国元证券 GUOYUAN SECURITIES

计算机行业

投资评级

推荐 维持

生成式AI加速创新，行业迎历史性机遇

——生成式人工智能行业专题研究：海外大模型篇

证券研究报告 2024年3月29日

分析师：耿军军

邮箱：gengjunjun@gyzq.com.cn

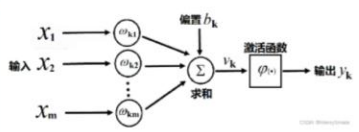
SAC执业资格证书编码：S0020519070002

联系人：王朗

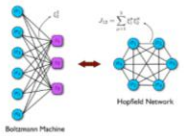
邮箱：wanglang2@gyzq.com.cn

目录

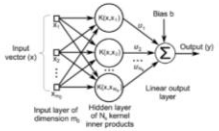
1.1 发展历程：算法模型持续迭代，AI行业快速发展



1943年，Warren McCulloch和Water Pitts提出神经元网络的数学模型。



1982年，John Hopfield 发明了霍普菲尔德网络，这是最早的RNN的雏形，它的出现振奋了神经网络领域。



1995年，Cortes和Vapnik提出联结主义经典的支持向量机。

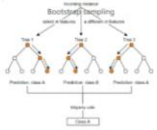
2012年，深度学习模型AlexNet引发深度学习领域革命，由Alex Krizhevsky、Ilya Sutskever和Geoffrey Hinton共同研发

2016年，AlphaGo与围棋世界冠军李世石进行围棋人机大战，以4比1的总比分获胜。

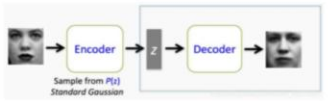
2022年11月，OpenAI发布ChatGPT。

1950年，图灵提出著名的 “图灵测试”，给出判定机器是否具有 “智能 “的方法。

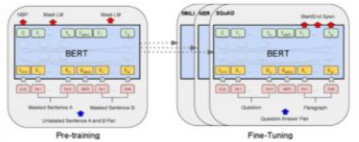
1985年，Pearl提出贝叶斯网络(Bayesian network)。



2001年，布雷曼博士提出随机森林（Random Forest）。



2013年，深度学习算法变分自编码器（VAE）被提出

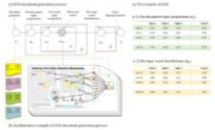


2018年，Google提出基于Transformer注意力机制的Bert模型。

2023年3月，OpenAI发布GPT-4。

1968年，Edward Feigenbaum提出首个专家系统，孕育第二次AI浪潮。

1986年，Geoffrey Hinton等人提出了多层感知器(MLP)与反向传播（BP）训练相结合理念，开启神经网络新一轮高潮。

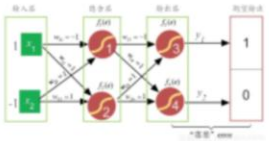


2003年，David Blei, Andrew Ng和Michael I. Jordan提出LDA（Latent Dirichlet Allocation）。

2014年，Ian J.Goodfellow提出生成式对抗网络GAN

2020年，OpenAI发布大语言模型GPT-3。

2023年2月，Meta开源大语言模型LLama

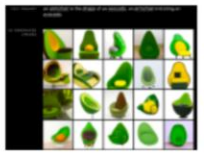


1974年，Paul Werbos首次提出了通过误差的反向传播(BP)来训练人工神经网络。

1989年，LeCun 发明了卷积神经网络（Convolutional Neural Network，CNN）。

2006年，杰弗里·辛顿及学生正式提出了深度学习的概念(Deeping Learning)。

2015年，Sam Altman等创建OpenAI。



2021年，OpenAI提出连接文本与图像的神经网络：DALL·E 和Clip。

2024年2月，OpenAI发布AI视频生成模型Sora。

20世纪50年代 — 90年代中期 — 21世纪10年代

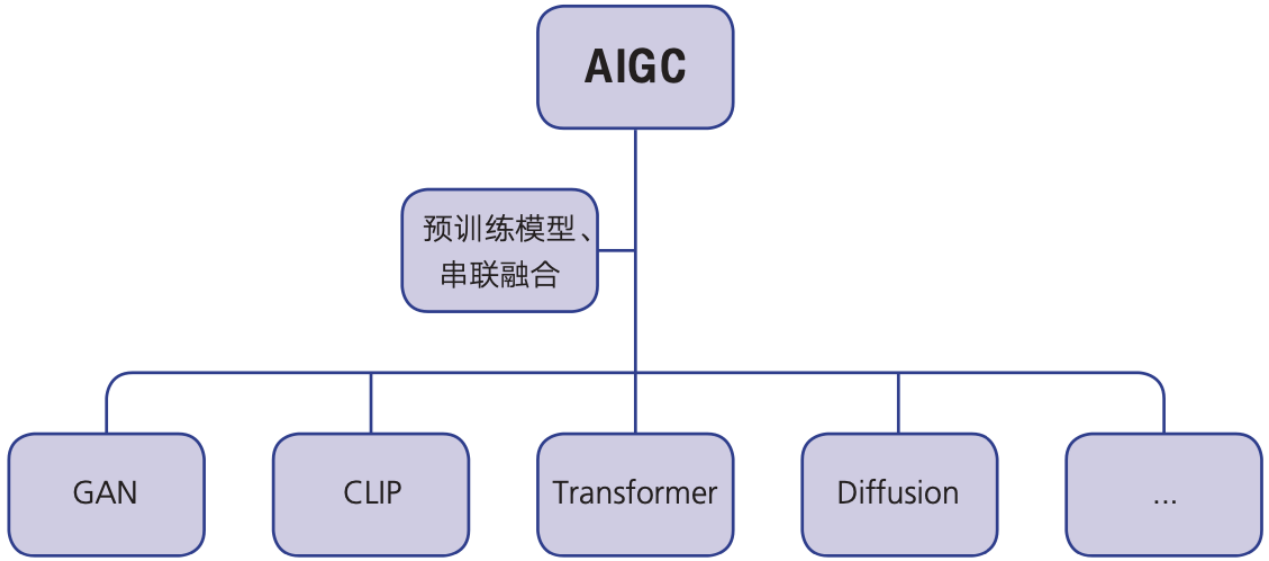
AI技术萌芽阶段　AI技术沉淀积累阶段　AI技术快速发展阶段

资料来源：信通院《人工智能生成内容(AIGC)白皮书》，CSDN官网，阿里云开发者社区，NIH Record官网，MIT官网，51CTO官网，机器之心官网，腾讯云开发者社区，科技行者官网，雷锋网，澎湃新闻网，winbuzzer官网，MBA百科，Geekwire官网，datamarketinglabs官网，安全客官网，AIGC开放社区公众号，IT之家官网，OpenAI官网，36氪官网，国元证券研究所

1 基础的生成算法模型是驱动AI的关键

- 2014年，伊恩·古德费洛(Ian Goodfellow)提出的生成对抗网络(Generative Adversarial Network, GAN)成为早期最为著名的生成模型。GAN使用合作的零和博弈框架来学习，被广泛用于生成图像、视频、语音和三维物体模型。随后，Transformer、基于流的生成模型(Flow-based models)、扩散模型(Diffusion Model)等深度学习的生成算法相继涌现。
- Transformer模型是一种采用自注意力机制的深度学习模型，这一机制可按输入数据各部分的重要性分配权重，可用于自然语言处理(NLP)、计算机视觉(CV)领域应用，后来出现的BERT、GPT-3、laMDA等预训练模型都是基于Transformer模型建立的。

图：AIGC技术累积融合



资料来源：腾讯研究院《AIGC发展趋势报告》，国元证券研究所

1 基础的生成算法模型是驱动AI的关键

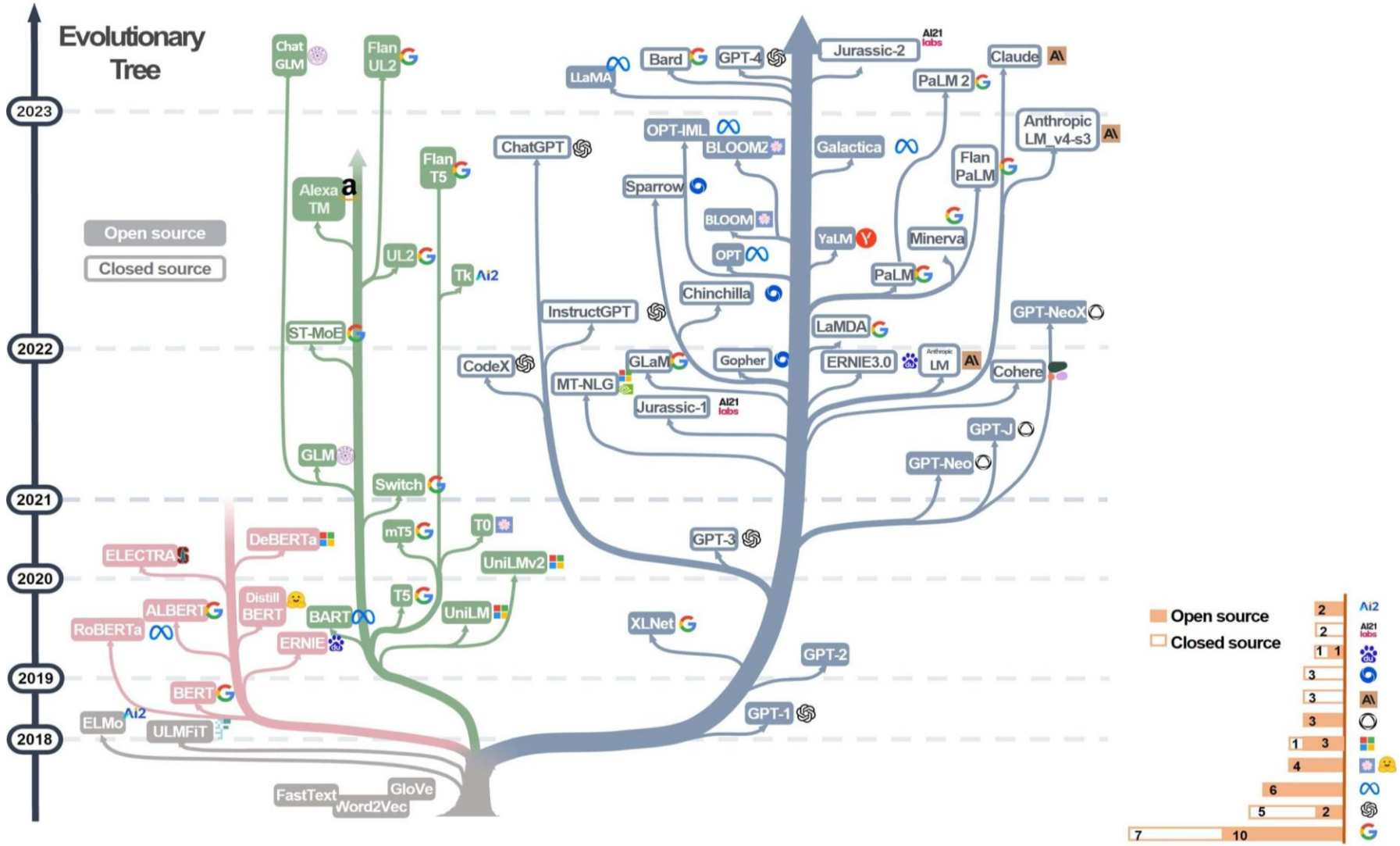
表：主流生成模型一览表

模型	提出时间	模型描述
变分自动编码(Variational Autoencoders, VAE)	2014年	基于变分下界约束得到的Encoder-Decoder模型对。
生成对抗网络(GAN)	2014年	基于对抗的Generator-Discriminator模型对。
基于流的生成模型(Flow-based models)	2015年	学习一个非线性双射转换(bijective transformation)，其将训练数据映射到另一个空间，在该空间上分布是可以因子化的，整个模型架构依靠直接最大化log-likelihood来完成。
扩散模型(Diffusion Model)	2015年	扩散模型有两个过程，分别为扩散过程和逆扩散过程。在前向扩散阶段对图像逐步施加噪声，直至图像被破坏变成完全的高斯噪声，然后在逆向阶段学习从高斯噪声还原为原始图像的过程。经过训练，该模型可以应用这些去噪方法，从随机输入中合成新的“干净”数据。
Transformer模型	2017年	一种基于自注意力机制的神经网络模型，最初用来完成不同语言之间的文本翻译任务，主体包含Encoder和Decoder部分，分别负责对源语言文本进行编码和将编码信息转换为目标语言文本。
神经辐射场(Neural Radiance Field, NeRF)	2020年	提出了一种从一组输入图像中优化连续5D神经辐射场的表示（任何连续位置的体积密度和视角相关颜色）的方法，要解决的问题就是给定一些拍摄的图，如何生成新的视角下的图。
CLIP(Contrastive Language-Image PreTraining)模型	2021年	1) 进行自然语言理解和计算机视觉分析； 2) 使用已经标记好的“文字-图像”训练数据。一方面对文字进行模型训练，一方面对图像进行另一个模型的训练，不断调整两个模型的内部参数，使得模型分别输出的文字特征和图像特征值确认匹配。
DiT(Diffusion Transformers)模型	2023年	用Transformer替换了传统的U-Net主干，在潜在空间中对图像进行建模，并通过Transformer的注意力机制学习图像的全局依赖关系，具有良好的可扩展性，可以训练到更高的分辨率和更大的模型容量。

资料来源：腾讯研究院《AIGC发展趋势报告》，经纬创投公众号，国元证券研究所

1 基础的生成算法模型是驱动AI的关键

- 通过梳理全球主流大语言模型(LLM)的发展脉络，2018年以来的GPT系列、LLaMA系列、BERT系列、Claude系列等多款大模型均发源于Transformer架构。



