计算机

多模态，AI 大模型新一轮革命

强于大市（维持评级）

投资要点：

- **多模态推动人工智能迈向 AGI，底层技术日臻成熟**

相比单模态，多模态大模型同时处理文本、图片、音频以及视频等多类信息，与现实世界融合度高，更符合人类接收、处理和表达信息的方式，与人类交互方式更加灵活，表现的更加智能，能够执行更大范围的任务，有望成为人类智能助手，推动 AI 迈向 AGI。就技术架构而言，多模态技术可拆解为编码、对齐、解码与微调等步骤，逐步挖掘多模态关联信息，输出目标结果。文生图 CLIP 模型为最先成熟的多模态技术，目前，多模态已不再局限于图文两层信息。例如，Meta-Transformer 可同时理解并处理 12 种模态信息。

- **OpenAI 谷歌开启多模态军备竞赛，Sora 和 Gemini 各领风骚**

海外龙头具备先发与技术优势，引领多模态大模型前进方向：1）OpenAI 近期密集剧透 GPT-5，相比 GPT-4 实现全面升级，重点突破语音输入和输入、图像输出以及最终的视频输入方向，或将实现真正多模态；此外，2 月发布文生视频大模型 Sora，能够根据文本指令或静态图像生成 1 分钟的视频，其中包含精细复杂的场景、生动的角色表情以及复杂的镜头运动，同时也接受现有视频扩展或填补缺失的帧，能够很好地模拟和理解现实世界。2）Google 推出原生多模态大模型 Gemini，可泛化并无缝地理解、操作和组合不同类别的信息；此外，2 月推出 Gemini 1.5 Pro，使用 MoE 架构首破 100 万极限上下文纪录，可单次处理包括 1 小时的视频、11 小时的音频、超过 3 万行代码或超过 70 万个单词的代码库。3）Meta 坚持大模型开源，建设开源生态巩固优势，已陆续开源 ImageBind、AnyMAL 等多模态大模型。国内大模型厂商有望沿着复制海外先进技术与发挥生态禀赋优势的两大路径，与海外大厂逐步缩小差距。

- **多模态提升大模型泛化能力，垂直领域应用场景广阔**

强调技术与业务的融合以推动业务的数字化转型和智能化升级，才能够最大化的发挥大模型价值同时激励大模型创新升级，实现业务效率提升与技术创新的良性循环。多模态大模型的应用场景和价值正在不断扩展和提升。从语音识别、图像生成、自然语言理解、视频分析，到机器翻译、知识图谱等，多模态大模型都能够提供更丰富、更智能、更人性化的服务和体验。在强大泛化能力基础上，大模型可以在不同模态和场景之间实现知识的迁移和共享，将大模型的应用扩展到不同的领域和场景。

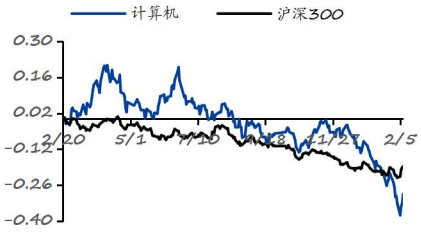
- **投资建议**

我们看好具有算法、数据等先发优势的国产大模型厂商，同时多模态提升大模型泛化能力，多元信息环境下实现“多专多能”，在垂直领域具有广阔的应用场景和市场价值。建议关注：**1）AI+多模态：**万兴科技、中科创达、虹软科技、当虹科技、大华股份、海康威视、漫步者、萤石网络、汉仪股份、美图公司、云从科技；**2）AI+办公：**金山办公、万兴科技、福昕软件、彩讯股份、金蝶国际、泛微网络、致远互联、鼎捷软件、汉得信息，用友网络；**3）AI+教育/电商/医疗：**科大讯飞、佳发教育、鸥玛软件、盛通股份、光云科技、值得买、焦点科技、小商品城、润达医疗、嘉和美康、创业慧康、迪安诊断等。

- **风险提示**

技术发展不及预期、产品落地不及预期、AI 伦理风险等。

一年内行业相对大盘走势



团队成员

分析师： **施晓俊(S0210522050003)**
SXJ3780@hfzq.com.cn
联系人： **李杨玲(S0210123100071)**
lyl30339@hfzq.com.cn
联系人： **王思(S0210123070006)**
ws30181@hfzq.com.cn

相关报告

1、计算机行业当前处于什么周期位置？——2024.02.05
2、AI 应用大幕徐徐展开——2024.01.28
3、计算机板块央国企控股公司梳理——2024.01.25

正文目录

图表目录

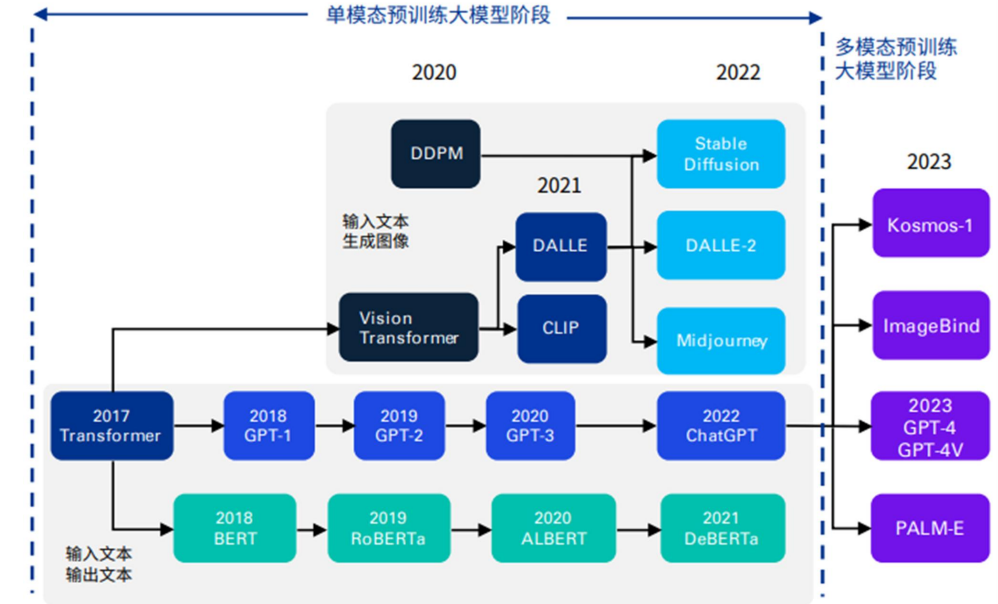
1 多模态推动人工智能迈向 AGI

1.1 多模态或成为 AI 大模型主流

相比单模态，多模态向通用人工智能（AGI）迈前一步。多模态大模型同时处理文本、图片、音频以及视频等多类信息，与现实世界融合度高，有望成为人类智能助手，推动 AI 迈向 AGI：1）多模态更符合人类接收、处理和表达信息的方式。人类能够感知多元信息，每一类信息均为一种模态，这些信息往往是相互关联的。2）多模态信息使得大模型更为智能。多模态与用户交互方式更便捷，得益于多模态输入的支持，用户可以更灵活的方式与智能助手进行交互和交流。3）多模态提升任务解决能力。LLM 通过可以执行 NLP 任务，而多模态通常可以执行更大范围的任务。

目前，多模态大模型已成为大模型发展前沿方向。2022 年及之前，大模型处于单模态预训练大模型阶段，主要探索文本模式的输入输出。2017 年，Transformer 模型提出，奠定了当前大模型的主流算法结构；2018 年，基于 Transformer 架构训练的 BERT 模型问世，参数规模首次突破 3 亿；随后 GPT 系列模型推出，2022 年底至今 ChatGPT 引爆全球大模型创新热潮。步入 2023 年，大模型发展从文本、图像等单模态任务逐渐发展为支持多模态的多任务，更为符合人类感知世界的方式。大模型公司的比拼重点转移为多模态信息整合和数据挖掘，精细化捕捉不同模态信息的关联。例如，2023 年 9 月，OpenAI 推出最新多模态大模型 GPT-4V，增强了视觉提示功能，在处理任意交错的多模态方面表现突出。

图表 1：大模型朝多模态方向发展



数据来源：中关村产业研究院，毕马威分析，华福证券研究所

1.2 多模态发展路径逐步清晰，底层技术日臻成熟

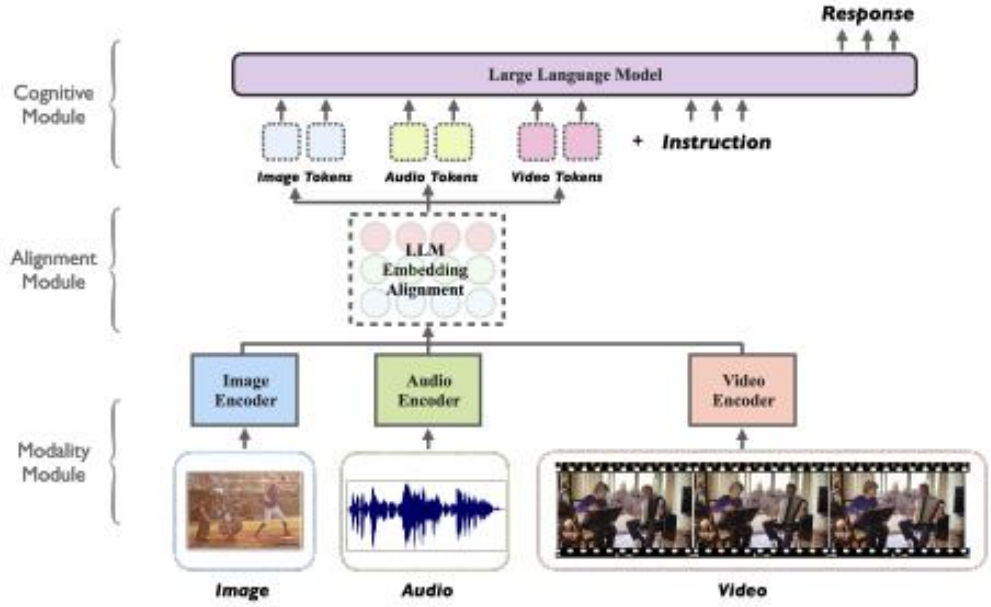
目前，多模态大模型发展路径逐步清晰。发展思路主要有三：1）利用单模态模型如 LLMs 来调动其他数据类型的功能模块完成多模态任务，典型代表有 Visual、ChatGPT、Hugging GPT 等；2）直接利用图像和文本信息训练得到多模态大模型，典型代表有 KOSMOS-1 等；3）将 LLMs 与跨模态编码器等有机结合，融合 LLMs 的推理检索能力和编码器的多模态信息整合能力，典型代表有 Flamingo、BLIP2 等。

多模态大模型底层技术日臻成熟，支撑实现多类信息融合与转换。

从技术架构来看，多模态大模型一般包括编码、对齐、解码和微调等步骤，逐步整合多模态关联信息，输出目标结果。1）编码：包括视觉、音频、文本等模态编码器，目的是有效处理多个模态信息，转化为可处理状态；2）对齐：解决不同模态编码器可能不能直接融合的问题，建立共同表示空间，将不同模态的表示统一，实现多个模态信息的有效整合；3）解码：编码的反向过程，把模型的内部表示转化为物理世界的自然信号，即输出人类可识别的信息；4）微调：针对个性化如垂直行业

