国海证券
SEALAND SECURITIES

证券研究报告
计算机
2023年03月20日

人工智能系列深度报告：AIGC行业综述篇
——开启AI新篇章

陈梦竹(证券分析师)
S0350521090003
chenmz@ghzq.com.cn

陈凯艺(联系人)
S0350121070080
chenky@ghzq.com.cn

核心提要

本篇报告主要解答了以下问题：AI、AIGC当下发展处于什么阶段？未来将呈现怎样的趋势？AIGC的核心生产要素是什么？各生产要素的发展趋势如何？NLP、CV、ASR、TTS算法及发展？ChatGPT为何“火爆出圈”？AIGC包括什么？已有哪些产品？应用现状及前景如何?有哪些企业进行了布局？商业模式如何？

- **行业发展：人工智能步入新发展阶段，逐步迈向AGI；AIGC拥抱人类，创造人机交互新变革，将迎来更多新机遇。**人工智能从理论发展分为四个阶段：规则导向、机器学习、深度学习、自主学习阶段，目前处于深度学习阶段；从应用成熟度可分为三个阶段：弱人工智能阶段（ANI）、强人工智能阶段（AGI）、超人工智能阶段（ASI），目前处于ANI阶段；从应用类型可分为四种：感知式AI与分析式AI应用较成熟，决策式AI近年来发展迅速，生成式AI迎来突破。生成式AI，即AIGC，较传统内容创作模式UGC、PGC可实现更大数量、更高质量、更低单位成本，未来将从辅助创作生成趋向高度自动化自主创造。此外，AIGC将赋能多领域，加速人机共生的建设，迎接更多机遇与挑战。
- **技术进步：算力是支撑，数据是瓶颈，算法迎来突破。**算力层，近年来大模型流行，模型参数量迅速膨胀，所需计算资源越来越大，算力是AIGC核心生产要素；而AI芯片全球短缺，美对华芯片制裁升级，我们认为国内短期算力充足，长期仍需要逐步实现AI芯片国产化替代。数据是机器学习的核心，AI发展的瓶颈,数据决定模型质量的上限；大模型训练需要海量且优质数据，AI对数据训练集的消耗量远大于人类数据生产的速度，专业领域、图像视频等数据获取和标注成本也将越来越高，我们认为加速商业化，实现数据反哺是对提高数据量、降成本的重要解决办法。算法层，近年来迎来不少突破，过去NLP领域以RNN及其变体为主，CV领域以CNN及其变体为主，但各有优劣，Transformer架构突破了RNN不能并行计算的限制，较CNN有更好的计算局部特征间的关联等，自2017年开始在NLP领域应用、变种升级，Transformer在多模态的发展和应用将让AI越来越多的向人类推理方式靠近，以实现AGI。AIGC包括文本/音频/图像/视频/代码/3D/数字人/跨膜态生成等，目前文本、音频和图像领域都迎来较大突破，图像生成的突破是Difussion的出现，文本生成的突破则是GPT的出现，AIGC基本采用GAN算法，算法及产品越来越丰富多元，AI因AIGC的蓬勃发展，已开启技术与应用的新篇章。
- **应用概览：技术突破实现应用创新。**AI小模型是过去主流的研究和应用方向，在B端部分行业、赛道已有不少企业布局，预计未来仍将依托其细分行业、细分赛道的先发优势和数据、项目实施经验、产品优势等壁垒仍将有较好的发展。但大模型尚未实现商业价值闭环，未来需要重点关注数据、算法层面的突破与变革，探索新的商业模式，目前已在影视、传媒、电商、C端娱乐规模应用，游戏领域逐步应用，金融、工业、医疗、法律、设计等专业领域还在持续拓展。
- **产业布局：科技巨头全面布局，中下游厂商百花齐放。**国外主要以微软、谷歌、Meta为主，国内以百度、腾讯、阿里、华为等为主，既拥有充足的算力支撑，又有优秀的人才团队，多年算法、数据积累，在大模型领域的发展及应用具备天然优势。上游除云厂商外，还有光通信厂商、数据服务商、算力相关设备厂商，将较大程度受益于大模型发展带来的更多计算资源和数据需求。中游有商汤、科大讯飞、旷视、拓尔思等企业多年细分领域布局，部分也有一定算力储备，垂直行业细分赛道深耕，相关技术、数据储备丰富。下游主要是受益于AIGC对业务的驱动、降本增效，空间较大，多行业公司均将逐步受益。
- **商业模式：商业化初启，期待产业生态、技术与产品发展完善。**小模型在B端已应用多年，大模型商业刚刚开始，主要是MaaS，包括大模型厂商自用，实现增量或降本增效；云厂商“MaaS+IaaS”打包输出；替代翻译、美工、原画师、程序员、分析师、设计师等繁琐重复的低端工作等。大模型商业价值闭环未成，国内SaaS生态、付费意识较差，商业落地还需要各行各业共同发展、相互奔赴，共建良好产业生态。
- **风险提示**：人工智能发展不及预期，AIGC发展不及预期；技术发展不及预期；商业化拓展不及预期；行业竞争加剧风险；中美科技竞争不确定性风险。

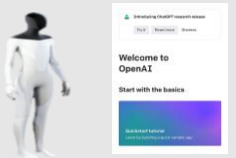
目录

目录

目录

核心分析框架

核心分析框架：每一轮人机交互的变革都会带来产业级投资机会

变革节点	PC操作系统 Macintosh、Windows	浏览器 IE浏览器、网景浏览器等	搜索引擎 Yahoo、Google等	智能手机 Iphone等	ARVR Oculus Quest、HTC Vive、Hololens等	人机共生 人形机器人、AIGC等
						
	1984年苹果推出划时代的Macintosh计算机，不仅首次采用图形界面的操作系统，并第一次使个人计算机具有了多媒体处理能力；1985年微软推出Windows系统	1993年NCSA中Mosaic项目的负责人辞职并建立了网景通讯公司，推出网景浏览器；1995年微软推出IE1.0浏览器，作为Windows 95的默认浏览器，改变了用户网上冲浪方式	1995年Yahoo公司正式成立，2002年收购Inktomi并将其网页搜索技术融入雅虎官网；1998年Google成立，后NetScape放弃Excite，开始使用Google的搜索数据，具备里程碑意义	2007年苹果发布自PC以来最具变革性的产品——iphone 2G，大部分操作都将由用户触控屏幕实现；iPhone 4在外观、显示、芯片均大幅改善，并提供六轴动作感应	2016年Facebook正式发售Oculus rift消费者版本，被称为消费级VR设备元年；2015年索尼推出PlayStation VR；2015年微软发布混合现实的智能眼镜Hololens	2013年，波士顿动力发布初代Atlas；2022年，Tesla预计发布Optimus原型机；2022年11月Open AI发布人工智能技术驱动的自然语言处理工具ChatGPT
人机交互模式	鼠标+键盘，可点击，但交互模式单一且不智能，人较为被动	鼠标+键盘，浏览器聚合功能改善交互成本	鼠标+键盘，搜索引擎的检索功能以人为中心，降低精准信息获取门槛	触屏+键盘，人机交互更加直观便捷，人处于主动地位	手势追踪、Inside-out、Outside-in、眼球追踪等，交互方式多元化，沉浸感强	人机共生，文字、音频、视频、3D、策略等交互模式融合，智能化程度显著提升
产业机会	操作系统、早期邮箱、早期超级计算中心等	光缆/运营商、浏览器、门户网站、通讯软件等	搜索引擎、众多PC互联网网页应用等	手机硬件产业链、应用商店、各大手机APP应用等	VRAR硬件产业链、云计算/边缘计算、视频&直播&游戏应用等	机器人硬件产业链、AI产业链（模型算力数据等）、下游应用等

资料来源：36氪，Companyhistory，digitaltrends，IDG news service，likecs，payititi，TechCPR，the verge，VRcompare，界面新闻，甲子光年，开源资讯，国海证券研究所