 资料来源：虎嗅APP公众号，国元证券研究所

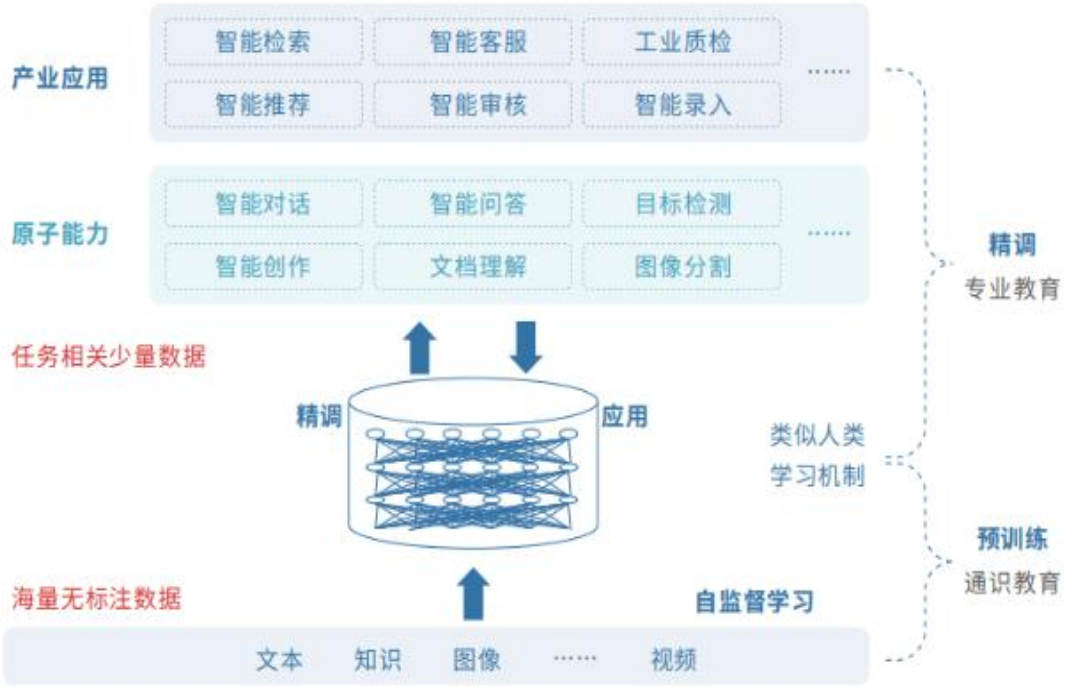
2 预训练模型引发了AI技术能力的质变

- 预训练模型是为了完成特定任务基于大型数据集训练的深度学习模型，让AI模型的开发从手工作坊走向工厂模式，加速AI技术落地。

2017年，Google颠覆性地提出了基于自注意力机制的神经网络结构——Transformer架构，奠定了大模型预训练算法架构的基础。

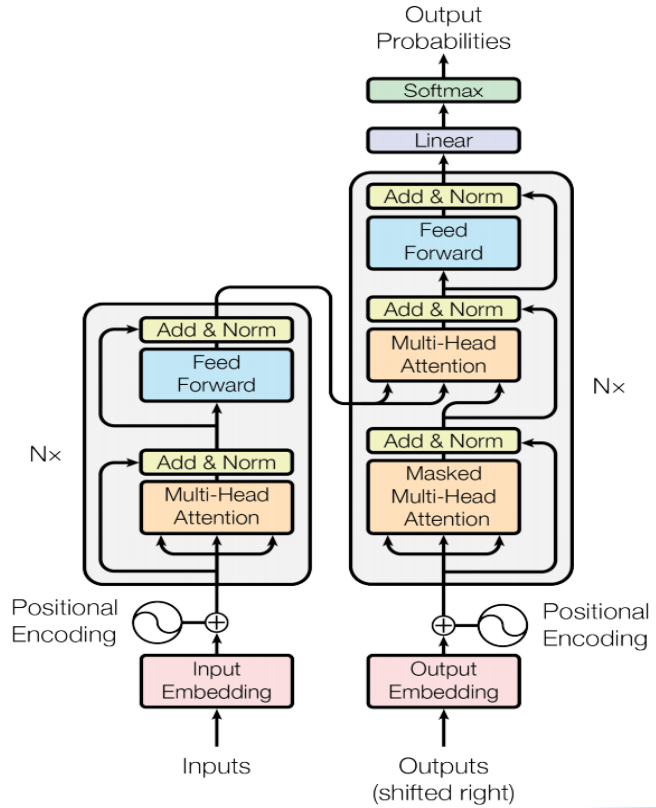
2018年，OpenAI和Google分别发布了GPT-1与BERT大模型，意味着预训练大模型成为自然语言处理领域的主流。

图：预训练相当于“通识教育”



资料来源：IDC《2022中国大模型发展白皮书》，国元证券研究所

图：Transformer模型结构



资料来源：CSDN官网，国元证券研究所

2 预训练模型引发了AI技术能力的质变

表：海外主要预训练大模型汇总

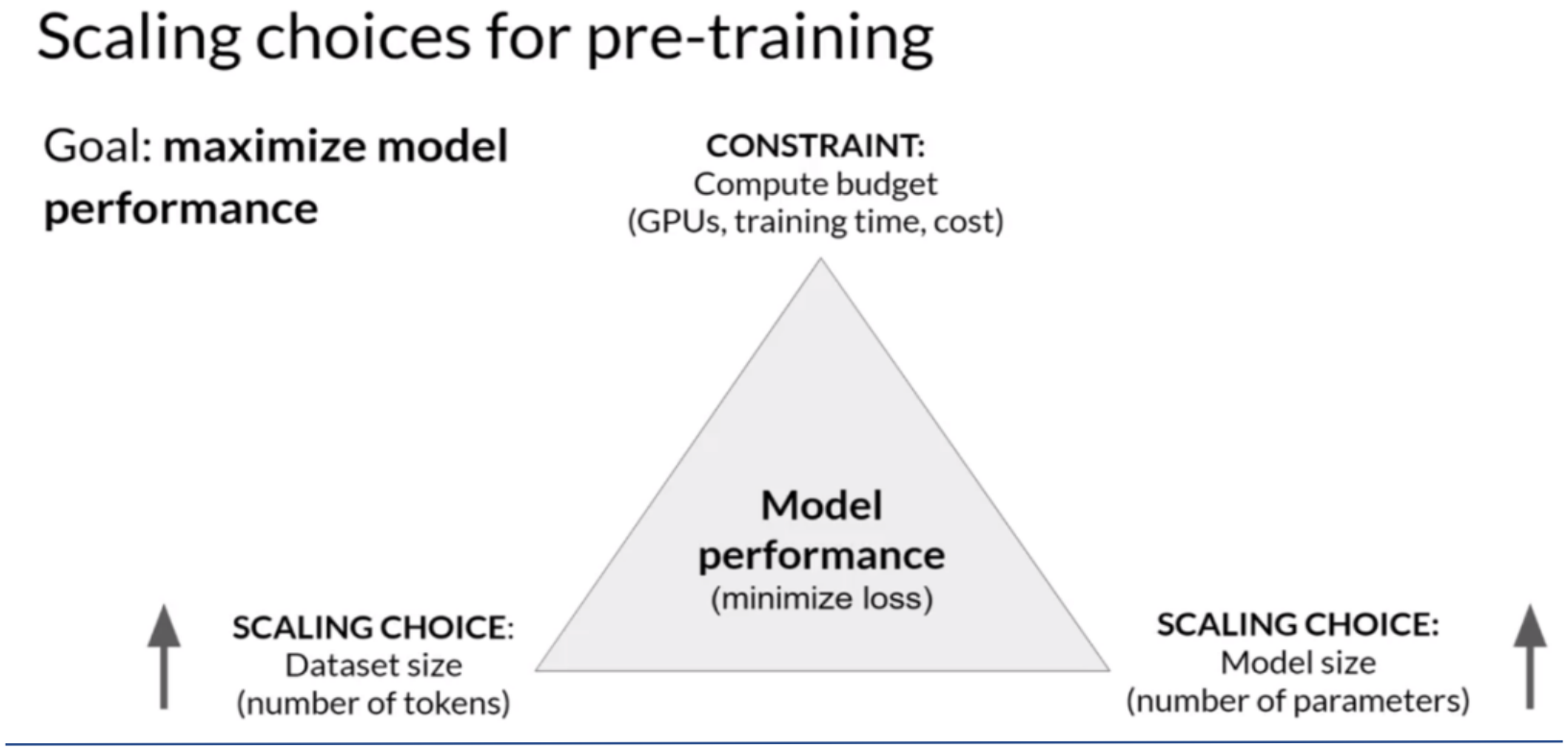
开发者	预训练模型	应用	参数量	领域	开发者	预训练模型	应用	参数量	领域
谷歌	Gemini 1.5	图像、文本、视频、音频和代码理解，生成文本等		多模态	DeepMind	Gato	多面手的智能体	12亿	多模态
	Gemini	图像、文本、视频、音频和代码理解，生成文本等		多模态		Gopher	语言理解与生成	2800亿	NLP
	BERT	语言理解与生成	4810亿	NLP		AlphaCode	代码生成	414亿	NLP
	LaMDA	对话系统		NLP	OpenAI	GPT4	图像与文本理解、文本生成等		多模态
	PaLM	语言理解与生成、推理、代码生成	5400亿	NLP		GPT3	语言理解与生成、推理等	1750亿	NLP
	Imagen	语言理解与图像生成	110亿	多模态		CLIP&DALL-E	图形生成、跨模态检索	120亿	多模态
	Parti	语言理解与图像生成	200亿	多模态		Codex	代码生成	120亿	NLP
微软	Florence	视觉识别	6.4亿	CV		ChatGPT	语言理解与生成、推理等		NLP
	Turing-NLP	语言理解、生成	170亿	NLP	英伟达	Megatron	语言理解与生成	5300亿	NLP
Facebook	OPT-175B	语言模型	1750亿	NLP		Turing NLP			
	M2M-100	100种语言互译	150亿	NLP	Stability AI	Stable Diffusion	语言理解与图像生成		多模态
Meta	LLaMA	语言理解与生成	70-650亿	NLP	Anthropic	Claude	语言理解与生成等		NLP
	LLaMA 2	语言理解与生成	70-700亿	NLP		Claude 2	语言理解与生成、编程、推理等		NLP
	SAM	图像分割	10亿	CV		Claude 3	语言理解与生成、编程、推理、图片理解等		多模态

资料来源：腾讯研究院《AIGC发展趋势报告》，智东西公众号，澎湃新闻网，Llama family官网，百度云智能官网，Datalearner官网，AIGC开放社区公众号，机器之心公众号，CSDN官网，国元证券研究所

3 预训练数据直接决定AI大模型性能

- Scaling Laws：模型容量、数据量、训练成本共同构成了大模型训练的不可能三角。大模型训练的目标是最大化模型性能，模型训练成本（GPU的数量和训练时间等）是受限的，因此一般通过增加数据集大小和增加模型中的参数量两种途径来提升模型性能。

图：扩展大模型的三个选项：模型容量、数据量、训练成本



资料来源：神州问学公众号，国元证券研究所

3 预训练数据直接影响AI大模型性能

➢ 预训练数据从数据来源多样性、数据规模、数据质量三方面影响模型性能。以GPT模型为例，其架构从第1代到第4代均较为相似，而用来训练数据的数据规模和质量却有很大的提升，进而引发模型性能的飞跃。以吴恩达(Andrew Ng)为代表的学者观点认为，人工智能是以数据为中心的，而不是以模型为中心。“有标注的高质量数据才能释放人工智能的价值，如果业界将更多精力放在数据质量上，人工智能的发展会更快”。

图：预训练数据直接影响模型性能

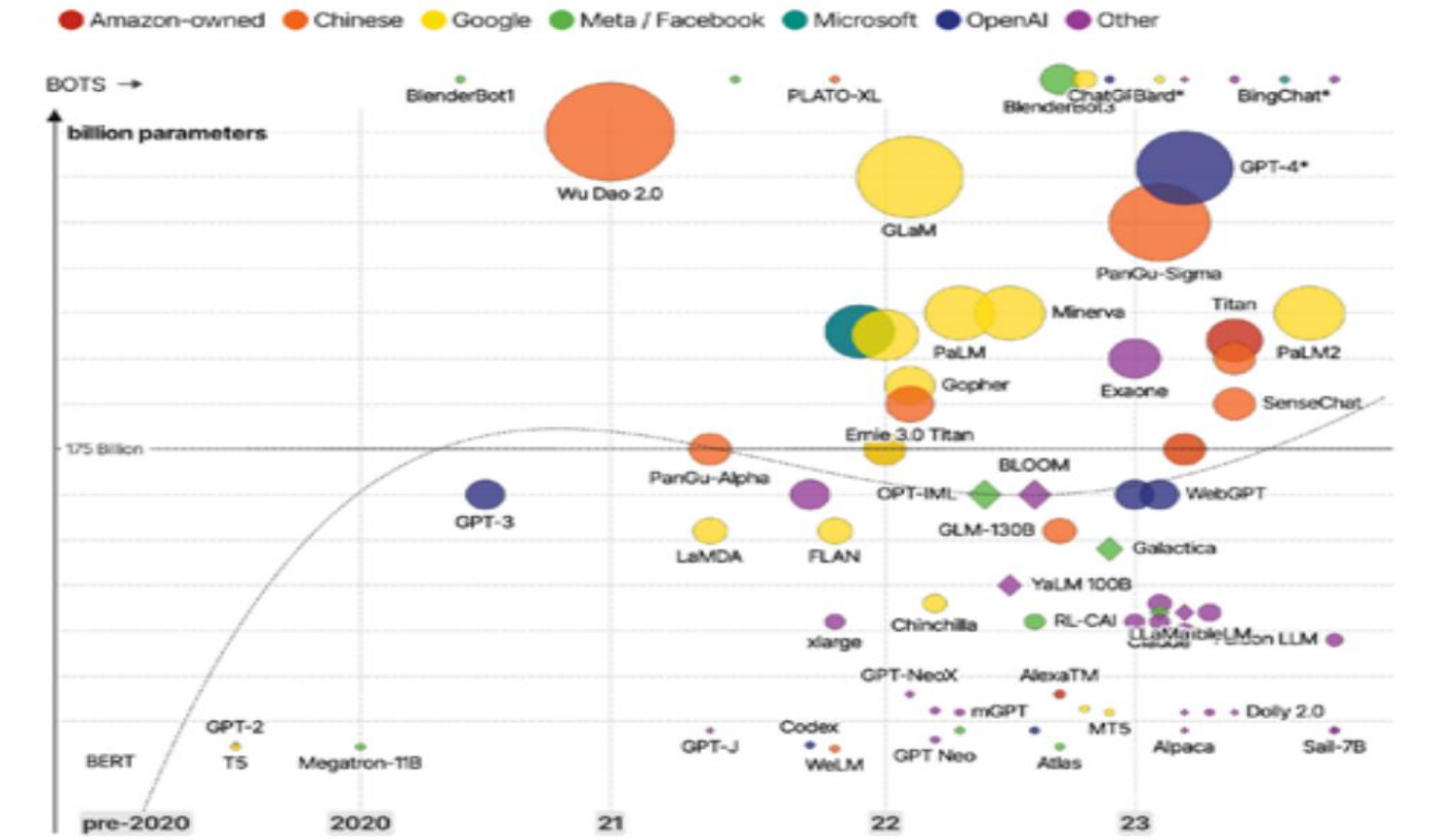
来源多样性	数据规模	数据质量
◆ 针对大模型需求制定配比 不同场景/领域的数据具有不同的语言特征，对模型能力提升的点也不同。如：书籍语料占比提升，可以提升上下文理解能力	◆ 足够规模的高质量语料 随着模型参数量的增加，也需要更多数据来训练。只有参数规模突破了100亿以上的大模型才具有“涌现能力”	◆ 高质量训练集能提高模型精度，减少训练时长 有重复、噪声、错误数据等低质量语料会损害模型性能。如：训练语料有重复，会影响模型对上下文的理解能力

资料来源：阿里研究院公众号，国元证券研究所

3 预训练数据直接影响AI大模型性能

- 为了追求更好的模型性能，模型参数规模也与训练数据量同步快速增长，模型参数量大约每18个月时间就会增长40倍。例如2016年最好的大模型ResNet-50参数量约为2000万，2020年的GPT-3模型参数量达1750亿，2023年的GPT-4参数规模则更加庞大。