大模型，重新训练大模型消耗算力成本较高，便可在预训练模型的基础上，通过在自有小数据集上的训练来适应新的任务，更好地提升大模型在下游特定领域能力。

图表 2：多模态大模型一般架构

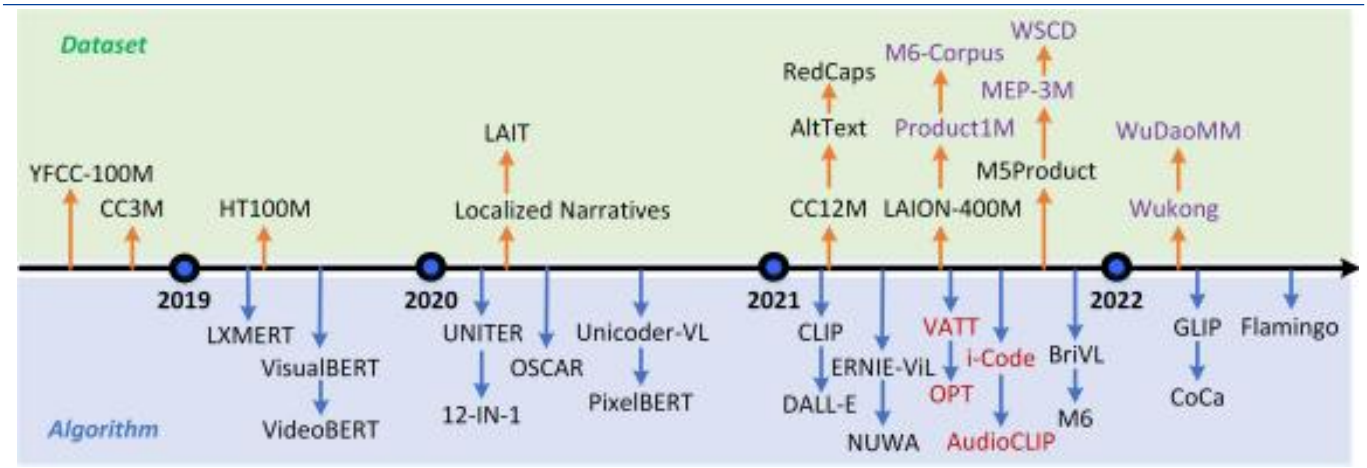


数据来源：Lyu et al.《MACAW-LLM: MULTI-MODAL LANGUAGE MODELING WITH IMAGE, AUDIO, VIDEO, AND TEXT INTEGRATION》，华福证券研究所

文生图为最先成熟的多模态技术领域，其代表技术为 OpenAI 于 2021 年推出的 CLIP 模型。CLIP 使用约 4 亿从网页中爬取的图像-文本对数据进行对比学习，采用图像和文本双编码器，用于评估给定图像与给定文本描述的匹配程度，成为连接文本和图像的桥梁。

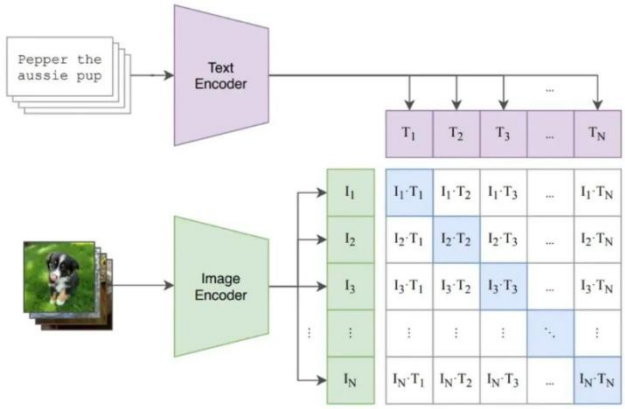
目前，多模态底层技术不再局限于文本与图像两层信息，**Meta-Transformer 可同时理解 12 种模态信息。**2023 年 7 月，香港中文大学多媒体实验室联合上海人工智能实验室的 OpenGVLAB 研究团队提出一个统一多模态学习框架 Meta-Transformer，实现骨干网络的大一统，具有一个模态共享编码器，并且无需配对数据，即可理 12 种模态信息，并提供了多模态无边界融合的新范式。相比 CLIP、BEiT-3、Imagebind，模态数目大幅增加，并且摆脱了多模态训练过程中对于配对数据的依赖性，为多模态学习提供了新路径。

图表 3：2019 年至今多模态预训练大模型重要算法与数据集



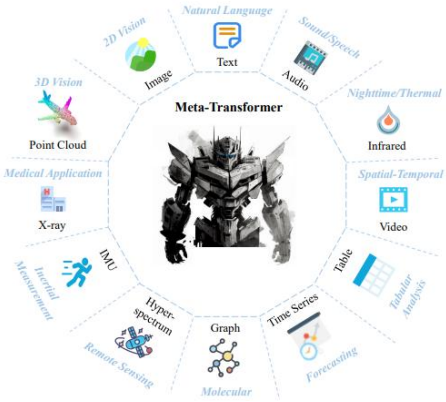
数据来源：Wang et al.《Large-scale Multi-modal Pre-trained Models: A Comprehensive Survey》，华福证券研究所（注：紫色字体表示该数据集包含中文文本，其他数据集包含英文文本；红色突出显示的模型是使用两个以上的模态进行训练的）

图表 4：CLIP 为连接文本与图像的桥梁



数据来源：Radford et al.《 Learning Transferable Visual Models From Natural Language Supervision》，华福证券研究所

图表 5：Meta-Transformer 可同时处理 12 种模态



数据来源：Zhang et al.《 Meta-Transformer: A Unified Framework for Multimodal Learning》，华福证券研究所

2 国内外大模型陆续更新，瞄准多模态方向升级

2.1 OpenAI 谷歌引战多模态，视频为重要角力点

2.1.1 OpenAI 密集剧透 GPT-5，或将实现真正多模态

2024 年 1 月，OpenAI 首席执行官奥特曼在与比尔·盖茨的对话中以及参加达沃斯论坛时频繁提及新一代大模型 GPT-5。**据奥特曼介绍，GPT-5 相比 GPT-4 实现全面升级，如果 GPT-4 目前解决了人类任务的 10%，GPT-5 应该是 15%或者 20%。**GPT-5 将是一个多模态模型，支持语音、图像、代码和视频，并在个性化和定制化功能方面实现重大更新，具备更强的推理能力和更高的准确性。当前大模型的通病——幻觉问题也将在 GPT-5 中得到解决。

1）个性化与定制化功能重大更新。GPT-5 最关键的增强部分将围绕个人偏好的理解，比如整合用户信息、电子邮件、日历、约会偏好，以及与外部数据源建立联系，由此实现个性化的风格。

2）更强的推理能力和更高的准确性。当代大模型存在最大“幻觉”问题将在 GPT-5 中得到解决，提升大模型可靠性。例如，如果向 GPT-4 中询问 1 万次问题，这一万次回答中可能只有一次是好的，但 GPT-4 无法判断，这一点在 GPT-5 或许有所改善。

3）实现真正的多模态。GPT-5 **不仅支持文本输入，还支持语音、图像、代码和视频，处理更加复杂和多样的信息，多模态处理能力实现飞跃。**在与比尔盖茨交谈 OpenAI 下一阶段最重要发展方向时，奥特曼表示语音输入和输出、图像输出以及最终的视频输入将成为公司重点发力方向。

早在 2023 年 7 月，GPT-5 商标处于注册流程中，新一代大模型发布箭在弦上。

图表 6：2023 年 7 月，GPT-5 商标处于注册流程中

数据来源：量子位，华福证券研究所

梳理 GPT 历次更新，多模态能力升级成为重要看点。2018-2022 年，OpenAI 基于 Transformer 架构先后推出 GPT-1 至 GPT-3.5，在训练数据集上主要考虑文本数据，能够实现上下文理解和多轮对话，而在多模态能力上存在欠缺。2023 年 3 月，OpenAI 推出 GPT-4，增加了额外的视觉语言模块，在 GPT-3 和 GPT3.5 训练数据集上增加了多模态数据集，能够实现图生文。之后更新的 GPT-4V 以及 GPT4-Turbo 进一步突破音频输入技术，使得文本转语音（TTS）成为可能。近期，OpenAI 剧透 GPT-5，能够同时支持文本、图片、语音、视频等多元信息，多模态能力实现跨越。

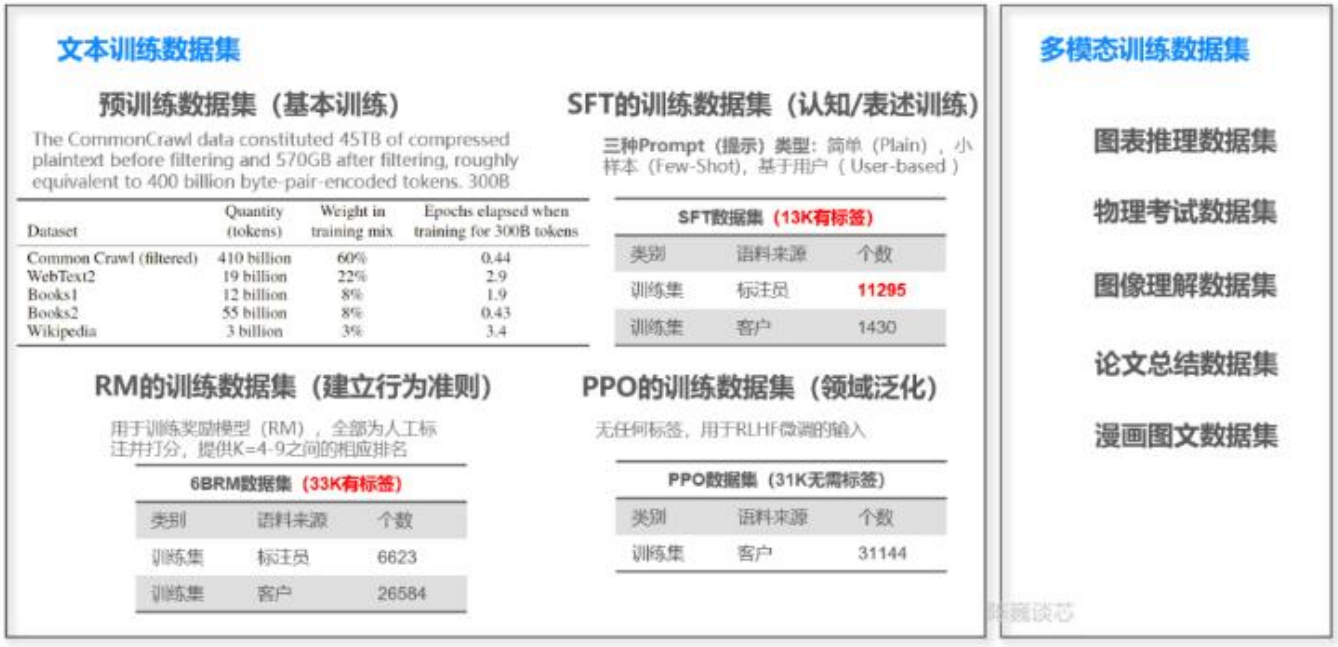
我们认为，OpenAI 作为全球领先科技企业，在大模型的技术方向可作为其他公司研发方向标，GPT 历次更新着重多模态能力，以及近期奥特曼频繁剧透 GPT-5 关键信息，或将掀起国内外大模型新一轮军备竞赛，进一步提升 AI 领域景气度。

图表 7：GPT 历次更新梳理

大模型	发布时间	参数量	数据集	输入				输出			
				文本	图片	语音	视频	文本	图片	语音	视频
GPT-1	2018 年	117M	BookesCorpus	√				√			
GPT-2	2019 年	1.5B	WebText	√				√			
GPT-3	2020 年	175B	Common Cral, WebText2,Books1,Booke2,Wikipedia	√				√			
GPT-3.5	2022.11	-	类似 GPT-3，但可能有更新	√				√			
GPT4	2023.03	-	更大规模和多样化	√	√			√			
GPT-4V	2023.09	-	-	√	√	√		√		√	
GPT4-Turbo	2023.11	-	-	√	√	√		√	√	√	
GPT-5	待定	-	-	√	√	√	√	-	-	-	-

