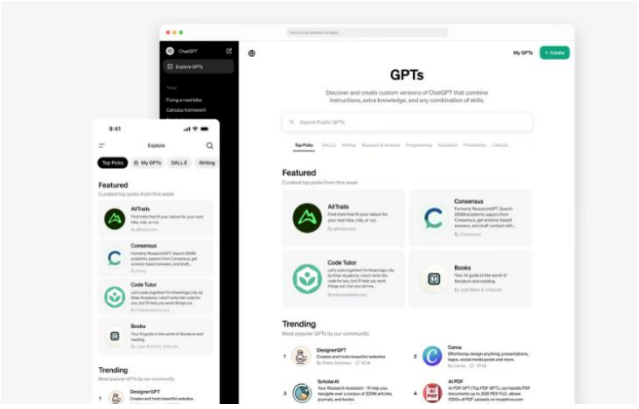
4. 应用端 OpenAI 正式推出 GPT store，生态体系正式建立

应用端持续增长，超过 300 万 GPT 应用在短时间迸发。ChatGPT 推出以来，人们产生了多样化的需求。为了满足这些需求，许多用户积累了精心设计、可复用的指令列表，手动将其复制到 ChatGPT 中。“指令工程师”（prompt engineer）的概念应运而生。2023 年 11 月 6 日，OpenAI 推出 ChatGPT 自定义版本，无需编程，用户就可以将积累的指令列表封装至自定义的 GPT 中。基于这一功能，用户 2 个月的时间就创建了超过 300 万个自定义版本的 ChatGPT。2024 年 1 月 10 日，OpenAI 推出 GPT 商店，整合这些用 GPT 搭建的应用程序。在第一季度，OpenAI 会推出 GPT 构建者激励计划，根据开发参与度提供报酬。此外，GPT 商店即将推出企业版，内置更细致的分享机制，以迎合企业客户的管理需求。

图 50：ChatGPT 自定义版本

资料来源：OpenAI 官网，天风证券研究所

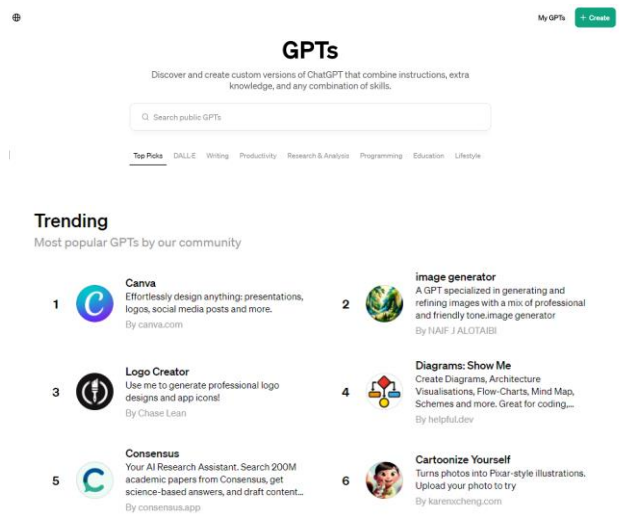
图 51：OpenAI GPT 商店



资料来源：OpenAI 官网，天风证券研究所

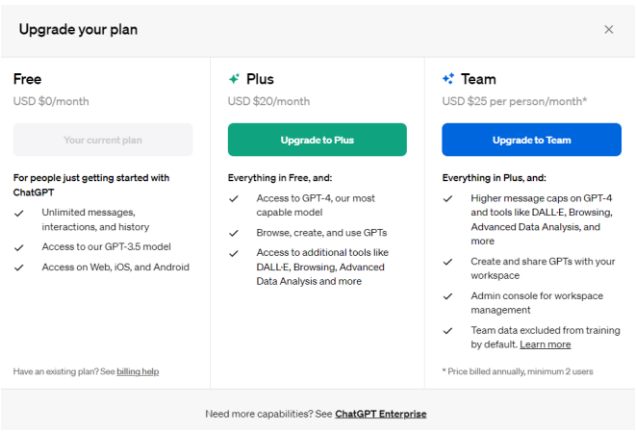
GPTs 覆盖领域广泛，形成付费生态。GPTs 应用覆盖绘画、写作、生产力、研究分析、编程、教育和生活领域，用户既可以使用其他 GPTs 应用，也可以自己创建应用。在 GPTs 社区中，趋势榜（Trending）最受欢迎的应用（2024 年 2 月 5 日）包括设计领域的 Canva、image generator、Logo Creator 和 Cartoonize Yourself，用于绘制流程图的 Diagrams: Show Me，以及用于学术问答的 Consensus。GPTs 并非免费提供，购买了 PLUS 和团队版的用户可以使用。