图：大模型参数规模快速增长

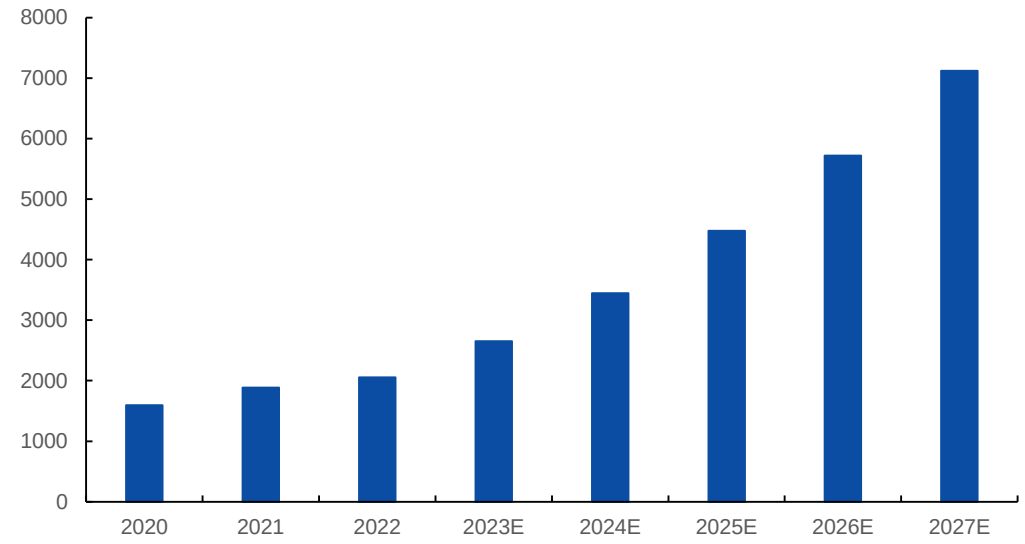


 资料来源：中国信通院《大模型智算服务白皮书》，国元证券研究所

4 市场规模

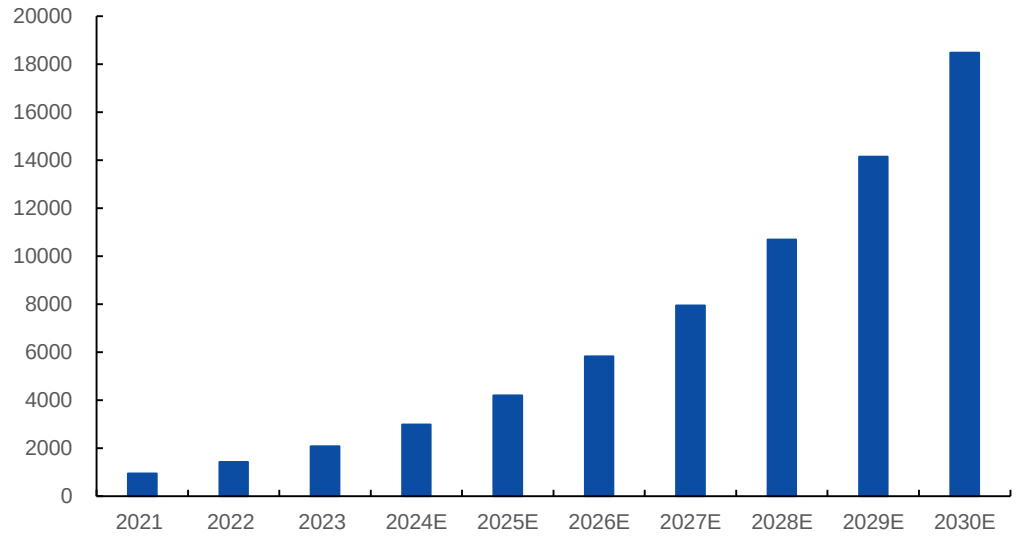
- 随着人工智能技术的不断发展，其应用场景日益丰富，各行各业所汇聚的庞大数据资源为技术的实际应用和持续完善提供了坚实基础。根据第三方咨询机构格物致胜的统计数据，2022年中国人工智能市场规模达到2058亿元，预计2023-2027年市场规模将保持28.2%的复合增长率，2027年中国人工智能市场规模将达到7119亿元。根据statista的统计数据，2023年全球人工智能市场规模达2079亿美元，预计2030年将增至18475亿美元。

图：中国人工智能市场规模及预测（单位：亿元人民币）



资料来源：格物致胜公众号，国元证券研究所

图：全球人工智能市场规模及预测（单位：亿美元）

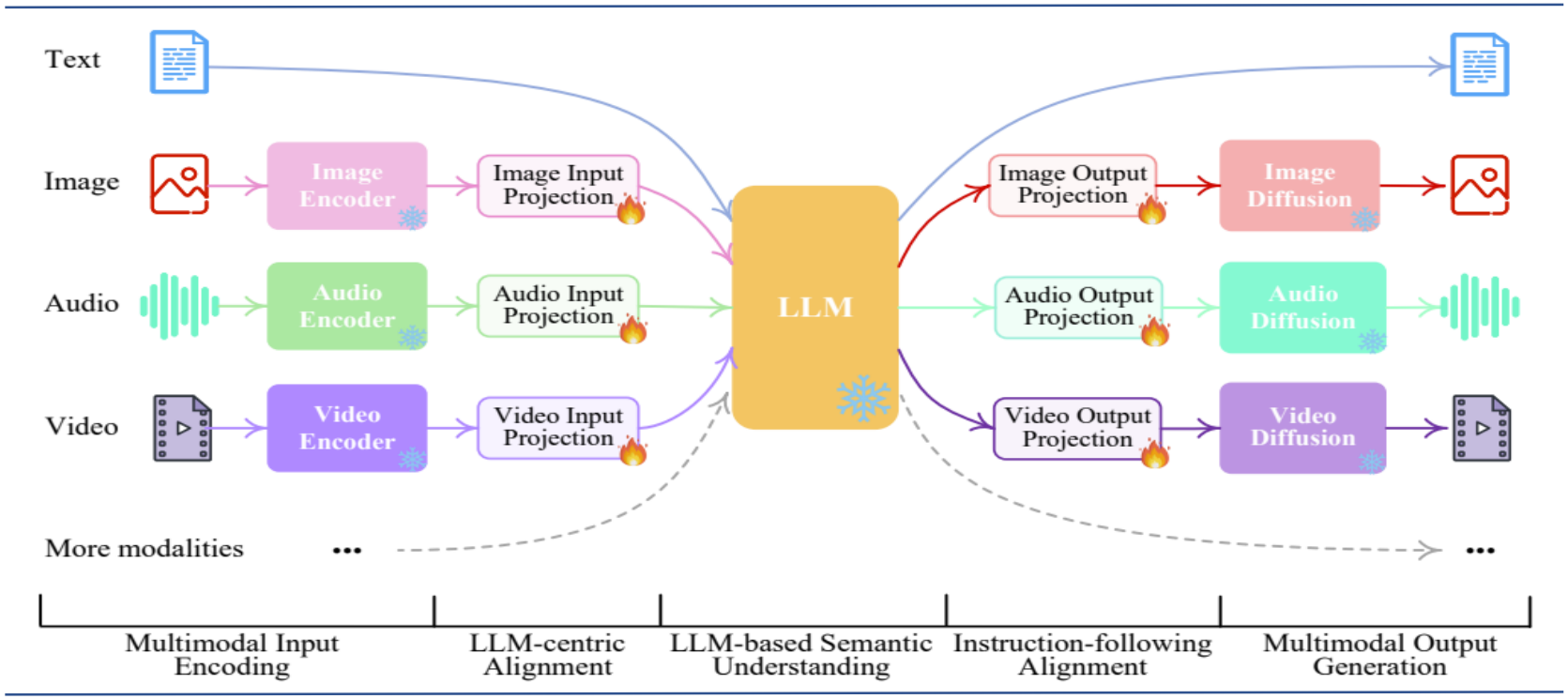


资料来源：Statista官网，国元证券研究所

1 多模态技术成为大模型主战场

- 多模态较单一模态更进一步，已经成为大模型主战场。人类通过图片、文字、语言等多种途径来学习和理解，多模态技术也是通过整合多种模态、对齐不同模态之间的关系，使信息在模态之间传递。2023年以来，OpenAI发布的GPT-4V、Google发布的Gemini、Anthropic发布的Claude 3均为多模态模型，展现出了出色的多模态理解及生成能力。未来，多模态有望实现any to any模态的输入和输出，包括文本、图像、音频、视频、3D模型等多种模态。

图：多模态模型实现any to any模态的输入和输出



 资料来源：Shengqiong Wu等《NExT-GPT: Any-to-Any Multimodal LLM》，国元证券研究所

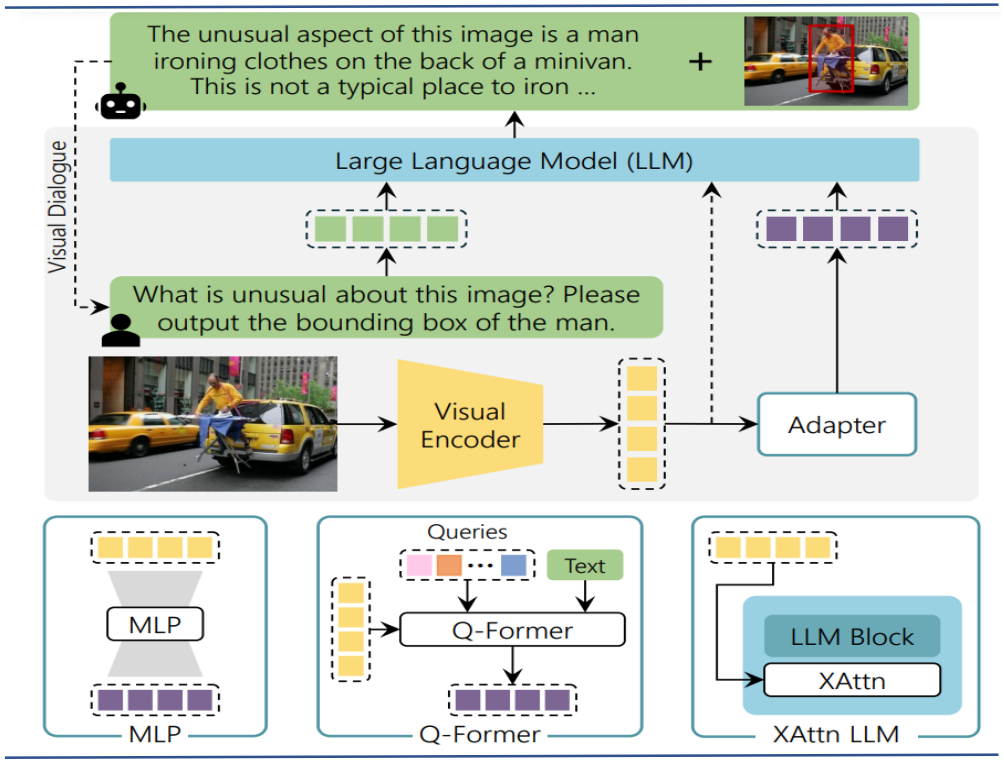
1 多模态技术成为大模型主战场

- 多模态大型语言模型(MLLMs)的通用架构，由1) 视觉编码器(Visual Encoder)、2) 语言模型(Language Model)和3) 适配器模块(Adapter Module)组成。1) 负责处理和理解输入的视觉信息，通常使用预训练的视觉模型，如Vision Transformer(ViT)或其他卷积神经网络(CNN)架构，来提取图像特征；2) 负责处理文本输入，理解和生成自然语言，语言模型基于Transformer架构，如BERT或GPT系列模型；3) 负责在视觉和语言模态之间建立联系。

图：多模态模型GPT-4V的问答展示

资料来源：机器之心公众号，国元证券研究所

图：多模态模型架构图

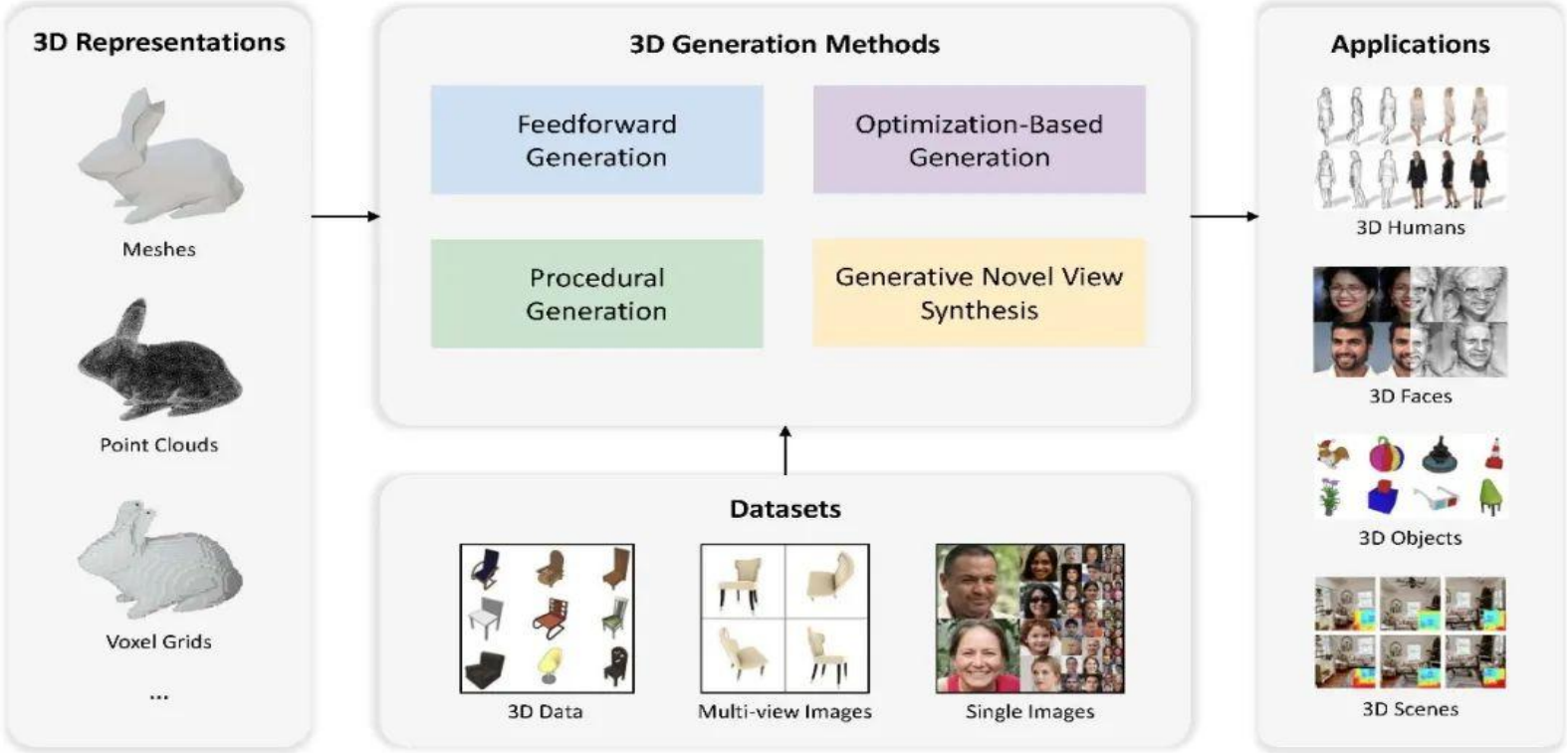


资料来源：Davide Caffagni等《The Evolution of Multimodal Large Language Models: A Survey》，国元证券研究所

2 3D生成：AI生成技术的下一个突破口

➢ 3D生成技术应用广阔，但仍处在技术临界点以前。3D生成技术可广泛应用于3D虚拟人、3D人脸、3D场景等领域，目前3D生成的主流技术路径大致可分为：1) text-to-2D，再通过NeRF或Diffusion模型完成2D-to-3D，或直接通过2D素材完成3D建模；2) 直接text-to-3D，该路径直接使用3D数据进行训练，从训练到微调到推理都基于3D数据。