数据来源：智东西，新智元，OpenAI，机器之心，华福证券研究所

图表 8：GPT-4 数据集构成（预测）



数据来源：智东西，华福证券研究所

2.1.2 OpenAI 推出首款视频生成模型 Sora，视频更加接近真实世界

美国当地时间 2 月 15 日，**OpenAI 发布视频生成模型 Sora，是一种通用的视觉**

数据模型，可以生成持续时间、宽高比和分辨率各异的视频和图像，长达一分钟的高清视频更加接近真实世界。Sora 是一种扩散模型，生成的视频一开始像静态噪音，之后通过多个步骤去除噪音，逐步转换视频。与 Midjourney 和 Stable Diffusion 同样基于扩散模型相比，Sora 生成视频的质量更高，更像是创建了真实的视频。而与 Gen-2、SVD-XT、Pika 等主流产品相比，Sora 可生成最长一分钟的视频，具备更强的构建物理世界的模拟能力。

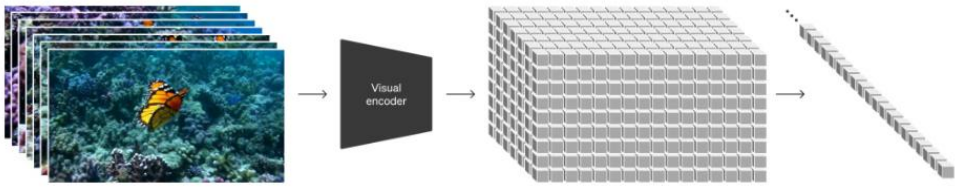
图表 9：Sora 可生成一分钟长视频

数据来源：OpenAI，华福证券研究所

从技术上来讲，Sora 充分吸收了 OpenAI 前期在大模型积累的技术能力。

1）Sora 与 GPT 模型类似，使用了 Transformer 架构，从而实现了卓越的扩展性能。OpenAI 将视频和图像表示为称为 patch 的较小数据单元的集合，每个 patch 类似于 GPT 中的 token。通过统一数据表示方式，OpenAI 能够在比以往更广泛的视觉数据上训练扩散 transformer，包括不同的持续时间、分辨率和宽高比。

图表 10：Sora 将视觉数据转换为 patch



数据来源：OpenAI，华福证券研究所

2）Sora 建立在过去 DALL·E 和 GPT 模型的研究基础之上。它采用了 DALL·E 3 中的重述技术，即为视觉训练数据生成高度描述性的字幕。因此，该模型能够在生成的视频中更忠实地遵循用户的文字提示。

就模型能力而言，Sora 文生视频大模型具有如下特点：

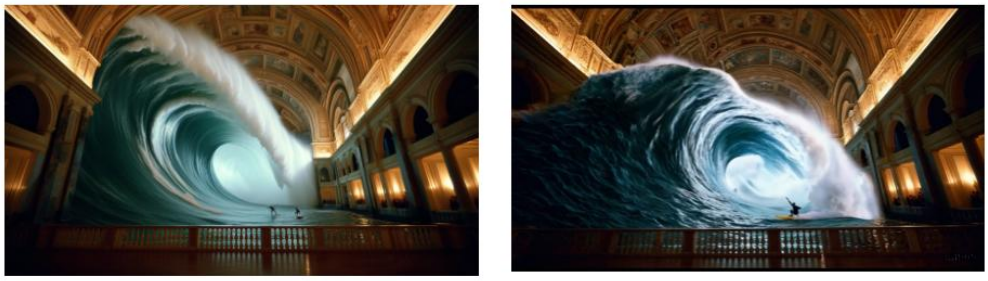
强大的语言理解能力：训练文本到视频生成系统需要大量带有相应文本说明的视频。OpenAI 将 DALL · E 3 中介绍的字幕重配技术（Recaptioning）应用到视频中，首先训练一个高度描述性的字幕模型，然后使用它为其训练集中的所有视频生成文本字幕。OpenAI 发现，对高度描述性的视频字幕进行训练可提高文本保真度以及视频的整体质量。

图表 11：Sora 根据文本说明生成高质量视频

数据来源：OpenAI，华福证券研究所

支持图片与视频输入：Sora 能够执行广泛的图像和视频编辑任务——创建完美的循环视频、动画静态图像、向前或向后扩展视频等。比如，基于 DALL · E 3 图像生成视频，从一个生成的视频片段开始向前/向后扩展视频，编辑转换视频的风格/环境，将两个输入视频无缝衔接在一起。

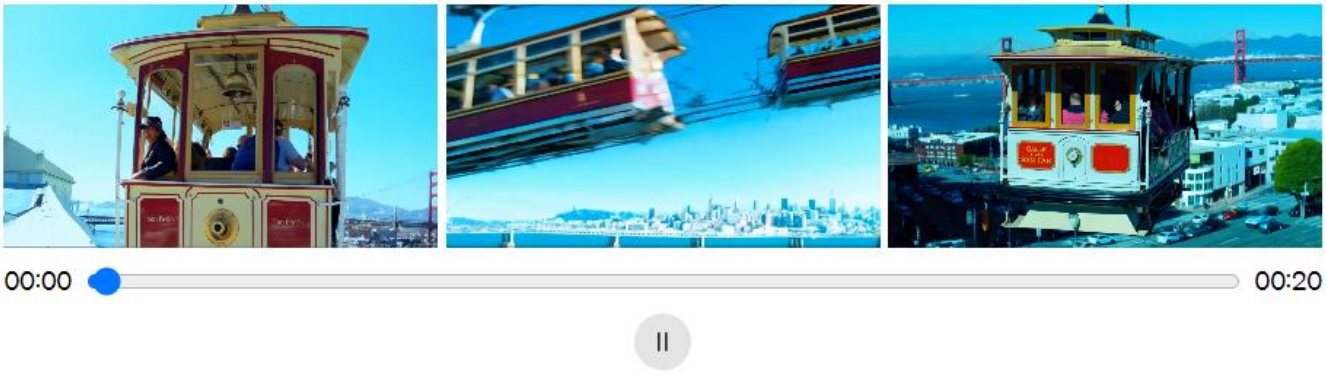
图表 12：Sora 根据冲浪图片（左）生成冲浪动态视频（右）



In an ornate, historical hall, a massive tidal wave peaks and begins to crash. Two surfers, seizing the moment, skillfully navigate the face of the wave.

数据来源：OpenAI，华福证券研究所

图表 13：Sora 从视频片段开始向前/向后扩展视频



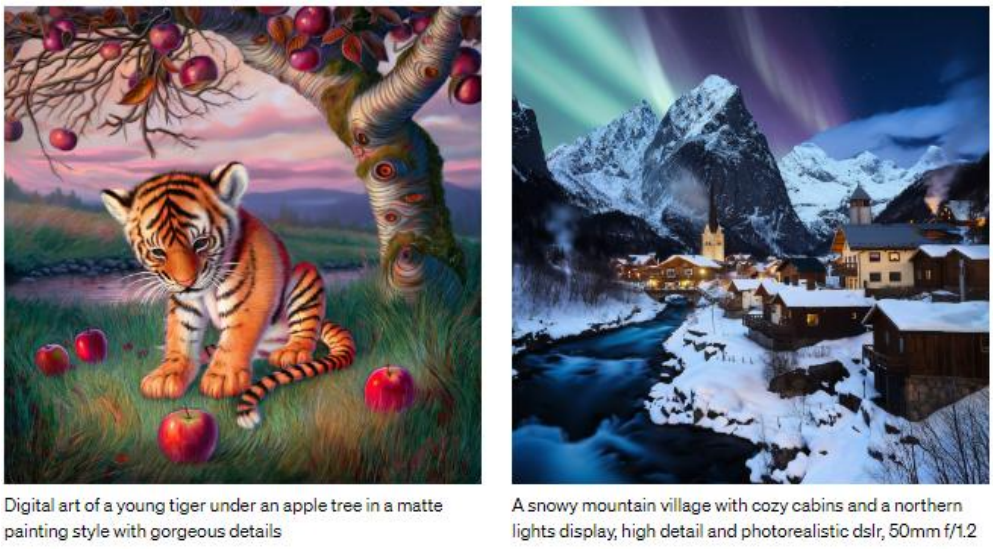
数据来源：OpenAI，华福证券研究所

图表 14：Sora 能够编辑视频风格

数据来源：OpenAI，华福证券研究所

图像生成功能：研究团队通过在一个时间范围为一帧的空间网格中排列高斯噪声块来实现这一点。该模型可以生成可变大小的图像，最高可达 2048 × 2048 分辨率。

图表 15：Sora 生成可变大小的图像



数据来源：OpenAI，华福证券研究所

新兴的仿真能力：OpenAI 发现视频模型在大规模训练时表现出许多有趣的突发能力。这些功能**使 Sora 能够从现实世界中模拟人、动物和环境的某些方面。**Sora 可以生成带有动态摄像机运动的视频。随着摄像机的移动和旋转，人物和场景元素在三维空间中始终如一地移动。

图表 16：Sora 生成带有动态摄像机运动的视频

数据来源：OpenAI，华福证券研究所

总体而言，不管是在视频的保真度、长度、稳定性、一致性、分辨率、文字理解等方面，Sora 都做到了业内领先水平。**我们认为，Sora 能够很好地理解和模拟现**

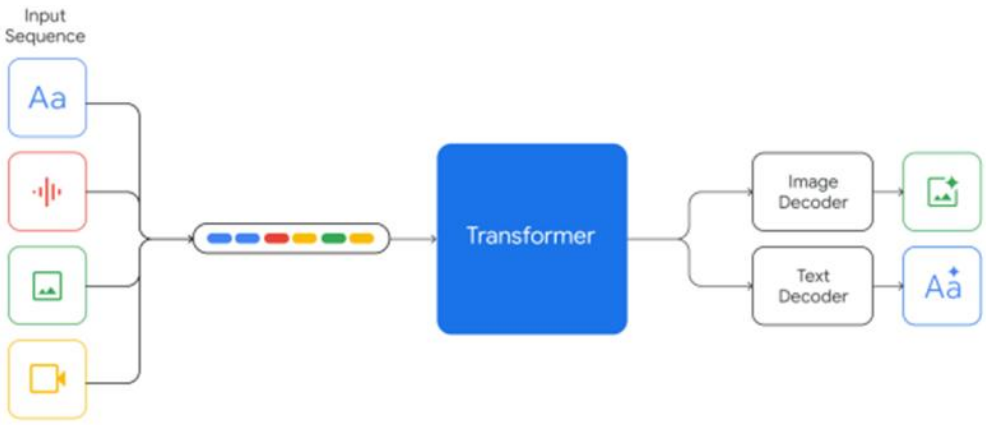
实世界，强大的视频生成多模态功能使得 OpenAI 朝着 AGI 迈向一大步，同时考虑到之前已发布的文生图模型 DELL·E3、语音模型 Whisper，OpenAI 已经具备了文本、图像、视频、音频 4 大多模态能力，进一步夯实在大模型领域的龙头地位。

2.2 谷歌推出 Gemini，实现大模型多模态原生

2.2.1 Gemini 正式对外发布，多模态理解优势突出

美国时间 2023 年 12 月 6 日，谷歌大语言模型 Gemini 在正式对外发布。**值得关注的是，Gemini 是一个原生多模态大模型，可以同时识别和理解文本、图像、音频、视频和代码五种信息，即可以泛化并无缝地理解、操作和组合不同类型的信息。**这意味着用户可以自然地交错输入：说几句话，添加图像、文本，或是短视频。同样，模型也会自然地交错文本和图像作为输出。