图 52：GPTs 覆盖领域和趋势榜（2024 年 2 月 5 日）



资料来源：GPTs 网站，天风证券研究所

图 53：GPTs 解锁费用

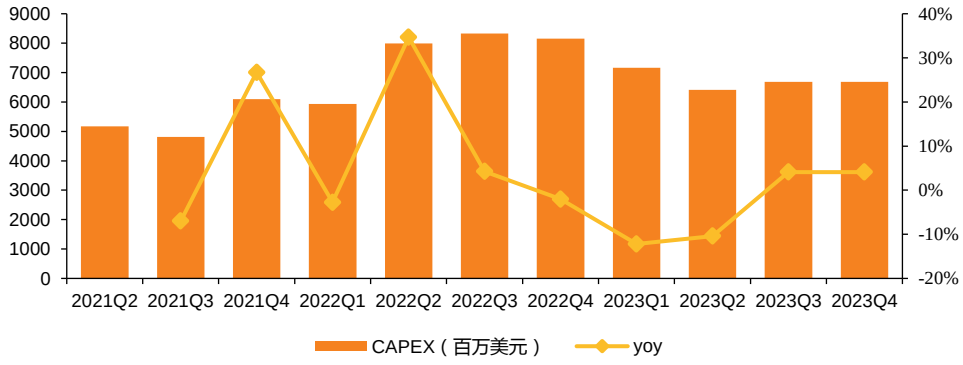


资料来源：GPTs 网站，天风证券研究所

5. 海外模型大厂算力需求持续增加，模型 Scaling 趋势仍在继续

Meta2024 年扩大 GPU 投资，预测未来计算需求将进一步增加。Meta 预计，到 2024 年底会拥有约 35 万台 H100，将其他 GPU 纳入计算，Meta 将拥有约 60 万台 GPU。今年会加大 AI 基础设施的投资，2024 年全年的资本支出将在 300 亿至 370 亿美元之间，比之前的上限增加 20 亿美元。

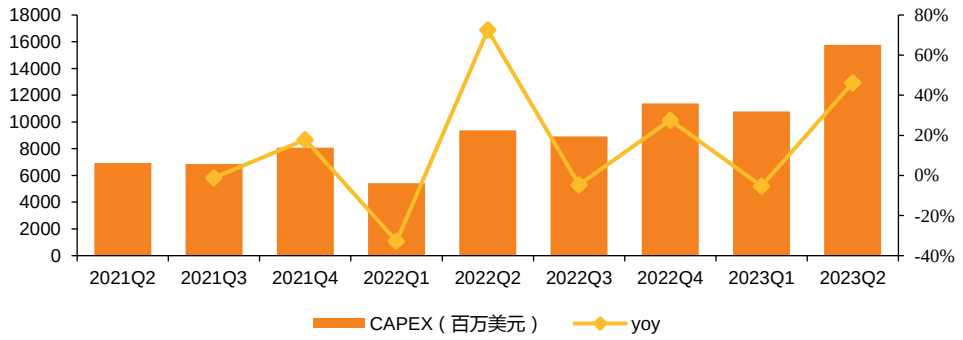
图 54：Meta 2021 财年 Q2 到 2023 财年 Q4 的资本性支出



资料来源：Choice，天风证券研究所

2023Q2 微软 CAPEX 创 2021 财年以来新高，预计 Q3 资本支出将环比大幅增加。微软 Azure 和其他云服务收入分别增长了 30%和 28%（固定汇率），其中 6%的增长来自于 AI 服务。目前共有 53,000 个 Azure AI 客户，超过三分之一的人是过去 12 个月中刚接触。微软预计资本支出将环比大幅增加，针对数据中心的投资将被用于满足云和 AI 基础设施的需求。