核心分析框架： 期待算力、数据、算法的突破，迈向强人工智能AGI阶段

国海证券 SEALAND SECURITIES

规则导向
1950s-1980s
感知机、专家系统

机器学习
1980s-21世纪初期
支持向量机、神经网络、决策树

深度学习
21世纪初期至今
多层神经网络模型

自主学习
未来发展方向
基于大规模数据识别、预测、学习、决策

理论发展

硬件奠基，算法发展

核心硬件发展，算法突破

算力、数据、算法全面突破

运营营销决策、金融风险评估、医疗诊断等。

始于1950s，让计算机感知和理解现实世界，**兴盛于计算机硬件发展与大数据出现。**

感知式AI

分析式AI

始于1960s，利用逻辑推理等方法进行分析和解决；**兴盛于机器学习、数字挖掘技术兴起。**

决策式AI

始于1970s，分析之后，进行智能决策；**近年来才开始真正兴盛。**

始于1980s，起源于分析式AI，生成新的内容； **2022年迎来突破。**

生成式AI

生成式AI生成各种数据、图像、语音等内容后，决策式AI对生成内容进行分析和决策，实现更加全面和智能的应用。

文本识别、语音识别、图像识别等。

自然语言生成、图像生成、音乐生成等。

弱人工智能阶段
ANI

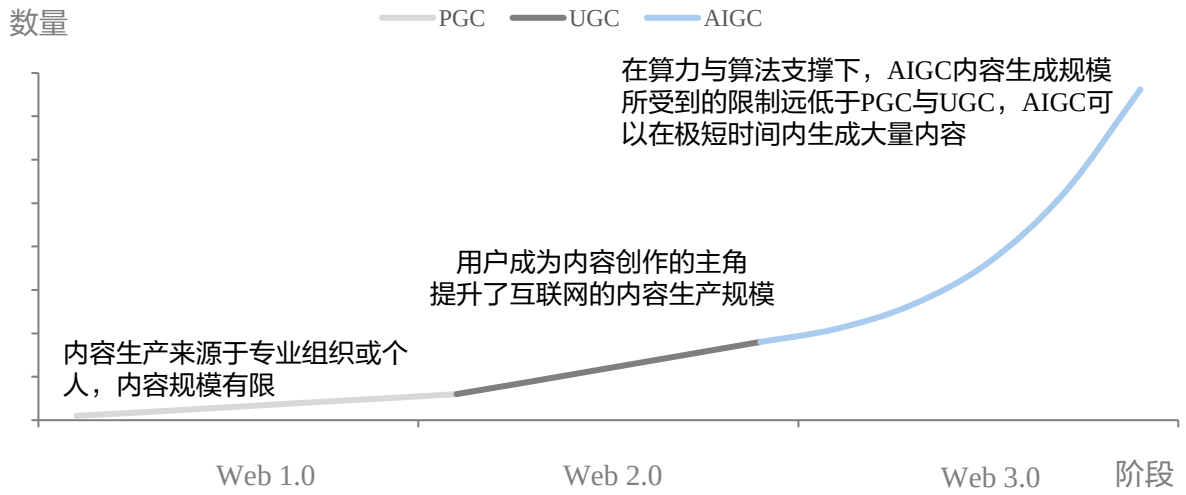
强人工智能阶段
AGI

超人工智能阶段
ASI

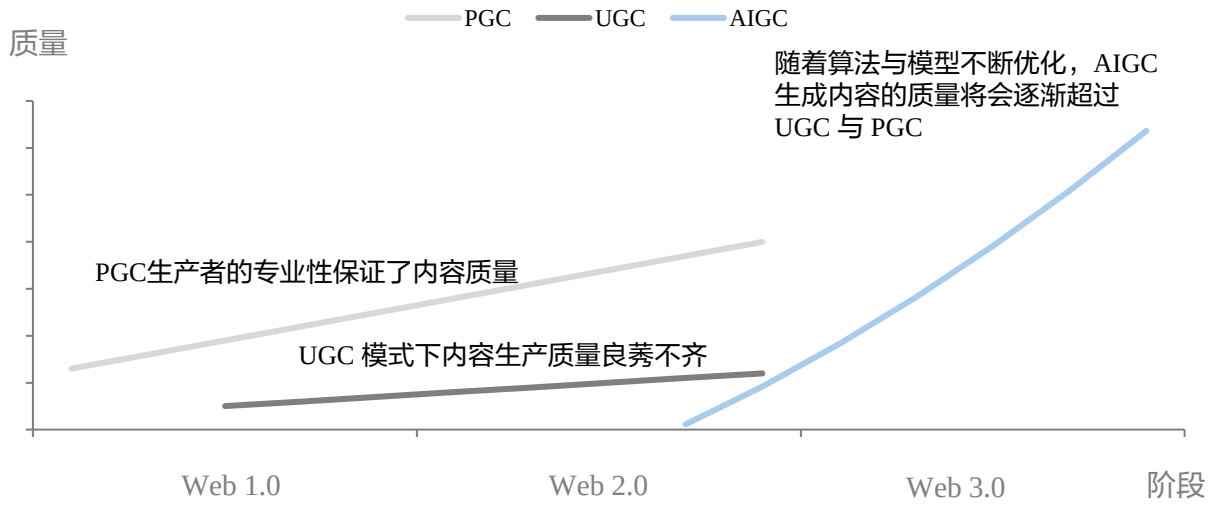
核心分析框架： AIGC与PGC、UGC内容创作模式对比

传统的 PGC 与 UGC 模式受到规模、质量和成本的制约，而AIGC 则能够有效地弥补 PGC 与 UGC 模式的不足，具有生成内容规模大、质量高、单位成本低的优势，将会成为元宇宙场景下的主要内容生成模式，从而为元宇宙建设提供内容支撑。

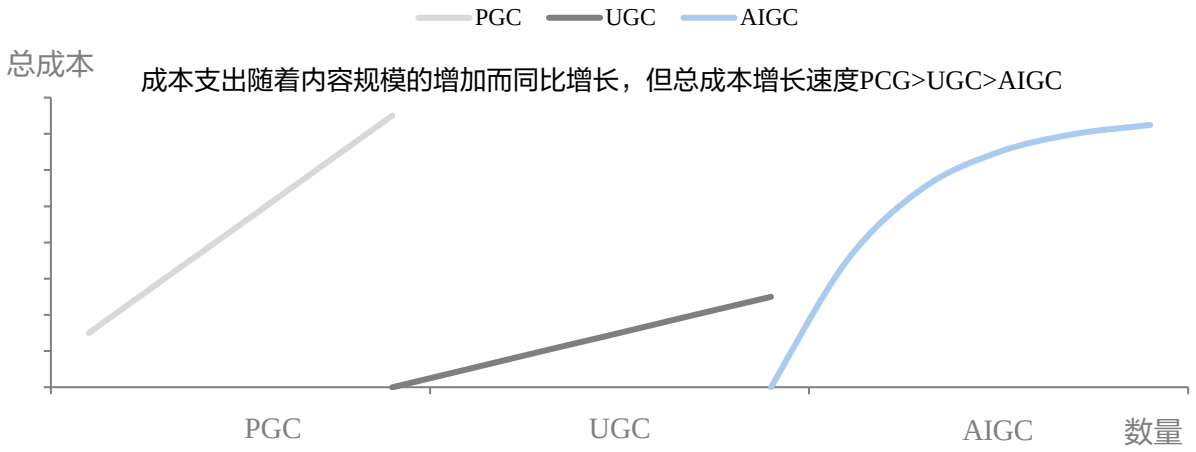
内容生成的数量



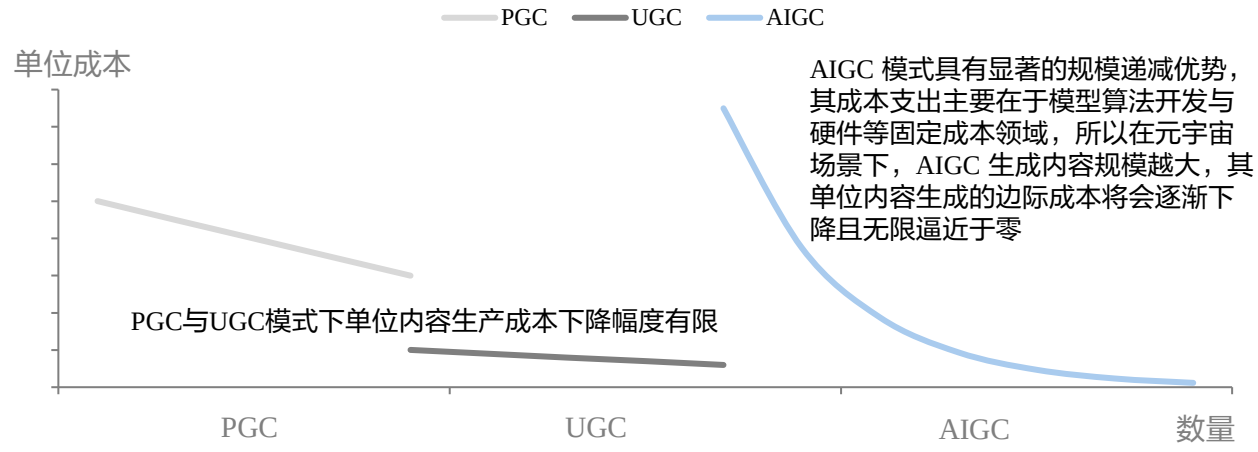
内容生成的质量



内容生成的总成本



内容生成的单位成本



核心分析框架：机器学习分为训练和推理，数据决定上限，算法逼近上限

国海证券 SEALAND SECURITIES

- **机器学习可以分为训练和推理两个阶段**，训练是指使用已知数据集训练机器学习模型；推理是指使用已训练好的模型对新的数据进行预测、分类等任务。
- **数据和特征决定了机器学习的上限，模型和算法逼近上限。**

训练

收集数据 → **数据准备** → **特征工程** → **模型选择** → **模型训练** → **模型评估** → **模型调优** → **模型部署和应用**

- 收集数据：包括数据收集、预处理、存储；数据的质量和梳理对算法效果至关重要；也是机器学习的瓶颈之一。
- 数据准备：从原始数据提取有代表性、可解释的特征。
- 特征工程：自己写代码实现模型或使用现成的框架。
- 模型选择：
  - 无监督学习
  - 监督学习
  - 强化学习
- 模型训练：根据分类、回归、聚类等各种问题有相应评估指标。
- 模型评估：网格搜索、随机搜索、贝叶斯优化、梯度优化、集成方法等选取最优参数组合，提高模型性能。
- 模型调优：模型封装后，导入生产环境进行推理。
- 模型部署和应用：API部署、边缘部署、集成部署、批处理部署、容器化部署

推理

数据预处理 → **模型加载** → **输入转换** → **模型推理** → **输出转换** → **结果后处理** → **结果展示**

深度学习框架	开发者	发布/开源时间	GitHub Star	功能	特点	受众
Tensor Flow	Google	2015.11	172k	端到端开源机器学习平台，拥有全面而灵活的生态系统，其中包含各种工具、库和社区资源，包括自定义、分布式训练、图像、文本、音频、结构化数据、生成式、模型理解、强化学习、tfEstimator等。	TensorFlow是工业型框架，自成立以来一直是面向部署的应用程序的首选框架，TensorFlow Serving和TensorFlow Lite可让用户轻松地在云、服务器、移动设备和 IoT 设备上进行部署。	谷歌、英特尔、ARM、GE医疗、PayPal、推特、联想、中国移动、WPS等。
Pytorch	Meta	2016.9	63.6k	基于Torch的Python开源机器学习库，包括分类器模型、计算机视觉模型、自然语言处理模型（聊天机器人，文本生成）等。还提供了两个高级功能： 1.具有强大的GPU加速的张量计算（如Numpy） 2.包含自动求导系统的深度神经网络 。	不仅能够实现强大的GPU加速，同时还支持动态神经网络，这一点是现在很多主流框架如TensorFlow都不支持的；简单易用可以实现快速验证，因此科研人员更为偏爱，各大期刊发表论文约80%使用Pytorch。	Meta、Amazon、Salesforce、Stanford University等。
Paddle Paddle	百度	2016.8	19.8k	集深度学习核心框架、基础模型库、端到端开发套件、工具组件和服务平台于一体，包括开发与训练框架、模型库、模型预训练/压缩工具及部署框架和引擎。	源于产业实践，始终致力于与产业深入融合目前飞桨已广泛应用于工业、农业、服务业等，服务 406万开发者。	英特尔、英伟达、浪潮、华为、寒武纪、中国联通、中信银行、中国南方电网、比特大陆、深交所、千千音乐等。