图表 17：Gemini 支持输入文本、图像、语音和视频输出文本和图像

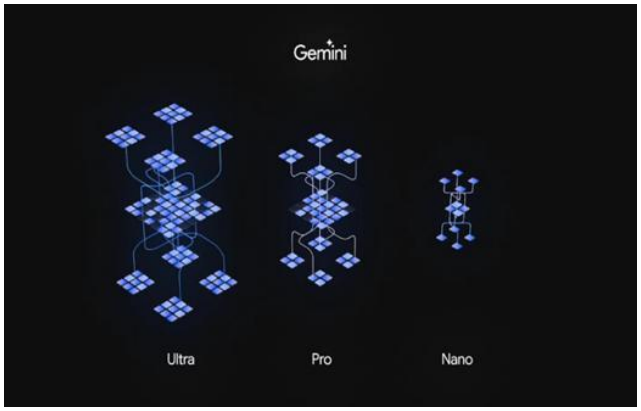


数据来源：Gemini Team《Gemini: A Family of Highly Capable Multimodal Models》，华福证券研究所

谷歌当天发布的 Gemini 1.0 共分为 Ultra, Pro 和 Nano 三个版本，其中 Ultra 的能力最强，复杂度最高，能够处理最为困难的任务，Pro 能力稍弱，可以用来处理多任务，Nano 则更注重于端侧的处理能力。

Gemini Ultra 在 MMLU（大规模多任务语言理解数据集）中的得分率高达 90.0%，首次超越了人类专家。该模型不仅可以综合使用数学、物理、历史、法律、医学和伦理学等 57 个科目来测试世界知识和解决问题的能力可以理解复杂主题中的细微差别和推理，且有能力处理图像、视频和音频，甚至会变得更加通用，有动作和触摸之类更像机器人类型的功能。

图表 18：Gemini 包括三种不同规模的模型



数据来源：澎湃新闻，华福证券研究所

图表 19：Gemini 识别蓝色小鸭子素材

数据来源：新智元，华福证券研究所

在对比 Gemini 和 GPT-4 的基准测试中，**Gemini 最明显的优势来自于它理解视频和音频并与之交互的能力，从测评数据来看，其在多模态任务上全面超越 GPT-4V，**在最常规的文本基准测试中， Gemini 与其他大模型的差距并不大。

图像、语音任务分析：通过图像和语音输入，Gemini 可以指导做菜，并在不同

阶段提出相应的建议。

图表 20：Gemini 处理做菜任务

Input Image	Input Audio (transcribed)	Model Response: Text
	What's the first step to make a veggie omelet with these ingredients?	Crack the eggs into a bowl and whisk them.
	Thank you for the instructions. I started making my omelet, does it look ready now?	It looks like it's almost ready. You can flip it over to cook the other side.
(No image - it's a follow up on the previous question)	Why is it not ready?	It's not ready because the eggs are still runny.
	What about now?	It looks ready now. You can take it off the heat and serve it.

数据来源：智东西，华福证券研究所

视频任务分析：通过一段几秒钟的视频，Gemini 可以推断出这出自哪一部电影，并且还给出了具体涉及的剧情桥段。

图表 21：Gemini 处理视频任务

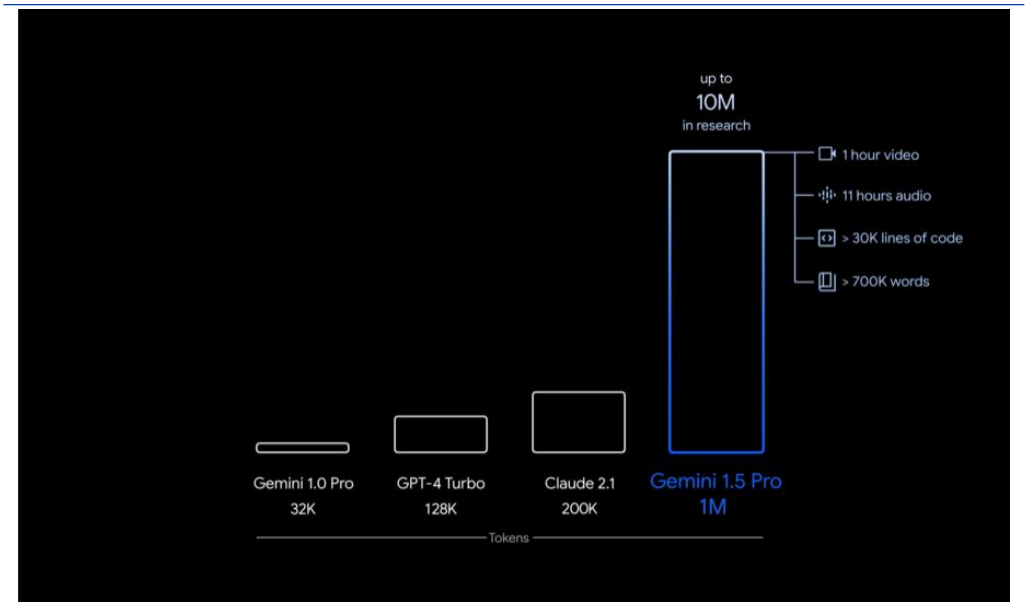
数据来源：智东西，华福证券研究所

2.2.2 Gemini 1.5 突破 100 万 token，多模态能力实现飞跃

美国时间 2 月 15 日，**谷歌宣布推出下一代人工智能 Gemini 1.5，引入 MoE 架构极大地提高了模型效率**。Gemini 1.5 建立在谷歌基础模型开发和基础设施的研究与工程创新的基础上，包括通过新的专家混合 (MoE) 架构使 Gemini 1.5 的训练和服务更加高效。传统 Transformer 充当一个大型神经网络，而 MoE 模型则分为更小的“专家”神经网络。根据给定输入的类型，MoE 模型学会选择性地仅激活其神经网络中最相关的专家路径。这种专业化极大地提高了模型的效率。

谷歌目前推出的是用于早期测试的 Gemini 1.5 的第一个版本 ——**Gemini 1.5 Pro，突破性地可理解 100 万 token 的“上下文窗口”。**通过一系列机器学习创新，谷歌增加了 1.5 Pro 的上下文窗口容量，远远超出了 Gemini 1.0 最初的 32k 个 token。该大模型现在可以在生产环境中运行多达 100 万个 token。这意味着 1.5 Pro 可以一次性处理大量信息，包括 1 小时的视频、11 小时的音频、超过 30000 行代码或超过 700000 个单词的代码库。此外，在基础研究中，谷歌还成功测试了多达 1000 万个 token。

图表 22：Gemini 1.5 Pro 领先基础模型的上下文长度

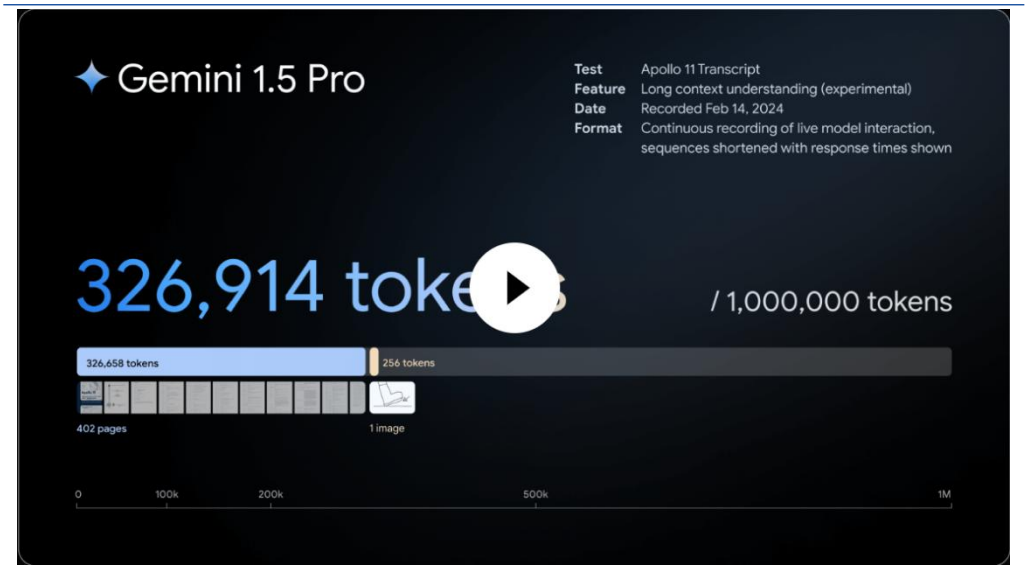


数据来源：Google，华福证券研究所

更长的上下文意味着更有用的功能，大模型视野被史诗级拓宽，在文本、音频、视频、图像、代码等多模态能力展现出更强的性能。

深入理解海量信息：它能够洞察文档中的对话、事件和细节，展现出对复杂信息的深刻理解。例如，当给出阿波罗 11 号登月任务的 402 页记录时，它可以推理整个文档中的对话、事件和细节。

图表 23：Gemini 1.5 Pro 分析和总结阿波罗 11 号登月任务的 402 页记录



数据来源：Google，华福证券研究所

更好地理解和推理跨模态：Gemini 1.5 Pro 还能够在视频中展现出深度的理解和推理能力。例如，当给定一部 44 分钟的巴斯特·基顿无声电影时，该模型可以准确分析各种情节点和事件，甚至推理出电影中容易被忽略的小细节。

图表 24：Gemini 1.5 Pro 分析和总结 44 分钟的巴斯特·基顿无声电影

数据来源：Google，华福证券研究所

高效处理更长代码：Gemini 1.5 Pro 可以跨较长的代码块执行更相关的问题解决任务。当给出超过 100000 行代码的提示时，它可以更好地推理示例、建议有用的修改并解释代码不同部分的工作原理。

图表 25：Gemini 1.5 Pro 高效处理 100000 行代码

数据来源：Google，华福证券研究所

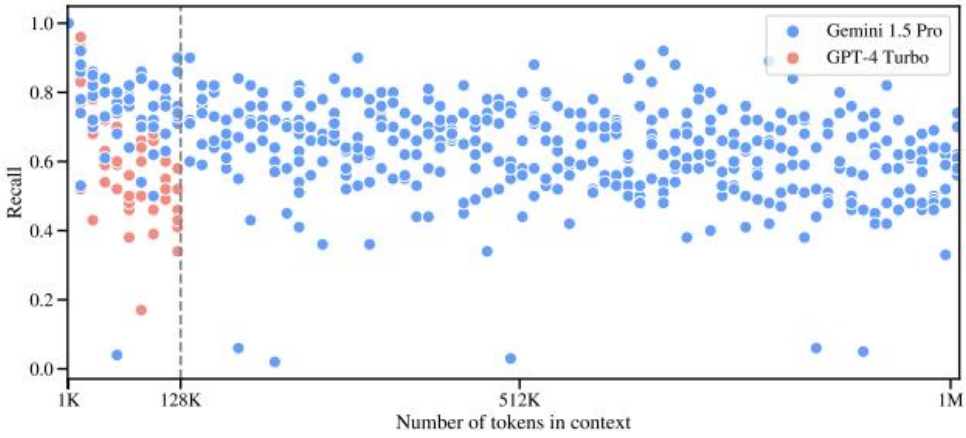
在基准测试中，Gemini 1.5 Pro 的 87%性能优于 1.0 Pro，与 1.0 Ultra 性能相当，并能够在上下文窗口增加时，模型依旧保持较高水平性能。具体而言，Gemini 1.5 Pro 在分别在 27 / 17 项测试中超越 1.0 Pro / 1.0 Ultra。谷歌研究人员开发了通用版本的“大海捞针”测试，在这个测试中，模型需要在一定的文本范围内检索到 100 个不同的特定信息片段。Gemini 1.5 Pro 在较短的文本长度上的性能超过了 GPT-4-Turbo，并且在整个 100 万 token 的范围内保持了相对稳定的表现。与之对比鲜明的是，GPT-4 Turbo 的性能则飞速下降，且无法处理超过 128000token 的文本。