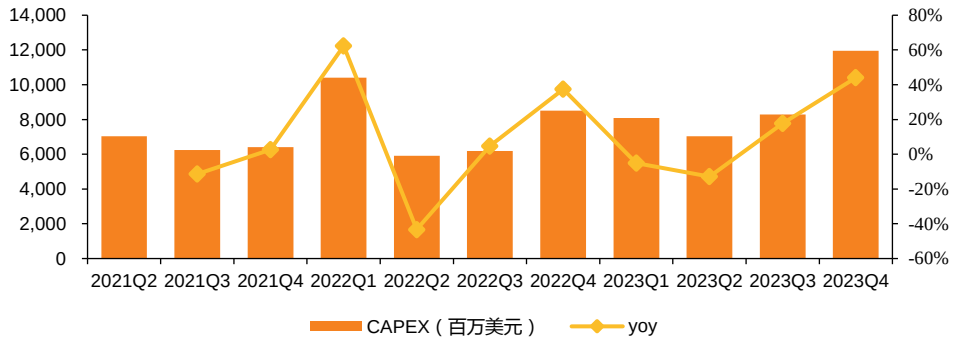
图 55：Microsoft 2021 财年 Q2 到 2023 财年 Q2 的资本性支出



资料来源：Choice，天风证券研究所

谷歌 2023Q4 的资本支出为 110 亿美元，2024 年有望进一步扩大。Vertex AI 已经得到了广泛的应用，2023 年上半年到下半年 API 请求量增长了近 6 倍。此外三星最近宣布推出搭载 Vertex AI 的 Galaxy S24 系列智能手机使用了 Gemini 模型。2023Q4 大部分 CAPEX 用于投资技术基础设施，首先是服务器其次是数据中心。谷歌云引入了一种突破性的超级计算机架构 AIHyper，结合了强大的 TPU 和 GPU、AI 软件和多主机技术，为模型的训练和部署提供了性能和成本优势。到 2024 年，谷歌预计 CAPEX 将比 2023 年显著增加。

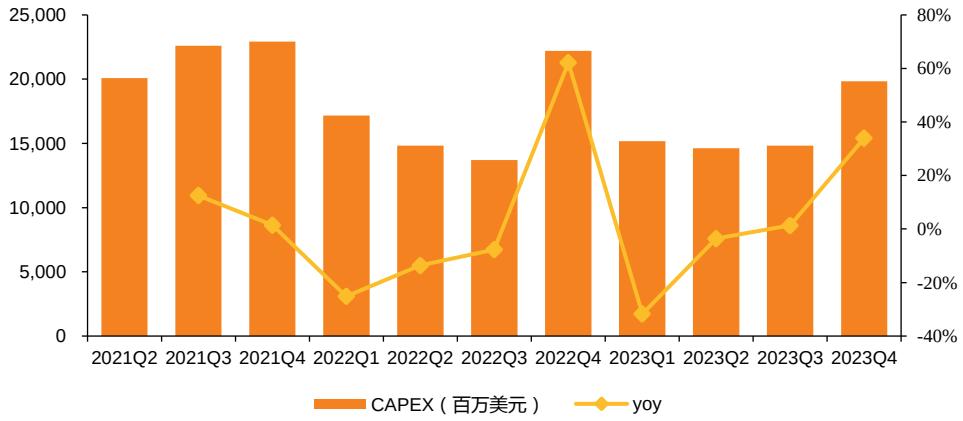
图 56：Google 2021 财年 Q2 到 2023 财年 Q4 的资本性支出



资料来源：Choice，天风证券研究所

亚马逊 2024 年资本支出将增加，主要用于基础设施以支持 AWS 业务增长。2023 年全年资本支出为 484 亿美元，同比下降 102 亿美元。展望 2024 年，亚马逊预计资本支出将同比增加，主要是由基础设施资本支出增加推动的，支持 AWS 业务的增长，包括对生成人工智能和大型语言模型的额外投资。亚马逊认为，生成式 AI 分为 3 个层次，每一个层次都是巨大的，需要深度投资，亚马逊提供了最广泛的 Nvidia 芯片计算实例结合。此外，亚马逊的客户在生成式 AI 早期阶段认识到，不断地迭代模型版本是有必要的。因此，为了适应客户需求，提高 AI 芯片的性价比，亚马逊构建了名为 Trainium 的定制训练芯片和名为 Inferentia 的推理芯片，2023 推出的 Trainium 2 训练性能是 Trainium 1 的 4 倍，内存容量是 Trainium 1 的 3 倍，用于服务包括 Anthropic、Airbnb 在内的客户。