核心分析框架：数据是机器学习的核心，也是机器学习的瓶颈

数据决定了机器学习算法的性能、泛化能力、应用效果；数据获取、标注、清洗、存储也是机器学习瓶颈之一。

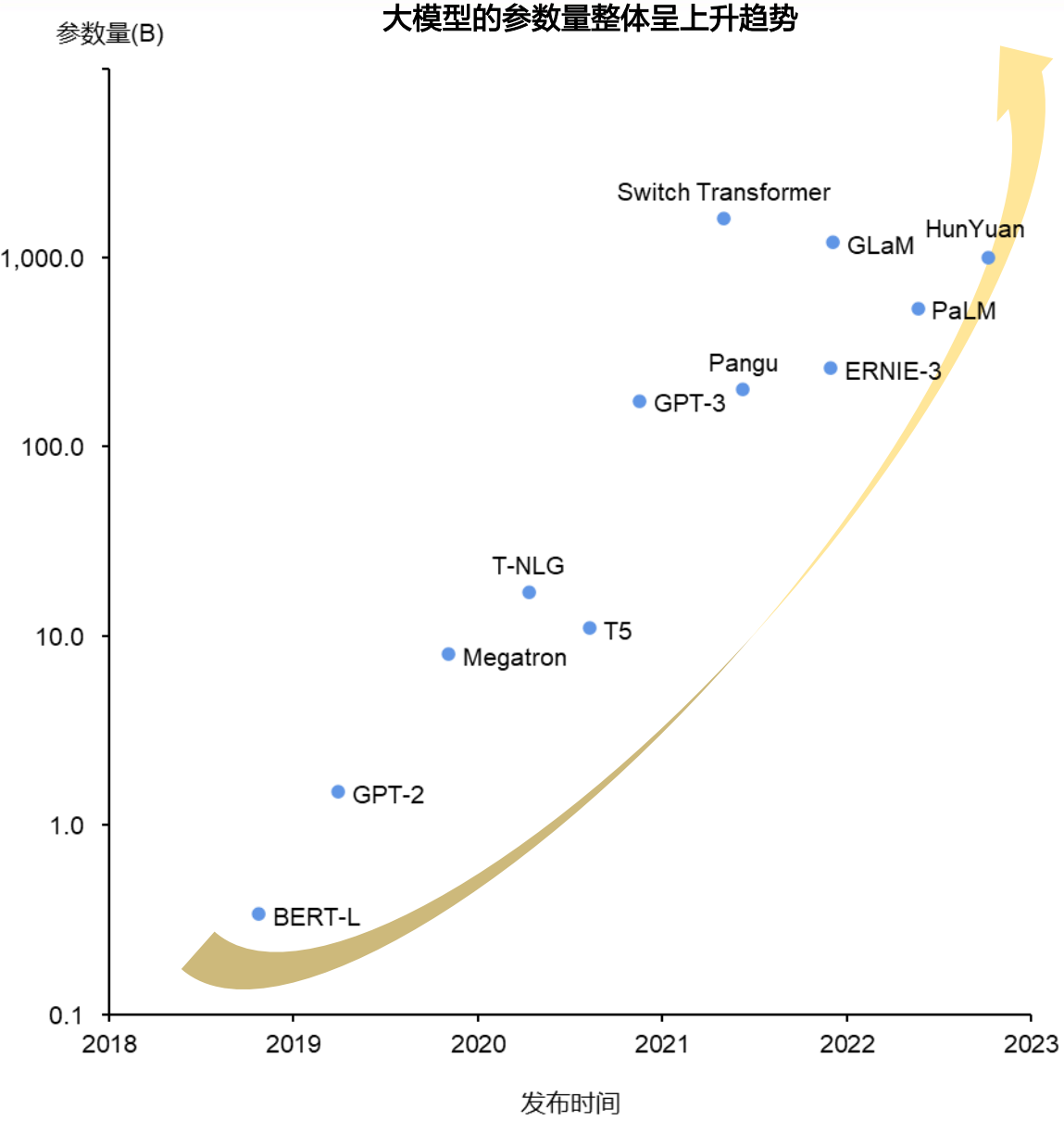
步骤	定义	成本占比	特点	展望
数据收集	通过爬虫、API接口、数据采购等方式，从不同的数据源中获取数据，例如文本、图像、视频、音频等。	30%	主要来源：1）公共数据库（API接口等）；2）企业自行收集（爬虫、问卷、访谈等）；3）第三方数据供应商采购；4）经授权的客户数据；5）平台模拟生成数据	随着AI商用化提速加码，数据反哺，可用数据将越来越多，数据获取边际成本将逐步降低。
数据标注	人工或半自动对原始数据进行标注，包括分类、语义分割（图像背景，物、人）、目标检测标注（边界框、关键信息）、序列标注（序列数据文本音频中，类别、实体、关键字等）。	40%-50%	1）无监督学习无需数据标注，部分简单数据，机器学习平台可自动化标注； 2）监督学习仍需标注数据； 3）专业领域、图像等复杂数据基本仍需人工标注。	无监督学习逐渐流行，自动化程度逐步升高，对于简单数据集标注需求下降；但专业领域和复杂数据集仍需要人工标注，且人工单位成本更高；随着人工智能快速发展，智能化程度的提升，数据标注全面自动化也是有可能的。
数据清洗	根据数据类型和需求，进行缺失值处理、异常值处理、噪声处理、重复数据处理、数据格式转换等。	20%-30%	减少错误和不准确数据对模型的干扰，提高模型准确性和可靠性。	目前数据清洗仍以手动为主，但在某些数据较为标准化的场景中（如日志数据、网络流量分析），一般可以通过编写自动化的脚本或者使用一些现成的工具来实现，以去除无效或者重复的数据；随着人工智能快速发展，智能化程度的提升，数据清洗全面自动化也是有可能的。
数据存储	将机器学习算法需要用到的数据保存到磁盘或内存中，以便后续的训练、测试和预测。		数据分为训练集（约60%）、验证集（约20%）、测试集（约20%）；需要选择合适的数据格式存储，不同格式会影响读取速度、空间占比等；大规模数据集需要进行分割后存储。	需要选择合适的数据格式存储，不同格式会影响读取速度、空间占比等；大规模数据集需要进行分割后存储。

资料来源：专家调研，Appen，Defined.ai，Mindflow，PhoenixNAP，EasyAI，国海证券研究所

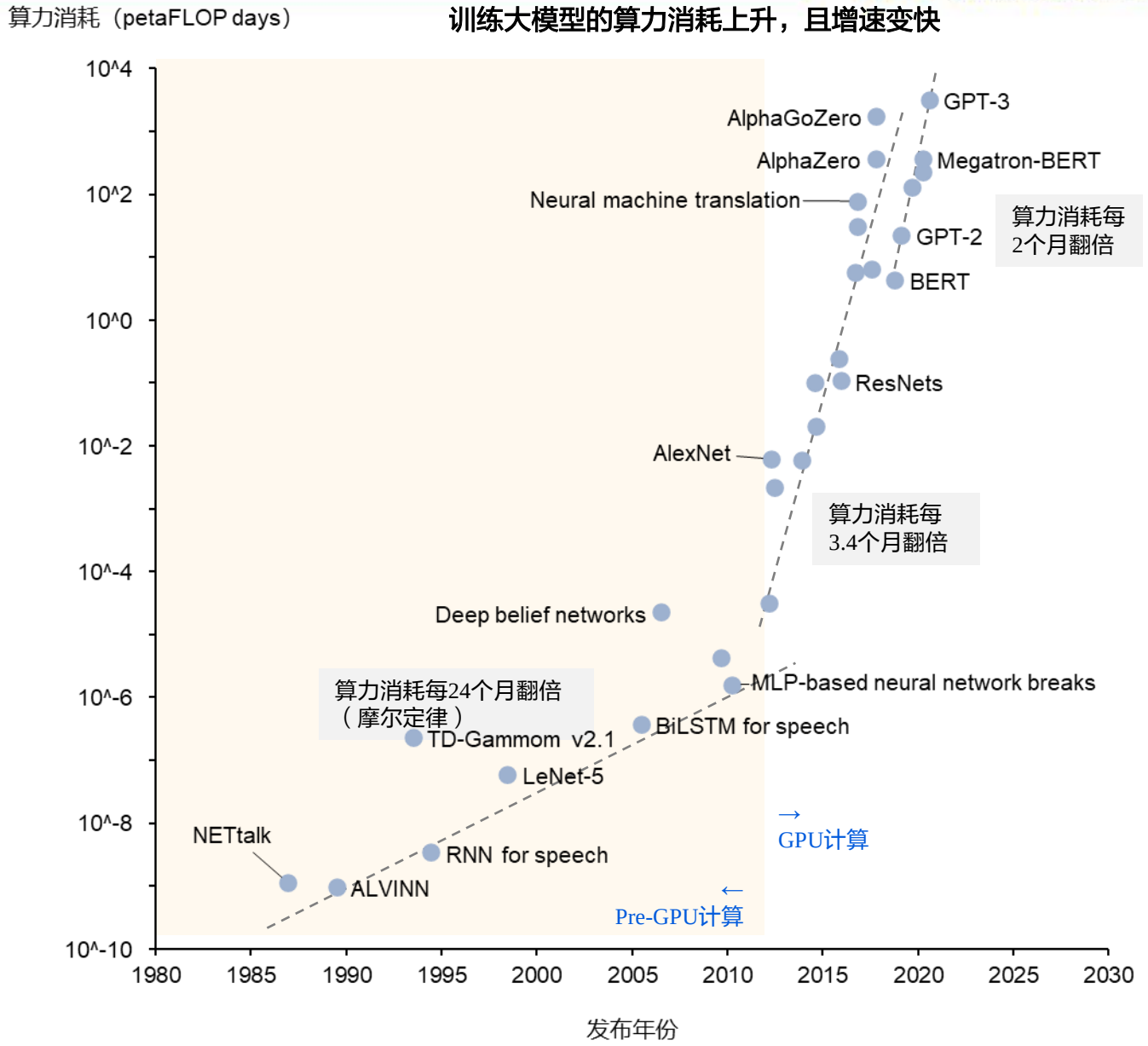
核心分析框架：随着模型参数量的提升，算力需求显著增加

国海证券 SEALAND SECURITIES

大模型的参数量整体呈上升趋势



训练大模型的算力消耗上升，且增速变快



核心分析框架： AIGC——生产力的革命

国海证券 SEALAND SECURITIES

类型	任务	应用
文本生成	• 交互文本：闲聊机器人、文本交互游戏； • 非交互文本：结构化/非结构化、辅助性写作。	ChatGPT、Writesonic、Conversion..ai、Snazzy Al、Copysmith、Copy.ai、彩云小梦等。
音频生成	语音克隆、文本生成特定语音、音乐生成等。	Deepmusic、AIVA、Landr、IBM Watson Music、Magenta、网易-有灵智能创作平台等。
图像生成	图像编辑/修复、风格转化、图像生成（AI绘画）等。	GLIDE、DiscoDiffusion、 Big Sleep、StarryAl、VOMBO Dream、百度文心ERNIE-VLG等。
视频生成	视频编辑（AI换脸、特效、删除特定主体、跟踪剪辑等）、自动剪辑等。	Gliacloud、Pencil、VideoGPT、百度智能视频合成平台VidPress、慧川智能等。
3D生成	目前主要是基于图像、文本生成3D建模；AR、VR；3D打印等。	DreamFusion、GET3D、3DiM等。
数字人生成	视频生成、实时交互	腾讯、网易、影谱科技、硅基智能、倒映有声等。
游戏生成	• 元素生成：游戏场景、剧情、NPC生成； • 策略生成：对战策略等。	rct AI、超参数、腾讯AI Lab、网易伏羲等。
代码生成	代码补全、自动注释、根据上下文/注释自动生成代码等。	Codex、Tabnine、CodeT5、Polycoder、Cogram等。
跨模态生成	目前主要是文本生成图像、视频，根据图像视频生成文本等；未来将有更多跨模态应用。	ChatGPT4、百度文心、阿里M6等。