图表 26：Gemini 1.5 Pro 在基准测试中性能领先

Gemini 1.5 Pro	Relative to 1.0 Pro	Relative to 1.0 Ultra
Long-Context Text, Video & Audio	from 32k up to 10M tokens	from 32k up to 10M tokens
Core Capabilities	Win-rate: 87.1% (27/31 benchmarks)	Win-rate: 54.8% (17/31 benchmarks)
Text	Win-rate: 100% (13/13 benchmarks)	Win-rate: 77% (10/13 benchmarks)
Vision	Win-rate: 77% (10/13 benchmarks)	Win-rate: 46% (6/13 benchmarks)
Audio	Win-rate: 60% (3/5 benchmarks)	Win-rate: 20% (1/5 benchmarks)

数据来源：Google《Gemini 1.5: Unlocking multimodal understanding across millions of tokens of Context》，华福证券研究所

图表 27：Gemini 1.5 Pro 在长 token 理解上性能超越 GPT-4 Turbo



数据来源：Google《Gemini 1.5: Unlocking multimodal understanding across millions of tokens of Context》，华福证券研究所

2.3 Meta 坚持模型开源，建设生态巩固优势

2022 年 5 月，Meta AI 全面开放 1750 亿参数大模型 OPT-175B，首次毫无保留公开训练代码及使用代码、日志记录，从此铺垫了 Meta 大模型开源之路。2023 年，Meta 致力于研发多模态大模型，陆续发布 SAM、DINO v2、ImageBind、AnyMAL，并坚守模型开源，形成独特开源大模型生态，巩固行业竞争力。

图表 28：Meta 主要开源大模型梳理

模型名称	发布时间	模型类型	简介
OPT-175B	2022 年 5 月	大语言模型	已在生成创意文本、解决基础数学试题、回答阅读理解问题等方面表现出了令人惊讶的能力。
OPT-IML	2022 年 12 月	大语言模型	微调升级版 OPT。
SAM	2023 年 4 月	图像分割模型	可以根据任何用户提示分割任何图像中的任何对象。
DINOv2	2023 年 4 月	视觉大模型	不需要微调训练高性能计算机视觉模型的新方法。
ImageBind	2023 年 5 月	多模态模型	能够将文本、音频、视觉、热量（红外），还有 IMU 数据，嵌入到一个向量空间中。
AnyMAL	2023 年 9 月	多模态模型	可以对不同模态输入内容（文本、图像、视频、音频、IMU 运动传感器数据）实现理解，并生成文本响应。

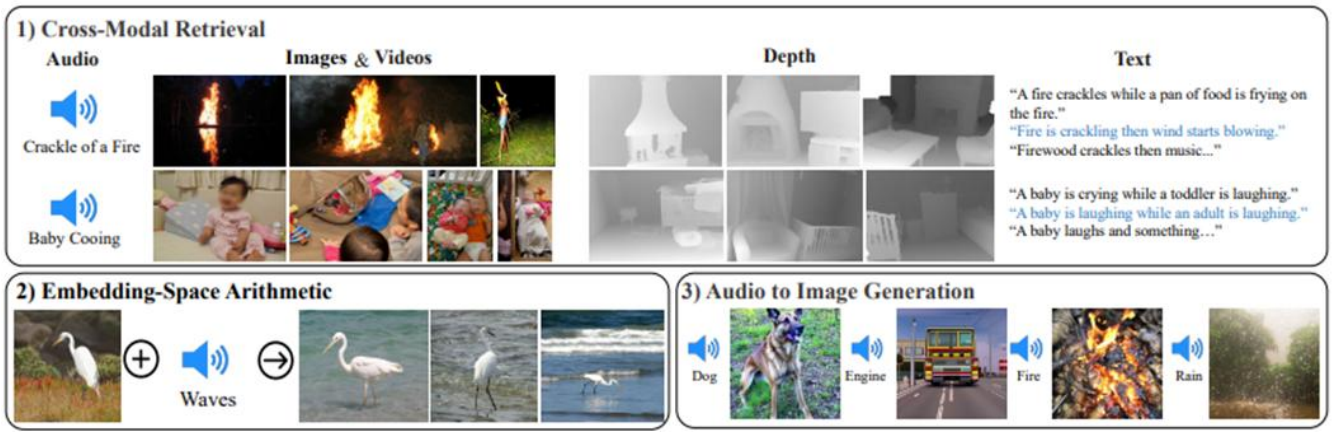
数据来源：Meta，阿里云，新智元，华福证券研究所

2023 年 5 月，Meta 开源多模态大模型 ImageBind，能够将文本、音频、视觉、热量（红外）还有 IMU 数据嵌入到一个向量空间，调动 6 种不同的感知区域进行联

动交流。具体而言，ImageBind 可以实现跨 6 种模态的检索，把不同的模式嵌入叠加，可以自然地构造出它们的语义。比如，ImageBind 可以与 DALL-E 2 解码器和 CLIP 文本一起嵌入，生成音频到图像的映射。从实现方式来看，Meta 将系列开源视觉模型 DINOv2、分割模型 SAM、动画模型 Animated Drawings 结合，目的是给不同模式的学习提供统一的特征空间，并可利用 DINOv2 的强大视觉特征进一步提高其能力，训练图像匹配数据能力，增强六种模态绑定关系。

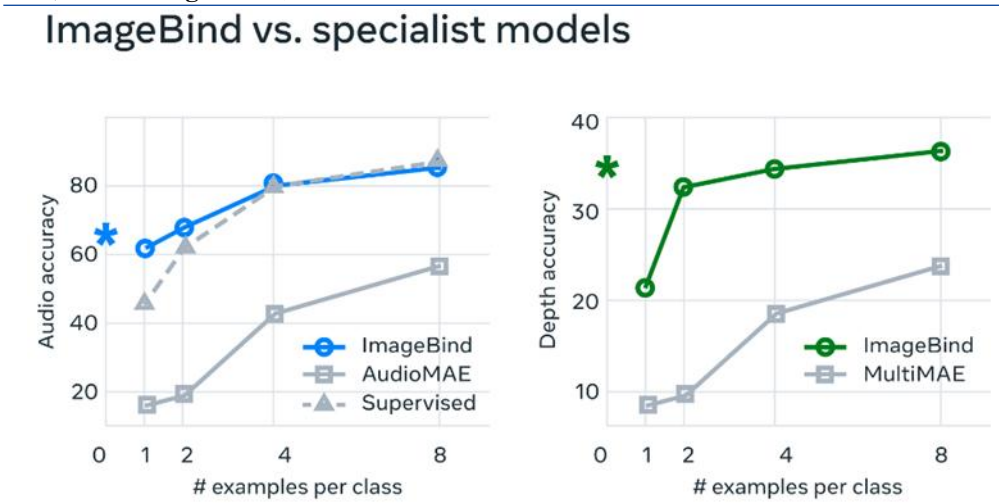
ImageBind 的特征可以用于少样本音频和深度分类任务，并且可以胜过专门针对这些模态的先前方法。对比专家模型，ImageBind 在少于四个样本分类的 top-1 准确率上，要比 Meta 的自监督 AudioMAE 模型和在音频分类 fine-tune 上的监督 AudioMAE 模型提高了约 40%的准确率，ImageBind 还在跨模态的新兴零样本识别任务上取得了新的最先进性能。

图表 29：ImageBind 为跨越六种模态的大模型



数据来源：Meta AI《IMAGEBIND: One Embedding Space To Bind Them All》，华福证券研究所

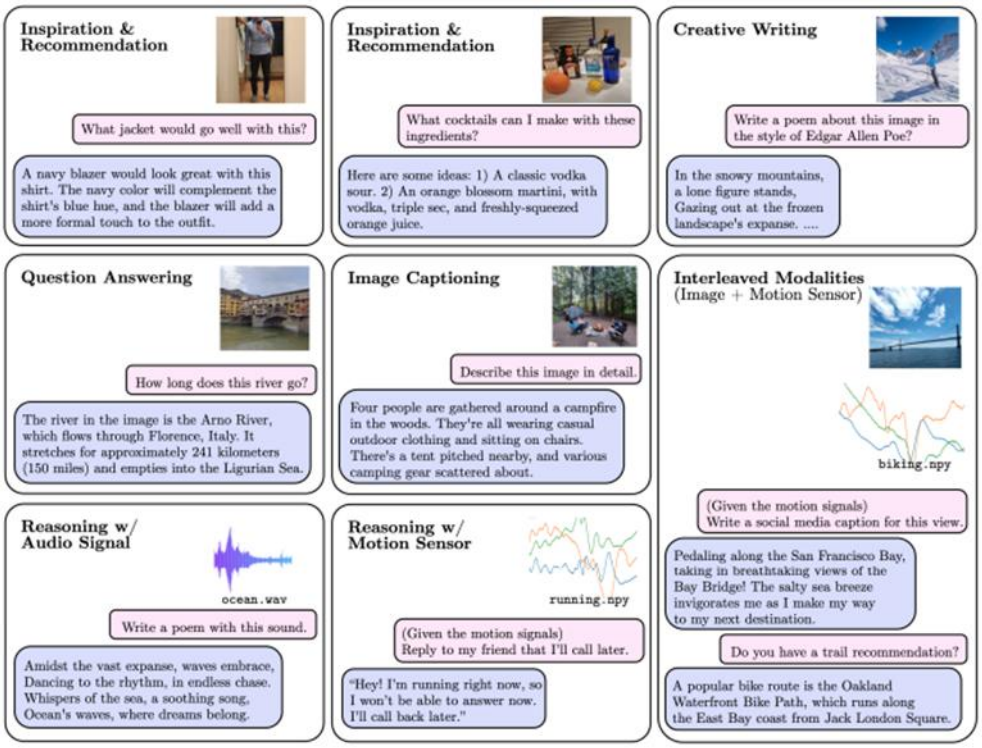
图表 30：ImageBind 在音频和深度方面优于专家模型



数据来源：Meta，华福证券研究所

Meta 发布多模态大模型 AnyMAL，进一步巩固开源大模型竞争力。2023 年 9 月，Meta 发布 AnyMAL，模型基于多模态版 Llama 2 技术底座，是一个经过训练的多模态编码器集合，可将来自各种模态（包括图像、视频、音频和 IMU 运动传感器数据）的数据转换到 LLM 的文本嵌入空间。与现有文献中的模型相比，该模型在各种任务和模式的自动和人工评估中都取得了很好的零误差性能，在 VQAv2 上提高了 7.0%的相对准确率，在零误差 COCO 图像字幕上提高了 8.4%的 CIDEr，在 AudioCaps 上提高了 14.5%的 CIDEr，创造了新的 SOTA。**我们认为，AnyMAL 发布将完善 Meta 多模态大模型生态，并加快多模态大模型在开源社区的开发节奏。**

图表 31：AnyMAL 多模态输出示例



数据来源：FAIR, Meta & Meta Reality Labs《AnyMAL: An Efficient and Scalable Any-Modality Augmented Language Model》，华福证券研究所

2.4 国内加速对齐海外龙头，细分领域或有优势

2022 年底，OpenAI 发布 ChatGPT，掀起全球大模型关注热度，2023 年 3 月，OpenAI 再次发布多模态大模型 GPT-4，将大模型竞争切入多模态赛道。国内科技公司积极研发国产大模型，互联网大厂在数据积累与算法水平兼具优势，率先切入多模态大模型赛道，其后不断涌现大模型科技公司与初创公司，在多模态大模型领域持续投入同时陆续更新大模型能力。例如，百度 2023 年 3 月发布文心一言，成为全球大厂中第一个对标 ChatGPT 甚至是 GPT-4 的大模型，同时具备文字生成图片、音频（方言）、视频等多模态能力。其后，阿里巴巴、腾讯等互联网大厂，商汤科技等大模型公司以及智源研究院、智谱等初创公司或研究所均发布了国产多模态大模型，并通过不断迭代实现能力突破，逐步缩小与海外大模型差距。