图 57：Amazon 2021 财年 Q2 到 2023 财年 Q4 的资本性支出



资料来源：Choice，天风证券研究所

6. 投资建议

考虑到国内外在模型能力和算力支出上的亮眼表现，我们在此推荐 AI 应用与算力板块的机会，建议关注：

应用：（1）办公软件：金山办公、福昕软件、彩讯股份（通信团队覆盖）

（2）多模态：万兴科技、美图公司（与海外团队联合覆盖）、虹软科技、光云科技

（3）B 端应用：用友网络、金蝶国际、致远互联、泛微网络、鼎捷软件、汉得信息

（4）金融、教育、医疗：同花顺、恒生电子、新致软件、科大讯飞、视源股份（与电子组联合覆盖）、润达医疗

基础设施：神州数码、烽火通信、拓维信息、高新发展、海光信息、星环科技、寒武纪、

景嘉微（与电子团队联合覆盖）

7. 风险提示

（1） 国内大模型效果提升不及预期

若国内大模型能力提升至 GPT-4 的时间不及预期，应用可能落地进度放缓

（2） 国产算力供应不及预期

若国产算力性能提升不及预期，则可能模型能力提升速度不及预期

（3） 国内应用场景落地不及预期

国内可能会存在应用场景逻辑商业化难以闭环的情况

分析师声明

本报告署名分析师在此声明：我们具有中国证券业协会授予的证券投资咨询执业资格或相当的专业胜任能力，本报告所表述的所有观点均准确地反映了我们对标的证券和发行人的个人看法。我们所得报酬的任何部分不曾与，不与，也将不会与本报告中的具体投资建议或观点有直接或间接联系。

一般声明

本报告是机密的，仅供我们的客户使用，天风证券不因收件人收到本报告而视其为天风证券的客户。本报告中的信息均来源于我们认为可靠的已公开资料，但天风证券对这些信息的准确性及完整性不作任何保证。本报告中的信息、意见等均仅供客户参考，不构成所述证券买卖的出价或征价邀请或要约。该等信息、意见并未考虑到获取本报告人员的具体投资目的、财务状况以及特定需求，在任何时候均不构成对任何人的个人推荐。客户应当对本报告中的信息和意见进行独立评估，并应同时考量各自的投资目的、财务状况和特定需求，必要时就法律、商业、财务、税收等方面咨询专家的意见。对依据或者使用本报告所造成的一切后果，天风证券及/或其关联人员均不承担任何法律责任。

本报告所载的意见、评估及预测仅为本报告出具日的观点和判断。该等意见、评估及预测无需通知即可随时更改。过往的表现亦不应作为日后表现的预示和担保。在不同时期，天风证券可能会发出与本报告所载意见、评估及预测不一致的研究报告。

天风证券的销售人员、交易人员以及其他专业人士可能会依据不同假设和标准、采用不同的分析方法而口头或书面发表与本报告意见及建议不一致的市场评论和/或交易观点。天风证券没有将此意见及建议向报告所有接收者进行更新的义务。天风证券的资产管理部门、自营部门以及其他投资业务部门可能独立做出与本报告中的意见或建议不一致的投资决策。

特别声明

在法律许可的情况下，天风证券可能会持有本报告中提及公司所发行的证券并进行交易，也可能为这些公司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。因此，投资者应当考虑到天风证券及/或其相关人员可能存在影响本报告观点客观性的潜在利益冲突，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	说明	评级	体系
股票投资评级	自报告日后的 6 个月内，相对同期沪深 300 指数的涨跌幅	买入	预期股价相对收益 20%以上
		增持	预期股价相对收益 10%-20%
		持有	预期股价相对收益-10%-10%
		卖出	预期股价相对收益-10%以下
行业投资评级	自报告日后的 6 个月内，相对同期沪深 300 指数的涨跌幅	强于大市	预期行业指数涨幅 5%以上
		中性	预期行业指数涨幅-5%-5%
		弱于大市	预期行业指数涨幅-5%以下

天风证券研究

北京	海口	上海	深圳
北京市西城区德胜国际中心 B 座 11 层	海南省海口市美兰区国兴大道 3 号互联网金融大厦	上海市虹口区北外滩国际客运中心 6 号楼 4 层	深圳市福田区益田路 5033 号平安金融中心 71 楼
邮编：100088	A 栋 23 层 2301 房	邮编：200086	邮编：518000
邮箱：research@tfzq.com	邮编：570102	电话：(8621)-65055515	电话：(86755)-23915663
	电话：(0898)-65365390	传真：(8621)-61069806	传真：(86755)-82571995
	邮箱：research@tfzq.com	邮箱：research@tfzq.com	邮箱：research@tfzq.com