算法

生成式对抗网络（GAN）

- 2014年提出，由**生成器网络（Generator）**和**判别器网络（Discriminator）**组成，相互博弈、对抗，不断提高生成样本真实性和判别器准确性。
- **优点**：生成样本质量高，无需大量数据标注，适用于多种数据类型，可用于数据增强。
- **缺点**：训练不稳定、容易崩溃，生成样本难控制，需要大量计算资源，容易过拟合。

多模态预训练模型

- 2019年提出，多模态数据预训练，实现多种模态数据的联合表示。
- **优点**：泛化能力、数据利用率、模型鲁棒性和可迁移性高。
- **缺点**：数据、算力需求大，特定任务需调参。

扩散模型（Diffusion）

- 2021年提出，相较于GAN，是图像生成领域的一大进步，不需要明确地计算数据的先验概率分布。通过“扩散”来执行隐空间中的推断。
- **优点**：更加灵活建模，样本多样性、可控性更高，训练过程简单、可扩展。
- **缺点**：数据、算力需求大，过程复杂，模型鲁棒性较低。

资料来源：量子位，wuhu动画人空间，数据派THU，Forbes，《Generative Adversarial Networks》Goodfellow等，《Using Pre-Training Can Improve Model Robustness and Uncertainty》Hendrycks等，《Diffusion Models Beat GANs on Image Synthesis》Dhariwal等，国海证券研究所

核心分析框架： ChatGPT史上用户数增长最快

国海证券 SEALAND SECURITIES

2015-11-11	2018-6	2019-2-14	2020-5-28	2022-1-27	2022-11-30	2023.3
Open AI宣布成立 • 初期为非盈利AI研究公司性质； • 启动资金**10亿美元**。	**GPT-1发布** • **参数量1.17亿**；BooksCorpus数据集作为语料库，Tokens为**1.3B**； • 结合无监督学习及有监督的微调。	**GPT-2发布** • **参数量15亿**；Tokens为**15B**； • 学习在无明确监督情况下执行多种任务。	**GPT-3发布** • **参数量1750亿**；Tokens为**499B**； • 结合少样本学习及无监督学习。	**InstructGPT发布** • **参数量13亿**； • 运用RHLF，利用奖励模型训练学习模型； • 在遵循指令及输出内容等性能方面优于GPT-3。	**ChatGPT发布** • 基于**GPT-3.5**预训练模型； • 截至2023年1月末，**活跃用户超过1亿**，成为**史上用户增长最快的应用**。	**预计推出ChatGPT4** • 预计为**多模态大模型**（语音、图像、视频）； • 新必应已集成ChatGPT4。

ChatGPT发布后市场反应热烈

表象：杰出的用户体验

- **ChatGPT功能**：回答后续问题、承认错误、质疑不正确的要求以及拒绝不适当的请求。
- **理解用户输入信息意图**，回答内容完整有逻辑、有条理，重点清晰；
- 真正做到**多轮沟通**，对上下文有理解和记忆，对话能力更强。

主因：算法的突破

- **LLM（large language model）**：当模型规模超过某个阈值之后，对于通用任务的效果会显著提升；
- **无监督学习（Unsupervised pre-training）**：又称可预测学习，该学习方式使得ChatGPT在无人工标注数据的条件训练，数据更多、数据成本更低，模型泛化能力更强；
- **CoT(Chain of Thought)思维链**：该算法使得模型生成推理路径，并在敏感话题方面避免了无法回答的问题；
- **RLHF(Reinforcement Learning From Human Feedback)人类反馈强化学习**：ChatGPT能够凭借强化学习的方式不断优化人类反馈的语言模型。

背后：算力、数据、人才、资金的投入

- **算力、资金的投入**：GPT3.5训练阶段总算力消耗约3640PF-days，约使用10000个GPU+285000个CPU，OpenAI耗费10亿美元租用Azure；截止202301，每个月,ChatGPT预计花费公司1200万美元；
- **数据**：ChatGPT1训练数据来自公开的电子书；ChatGPT2训练数据来自Reddit； ChatGPT2训练数据来自82.2%预处理的CommonCrawl语料库、13.5%的线上图书（GPT1数据集及Bibliotik,）、3.8% Reddit 。
- **OpenAI创始人、技术团队**：OpenAI创始人均为科技人才，现有375名正式员工，OpenAI一年人员支出高达8931万美元，人均约为23.8万美元/年/人。

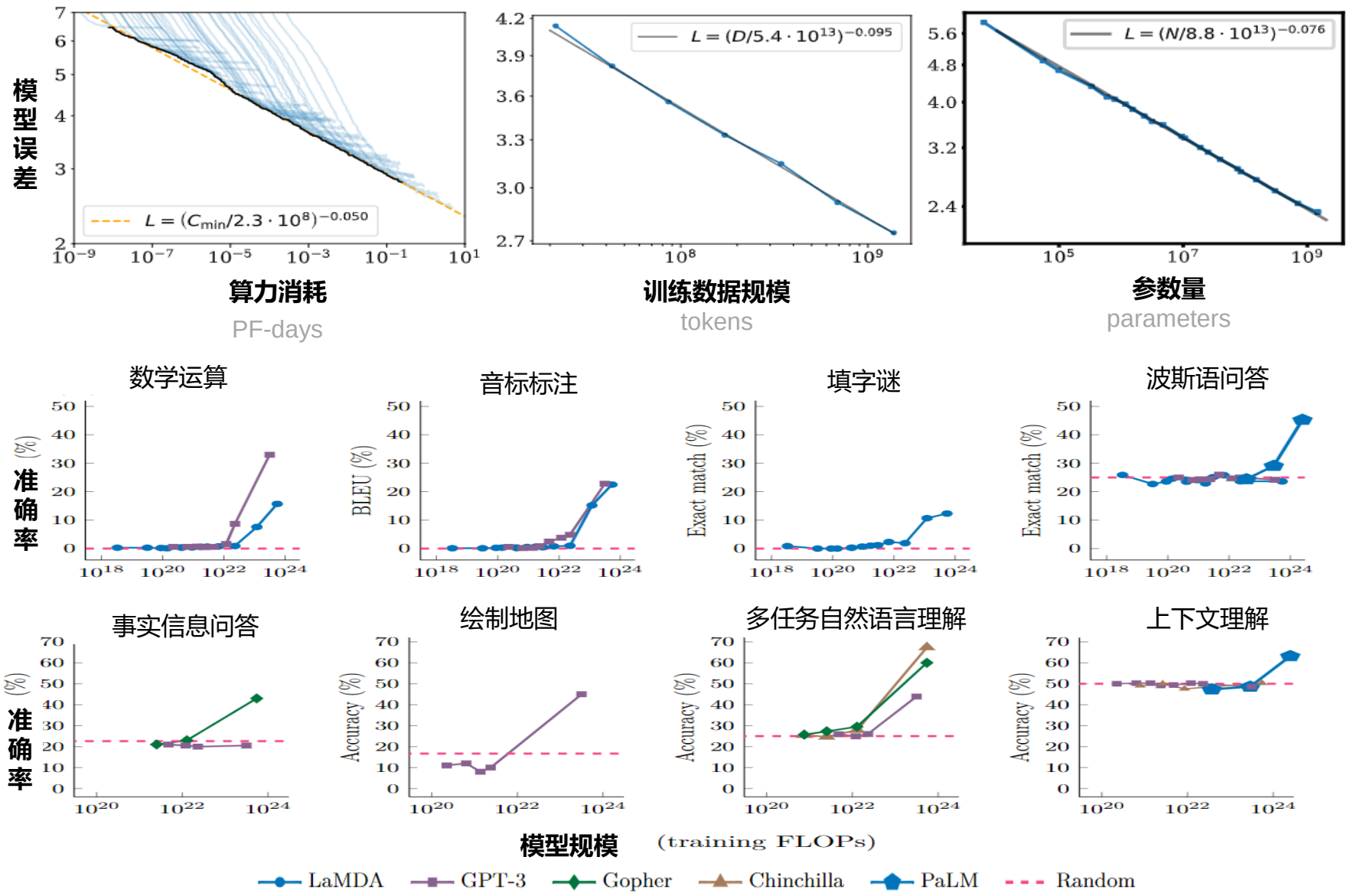
核心分析框架：当模型规模达到某个阈值时，模型出现涌现能力

2020年1月，OpenAI发表论文《Scaling Laws for Neural Language Models》，探讨模型效果与模型规模之间的关系。

结论是：模型的表现与模型的规模之间服从Power Law，即随着模型规模指数级上升，模型性能实现线性增长。

而在**2022年8月**，Google发表论文《Emergent Abilities of Large Language Models》，重新探讨了模型效果与模型规模之间的关系。

结论是：当模型规模达到某个阈值时，模型对某些问题的处理性能突然呈现快速增长。作者将这种现象称为Emergent Abilities，即涌现能力。



资料来源：《Scaling Laws for Neural Language Models》Kaplan等，《Emergent Abilities of Large Language Models》Wei等，国海证券研究所

核心分析框架： ChatGPT采用RLHF学习机制，效果优于GPT-3的无监督学习

GPT-3采用无监督学习机制，优点在于无需人工进行数据标注，可以节省模型训练成本，模型泛化能力更强。

而ChatGPT采用RLHF学习机制，即人工反馈的强化学习，属于强化学习。不同于传统的相比于传统的有监督学习机制，ChatGPT无需提前对训练数据进行标注，而只需要对人工对模型输出的结果进行评分，从而可以节省人力。虽然相比于GPT-3，ChatGPT需要消耗一定的人力，但是模型结果会更加符合人类偏好。