然而，总体而言，**我们认为，由于国产训练数据集、算力支持和应用场景等与海外大厂仍存在较大差距，国产大模型仍在向海外大厂靠齐过程中。**根据 SuperCLUE 测评数据，截至 2023 年 12 月，海外 GPT-4 Turbo、GPT-4 依旧为全球性能最优大模型。

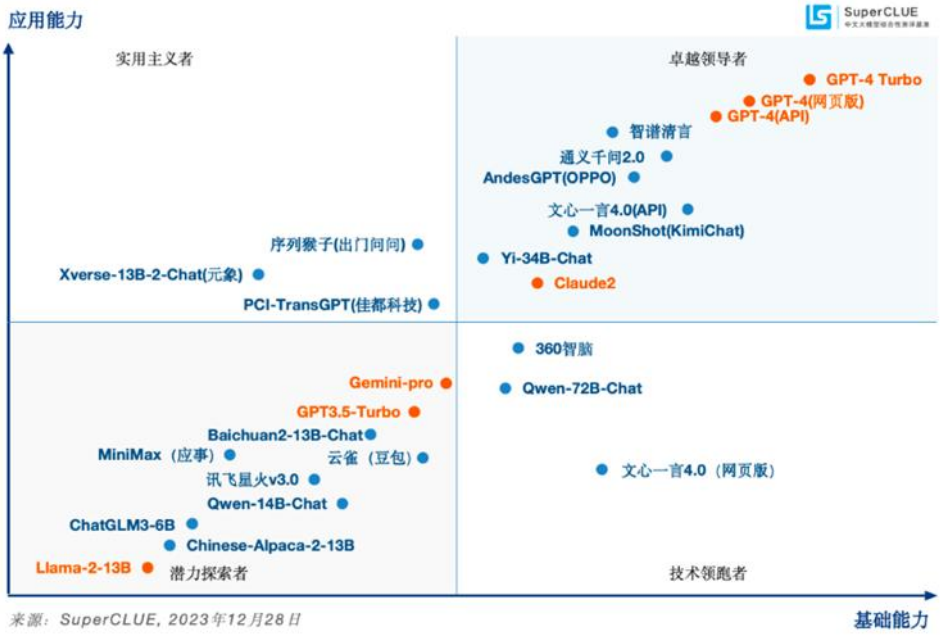
图表 32：我国部分多模态大模型梳理

类别	公司名称	发布时间	代表模型
互联网公司	百度	2023 年 3 月	文心一言
	阿里巴巴	2023 年 4 月	通义千问
		2023 年 11 月	mPLUG-Owl2
	腾讯	2023 年 9 月	混元大模型
	字节跳动	2023 年 9 月	BuboGPT
		2023 年 12 月	PixelLM
大模型公司	商汤科技	2023 年 3 月	书生 2.5
	昆仑万维	2023 年 4 月	天工大模型
	科大讯飞	2023 年 5 月	讯飞星火
其他	智源研究院	2023 年 12 月	Emu2
	智谱 AI	2024 年 1 月	GLM-4

零一万物	2024 年 1 月	Yi-VL

数据来源：界面新闻，阿里云，腾讯网，电商报 Pro，Alibaba Group，澎湃，光锥智能，量子位，新浪 XR，智东西，新浪新闻，商汤科技，智源研究院，零一万物，华福证券研究所

图表 33：国产大模型与海外龙头厂商仍有差距



数据来源：SuperCLUE，元宇宙新声，华福证券研究所（注：其中潜力探索者代表模型正在技术探索阶段拥有较大潜力；技术领跑者代表模型聚焦基础技术研究；实用主义者代表模型在场景应用上处于领先定位；卓越领导者代表模型在基础和场景应用上处于领先位置，引领国内大模型发展）

我们认为，国产多模态大模型目前存在两类发展路径：

1）海外龙头厂商具有示范效应，Meta 等厂商算法开源显著降低国产大模型学习成本，国产大模型可通过复制海外龙头厂商先进技术快速成长，通过逐步超越海外龙头上代产品，并摸索最新技术的方式升级迭代：

阿里巴巴最新通义千问可媲美 GPT-4V 和 Gemini。2023 年 8 月，阿里发布 Qwen-VL 模型的第一个版本，并很快对通义千问进行了升级。Qwen-VL 支持以图像、文本作为输入，并以文本、图像、检测框作为输出，让大模型真正具备了看世界的能力。在多模态大模型性能整体榜单 OpenCompass 中，Qwen-VL-Plus 紧随 Gemini Pro 和 GPT-4V，占据了前三名的位置。2024 年 1 月，阿里巴巴新升级的通义千问视觉语言大模型 Qwen-VL-Max 发布，在多个测评基准上取得较好成绩，并实现了强大的图像理解能力，整体能力达到了媲美 GPT-4V 和 Gemini 的水平，在多模态大模型领域实现了业内领先。

图表 34：阿里通义千问多模态大模型测试性能媲美 GPT-4V 和 Gemini

模型	文档理解	图标理解	科学图例	文字阅读	多学科问题	数学推理	中文问答
Other Best Open-source LVLM	81.6% (CogAgent)	68.4% (CogAgent)	73.7% (Fuyu-Medium)	76.1% (CogAgent)	45.9% (Yi-VL-34B)	36.7% (SPHINX-V2)	72.4% (InternLM-Xcomposer-VL)
Gemini Pro	88.1%	74.1%	73.9%	74.6%	47.9%	45.2%	74.3%
Gemini Ultra	90.9%	**80.8%**	**79.5%**	**82.3%**	**59.4%**	**53.0%**	-
GPT-4V	88.4%	78.5%	78.2%	78.0%	56.8%	49.9%	73.9%
Qwen-VL-Plus	91.4%	78.1%	75.9%	78.9%	44.0%	43.3%	68.0%
Qwen-VL-Max	**92.5%**	**79.8%**	**79.3%**	**79.5%**	**50.8%**	**50.0%**	**75.1%**

数据来源：机器之心，华福证券研究所（注：红色、蓝色与黄色数字分别表示排名第一、第二、第三）

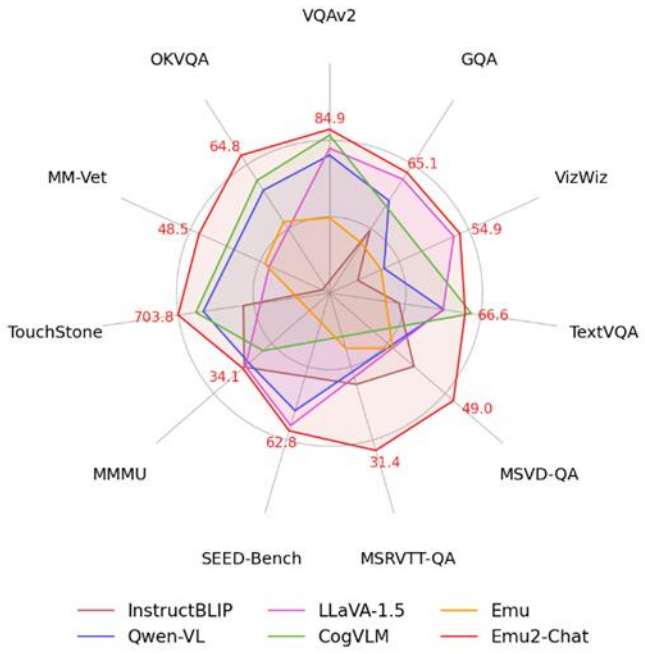
智谱 AI 发布多模态大模型 GLM-4，模型性能均达 GPT-4 九成以上。作为国内唯一一个产品线全对标 OpenAI 的大模型公司，GLM-4 性能相比 GLM-3 提升 60%，逼近 GPT-4（11 月 6 日最新版本效果）。多模态能力方面，GLM-4 则是把原本就有的文生图（CogView3）、代码能力做了升级，CogView3 效果超过开源最佳的 Stable Diffusion XL，逼近 DALLE·3。

图表 35：智谱 CogView3 效果逼近 DALLE·3

数据来源：大模型之家，华福证券研究所

2023 年 12 月，智源研究院开源发布新一代多模态基础模型 Emu2，**成为目前最大的开源生成式多模态模型，通过大规模自回归生成式多模态预训练，显著推动多模态上下文学习能力的突破。**Emu2 在少样本多模态理解任务上大幅超越 Flamingo-80B、IDEFICS-80B 等主流多模态预训练大模型，在包括 VQAv2、OKVQA、MSVD、MM-Vet、TouchStone 在内的多项少样本理解、视觉问答、主体驱动图像生成等任务上取得最优性能。Emu2-Chat 可以精准理解图文指令，实现更好的信息感知、意图理解和决策规划。Emu2-Gen 可接受图像、文本、位置交错的序列作为输入，实现灵活、可控、高质量的图像和视频生成。

图表 36：Emu2 在十余个图像和视频问答评测集上取得最优性能



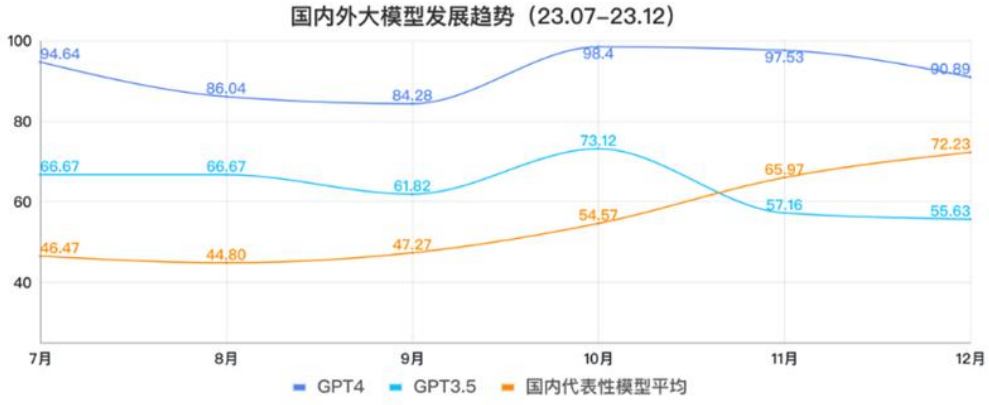
数据来源：智源研究院，华福证券研究所

2）国产大模型有望凭借独特生态优势在细分领域取得差异化竞争优势。

百度 2023 年 3 月发布的文心一言，其训练数据包含万亿级网页数据、数十亿的搜索数据和图片数据、百亿级的语音日均调用数据，以及 5500 亿事实的知识图谱等，在搜索领域或具有技术与数据优势；阿里巴巴 2023 年 4 月发布的通义千问训练数据包括大量文本、专业书籍、代码等，生成的大模型或在电商领域具有较强竞争力。

总体而言，通过向海外技术对齐和利用独特生态禀赋，国产大模型与海外大厂差距逐步缩小。根据 SuperCLUE 测评数据，在 2023 年下半年，国内领军大模型企业实现了大模型代际追赶的奇迹，从 7 月份与 GPT3.5 的 20 分差距，每个月都有稳定且巨大的提升，到 11 月份测评时已经完成总分上对 GPT3.5 的超越。