图：3D生成技术的方法、数据集和应用

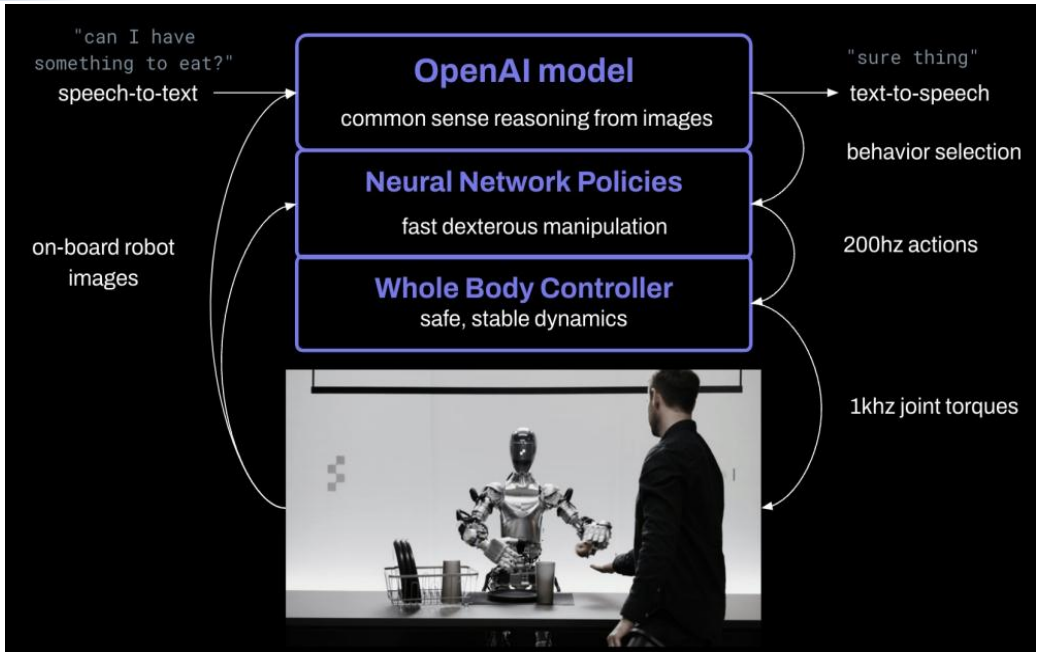


资料来源：Xiaoyu Li等《Advances in 3D Generation: A Survey》，国元证券研究所

3 具身智能：智能涌现从虚拟世界走向物理世界

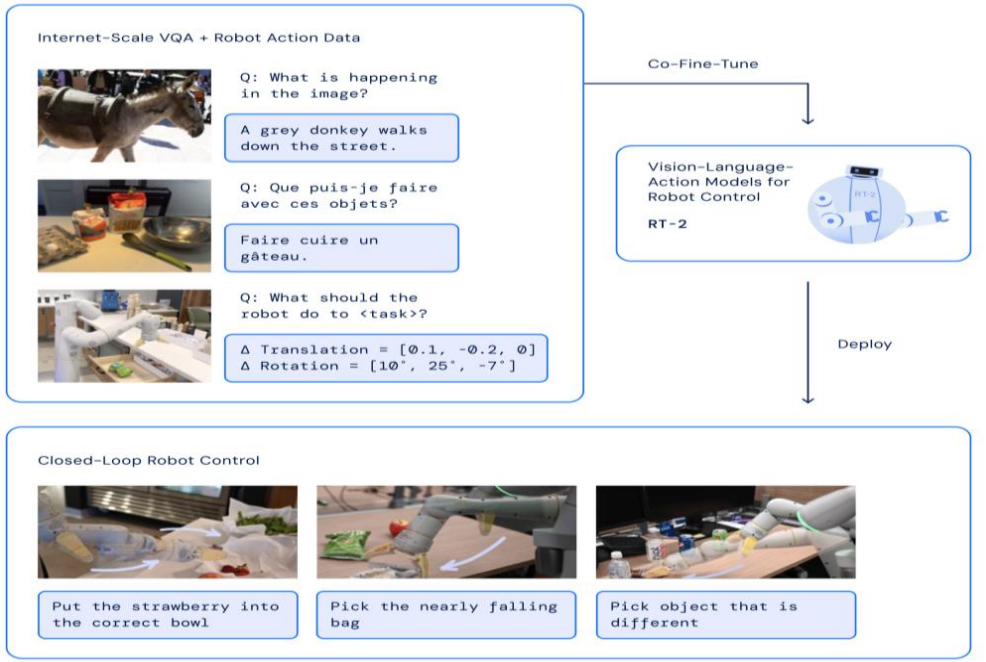
- 当大模型迁移到机器人身上，大模型的智能和泛化能力有望点亮通用机器人的曙光。2023年7月，谷歌推出机器人模型Robotics Transformer 2(RT-2)，这是一个全新的视觉-语言-动作(VLA)模型，从网络和机器人数据中学习，并将这些知识转化为机器人控制的通用指令。2024年3月，机器人初创企业Figure展示了基于OpenAI模型的全尺寸人形机器人Figure 01，机器人动作流畅，所有行为都是学到的（不是远程操作），并以正常速度(1.0x)运行。

图：Figure 01的技术原理



资料来源：AI前线公众号，国元证券研究所

图：RT-2的技术原理



资料来源：甲子光年公众号，国元证券研究所

4 通用人工智能还有多远

- 通用人工智能(Artificial General Intelligence, AGI)是一种可以执行复杂任务的人工智能，能够完全模仿人类智能的行为。DeepMind提出了一个衡量“性能”和“通用性”的矩阵，涵盖从无人工智能到超人类AGI（一个在所有任务上都优于所有人的通用人工智能系统）的五个级别。性能是指人工智能系统的能力与人类相比如何，而通用性表示人工智能系统能力的广度或其达到矩阵中指定性能水平的任务范围。

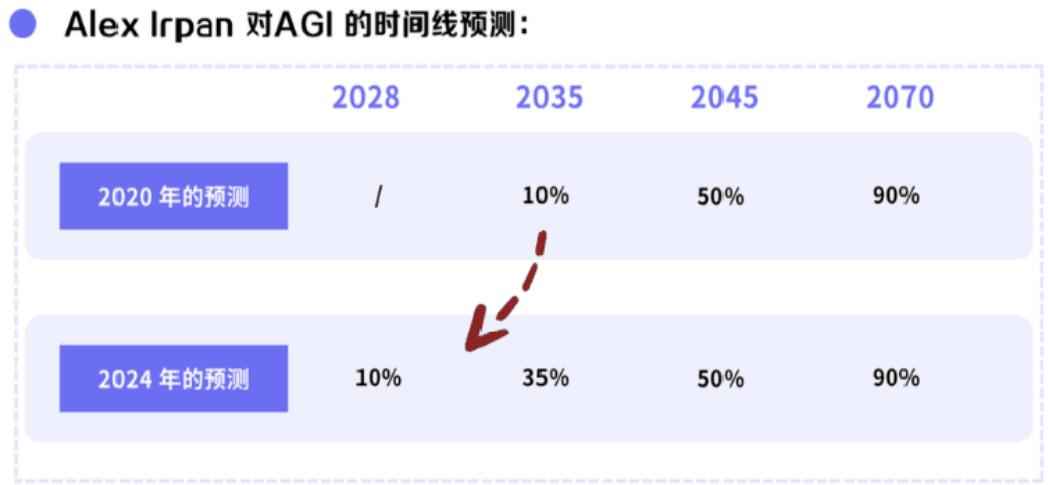
性能	狭义（明确范围的任务或任务集）	广义（广泛的非体力任务，包括元认知能力，如学习新技能）
0级：No AI	Narrow Non-AI（计算机软件、翻译器）	General Non-AI（human-in-theloop计算）
1级：Emerging（等于或略优于人类）	Emerging Narrow AI(GOFAI4:简单基于规则的系统，例如SHRDLU(Winograd，1971))	Emerging AGI(ChatGPT(Open AI, 2023)、Bard、Llama 2)
2级：Competent（至少50百分位的熟手）	Competent Narrow AI (Jigsaw, Siri, Alexa, Google Assistant, PaLI)	Competent AGI 尚未实现
3级：Expert（至少90百分位的熟手）	Expert Narrow AI（拼写和语法检查器，如Grammarly；生成图像模型，如Imagen）	Expert AGI 尚未实现
4级：Virtuoso（至少99百分位的熟手）	Virtuoso Narrow AI (Deep Blue(Campbell et all)AlphaGo)	Virtuoso AGI 尚未实现
5级：Superhuman（超过100%的人类）	Superhuman Narrow AI (AlphaFold, AlphaZero, StockFish)	Artificial Superintelligence(ASI) 尚未实现

资料来源：DeepMind《Levels of AGI: Operationalizing Progress on the Path to AGI》，国元证券研究所

4 通用人工智能还有多远

- 2023年12月，黄仁勋表示，如果把通用人工智能(AGI)定义为能以“相当有竞争力”的方式完成人类智能测试的计算机，那么在未来五年内，我们将看到AGI。
- 2023年11月，DeepMind联合创始人兼首席AGI科学家Shane Legg在访谈中表示，2028年，人类有50%的概率开发出第一个AGI，并且带领的DeepMind研究团队在Arxiv上公布了一篇名为《AGI的水平：实现AGI道路上的操作进展》论文，具体阐述了AGI的路线图和时间表。
- 2020年，谷歌机器人团队的软件工程师Alex Irpan认为，到2035年我们有10%的概率实现AGI，但到了2024年，他认为在2028年就有10%的概率接近AGI，到2035年则有25%的概率实现AGI。

图：对AGI时间线的预测变得更乐观



资料来源：海外独角兽公众号，国元证券研究所

图：DeepMind关于AGI论文

资料来源：DeepMind《Levels of AGI: Operationalizing Progress on the Path to AGI》，国元证券研究所

1 OpenAI创立：以实现安全的AGI为主旨

- OpenAI由Sam Altman、Elon Musk等在2015年创办，主旨是努力在安全的前提下创建通用人工智能(AGI)并让全人类共同受益；2020年发布GPT-3模型，2022年11月发布GPT-3.5模型，能够与人类进行多轮连续的各种对话，给出较为合理的回答；2023年3月发布GPT-4模型；2024年2月发布AI视频生成模型Sora，AI视频生成领域迎来ChatGPT时刻。

图：OpenAI发展历程

2015	2019	2020	2022	2023	2024
OpenAI宣布成立；公司定位为“非盈利组织”，主旨是努力在安全的前提下创建通用人工智能（AGI）并让全人类共同受益。	OpenAI从非盈利过渡到“封顶盈利”OpenAI接受微软10亿美元投资，双方合作为微软Azure云端平台服务开发AI技术。	OpenAI于6月发布GPT-3模型，9月微软获得该模型独家许可。	OpenAI于11月发布聊天机器人模型ChatGPT，能够与人类进行多轮连续的各种对话，给出较为合理的回答，引发全球关注。	OpenAI于3月发布GPT-4；OpenAI的2023 ARR年收入已达16亿美元，相比去年增长56倍，公司估值达1000亿美元。	OpenAI于2月发布AI视频生成模型Sora，能根据提示词生成长达一分钟的高清视频。

资料来源：AI前线公众号，MBA百科，腾讯研究院公众号，机器之心官网，华尔街见闻官网，腾讯网，国元证券研究所

2 GPT发展回顾：模型性能随结构、规模的提升不断优化

- GPT-1通过无监督预训练和有监督微调两个步骤训练；GPT-2无需有监督微调，而是通过更大规模的模型参数和训练数据集进行无监督预训练，模型参数量达到15亿；GPT-3的模型参数和数据集进一步扩大，模型参数量增加到1750亿，上下文窗口宽度增加到2048个token。

图：GPT模型经历多轮迭代

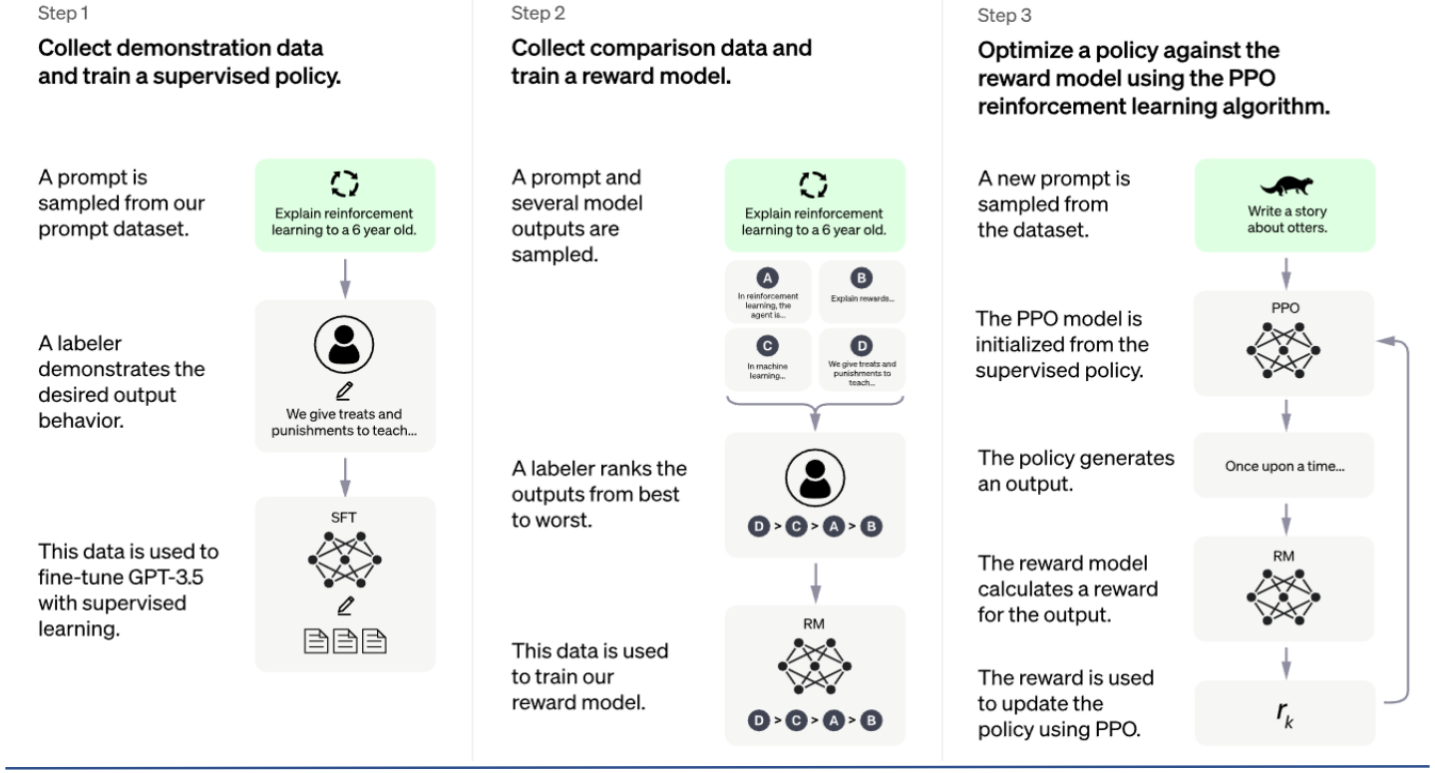
	GPT-4（多模态）	ChatGPT / GPT-3.5	GPT-3	GPT-2（开源）	BERT（开源）	GPT-1
发布时间	2023年3月14日	2022年11月30日	2020	2019	2018	2018
文字逻辑推理能力	■■■■■■■	■■■■■	■■■■	■■	■■■	■
应答速度	■■	■■■■■	■■■■■	■■■■■■■	■■■■■■■	■■■■■■■■
应答简洁度	■■■■■■	■■■■	■■■■	■■	■■	■■
模型技术结构特点与安全性措施	- 预期参数数量远大于1750亿 - 视觉语言模型组件（VLM） - 对抗性测试和红队测试 - RLHF 训练提示 - 基于规则的奖励模型（RBRM） GPT-4	- 基于GPT-3的架构，进行了针对性的优化和调整 - -参数数量、层数和词表与GPT-3相近，但具体参数可能有所不同 - RLHF 训练 ChatGPT	- 96层，参数数量增加到约1750亿（175 B） - 上下文窗口宽度增加到2048个tokens - 训练数据更大 - 采用交替密度和局部带状稀疏注意模式	- 从GPT-1基础上增加到48层，使用1600维向量进行词嵌入 - 修改初始化的残差层权重，缩放为原来的1/√N（N为残差层的数量 - 交替密度和局部带状稀疏注意模式	- 基于Transformer架构，12层 - 双向编码器（全文本） - 词嵌入向量维度为768或1024使用掩码语言模型（MLM）进行预训练 - 下游任务微调	- 基于Transformer架构 - 12层，768维词嵌入向量 - 无监督预训练和有监督微调
模型参数	未知	大概175B-6B-1.3B	175B	1.5B	340M	117M
上下文窗口	32,000 token	4096 token	2048 token	1024 token	512 token	512 token
训练方法	1）有监督微调 2）RLHF训练奖励模型 3）构造基于规则的奖励模型（RBRM） 4）PPO强化学习	1）有监督微调 2）RLHF训练奖励模型 3）PPO强化学习	无监督预训练	无监督预训练	掩码语言模型预训练	无监督预训练和有监督微调