有监督学习vs无监督学习

			数据标注成本	模型精度
完全监督学习	所有数据都有标签	模型	高	高
半监督学习	含标签数据 无标签数据	模型	中	中
无监督学习	无标签的数据	模型	低	低

RLHF学习机制

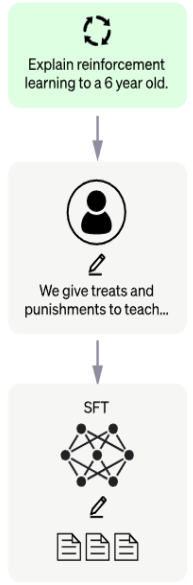
阶段1

收集演示数据并训练

随机从信息库抽取指令

专业的标注者对制定的提示给出高质量回答

专业人员用标注数据来调优GPT-3.5



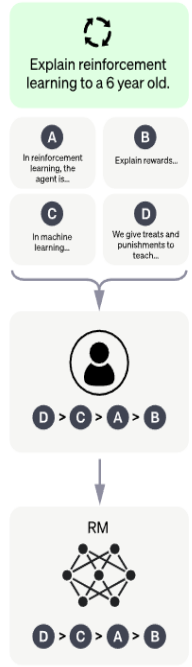
阶段2

通过人工标注训练数据来训练回报模型

标注一批模型产出及提示

标注人员根据多种标准对许多答案从优到差进行排序

利用以上排序结果来训练回报模型



阶段3

使用PPO强化学习法优化回报模型 - 根据RM评分结果更新预训练模型的参数

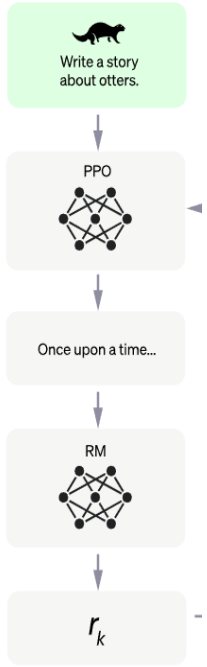
从用户提交的指令/问题中随机抽取一批新的命令

由监督模型初始化PPO模型的参数

PPO模型生成回答

用回报模型计算前一阶段训练好的模型给出的回答，得到分数

回报分数/策略梯度可以更新PPO模型参数



核心分析框架： AIGC何时突破工业红线？关注数据、算法和商业模式破局

国海证券 SEALAND SECURITIES

	大模型：通用型、任务型、行业级	小模型：专业领域，细分行业
参与方	**大模型技术巨头+第三方服务商** 巨头：微软、谷歌、meta、百度、阿里、华为、腾讯等； 第三方服务商：SaaS厂商、其他技术厂商等。	**AI企业** 商汤、科大讯飞、旷视、云从、依图、虹软、格灵深瞳、拓尔思等。
解决问题	**数据是瓶颈：数据增强、迁移学习、数据合成、数据要素市场实现数据共享、数据反哺加速商业化飞轮。** 1）数据获取：大模型所需数据量较大，而现实世界缺乏大量且优质数据； 2）数据存储、传输、管理：海量数据训练，读取和处理速度非常关键。	专业领域、长尾场景数据较少。
	算力是支撑：短期-国内云厂商等均早有囤货布局；长期-硬件进步、算法优化、并行计算、量子计算。 大模型往往需要大量计算资源，且模型参数仍在快速膨胀；但AI芯片全球短缺，英伟达A100、H100被禁止向中国供货。	**商业价值闭环：技术突破、AI企业深耕垂直细分行业（know-how、先发优势）、规模效应+飞轮效应双轮驱动。**
	人才是关键：“挖角”、企业高校合作。 美国人工智能一直领先，国内顶尖技术人才从数量、质量都存在较大差距，AI领域（尤其是CV）优秀的华人很多，但更多的在谷歌、微软、Meta等企业；北京的微软亚洲研究院的人才输出几乎撑起中国AI“半壁江山”。	技术成本（前期训练成本、数据成本、人才成本，后期使用的推理成本），与带来的增量或给企业实现降本增效相比，还不足以驱动企业投入AI。
	商业价值闭环：技术进步、国家支持、巨头推动、生态建设、市场化教育。 • 海外软件生态成熟，企业/个人用户付费意愿更高；国内市场无论是生态和市场都存在较大差距。 • 国内外目前商业模式、付费逻辑尚未跑通。	
	价值观、伦理、政治风险等：从技术层面让AI更可控，不要发展的那么快。	

核心分析框架：互联网大厂全面布局，中小厂商主要发力中下游环节

国海证券 SEALAND SECURITIES

AIGC产业链图谱

上游

■ 云计算
（000977.SZ）浪潮信息
（9988.HK）阿里
（9888.HK）百度集团
（0700.HK）腾讯
（未上市）华为

■ 芯片
（300474.SZ）景嘉微
（9888.HK）百度集团
（NVDA.O）英伟达
（9988.HK）阿里巴巴
（688256.SH）寒武纪
（002405.SZ）四维图新
（688981.SH）中芯国际
（未上市）地平线

■ IDC
（300738.SZ）奥飞数据
（603019.SH）中科曙光
（9698.HK）万国数据
（CD.US）秦淮数据

■ 光模块
（300308.SZ）中际旭创
（300502.SZ）新易盛
（220081.SZ）光迅科技

■ 服务器液冷
（600756.SH）浪潮信息
（300017.SZ）网宿科技
（000938.SZ）紫光股份

■ 数据供给方
（688787.SH）海天瑞声

中游

■ 多模态
（9888.HK）百度
（9988.HK）阿里巴巴
（0700.HK）腾讯控股
（300612.SZ）宣亚国际
（300418.SZ）昆仑万维
（603466.SH）风语筑
（688327.SH）云从科技
（2121.HK）创新奇智
（MSFT.O）微软
（GOOGL.O）谷歌
（NVDA.O）英伟达
（META.O）Meta
（未上市）珍岛
（未上市）中科闻歌
（未上市）澜舟科技

■ 策略生成
（未上市）rct AI
（未上市）超参数科技

■ NLP
（9988.HK）阿里巴巴
（002230.SZ）科大讯飞
（9888.HK）百度集团
（002230.SZ）科大讯飞
（688111.SH）金山办公
（300058.SZ）蓝色光标
（002292.SZ）奥飞娱乐
（学术机构）清华大学

■ 3D生成
（未上市）聚力维度

■ 代码生成
（MSFT.O）微软
（学术机构）清华大学
（学术机构）中国科学技术大学（学术机构）哈尔滨工业大学

■ 虚拟人
（300229.SZ）托尔思
（002467.SZ）二六三
（688088.SH）虹软科技
（002362.SZ）汉王科技
（300113.SZ）顺网科技
（未上市）小冰公司
（未上市）倒映有声
（未上市）相芯科技
（未上市）心识宇宙

■ 视频生成
（688039.SH）当虹科技
（0020.HK）商汤
（未上市）迈吉客
（未上市）影谱科技

下游

■ 电商
（300785.SZ）值得买

■ 传媒
（301270.SZ）汉仪股份
（300364.SZ）中文在线
（000681.SZ）视觉中国
（300781.SZ）因赛集团
（300624.SZ）万兴科技

■ 营销
（301052.SZ）果麦文化
（002803.SZ）吉宏股份
（301171.SZ）易点天下

■ 教育
（300081.SZ）恒信东方

■ 虚拟人
（300182.SZ）捷成股份
（002354.SZ）天娱数科

■ 游戏
（002624.SZ）完美世界
（0700.HK）腾讯控股
（300459.SZ）汤姆猫

■ 政务
（300075.SZ）数字政通
（002530.SZ）金财互联

■ C端应用
（MSFT.O）微软
（GOOGL.O）谷歌
（未上市）写作猫
（未上市）写作狐
（未上市）盗梦师
（未上市）诗云科技
（未上市）ZMO.ai
（未上市）影谱科技
（未上市）帝视科技
（未上市）不咕剪辑

核心分析框架：产业链各环节发展趋势

类型	代表机构	上游		中游		下游		竞争优势
		算力	数据	大模型	小模型	行业合作	内部赋能	
互联网大厂（全面布局）	百度	百度云 昆仑芯片	百度各产品数据 行业合作伙伴数据	文心大模型	包括在文心大模型中的各类行业模型	与B端企业有广泛合作	百度搜索 百度各类产品的内容推荐	先发优势 具有较多的行业数据和专业知识数据
	阿里	阿里云 平头哥芯片	淘宝、天猫电商数据 阿里云B端数据	阿里M6大模型	-	合作较多	电商搜索 阿里云和企业服务	在大模型研发上具有资金和人才优势
	腾讯	腾讯云	微信用户数据 腾讯视频、新闻数据 腾讯游戏数据	混元大模型	腾讯游戏AI	合作较少	腾讯游戏AI NPC 微信等产品的自媒体创作、内容推荐	具有较多的用户数据和娱乐内容数据
	华为	华为云 海思芯片	手机用户数据	盘古大模型	盘古大模型中的各类行业模型	合作较少	较少	深耕上游和中游 赋能下游厂商
	谷歌	谷歌云	搜索数据 谷歌学术 Youtube数据	Imagen、ExTS、PaLM等	-	合作较少	Bard+Google	AI赋能搜索业务，同时快速积累新用户
	微软	Azure云	Office用户数据 Bing搜索数据	LayoutLM、DiT以及OpenAI旗下的大模型	-	较多企业接入chatGPT接口	chatGPT+Bing chatGPT+Office	AI赋能搜索和办公业务，同时快速积累新用户
学术机构（中游为主）	清华大学 中国科学技术大学 哈尔滨工业大学等	主要通过外购	互联网公开数据	√	√	合作方向主要为学术研究	-	政府支持 人才储备
中小厂商（中下游为主）	中游小模型厂商	主要通过外购	垂直行业数据	-	垂直行业模型	√	√	行业know-how 积累行业数据
	下游应用厂商	主要通过外购	垂直行业数据	-	-	√	√	客户粘性 用户粘性
产业链核心竞争要素		规模效应 政府补助 前期研发投入	数据规模 数据质量 数据获取成本	资金能力 技术能力 人才储备	行业Know-how 行业数据	先发优势 行业know-how	内部用户规模和业务数据积累；业务和AI技术结合的可行性	
产业链未来发展方向		头部效应↑ 边际成本↓	通用类数据集中于大厂，而垂直行业数据分散	头部效应↑	百花齐放	通用型内容生成集中于大厂，而垂直行业解决方案百花齐放	大厂对外提供服务的同时内部赋能，小厂采取外购的方式更加经济	

资料来源： 澜舟科技，AspenCore，海思，前瞻产业研究院，IDC，机器之心，国海证券研究所

核心分析框架：大模型商业化初启，小模型在部分领域已实现商业价值闭环

国海证券 SEALAND SECURITIES

	大模型			小模型	
商业模式	MaaS（Model as a Service）			垂直行业解决方案	
	1)按调用次数或调用量（Tokens等）收费；	2）按年/月订阅套餐收费；	3）定制服务，特定领域再开发，将大模型和数据库打包，按项目收费。	1）纯软件及平台；	1）一站式解决方案
面向用户	企业、机构、个人		企业、机构	细分行业企业	
毛利率	推理算力成本，毛利率可达80%+。		含再开发项目实施费用。	标准化产品，毛利率可达90%+。	含外购硬件，毛利率30%-70%。
提供商	OpenAI、微软、谷歌、Meta、百度、阿里、华为、腾讯、商汤、科大讯飞、字节、京东等。			科大讯飞、商汤、旷视、云从、依图、虹软、格灵深瞳、云天励飞、拓尔思、海康威视等。	
商业模式	大模型厂商自用，实现增量或降本增效。	云厂商“MaaS+IaaS”打包输出，实现IaaS收入增长和增量服务收入。	替代翻译、美工、原画师、程序员、分析师、设计师等繁琐重复的低端工作。	垂直行业解决方案，包括SDK产品、一站式落地解决方案。	
付费逻辑	谷歌、微软必应搜索引擎，YouTube视频创作等，阿里电商营销产品，腾讯企业微信、腾讯会议相关产品等，字节内容创作等；基于C端用户使用量内部付费。	大模型厂商+SaaS厂商，打造更多可直接面向C端的产品，SaaS厂商根据调用情况付费。	1）企业开发者调用后自用或个人用户自行调用，基于自身需求调用付费；2）为SaaS厂商提供产品付费。	智慧城市、智慧交通、智慧楼宇、智慧园区、智慧医疗、智慧金融、智慧生活、智能制造等多领域均有企业布局，在过去主要是感知、分析、决策式AI，部分存在生成式AI，已有部分行业实现商业价值闭环，主要是传统软件收费逻辑，不同行业略有不同。	
中美差距	差距不大且均有较大需求，甚至国内厂商的产品更加丰富多元。	生态差距较大，美国SaaS厂商面向全球，中国SaaS行业尚在快速发展中。	海外付费意识更高。	中美企业格局略有差异，美国头部效应更为明显，主要由细分行业龙头或者科技巨头提供相关AI驱动的解决方案；中国不局限于科技巨头和行业龙头，还有众多AI企业在众多细分行业、领域布局。	