图表 37：国产大模型与海外大模型差距逐步缩小



数据来源：SuperCLUE，元宇宙新声，华福证券研究所

3 多模态提升大模型泛化能力，应用场景拓展性强

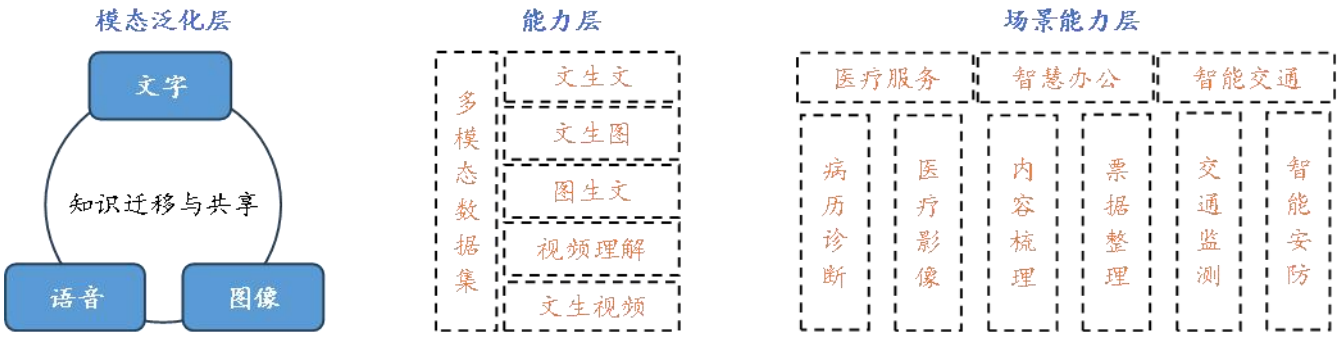
我们认为，大模型不仅仅是技术手段，更是推动业务创新和提升竞争力的有力工具。强调技术与业务的融合以推动业务的数字化转型和智能化升级，才能够最大化的发挥大模型价值同时激励大模型创新升级，实现业务效率提升与技术创新的良性循环。

多模态提升大模型泛化能力，在多元信息环境下实现“多专多能”，极大地解放了生产力。大模型可以用多模态数据进行跨模态学习，从而提升其在多个感知任务

上的性能和表现。多模态大模型将单一的文本拓展至图像、音频、视频等丰富信息环境中，通过多模态信息整合与数据挖掘，解析世界的本来面貌，具备更强的泛化能力。

多模态大模型在垂直领域具有广阔的应用场景和市场价值。多模态大模型的应用场景和价值正在不断扩展和提升。从语音识别、图像生成、自然语言理解、视频分析，到机器翻译、知识图谱、对话系统、内容创作，多模态大模型都能够提供更丰富、更智能、更人性化的服务和体验，在垂直领域的商业应用前景广阔。在强大泛化能力基础上，大模型可以在不同模态和场景之间实现知识的迁移和共享，将大模型的应用扩展到不同的领域和场景。例如，在办公领域，可以借助多模态文生文、音频理解、图生文等能力整理会议纪要与提升报销效率等。

图表 38：多模态大模型可灵活部署于垂直场景



数据来源：大模型之家，华福证券研究所

3.1 通用多模态大模型积极开放，挖掘垂直场景广阔空间

通用多模态大模型可通过开源与放开 API 入口两种方式推动垂直场景数智化转型升级。一方面，以 Meta 为代表的海外龙头坚持多模态大模型开源，以此助力开发者快速获取多模态大模型能力；另一方面，以 OpenAI 为代表的厂商向开发者开放通用大模型 API 调用入口。目前，OpenAI 已开放了通用大模型 GPT 系列、图文转换 DALL·E、音频转文本 Whisper 等产品的 API 调用入口，大幅降低垂直应用多模态能力开发门槛，推动垂直场景应用开发。

图表 39：调用 GPT API 客户梳理

公司名称	垂直领域	调用模型	时间	业务介绍	API 用途
Inworla AI	游戏	GPT3	2023 年 1 月	快速生成游戏 NPC	创建下一代 AI 驱动的角色
Stripe	电商	GPT4	2023 年 3 月	为电子商务网站和移动应用程序提供支付处理软件及相关产品功能集成	简化用户体验并打击欺诈行为
Duolingo	教育	GPT-4	2023 年 3 月	开发学习外语的应用软件	GPT-4 加强了 Duolingo 上的对话。
Khan Academy	教育	GPT-4	2023 年 3 月	开发人工智能学习助手，既可以充当学生的虚拟导师，也可以充当教师的课堂助手	能够理解自由形式的问题和提示，向每个学生提出个性化的问题以促进更深入的学习。
Morgan Stanley	金融	GPT4	2023 年 3 月	财富管理领域的领导者	财富管理部门部署 GPT-4 来组织其庞大的知识库。
Ironclad	办公	GPT4	2023 年 11 月	合同生命周期管理平台	简化合同审查流程
Wix	IT	GPT	2024 年 1 月	为用户提供专业的网络建站工具	简化网站内容创建

数据来源：OpenAI 官网，华福证券研究所

OpenAI 宣布 API 调用降价，垂直领域 AI 应用公司大模型训练成本显著下降。2023 年 11 月，OpenAI 召开首次开发者大会，现场发布 GPT-4 Turbo，重点围绕升级、

降价与生态三大内容展开。GPT-4 Turbo 拥有 128K 更长的上下文窗口，在多模态能力、放宽速率限制等方面均实现了升级。降价方面，输入/输出端降价 2-3 倍。以 1000tokens 为例，GPT-4 Turbo 的输入比 GPT-4 便宜 3 倍，为 0.01 美元，输出便宜 2 倍，为 0.03 美元；GPT-3.5 Turbo 输入比之前的 16K 型号便宜 3 倍，为 0.001 美元，输出便宜 2 倍，为 0.002 美元。我们认为，大模型更强的功能与更低的算力成本，将提高下游应用端开发功效，垂直场景将极大程度上受益。

图表 40：GPT 大模型降价前后对比

	旧模型	新模型	成本降幅
GPT-4 Turbo	GPT-4 8K Input: $0.03 Output: $0.06	GPT-4 Turbo 128K Input:$0.01 Output: $0.03	Input: 67% Output: 50%
	Gpt-4 32K Input: $0.06 Output: $0.12		
GPT-3.5 Turbo	GPT-3.5 Turbo 4K Input: $0.0015 Output: $0.002	GPT-3.5 Turbo 16K Input:$0.001 Output: $0.002	Input:33% Output: 0%
	Gpt-3.5 Turbo 16K Input: $0.003 Output: $0.004		
GPT-3.5 Turbo fine-tuning	GPT-3.5 Turbo 4K fine-tuning Training: $0.008 Input:$0.012 Output: $0.016	GPT-3.5 Turbo 4K and 16K fine-tuning Training: $0.008 Input:$0.003 Output: $0.006	Training: 0% Input: 75% Output: 63%

数据来源：OpenAI 官网，华福证券研究所

OpenAI 正式开放 GPT Store，有望形成 AIGC 应用生态，加速垂直领域应用开发。1 月 10 日，Open AI 官宣 GPT Store 正式上线，目前已有超 300 万个自定义 ChatGPT，奥特曼称其为“AI 领域的 Apple Store”，这些 GPTs 涉及众多领域，从开发工具、生产力、做图、语言学习，到客户支持、市场分析、医疗健康等应有尽有，能够满足工作生活中绝大多数场景。

3.2 AI+办公：重塑办公模式，解放员工生产力

微软全面接入多模态大模型 GPT-4 系列，借助多模态能力大幅提高用户办公效率。

2023 年 3 月，GPT-4 全面接入微软 Office 全家桶。

2023 年 9 月微软召开新品发布会，重磅更新 AI Copilot 功能。B 端：计划在 2023 年 11 月 1 日推出全球版 Microsoft 365 Copilot，适用于 Microsoft 365 E3、E5、商业标准版和商业高级版等 B 端客户，MS 365 Copilot 提高了 Office 办公、Outlook 邮件处理、Teams 会议记录等能力；C 端：9 月 26 日 Windows11 更新将提供 150 多项新功能，Copilot 随更新免费推出，将把 Copilot 强大的 AI 功能与体验带入 PC 中，并融入到画图、照片、Clipchamp 等应用程序中，同时可结合 Edge、Bing、MS365 等丰富操作系统功能。

今年 1 月，微软正式发布面向个人版的 Copilot Pro，助力 AI+办公规模化推广。用户可访问最新 GPT-4 Turbo，并支持在 PC、Mac、iPad 及手机端的 PPT、Word、Excel 等软件中使用 Copilot，同时对 B 端客户，微软取消 300 人以上大型企业限制，允许中小企业以$30/人/月价格订阅服务。

我们认为，GPT-4 的技术架构帮助提升模型在多种语言表达能力、图像理解能

力，复杂任务处理能力，改善幻觉、安全等局限性等问题，文本的加工和总结能力更强，模型更具备拟人化的能力理解语义，这些能力帮助 AI 在文本生成、文章总结等办公应用领域加速落地，大幅提升办公效率，解放员工生产力。

图表 41：MS365 Copilot 解放员工生产力、提高技能

数据来源：Linked in，华福证券研究所

基础办公领域：微软 Copilot 接入 Microsoft 365 全家桶，帮助提高使用者工作效率；Business Chat 可跨越用户所有的数据和应用程序，通过 AI 技术释放协同办公效率。

图表 42：Microsoft 365 Copilot 应用领域

Copilot 应用领域	功能
Copilot in Word	在人们工作时与他们一起撰写、编辑、总结和创作
Copilot in PowerPoint	在创作过程中，通过自然语言命令将想法转化为设计好的演示文稿
Copilot in Excel	帮助用户释放洞察、识别趋势，或在短时间内创建专业型式的数据可视化
Copilot in Outlook	帮助用户整合并管理收件箱，从而节约出更多时间用于实际沟通
Copilot in Teams	直接从对话上下文中提供实时摘要和待办事项，提高会议效率
Business Chat	汇集了来自文档、演示文稿、电子邮件、日历、笔记和联系人的数据，能够帮助用户总结聊天内容、撰写电子邮件、查找关键日期，甚至根据其他项目文件制定计划

数据来源：微软科技，华福证券研究所

协同办公领域：微软推出全球首个应用于 CRM / ERP 系统的 Dynamics 365 Copilot，让新一代人工智能全方位服务商业应用创新。

图表 43：Dynamics 365 Copilot 在 CRM/ERP 的应用

商业活动	产品	Copilot 功能
营销	Dynamics 365 Sales and Viva Sales	能够帮助销售显著减少花在案头工作上的时间，花更多时间接触客户
客服	Dynamics 365 Customer Service	能帮助服务专员提供更好的客户体验，智能化回答客户问题
市场	Dynamics 365 Customer Insights 和 Dynamics 365 Marketing	能帮助市场推广人员简化数据发现、受众分析、内容创作等环节的工作流程，通过自然语言处理技术准确定位特定客户群体，检索客户类别等
供应链	Dynamics 365 Supply Chain Management	对影响供应链流程事件发出警告；预测受到影响的订单；撰写邮件向合作伙伴发出警告
销售	Dynamics 365 Business Central	能够为电子商务梳理和创建产品列表，几秒钟就能自动生成用于在线商店的商品简介。

数据来源：微软官网，华福证券研究所

微软发挥龙头示范效应，海外 AI 办公产品竞相涌现。Salesforce、谷歌、Adobe 相继发布 AI+办公“拳头产品”。

图表 44：2023 年海外 AI+办公产品梳理

公司	推出时间	产品	功能
Notion	2 月 22 日	Notion AI	帮助用户自动整理笔记、改正错别字、列出文章重点、翻译、制作表格等。
Salesforce	3 月 8 日	Einstein GPT	接入 OpenAI，发布首个 AIGC CRM 产品，可回答客户提问或撰写营销邮件。
微软	3 月 17 日	MS 365 Copilot	为办公场所引入新一代 AI 强大功能，帮助用户释放创造力、解放生产力。
谷歌	5 月 11 日	Duet AI	Workspace 办公套件
Adobe	9 月 13 日	Firefly/Creative Cloud	由 AI 驱动的版本，为用户提供更多创造力。
ChatPDF	-	ChatPDF	PDF AI ，类似于 ChatGPT，但适用于 PDF，免费总结并解答问题。