 资料来源：智东西微信公众号，国元证券研究所

2 GPT 发展回顾：GPT3.5改进训练步骤实现性能跃升

- ChatGPT/GPT-3.5：2022年11月30日发布，在GPT-3的基础上进行有监督微调(Supervised Fine-Tuning)、奖励模型训练(Reward Modeling)和来自人类反馈的强化学习(Reinforcement Learning from Human Feedback, RLHF)。ChatGPT具有以下特征：主动承认自身错误、质疑不正确的问题、承认自身的无知和对专业技术的不了解以及支持连续多轮对话，极大提升了对话交互模式下的用户体验。

图：GPT-3.5训练过程

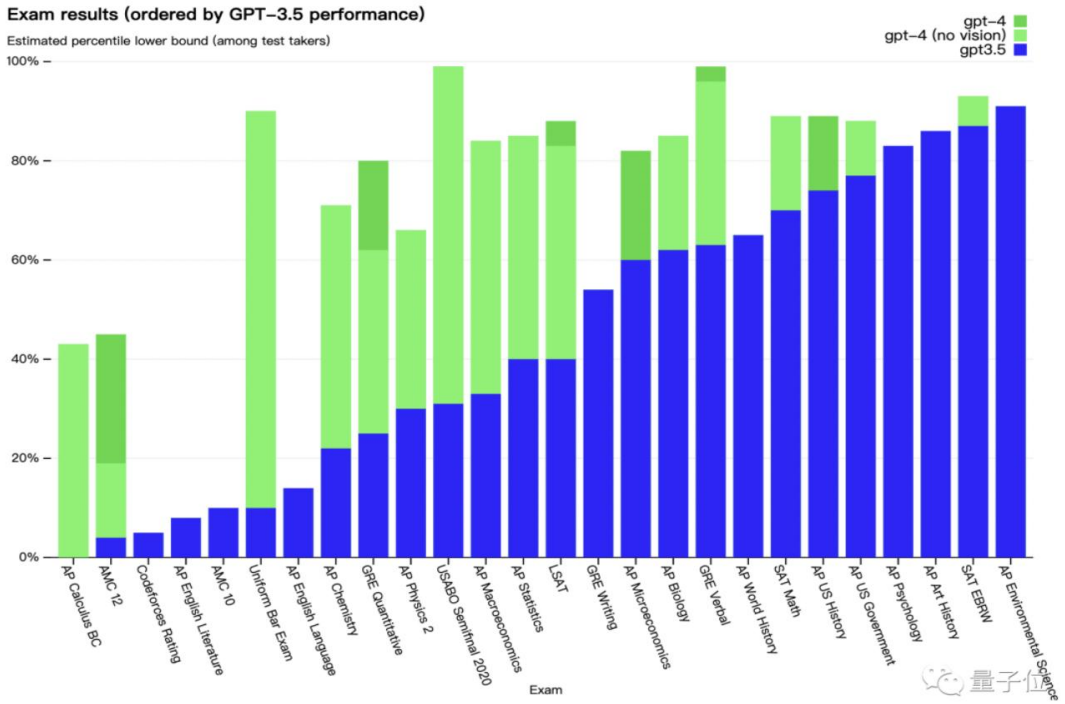


 资料来源：OpenAI官网，国元证券研究所

2 GPT发展回顾：多模态大模型GPT-4

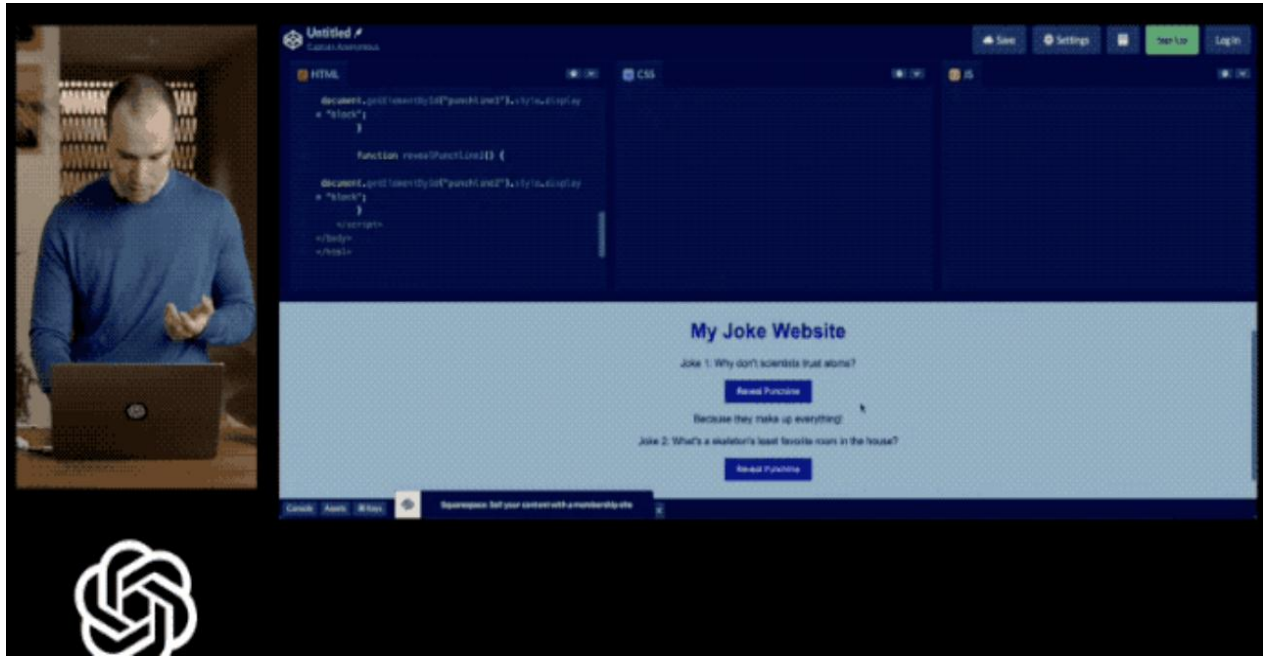
- 2023年3月14日，OpenAI宣布推出大型的多模态模型GPT-4，可以接收图像和文本输入。OpenAI称，GPT-4参加了多种基准考试测试，包括美国律师资格考试Uniform Bar Exam、法学院入学考试LSAT、“美国高考”SAT数学部分和证据性阅读与写作部分的考试，在这些测试中，它的得分高于88%的应试者。

图：GPT-4在各类学术水平测试中的成绩



资料来源：量子位微信公众号，国元证券研究所

图：GPT-4根据图片生成网站

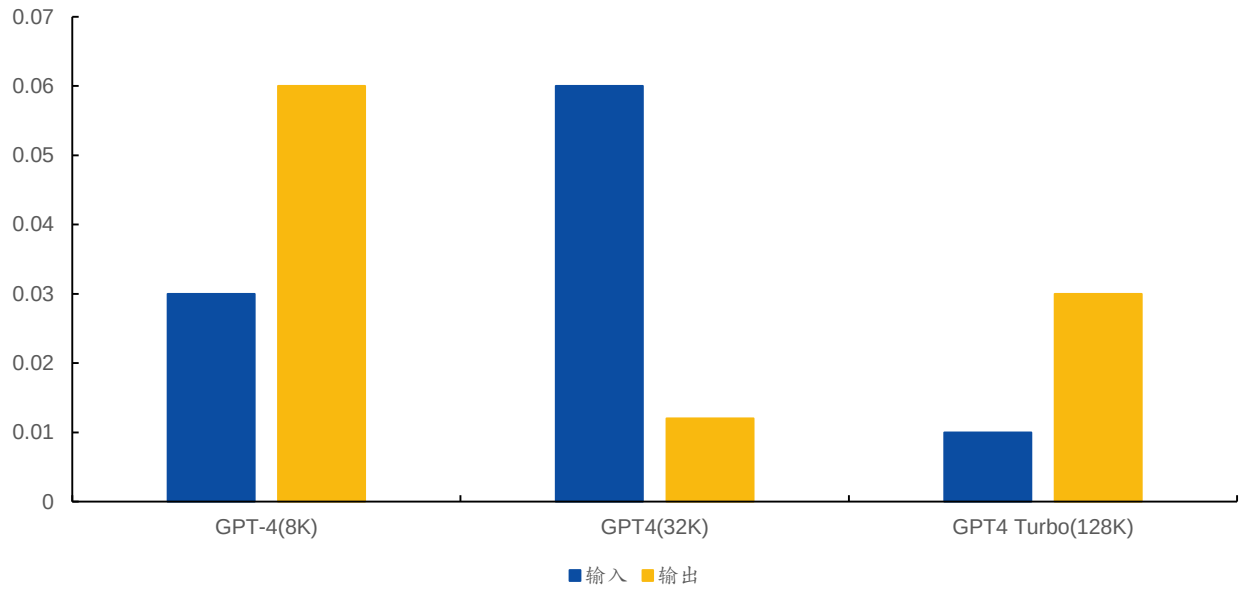


资料来源：量子位微信公众号，国元证券研究所

2 GPT发展回顾：更快更强更便宜的GPT-4 Turbo

- 2023年11月7日，OpenAI在开发者大会披露新版本具备：1) 更长的上下文长度：支持128K上下文窗口，相当于300页文本；2) 更便宜：新模型的价格是每千输入token 1美分，而每千输出token 3美分，输入和输出费用分别降至GPT-4(8K)的1/3和1/2，总体使用上降价约2.75倍；3) 更聪明：内部知识库更新至2023年4月，并支持上传外部数据库或文件；4) 视听多模态：支持文生图模型DALL·E3、文本转语音模型TTS，未来还将支持自动语音识别模型Whisper v3；5) 更快的速度：用户每分钟的Token速率限制将会翻倍，可通过API账户申请进一步提速。

图：GPT-4、GPT-4 Turbo模型价格对比（单位:美元/token）

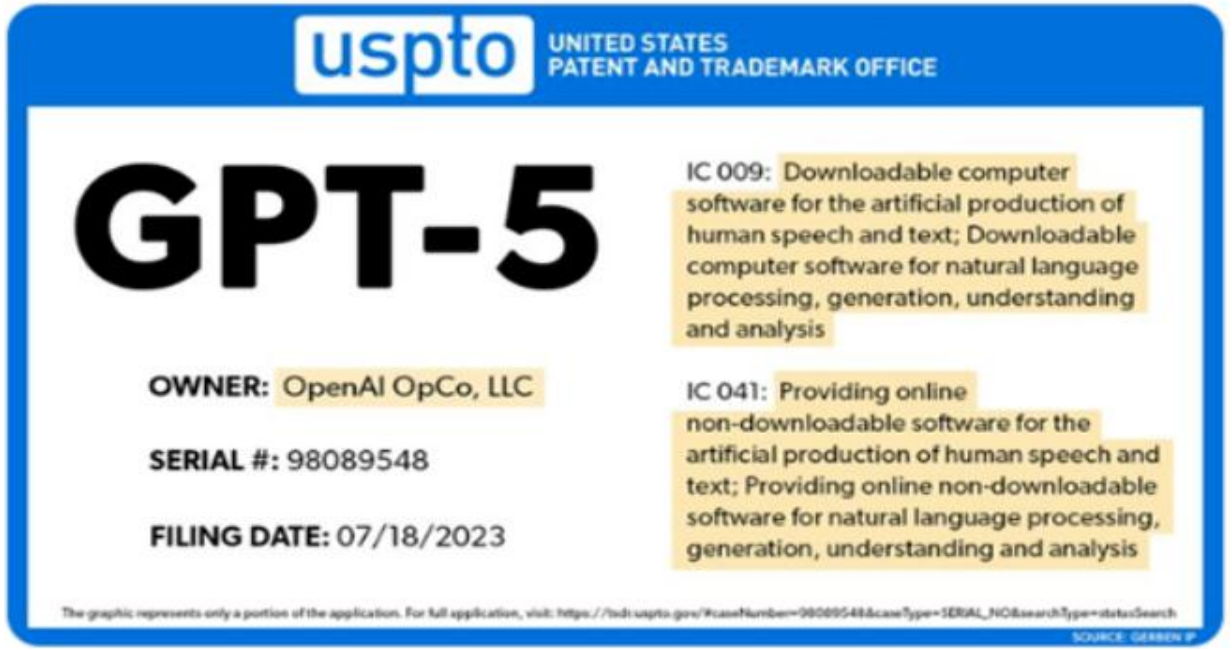


资料来源：爱范儿公众号，国元证券研究所

3 最新进展：GPT-5有望实现性能跃升

- 商标律师Josh Gerben在社交平台晒出OpenAI于2023年7月18日向美国专利商标局(USPTO)提交GPT-5商标的消息，GPT-5提供的功能包括自然语言处理、文本生成、理解、语音转录、翻译、预测和分析等，实际发布功能可能有变动。
- 根据OpenAI首席执行官Sam Altman的披露，GPT-5将具备三大升级点：1) 多模态：支持文本、语音、图像、代码和视频输入；2) 个性化：理解个人偏好的能力，如整合用户信息、电子邮件、日历、约会偏好，并与外部数据源建立联系；3) 推理能力和准确性：如果GPT-4目前解决了人类任务的10%，GPT-5应该是15%或者20%，当前大模型的通病——幻觉问题也将在GPT-5中得到解决。

图：GPT-5商标申请



 资料来源：AIGC开放社区公众号，国元证券研究所

4 图片生成模型：OpenAI发布DALL.E 3

- 2023年9月，OpenAI发布DALL.E 3，比以往系统更能理解细微差别和细节，能够让用户更加轻松地将自己的想法转化为非常准确的图像；该模型原生构建在ChatGPT之上，用ChatGPT来创建、拓展和优化prompt，用户无需在prompt上花费太多时间。
- DALL.E 3的技术架构主要分为图像描述生成和图像生成两大模块。图像描述生成模块使用了CLIP图像编码器和GPT语言模型(GPT-4)，可为每张图像生成细致的文字描述；图像生成模块先用VAE将高分辨率图像压缩为低维向量，降低学习难度。然后使用T5 Transformer将文本编码为向量，并通过GroupNorm层将其注入diffusion模型，指导图像生成方向。

图：DALL.E 3生成高质量图像

资料来源：OpenAI官网，国元证券研究所

5 视频生成模型：OpenAI发布“物理世界模拟器”Sora

- 2024年2月16日，OpenAI发布AI生成视频模型Sora，其卓越之处在于能够生成跨越不同持续时间、纵横比和分辨率的视频和图像，甚至包括生成长达一分钟的高清视频，“碾压”了行业目前平均约”4s”的视频生成长度，AI视频生成领域迎来ChatGPT时刻。
- OpenAI在Sora技术报告中写道: ”Our results suggest that scaling video generation models is a promising path towards building general purpose simulators of the physical world”.