核心分析框架：总成本持续提升，但同级别参数消耗量将显著下降

表：大模型训练成本中各成本占比概览

成本项	占比
算力成本	**40%-70%**
设备折旧	14%-24.5%
存储	10-17.5%
电费	12%-21%
宽带	4%-7%
数据成本	**20%-35%**
数据收集	6%-10.5%
数据标注	8%-17.5%
数据清洗	4%-7%
数据存储	
人力成本	**10%-25%**

注：参考ChatGPT、百度文心、阿里M6、华为盘古大模型数据

- 随着参数量快速膨胀，算力成本会持续上升；
- 但随着模型压缩、蒸馏等，同参数量级别的模型算力消耗量会显著下降。

- 数据获取：随着应用较快数据反哺，数据获取边际成本将下降；
- 数据标注：有两个方向，一是无监督学习流行、标注自动化提升，数据标注成本下降；而是对于专业领域、图像视频等复杂数据标注需求提升。
- 随着数据量快速膨胀，训练数据集需求越来越大，数据存储成本也将相应提升。

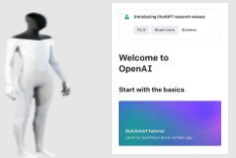
- AI资产复用、自动化程度提升，规模效应，单位人力成本下降。

表：各大模型全局训练成本概览

模型	算力成本占比	数据成本占比	人力成本占比	单次完整训练价格（万美元/次）	全年完整训练次数（次）	全年训练成本（万美元）	已投入金额（万美元）
ChatGPT3	70%	20%	10%	400-1000	1-2	2000左右	4300左右
ChatGPT3.5	60%	25%	15%	400-1000	1-2	不到2000	

一、行业篇：人工智能发展步入新阶段，AIGC创造新机遇

每一轮人机交互的变革都会带来产业级投资机会

变革节点	PC操作系统 Macintosh、Windows	浏览器 IE浏览器、网景浏览器等	搜索引擎 Yahoo、Google等	智能手机 Iphone等	ARVR Oculus Quest、HTC Vive、Hololens等	人机共生 人形机器人、AIGC等
						
	1984年苹果推出划时代的Macintosh计算机，不仅首次采用图形界面的操作系统，并第一次使个人计算机具有了多媒体处理能力；1985年微软推出Windows系统	1993年NCSA中Mosaic项目的负责人辞职并建立了网景通讯公司，推出网景浏览器；1995年微软推出IE1.0浏览器，作为Windows 95的默认浏览器，改变了用户网上冲浪方式	1995年Yahoo公司正式成立，1998年开始转向使用Inktomi的搜索数据；1998年Google成立，后NetScrpe放弃Excite，开始使用Google的搜索数据，具备里程碑意义	2007年苹果发布自PC以来最具变革性的产品——iphone 2G，大部分操作都将由用户触控屏幕实现；iPhone 4在外观、显示、芯片均大幅改善，并提供六轴动作感应	2016年Facebook正式发售Oculus rift消费者版本，被称为消费级VR设备元年；2015年索尼推出PlayStation VR；2015年微软发布混合现实的智能眼镜Hololens	2014年，波士顿动力发布初代Atlas (2021已实现跑酷)；2022年，Tesla预计发布Optimus原型机；2022年11月Open AI发布人工智能技术驱动的自然语言处理工具ChatGPT
人机交互模式	鼠标+键盘，可点击，但交互模式单一且不智能，人较为被动	鼠标+键盘，浏览器聚合功能改善交互成本	鼠标+键盘，搜索引擎的检索功能以人为中心，降低精准信息获取门槛	触屏+键盘，人机交互更加直观便捷，人处于主动地位	手势追踪、Inside-out、Outside-in、眼球追踪等，交互方式多元化，沉浸感强	人机共生，文字、音频、视频、3D、策略等交互模式融合，智能化程度显著提升
产业机会	操作系统、早期邮箱、早期超级计算中心等	光缆/运营商、浏览器、门户网站、通讯软件等	搜索引擎、众多PC互联网网页应用等	手机硬件产业链、应用商店、各大手机APP应用等	VRAR硬件产业链、云计算/边缘计算、视频&直播&游戏应用等	机器人硬件产业链、AI产业链（模型算力数据等）、下游应用等

资料来源：36氪，界面新闻，甲子光年，开源资讯，Companyhistory，digitaltrends，IDG news service，likecs，payititi，TechCPR，the verge，VRcompare，国海证券研究所

AI发展历程：期待算力、数据、算法的突破，迈向强人工智能AGI阶段

国海证券 SEALAND SECURITIES

规则导向
1950s-1980s
感知机、专家系统

理论发展

机器学习
1980s-21世纪初期
支持向量机、神经网络、决策树

硬件奠基，算法发展

深度学习
21世纪初期至今
多层神经网络模型

核心硬件发展，算法突破

自主学习
未来发展方向
基于大规模数据识别、预测、学习、决策

算力、数据、算法全面突破

运营营销决策、金融风险评估、医疗诊断等。

始于1950s，让计算机感知和理解现实世界，**兴盛于计算机硬件发展与大数据出现。**

感知式AI

分析式AI

始于1960s，利用逻辑推理等方法进行分析和解决；**兴盛于机器学习、数字挖掘技术兴起。**

决策式AI

始于1970s，分析之后，进行智能决策；**近年来才开始真正兴盛。**

始于1980s，起源于分析式AI，生成新的内容； **2022年迎来突破。**

生成式AI

生成式AI生成各种数据、图像、语音等内容后，决策式AI对生成内容进行分析和决策，实现更加全面和智能的应用。

文本识别、语音识别、图像识别等。

自然语言生成、图像生成、音乐生成等。

弱人工智能阶段
ANI

强人工智能阶段
AGI

超人工智能阶段
ASI

AIGC发展历程：文本、代码生成技术较成熟，图片、视频生成值得期待

国海证券 SEALAND SECURITIES

受限于科技水平，AIGC仅限于小范围实验

AIGC从实验性转向实用性，受限于算法瓶颈，无法直接进行内容生成

深度学习算法不断迭代，人工智能生成内容百花齐放，效果逐渐逼真

1950s~1990s 早期萌芽阶段

1950年，艾伦·图灵提出著名的“图灵测试”，给出判定机器是否具有“智能”的试验方法

1957年，第一支由计算机创作的弦乐四重奏《依利亚克组曲》完成

1966年，世界第一款可人机对话的机器人“Eliza”问世

1990s~2010s 沉淀积累阶段

2007年，世界第一部完全由人工智能创作的小说《I The Road》问世

2012年，微软展示全自动同声传译系统，可将英文演讲内容自动翻译为中文语音

2010s至今 快速发展阶段

2014年，Ian J. Goodfellow提出生成式对抗网络GAN

2017年，微软“小冰”提出世界首部100%由人工智能创作的诗集《阳光失了玻璃窗》

2018年，英伟达发布StyleGAN模型可以自动生成高质量图片

2018年，人工智能生成的画作在佳士得拍卖行得以43.25万美元成交，成为首个出售的人工智能艺术品

2019年，DeepMind发布DVD-GAN模型用以生成连续视频

2021年，OpenAI提出了DALL-E，主要用于文本与图像交互生成内容

	2020年前	2020	2022	2025E	2030E	2050E
文本生成	垃圾邮件识别 翻译 基础问答	基础文案写作 草稿撰写	长文本写作 草稿撰写与修改	专业文本写作（如科研、金融、医疗）	终稿写作 写作能力超越人类平均水平	终稿写作 写作能力超越人类专业人士
代码生成	单行代码生成	多行代码生成	长代码写作 准确率提升	支持更多代码语言 支持更多垂直行业	输入文本即可自动生成产品原型	输入文本即可自动生成最终产品
图片生成			艺术生成 Logo生成 照片编辑与合成	产品原型设计 建筑原型设计	产品最终设计定稿 建筑最终设计定稿	设计能力超越艺术家、专业设计师、专业摄影师
视频/3D生成				视频初稿	视频初稿与修改	AI机器人 个性化的电影
游戏AI						个性化的游戏

内容创作模式进化：去中心化↑连接数量↑创作速度↑创作规模↑

现在

未来

创作主体	生产关系	创作过程	创作规模
专业机构	内容消费者 / 内容消费者 / 内容消费者 / 内容创作者 / 内容消费者 / 内容消费者 / 内容消费者	沟通 / 加工 / 创意 / 设备 / 渲染 / 镜头所见 / 拍摄 / 演绎	
平台用户			
用户+AI	内容共创者 / 内容共创者 / 内容共创者 / 内容共创者 / 内容共创者 / 内容共创者		
纯AI		创意 → 镜头所想	