数据来源：Notion 官网，新浪科技，微软官网，IT 之家，Adobe 官网，ChatPDF 官网，华福证券研究所

3.3 AI+教育：助力教育行业应材施教，促进教育师资均衡

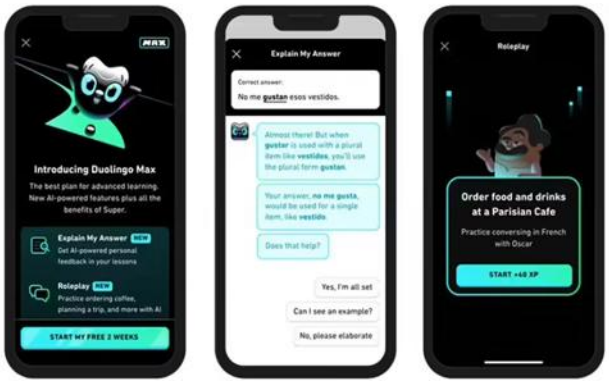
海外加速布局 AI 教育，掀起教育领域新变革。目前海外已形成五类 AI 教育软件：以 Question.AI，Nerd AI 与 Photomath 为主的拍照搜题软件；以 Headway 和 Quizlet 为主的垂直场景类软件；以多邻国为主的语言学习类软件；以 Clever、Toca World 和 Aha World 为主的游戏化学习类软件；以 Lingvano 和 PictureThis 为主的特殊学习类软件。

Duolingo：3 月 14 日，外语学习工具多邻国（Duolingo）宣布，将推出整合了 GPT-4 的语言学习增值服务 Duolingo Max。Duolingo Max 采用订阅制，包含 AI 角色扮演和问答功能，从而模拟真实对话语境或巩固学习成果。

Khan Academy：3 月 14 日，可汗学院（Khan Academy）宣布，它将使用 GPT-4 为其人工智能驱动的助手 Khanmigo 提供动力，Khanmigo 既是学生的虚拟导师，也是教师的课堂助手。

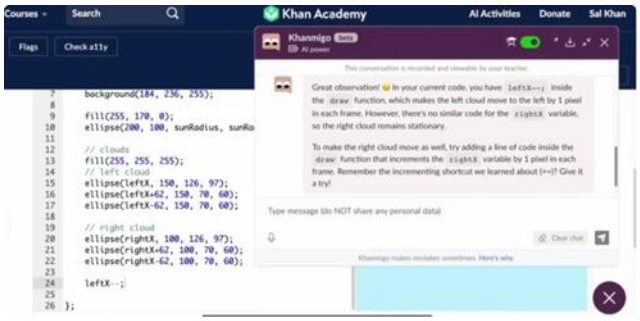
我们认为，AI 助力教育行业应材施教，有效解决教育师资均衡问题。用户同样存在较高粘性，龙头厂商加快布局 AI 产品有望提高自身竞争壁垒。

图表 45：Duolingo Max 产品介绍



数据来源：Duolingo 官网，华福证券研究所

图表 46：Khan Academy 引导学生解决问题



数据来源：多知网，华福证券研究所

3.4 AI+电商：AI 模特换装到 AIGC 赋能运营，全方位渗透电商产业链

AI 模特换装：Stable Diffusion / Midjourney 等借助文生图能力，使得 AI 模特适配服装，从而降低电商服装商品图片素材的生产成本和制作周期。

AIGC 赋能运营：AI 在营销、直播、社交、供应链管理等均存在广泛用途。例如：6 月，谷歌宣布将生成式人工智能引入在线购物领域；9 月，亚马逊生成式 AI 工具上线，帮助卖家更轻松地创建高质量的 listing，并向他们的用户提供更完整、一致、有吸引力的产品信息，从而进一步提升用户的购物体验。

图表 47：Stable Diffusion 应用 AI 对模特换装

数据来源：亚马逊云开发者，华福证券研究所

图表 48：2023 年海外公司利用 AIGC 赋能运营案例

数据来源：网经社，IT 之家，36Kr，AMZ123 跨境电商，华福证券研究所

3.5 AI+医疗：医疗领域数据模态丰富，大模型融入提升效能

医疗领域数据类别丰富，大模型的融合有望提升医疗产品创新能力和医疗健康服务水平。生命科学和医疗领域涵盖医学文本、医学图像、生命组学、蛋白质等多种模态数据。基于不同种类训练数据的大模型能够解决医疗领域多样化、复杂性问题，完成自然语言处理、计算机视觉和图学习等多种任务。

大语言模型：可用于生成医学文本、回答医学问题、提供医学建议等；

视觉大模型和视觉-语言大模型：可用于识别医学图像、生成图像注释、合成图像等；

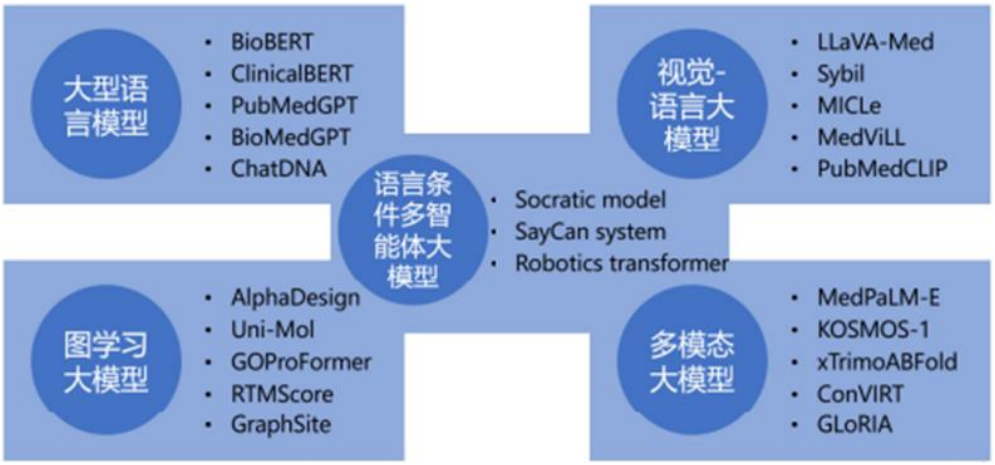
图学习大模型：可用于预测蛋白质结构、设计药物、分析基因组等；

语言条件多智能体大模型：可用于实现远程会诊、智能导诊、医疗机器人等；

多模态大模型：可用于融合多种医学数据、挖掘数据价值、辅助诊断等。以往单模态的模型只能分析疾病某一层面的信息，极大地限制了人工智能的医疗应用，而多模态大模型结合多种模态的医学信息，提高了医疗诊断的准确性，是人工智能

诊疗产品落地的关键。

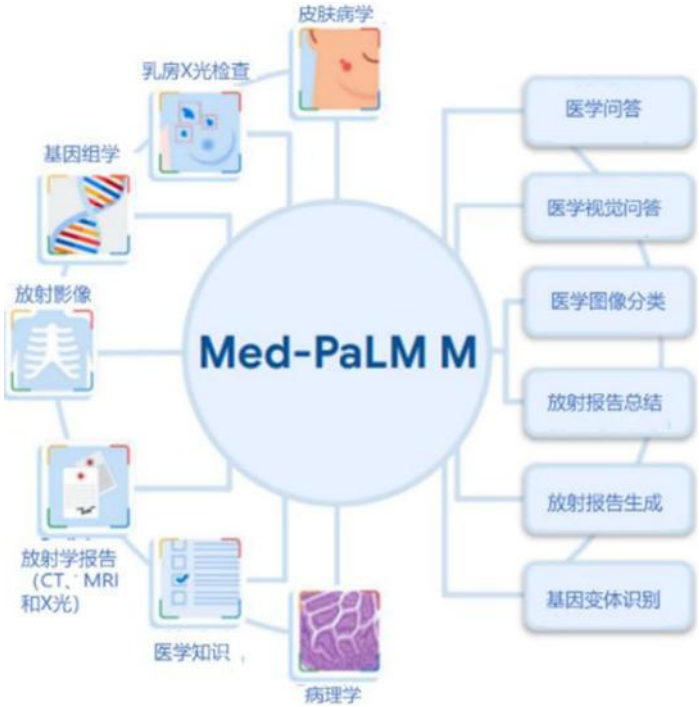
图表 49：医疗健康大模型的类别和实例



数据来源：中国信通院等《人工智能大模型赋能医疗健康产业白皮书（2023 年）》，华福证券研究所

国内外涌现了多个医疗多模态大模型案例，服务于患者诊断、手术导航、康复训练、影像报告生成等场景。例如，谷歌团队研发的 Med-PaLM-M 是一个多模态通用生物医学大模型，可以处理包括文学、医学图像和基因组学数据在内的多种健康数据。Med-PaLM-M 在 14 个不同医疗任务上接近或超过了现有的最先进模型，包括医疗问答、影像分类和基因预测等。

图表 50：Med-PaLM-M 所用基准数据集的模态和任务



数据来源：中国信通院等《人工智能大模型赋能医疗健康产业白皮书（2023 年）》，华福证券研究所

图表 51：国内外部分 AI 医疗大模型梳理

大模型名称	发布时间	企业名称	应用场景	数据类型
盘古药物分子大模型	2022.4	华为	药物研发	文本、图像、化学结构
文心生物计算大模型	2022.5	百度	生物研究	分子结构
BioMedLM	2022.12	斯坦福基础模型研究中心	医疗问答	文本
GatorTron	2023.3	佛罗里达大学	医学问答、病历识别	文本
OpenMEDLab 浦医	2023.6	上海人工智能实验室	文本、生物信息、蛋白质工程	多模态
灵医 Bot	2023.6	百度灵医智惠	文档理解、病历理解、医疗问答	多模态
华佗 GPT	2023.6	深圳市大数据研究院	健康资讯、就医导诊、情感陪伴	多模态
紫东太初	2023.6	中国科学院自动化研究所	手术辅助、辅助诊疗	多模态
京医千询	2023.7	京东健康	辅助诊疗、健康管理、文献挖掘、病例报告生成	
Med-PaLM	2023.7	Google	医疗问答、影像分类、基因预测	多模态

数据来源：郑琰莉等《人工智能大模型在医疗领域的应用现状与前景展望》，中国信通院等《人工智能大模型赋能医疗健康产业白皮书（2023 年）》，腾讯网，AI 新智界，华福证券研究所

4 投资建议

当前，多模态技术发展日臻成熟，国外龙头厂商引领多模态大模型技术前沿方向，国产大模型沿着技术复制与细分领域突破的发展路径逐渐靠齐海外龙头。我们看好具有算法、数据等先发优势的国产大模型厂商，同时多模态提升大模型泛化能力，多元信息环境下实现“多专多能”，在垂直领域具有广阔的应用场景和市场价值。

建议关注：

1）AI+多模态：万兴科技、中科创达、虹软科技、当虹科技、大华股份、海康威视、漫步者、萤石网络、汉仪股份、美图公司、云从科技；

2）AI+办公：金山办公、万兴科技、福昕软件、彩讯股份、金蝶国际、泛微网络、致远互联、鼎捷软件、汉得信息，用友网络；

3）AI+教育/电商/医疗：科大讯飞、佳发教育、鸥玛软件、盛通股份、光云科技、值得买、焦点科技、小商品城、润达医疗、嘉和美康、创业慧康、迪安诊断等。

5 风险提示

1）技术发展不及预期：AI 多模态技术发展未能取得新的突破。

2）产品落地不及预期：垂直领域产品推出速度缓慢，商业化进行较慢。

3）AI 伦理风险：AI 技术滥用导致的数据安全、隐私安全等问题。

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