图：Sora生成1分钟的连贯高清视频

资料来源：OpenAI官网，国元证券研究所

图：Sora官方简介

OpenAI Menu

Research

Video generation models as world simulators

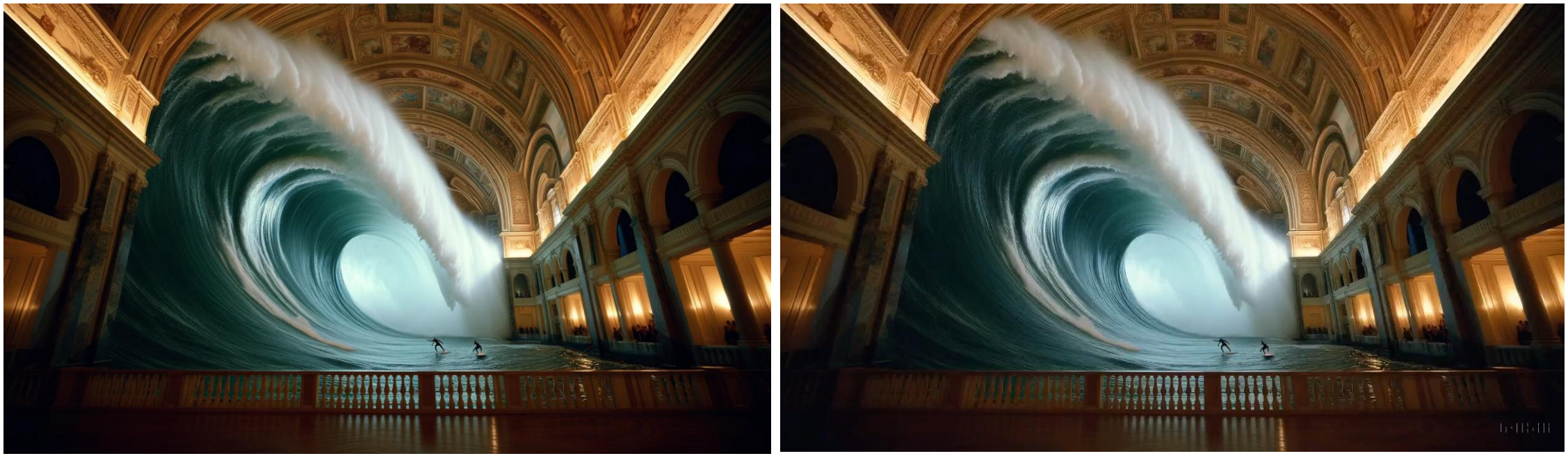
We explore large-scale training of generative models on video data. Specifically, we train text-conditional diffusion models jointly on videos and images of variable durations, resolutions and aspect ratios. We leverage a transformer architecture that operates on spacetime patches of video and image latent codes. Our largest model, Sora, is capable of generating a minute of high fidelity video. Our results suggest that scaling video generation models is a promising path towards building general purpose simulators of the physical world.

资料来源：OpenAI官网，国元证券研究所

5 视频生成模型：OpenAI发布“物理世界模拟器”Sora

- Sora不仅接受文字输入，还可根据图像和视频输入来生成视频。Sora能够执行各种图像和视频编辑任务——创建完美循环的视频、为静态图像制作动画、在时间维度上向前或向后扩展视频、在两个截然不同的输入视频之间实现无缝过渡、零输入转换输入视频风格和场景，展示了该模型在图像和视频编辑领域的强大能力和应用潜力，有望给产业端带来革命性的变革。

图：Sora根据图片输入生成视频



资料来源：OpenAI官网，国元证券研究所

5 视频生成模型：OpenAI发布“物理世界模拟器”Sora

- 模型尺度扩展带来惊人的涌现能力(emerging simulation capabilities)。1) 3D一致性：在3D一致性方面，Sora能够生成带有动态摄像头运动的视频。随着摄像头的移动和旋转，人物和场景元素在三维空间中始终保持一致的运动规律。2) 较长视频的连贯性和对象持久性：这是视频生成领域面对的一个重要挑战，而Sora能有效为短期和长期物体间的依赖关系建模，人和物被遮挡或离开画面后，仍能被准确地保存和呈现。3) 与世界互动：Sora能以简单的方式模拟影响世界状态的行为，例如画家可以在画布上留下新的笔触。4) 模拟数字世界：Sora能够模拟人工过程，比如视频游戏。

图：Sora生成的视频具备3D一致性

资料来源：OpenAI官网，国元证券研究所

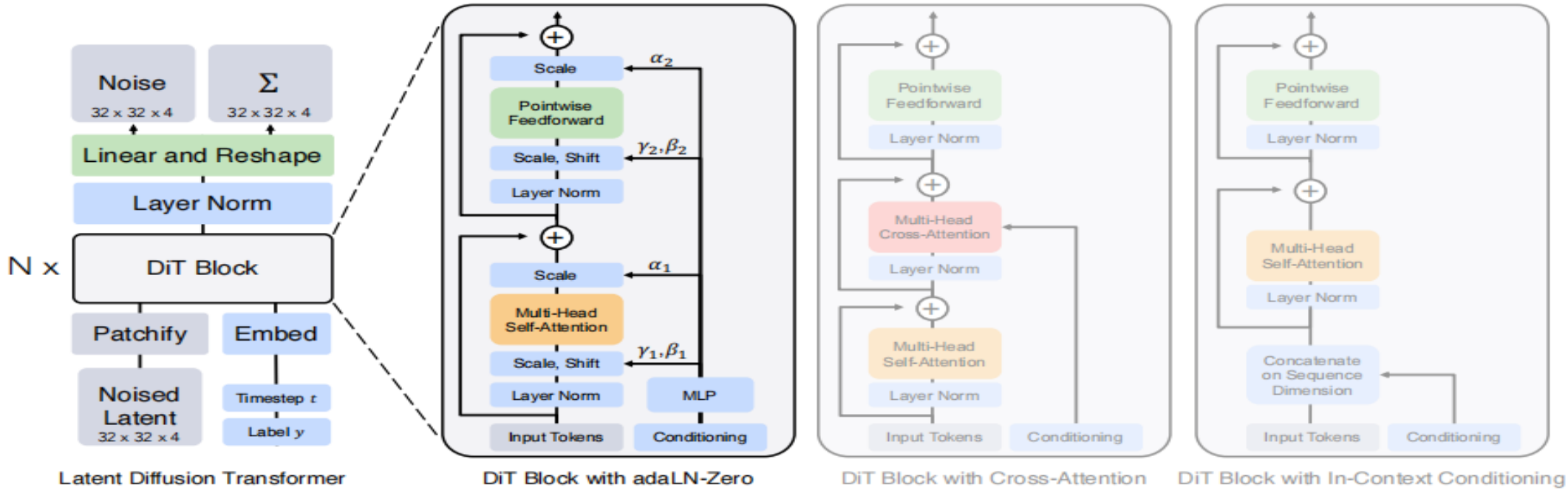
图：Sora生成《我的世界》游戏视频

资料来源：OpenAI官网，国元证券研究所

5 视频生成模型：OpenAI发布“物理世界模拟器”Sora

- Sora的本质是一种Diffusion transformer模型。Diffusion transformer (DiT)架构由William Peebles 和Saining Xie在2023年提出，使用Transformer来训练图像的潜在扩散模型，取代了通常使用的U-Net骨干网络，融合了扩散模型与自回归模型的双重特性。
- AI生成视频的技术路线主要经历了四个阶段：循环网络(RNN)、生成对抗网络(GAN)、自回归模型(autoregressive transformers)、扩散模型(diffusion models)。目前领先的视频模型大多数是扩散模型，比如Runway、Pika等。自回归模型由于更好的多模态能力与扩展性也成为热门的研究方向，如谷歌在2023年12月发布的VideoPoet。

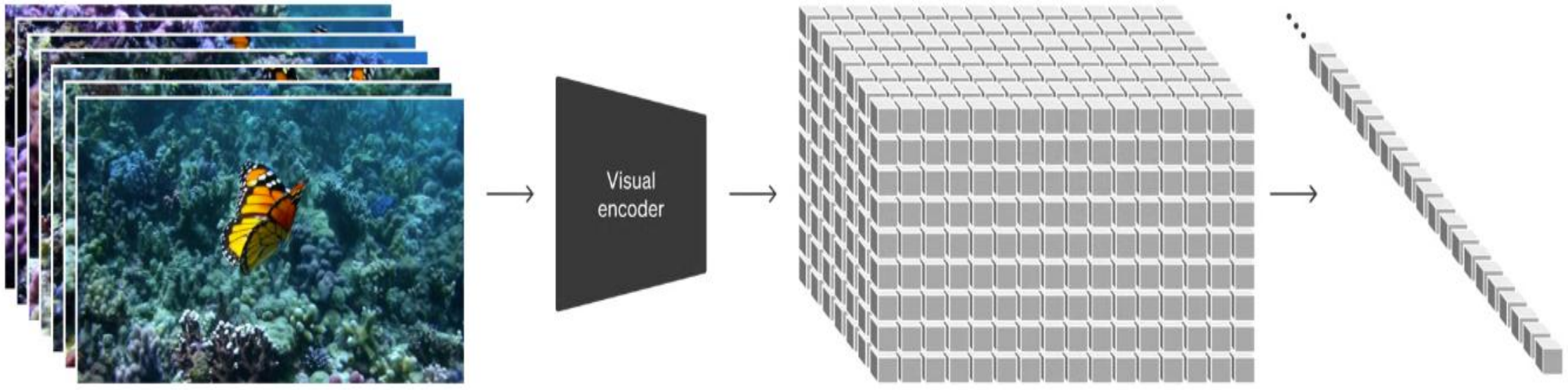
图：Diffusion transformer模型架构



5 视频生成模型：OpenAI发布“物理世界模拟器”Sora

- Sora模型训练范式：patch统一原始视觉数据。OpenAI提出了一种用patch作为视频数据来训练视频模型的方式，patch是将图像或视频帧分割成的一系列小块区域，是模型处理和理解原始数据的基本单元，这是从大语言模型的token汲取的灵感。Token统一了文本的多种模式——代码、数学和各种自然语言，而patch则统一了图像与视频。过去的图像和视频生成方法通常会将视频调整大小、裁剪或修剪为标准尺寸，而这损耗了视频生成的质量，将图片与视频数据patch化之后，无需对数据进行压缩，就能够对不同分辨率、持续时间和长宽比的视频和图像的原始数据进行训练。

图：OpenAI将视频转换为patch来统一视觉数据输入



资料来源：腾讯研究院公众号，国元证券研究所

5 视频生成模型：OpenAI发布“物理世界模拟器”Sora

- Sora模型训练范式：re-captioning标注技术带来优秀的语言理解能力。训练文本转视频生成系统需要大量带有相应文本字幕的视频，为此OpenAI借鉴了DALL·E3中的re-captioning技术，首先训练了一个高度描述性的转译员模型，然后使用它为训练集中的所有视频生成文本转译。通过这种方式对高度描述性的视频转译进行训练，可显著提高文本保真度和视频的整体质量。与DALL·E3类似，OpenAI利用GPT技术将简短的用户提示转换为更长的详细转译，并发送到视频模型，令Sora能精确按照用户提示生成高质量视频。

图：提示词“一个玩具机器人穿着绿色的连衣裙和太阳帽在美丽的日落期间在南极洲愉快地漫步”

 资料来源：OpenAI官网，国元证券研究所

1 大语言模型：开源LLaMA 2

- 2023年7月，Meta发布了开源大语言模型——LLaMA 2。LLaMA 2是在LLaMA 1基础之上构建而成，训练数据比上一版本多出40%，拥有70亿、130亿和700亿三种参数，并且允许商业化。技术方面，该预训练模型接受了2万亿个标记的训练，上下文长度是上一版本的两倍，能处理更长的文本内容；性能方面，LLaMA-13B在大多数基准上超过了参数量达1750亿的GPT-3。

图：LLaMA 2有三种参数规模

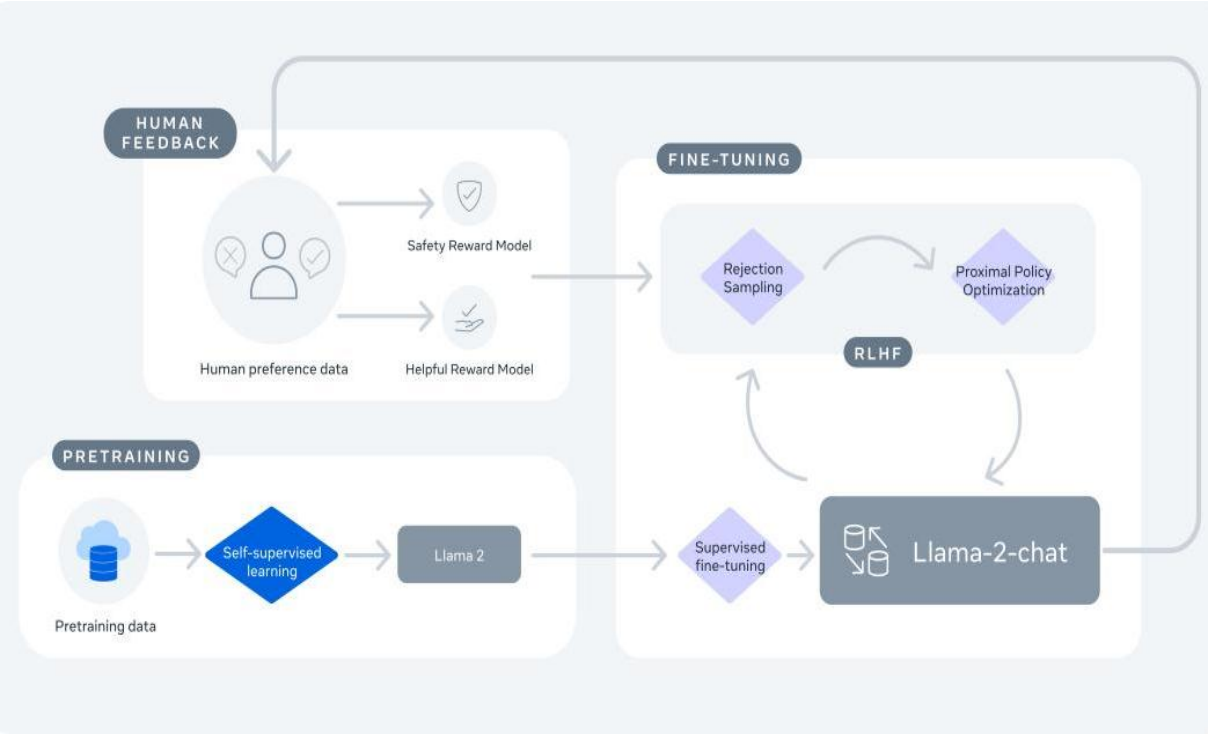
Llama 2 was trained on **40% more data** than Llama 1, and has double the context length.