内容创作模式进化：从供给转变为需求导向，从单次转变为多次生产

供给导向的真实世界→需求导向的虚拟世界

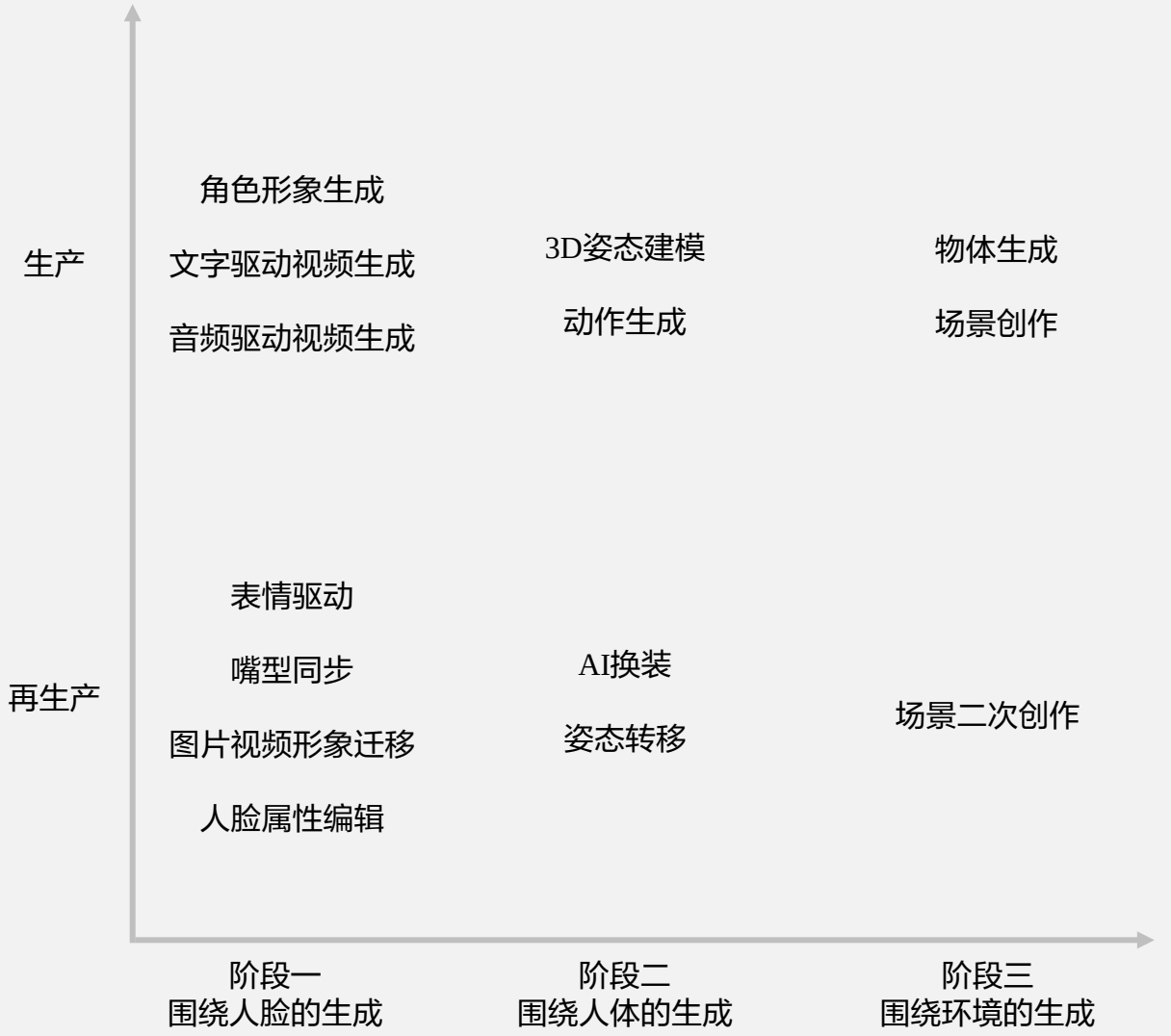
供给导向：真实世界信息的“数字化”+创意“数字化”

真实世界 → **虚拟世界**

需求导向：虚拟世界数字信息的传递

原生的真实世界	现实化的虚拟世界	数字化的虚拟世界	原生的虚拟世界
穿着	书籍	社交/社区	电子游戏
饮食	桌游	电商/交易	影视/动画
出行	手办	流媒体/媒体	设计/特效
住宿	Cosplay	协作/办公	训练/模拟
利用真实资源生产和消耗有形产品			**利用虚拟资源生产和消耗虚拟产品**

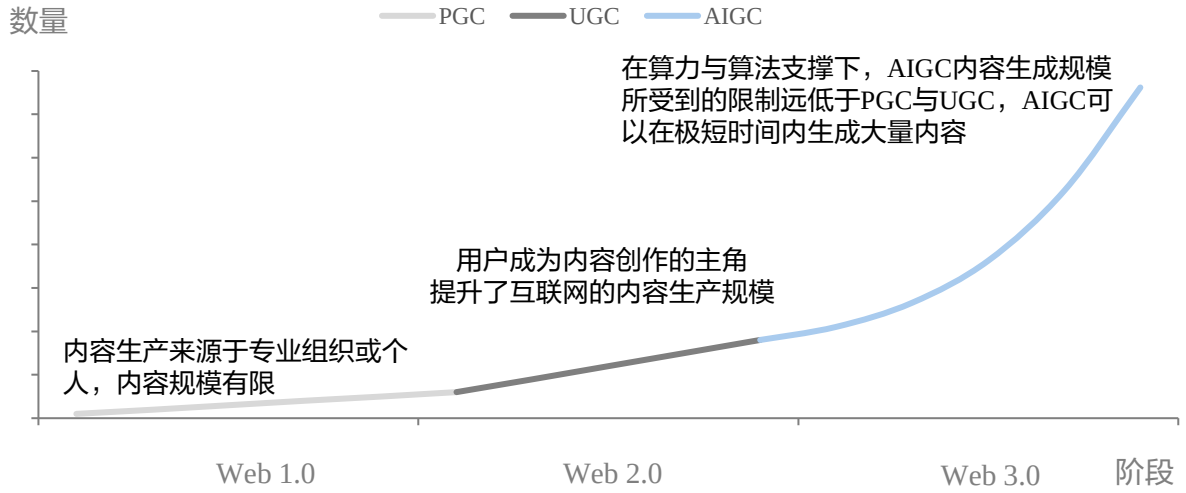
低效率的单次生产→高效率的多次生产



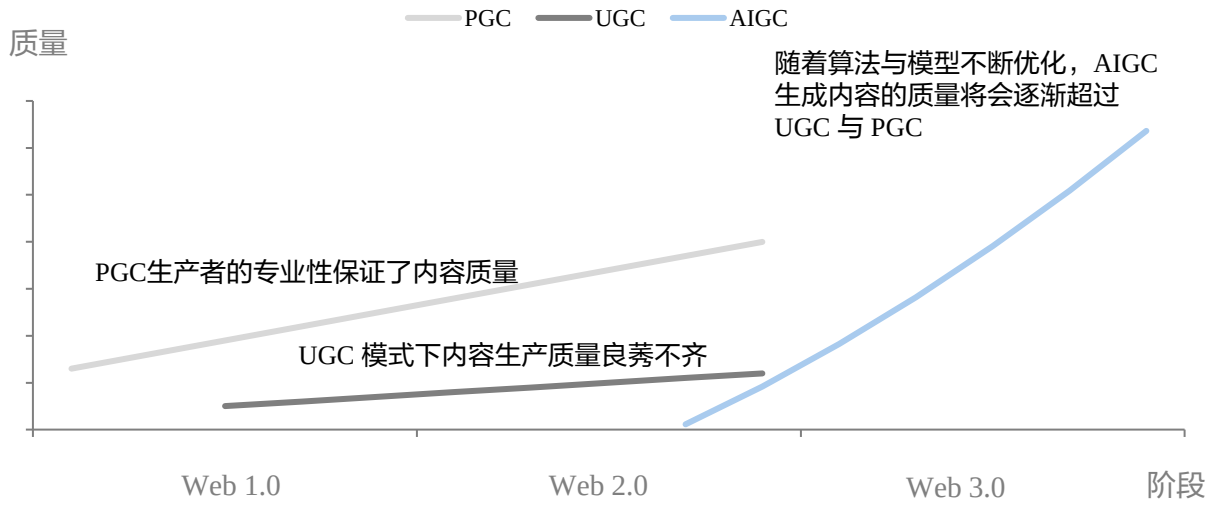
内容创作模式对比：AIGC实现内容创作呈高质量、大数量、低成本趋势

传统的 PGC 与 UGC 模式受到规模、质量和成本的制约，而AIGC 则能够有效地弥补 PGC 与 UGC 模式的不足，具有生成内容规模大、质量高、单位成本低的优势，将会成为元宇宙场景下的主要内容生成模式，从而为元宇宙建设提供内容支撑。

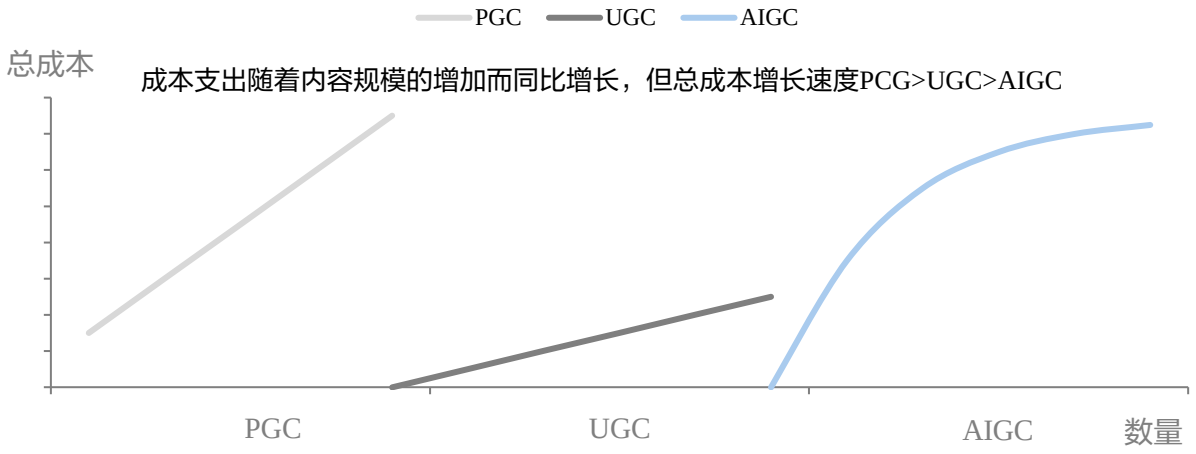
内容生成的数量



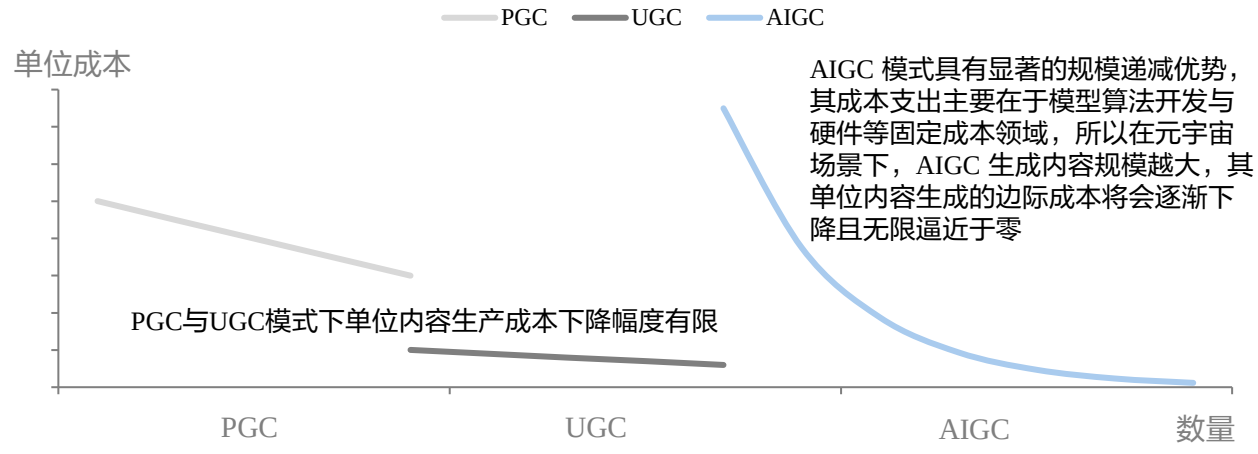
内容生成的质量



内容生成的总成本



内容生成的单位成本



AIGC演进趋势：辅助生产 → 自动化独立创作

随着人工智能算法的进步和算力的提升，AIGC 将逐步摆脱对 PGC 和 UGC 的依赖，从辅助内容生成转变完全独立创作，充分释放创作潜力，持续输出高质量、多样化、高自由内容，满足未来消费者对内容数量及质量的双重刚性需求。