Llama 2

MODEL SIZE (PARAMETERS)	PRETRAINED	FINE-TUNED FOR CHAT USE CASES
7B	Model architecture:	Data collection for helpfulness and safety:
13B	Pretraining Tokens: 2 Trillion	Supervised fine-tuning: Over 100,000
70B	Context Length: 4096	Human Preferences: Over 1,000,000

资料来源：36氪官网，国元证券研究所

图：LLaMA 2训练流程



资料来源：36氪官网，国元证券研究所

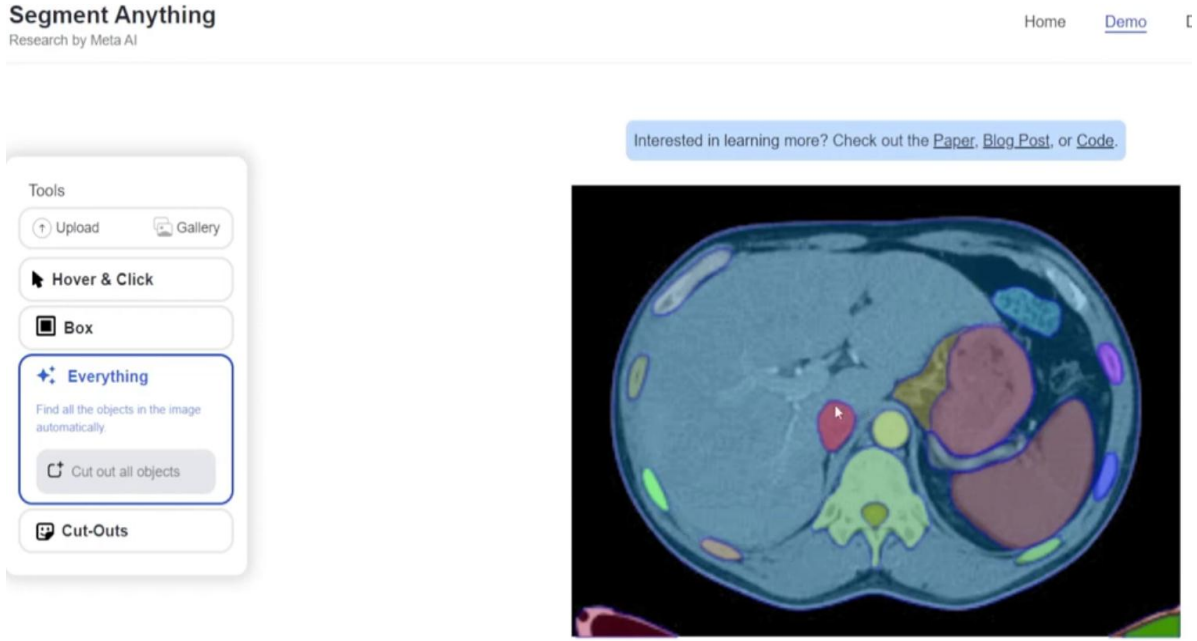
2 视觉大模型：开源图片分割基础模型SAM

- 2023年4月，Meta AI在官网发布了基础模型Segment Anything Model (SAM) 并开源。SAM已在1100万张图片和11亿个掩码的数据集上进行了训练，具有超强的自动识别、切割功能。SAM能感知超出数据训练的对象和图像，就算图片不在SAM训练范围内，它也能识别。这意味着，用户无需再收集自己的细分数据，并为用例模型进行微调。SAM可以集成在任何希望识别、切割对象的应用中，在医疗、农业、气象、天文、媒体等主流行业拥有广阔的应用空间。

图：SAM模型识别能力极强

资料来源：AIGC开放社区公众号，国元证券研究所

图：SAM可用于医疗领域

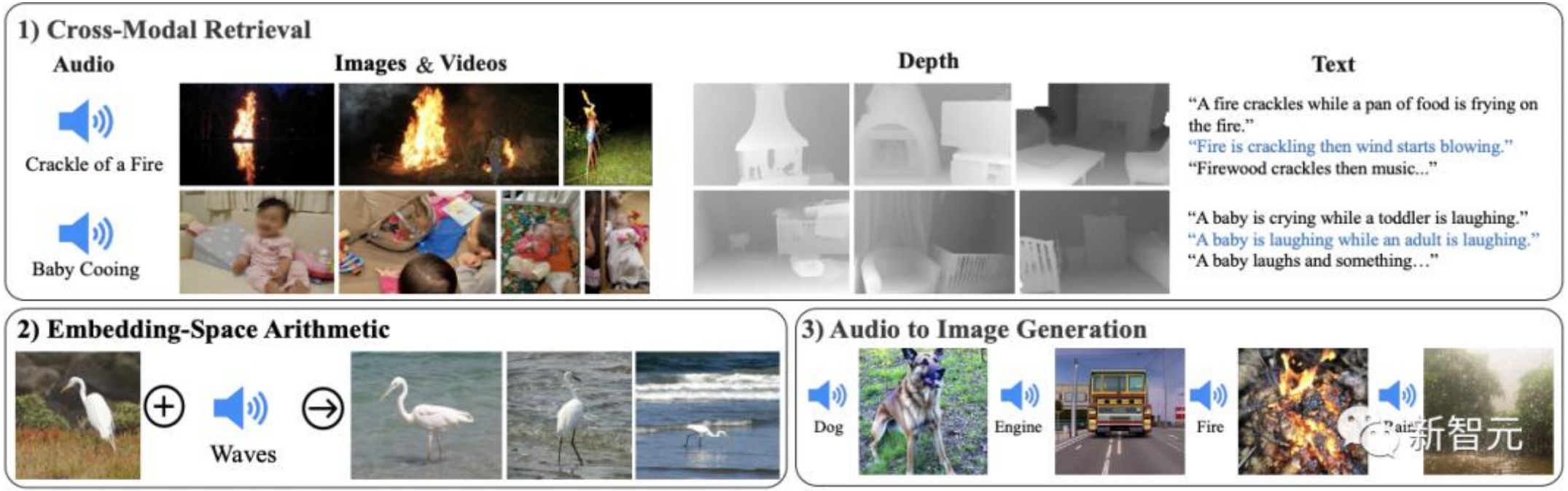


资料来源：AIGC开放社区公众号，国元证券研究所

3 多模态大模型：开源ImageBind，具备超强联想能力

- 2023年5月，Meta开源了多模态大模型ImageBind，可跨越图像、视频、音频、深度、热量和空间运动6种模态进行检索。例如，输入鸽子的图片，外加一个摩托音频，模型能够检索出一张摩托和鸽子的图片。ImageBind模型把不同模态数据串联在一个嵌入空间(Embedding Space)，从多维度理解世界，未来将引入更多模态增强对世界感知，比如如触觉、语音、嗅觉和大脑fMRI信号。

图：ImageBind模型可跨越6种模态进行检索



资料来源：新智元公众号，国元证券研究所

1 多年布局：理论基础深厚，发布多个基础架构

- 2016年，谷歌宣布公司战略从Mobile First转向AI First，此后陆续发布Transformer、BERT、T5等重要的基础模型（架构）；2023年4月，谷歌将Google Brain和DeepMind合并为Google DeepMind，全力冲刺AI，8个月后发布Gemini。

Google在AI领域的发展时间线

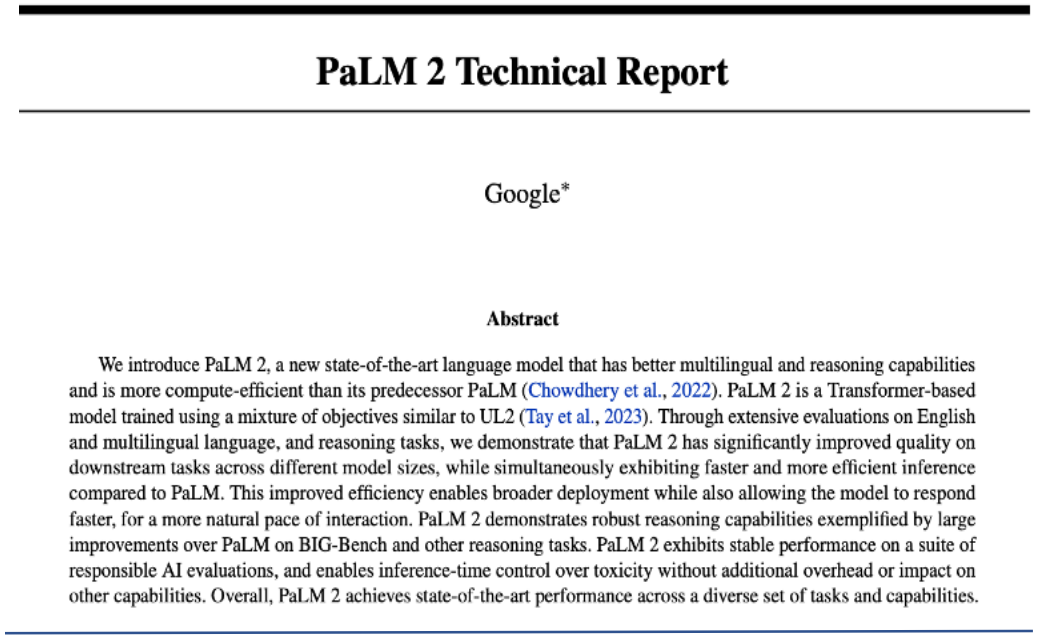
- **2014** Google收购DeepMind
- **2017.06** Google发布Transformer，成为后来所有LLM的基础框架；DeepMind提出RLHF的方法
- **2018.10** Google发布编码器的BERT（最大3.5亿参数），用于微调下游任务
- **2019.10** Google发布基于Transformer Decoder的T5，兼容BERT和GTP的下游任务
- **2021.10** Google发布FLAN，转向decoder-only，提出Instruction Tuning
- **2021.11** DeepMind发布Gopher（2800亿参数），加入LLM大战
- **2022.01** Google再发LaMDA（1370亿参数），称其具有“意识”
- **2022.04** Google发布PaLM（5400亿参数，decoder-only），提出神奇的思维链
- **2022.09** DeepMind发布Sparrow加入RLHF和Retrival（Google Search）
- **2023.02** Google官宣基于LaMDA的Bard
- **2023.04** Google表示将合并旗下两个主要的人工智能研究部门—Google Brain和DeepMind成立Google DeepMind
- **2023.12** Google发布Gemini
- **2024.02** Google发布Gemini1.5

资料来源：飞哥说AI微信公众号，澎湃新闻网，36氪官网，国元证券研究所

2 大语言模型：PaLM 2实现轻量化，可在移动设备上离线运行

- PaLM2性能升级，部分测试结果超过GPT-4，轻量版可运行在移动设备上：2023年5月，谷歌发布PaLM2，对于具有思维链prompt或自洽性的MATH、GSM8K和MGSM基准评估，PaLM 2的部分结果超越了GPT-4。PaLM2包含四种尺寸的模型，其中最轻量的“壁虎”版本能在移动设备上快速运行（包括离线状态）。
- 谷歌将PaLM2融入办公软件、搜索引擎等产品：AI聊天机器人Bard被整合到谷歌的办公软件“全家桶”中，为Gmail、Google Docs、Sheets以及Slides创造了名为”Duet AI”的办公助手；Bard还被整合到谷歌搜索优化搜索答案。

图：PaLM 2技术报告



PaLM 2 Technical Report

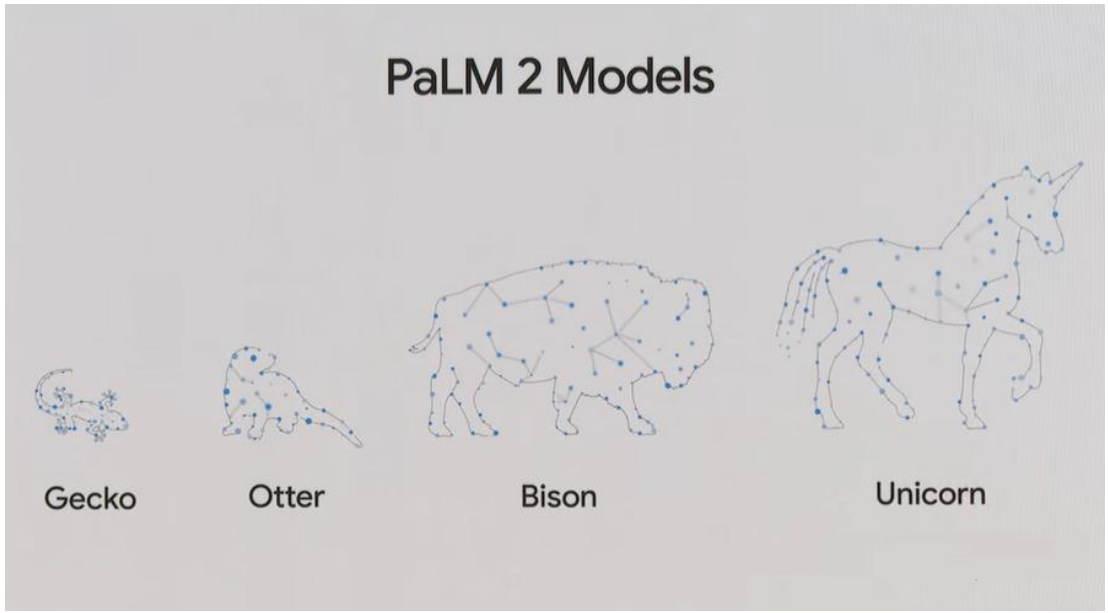
Google*

Abstract

We introduce PaLM 2, a new state-of-the-art language model that has better multilingual and reasoning capabilities and is more compute-efficient than its predecessor PaLM (Chowdhery et al., 2022). PaLM 2 is a Transformer-based model trained using a mixture of objectives similar to UL2 (Tay et al., 2023). Through extensive evaluations on English and multilingual language, and reasoning tasks, we demonstrate that PaLM 2 has significantly improved quality on downstream tasks across different model sizes, while simultaneously exhibiting faster and more efficient inference compared to PaLM. This improved efficiency enables broader deployment while also allowing the model to respond faster, for a more natural pace of interaction. PaLM 2 demonstrates robust reasoning capabilities exemplified by large improvements over PaLM on BIG-Bench and other reasoning tasks. PaLM 2 exhibits stable performance on a suite of responsible AI evaluations, and enables inference-time control over toxicity without additional overhead or impact on other capabilities. Overall, PaLM 2 achieves state-of-the-art performance across a diverse set of tasks and capabilities.

资料来源：澎湃新闻官网，国元证券研究所

图：PaLM 2包含四种尺寸的模型

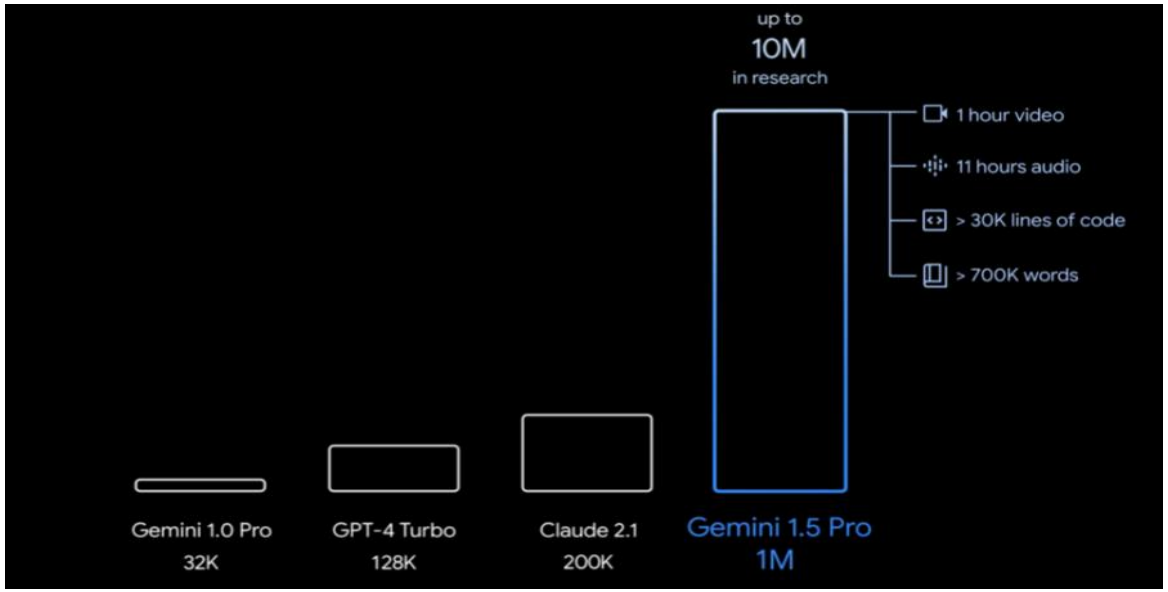


资料来源：科技最前线公众号，国元证券研究所

3 多模态模型：最新发布Gemini 1.5，支持超长上下文窗口

- 2024年2月，谷歌发布最新一代MoE多模态模型Gemini 1.5。MoE_(Mixture of Experts)是一种混合模型，由多个子模型（即专家）组成，核心思想是使用一个门控网络来决定每个数据应该被哪个模型训练，从而减轻不同类型样本之间的干扰。
- 支持超长的上下文窗口，信息处理能力进一步增强。谷歌增加了Gemini 1.5 Pro的上下文窗口容量，并实现在生产中运行高达100万个Token，远超32k的Gemini 1.0、128k的GPT-4 Turbo、200k的Claude 2.1，这意味着Gemini 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行代码的代码库或超过700000个单词。

图：Gemini 1.5的上下文窗口长度超过多个主流模型



资料来源：甲子光年公众号，国元证券研究所

图：Gemini 1.5 Pro与Gemini 1.0系列比较

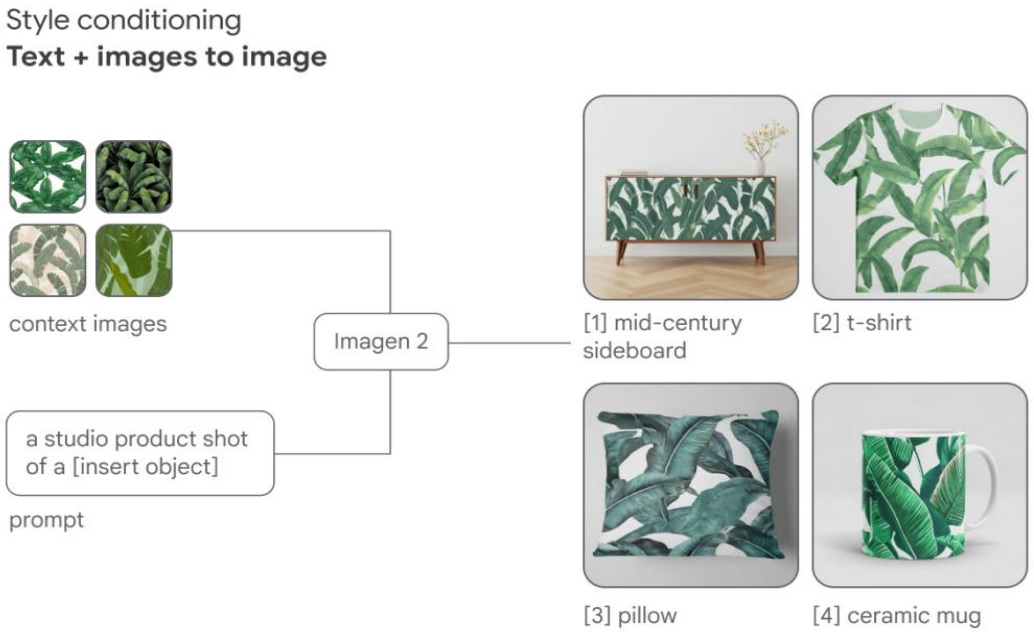
Gemini 1.5 Pro	Relative to 1.0 Pro	Relative to 1.0 Ultra
Long-Context Text, Video & Audio	from 32k up to 10M tokens	from 32k up to 10M tokens
Core Capabilities	Win-rate: 87.1% (27/31 benchmarks)	Win-rate: 54.8% (17/31 benchmarks)
Text	Win-rate: 100% (13/13 benchmarks)	Win-rate: 77% (10/13 benchmarks)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

资料来源：甲子光年公众号，国元证券研究所

4 图像生成模型：Imagen 2可生成高质量、更逼真的输出

- 2023年12月，Google发布最新的图像模型Imagen 2，在数据集和模型方面改善了文本到图像工具经常遇到的许多问题，包括渲染逼真的手和人脸，以及保持图像没有干扰视觉的伪影。
- Imagen 2基于扩散技术提供了高度的灵活性，使控制和调整图像风格变得更加容易。通过提供参考风格的图像并结合文字提示，使用者可以调节Imagen 2生成相同风格的新图像；此外，还支持修补(inpainting)和扩图(outpainting)等图像编辑功能。

图：Imagen 2通过使用参考图片和文本提示更容易地控制输出风格



资料来源：机器之心公众号，国元证券研究所

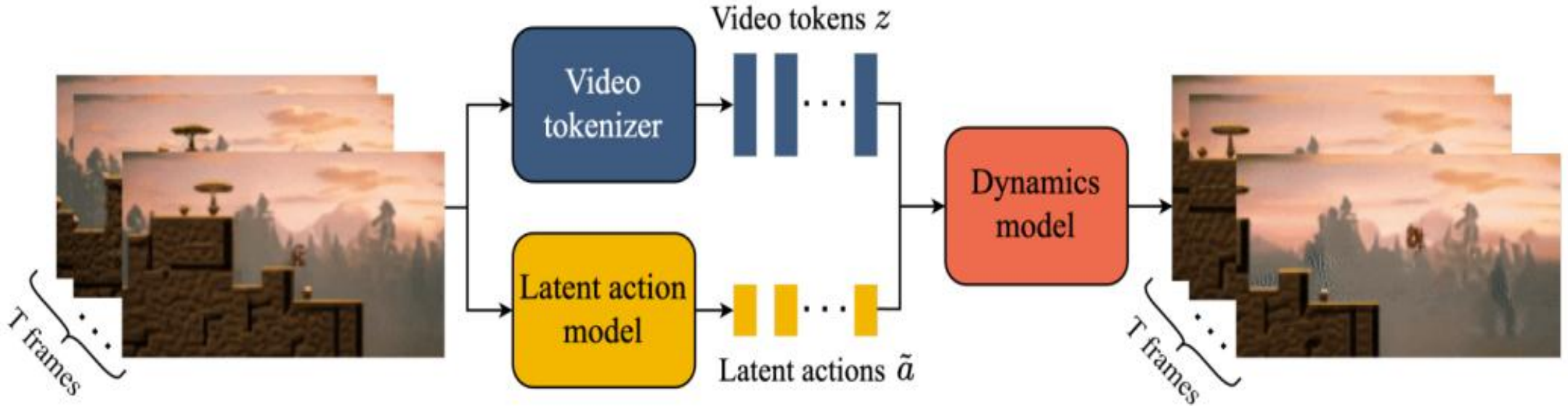
图：Imagen 2生成逼真手部和人脸的图像

资料来源：机器之心公众号，国元证券研究所

5 视频生成模型：Genie可通过单张图像生成交互式环境

- 2024年2月26日，谷歌发布Genie(Generative Interactive Environments)，它是一个110亿参数的基础世界模型，可通过单张图像提示生成可玩的交互式环境。谷歌认为Genie是实现通用智能体的基石之作，未来的AI智能体可以在新生成世界的无休止的curriculum中接受训练，从Genie学到的潜在动作可以转移到真实的人类设计的环境中。
- Genie包含三个关键组件：1) 潜在动作模型(Latent Action Model，LAM)，用于推理每对帧之间的潜在动作$\boldsymbol{a}$；2) 视频分词器(Tokenizer)，用于将原始视频帧转换为离散token $\boldsymbol{z}$；3) 动态模型，给定潜在动作和过去帧的token，用来预测视频的下一帧。

图：Genie模型训练过程

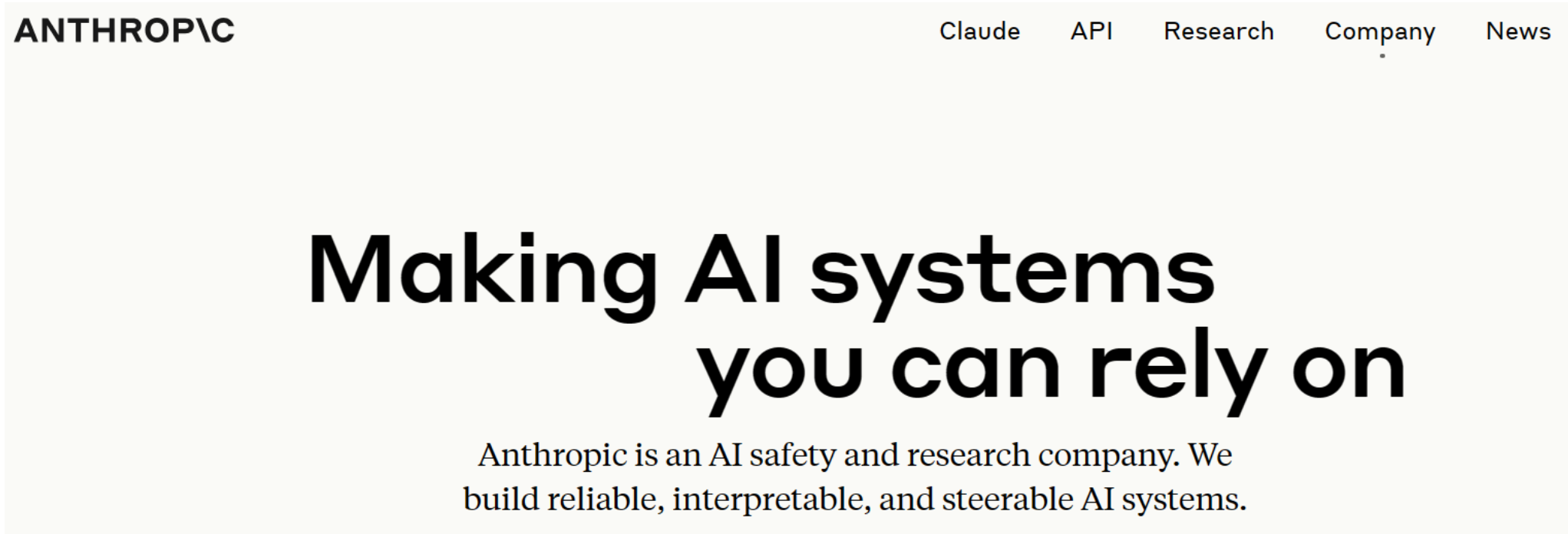


资料来源：机器之心公众号，国元证券研究所

1 AI独角兽Anthropic

- Anthropic是一家人工智能创业公司，由OpenAI前研究副总裁达里奥·阿莫迪(Dario Amodei)、大语言模型GPT-3论文的第一作者汤姆·布朗(Tom Brown)等人在2021年创立。2023年2月，获得Google投资3亿美元，Google持股10%；2023年3月，发布类似ChatGPT的大语言模型Claude；2023年7月，发布新一代Claude 2模型；2024年3月，发布Claude 3模型。

图：Anthropic官网简介



资料来源：Anthropic官网，国元证券研究所

2 多模态模型：Claude 3基准测试表现优秀

- 2024年3月，Anthropic发布最新的多模态模型Claude 3，该系列包含三个模型：Claude 3 Haiku、Claude 3 Sonnet和Claude 3 Opus。其中，能力最强的Opus在多项基准测试中得分都超过了GPT-4和Gemini 1.0 Ultra，在数学、编程、多语言理解、视觉等多个维度树立了新的行业基准。多模态方面，用户可以上传照片、图表、文档和其他类型的非结构化数据，让AI分析和解答。

图：Claude 3基准测试的表现结果

	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

资料来源：机器之心公众号，国元证券研究所

- 人工智能产业政策落地不及预期的风险；
- 人工智能相关技术迭代不及预期的风险；
- 商业化落地进展低于预期；
- 行业竞争加剧的风险。

（1）公司评级定义		（2）行业评级定义	
买入	预计未来 6 个月内，股价涨跌幅优于上证指数 20%以上	推荐	预计未来 6 个月内，行业指数表现优于市场指数 10%以上
增持	预计未来 6 个月内，股价涨跌幅优于上证指数 5-20%之间	中性	预计未来 6 个月内，行业指数表现介于市场指数±10%之间
持有	预计未来 6 个月内，股价涨跌幅介于上证指数±5%之间	回避	预计未来 6 个月内，行业指数表现劣于市场指数 10%以上
卖出	预计未来 6 个月内，股价涨跌幅劣于上证指数 5%以上		

分析师声明

作者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，以勤勉的职业态度，独立、客观地出具本报告。本人承诺报告所采用的数据均来自合规渠道，分析逻辑基于作者的职业操守和专业能力，本报告清晰准确地反映了本人的研究观点并通过合理判断得出结论，结论不受任何第三方的授意、影响。

证券投资咨询业务的说明

根据中国证监会颁发的《经营证券业务许可证》(Z23834000)，国元证券股份有限公司具备中国证监会核准的证券投资咨询业务资格。证券投资咨询业务是指取得监管部门颁发的相关资格的机构及其咨询人员为证券投资者或客户提供证券投资的相关信息、分析、预测或建议，并直接或间接收取服务费用的活动。证券研究报告是证券投资咨询业务的一种基本形式，指证券公司、证券投资咨询机构对证券及证券相关产品的价值、市场走势或者相关影响因素进行分析，形成证券估值、投资评级等投资分析意见，制作证券研究报告，并向客户发布的行为。

一般性声明

本报告仅供国元证券股份有限公司（以下简称“本公司”）在中华人民共和国内地（香港、澳门、台湾除外）发布，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。若国元证券以外的金融机构或任何第三方机构发送本报告，则由该金融机构或第三方机构独自为此发送行为负责。本报告不构成国元证券向发送本报告的金融机构或第三方机构之客户提供的投资建议，国元证券及其员工亦不为上述金融机构或第三方机构之客户因使用本报告或报告载述的内容引起的直接或连带损失承担任何责任。本报告是基于本公司认为可靠的已公开信息，但本公司不保证该等信息的准确性或完整性。本报告所载的信息、资料、分析工具、意见及推测只提供给客户作参考之用，并非作为或被视为出售或购买证券或其他投资标的的投资建议或要约邀请。本报告所指的证券或投资标的的价格、价值及投资收入可能会波动。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。本公司建议客户应考虑本报告的任何意见或建议是否符合其特定状况，以及（若有必要）咨询独立投资顾问。在法律许可的情况下，本公司及其所属关联机构可能会持有本报告中所提到的公司所发行的证券头寸并进行交易，还可能为这些公司提供或争取投资银行业务服务或其他服务。

国元证券研究所

合肥	上海
地址：安徽省合肥市梅山路18号安徽国际金融中心A座国元证券	地址：上海市浦东新区民生路1199号证大五道口广场16楼国元证券
邮编：230000	邮编：200135
传真：（0551）62207952	传真：（021）68869125
	电话：（021）51097188