分级		0级	1级	2级	3级	4级	5级	发展趋势
生产模式		生产人生产内容	机器辅助审核	机器辅助加工	机器有条件自动生产内容	机器高度自动生产内容	机器完全自动生产内容	AI 渗透率↑
生产主体	采集	生产人	生产人	生产人	生产人和机器	生产人和机器	机器	AI 渗透率↑
	加工	生产人	生产人	生产人和机器	生产人和机器	机器	机器	AI 渗透率↑
	审核	生产人	生产人和机器	生产人和机器	机器	机器	机器	AI 渗透率↑
生产力限制	采集	受限	受限	受限	部分受限	部分受限	不受限	生产力↑
	加工	受限	受限	部分受限	部分受限	不受限	不受限	生产力↑
	审核	受限	部分受限	部分受限	不受限	不受限	不受限	生产力↑
技术要求		• 素材上传、存储、分类、检索、权限设置 • 多媒体内容编辑，提供文字、图片、视频功能 • 内容在线批注、修改	• 支持内容审核，包括文字规范性核查，人物/机构/地域等实体属性核查等	• 自动标题、自动摘要、智能字幕、文本生成 • 在内容审核过程中自动屏蔽、剔除或修改内容	• 抓取线上数据 • 根据内容模板利用线上数据自动生成内容 • 采集素材的规范性与准确性审核	• 支持固定位置的线下设备进行数据采集 • 支持根据已设定的内容模板对原始数据进行加工后自动生成内容	• 支持可移动设备自动进行数据采集 • 分析原始数据，自动判断是否需要进一步采集，并根据素材挑选模板自动生成内容	技术能力↑

资料来源：中国新闻技术工作者联合会《机器生产内容自动化分级》团体标准，国海证券研究所

二、技术篇：算力是支撑，数据是核心，算法逐步迎来突破

机器学习：机器学习分为训练和推理，数据决定上限，算法逼近上限

- **机器学习可以分为训练和推理两个阶段**，训练是指使用已知数据集训练机器学习模型；推理是指使用已训练好的模型对新的数据进行预测、分类等任务。
- **数据和特征决定了机器学习的上限，模型和算法逼近上限。**

训练

收集数据 → **数据准备** → **特征工程** → **模型选择** → **模型训练** → **模型评估** → **模型调优** → **模型部署和应用**

- 收集数据：包括数据收集、预处理、存储；数据的质量和梳理对算法效果至关重要；也是机器学习的瓶颈之一。
- 数据准备：从原始数据提取有代表性、可解释的特征。
- 特征工程：自己写代码实现模型或使用现成的框架。
- 模型选择：
  - 无监督学习
  - 监督学习
  - 强化学习
- 模型训练：根据分类、回归、聚类等各种问题有相应评估指标。
- 模型调优：网格搜索、随机搜索、贝叶斯优化、梯度优化、集成方法等选取最优参数组合，提高模型性能。
- 模型部署和应用：模型封装后，导入生产环境进行推理。
  - API部署
  - 边缘部署
  - 集成部署
  - 批处理部署
  - 容器化部署

推理

数据预处理 → **模型加载** → **输入转换** → **模型推理** → **输出转换** → **结果后处理** → **结果展示**

深度学习框架	开发者	发布/开源时间	GitHub Star	功能	特点	受众
Tensor Flow	Google	2015.11	172k	端到端开源机器学习平台，拥有全面而灵活的生态系统，其中包含各种工具、库和社区资源，包括自定义、分布式训练、图像、文本、音频、结构化数据、生成式、模型理解、强化学习、tfEstimator等。	TensorFlow是工业型框架，自成立以来一直是面向部署的应用程序的首选框架，TensorFlow Serving和TensorFlow Lite可让用户轻松地在云、服务器、移动设备和 IoT 设备上进行部署。	谷歌、英特尔、ARM、GE医疗、PayPal、推特、联想、中国移动、WPS等。
Pytorch	Meta	2016.9	63.6k	基于Torch的Python开源机器学习库，包括分类器模型、计算机视觉模型、自然语言处理模型（聊天机器人，文本生成）等。还提供了两个高级功能： 1.具有强大的GPU加速的张量计算（如Numpy） 2.包含自动求导系统的深度神经网络 。	不仅能够实现强大的GPU加速，同时还支持动态神经网络，这一点是现在很多主流框架如TensorFlow都不支持的；简单易用可以实现快速验证，因此科研人员更为偏爱，各大期刊发表论文约80%使用Pytorch。	Meta、Amazon、Salesforce、Stanford University等。
Paddle Paddle	百度	2016.8	19.8k	集深度学习核心框架、基础模型库、端到端开发套件、工具组件和服务平台于一体，包括开发与训练框架、模型库、模型预训练/压缩工具及部署框架和引擎。	源于产业实践，始终致力于与产业深入融合目前飞桨已广泛应用于工业、农业、服务业等，服务 406万开发者。	英特尔、英伟达、浪潮、华为、寒武纪、中国联通、中信银行、中国南方电网、比特大陆、深交所、千千音乐等。

资料来源：各框架官网， Easy AI，GitHub，机器之心，国海证券研究所（注：GitHub Star为截止2023.3.13主体框架star数据）

数据：机器学习的核心，也是机器学习的瓶颈

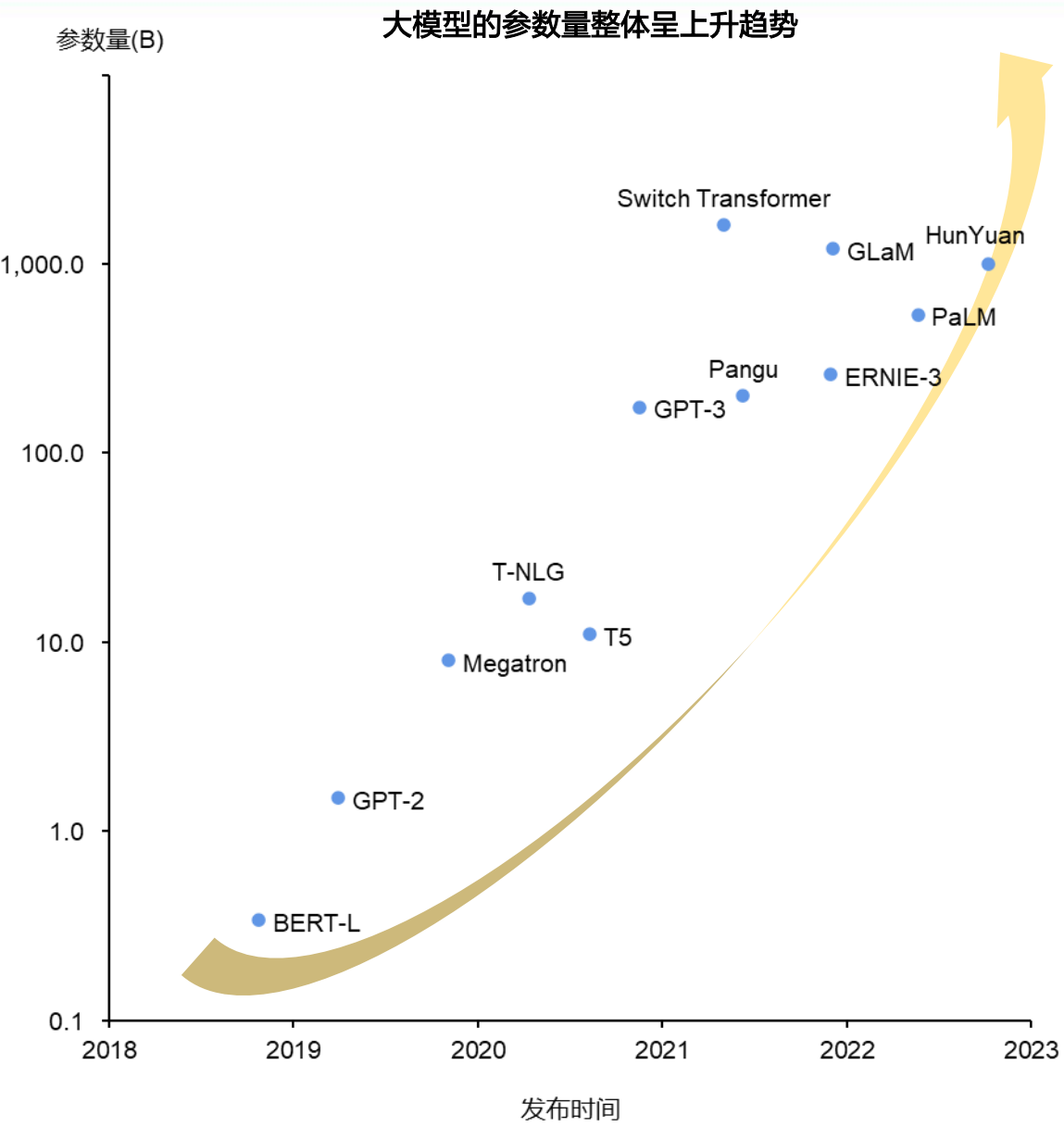
数据决定了机器学习算法的性能、泛化能力、应用效果；数据获取、标注、清洗、存储也是机器学习瓶颈之一。

步骤	定义	成本占比	特点	展望
数据收集	通过爬虫、API接口、数据采购等方式，从不同的数据源中获取数据，例如文本、图像、视频、音频等。	30%	主要来源：1）公共数据库（API接口等）；2）企业自行收集（爬虫、问卷、访谈等）；3）第三方数据供应商采购；4）经授权的客户数据；5）平台模拟生成数据	随着AI商用化提速加码，数据反哺，可用数据将越来越多，数据获取边际成本将逐步降低。
数据标注	人工或半自动对原始数据进行标注，包括分类、语义分割（图像背景，物、人）、目标检测标注（边界框、关键信息）、序列标注（序列数据文本音频中，类别、实体、关键字等）。	40%-50%	1）无监督学习无需数据标注，部分简单数据，机器学习平台可自动化标注； 2）监督学习仍需标注数据； 3）专业领域、图像等复杂数据基本仍需人工标注。	无监督学习逐渐流行，自动化程度逐步升高，对于简单数据集标注需求下降；但专业领域和复杂数据集仍需要人工标注，且人工单位成本更高；随着人工智能快速发展，智能化程度的提升，数据标注全面自动化也是有可能的。
数据清洗	根据数据类型和需求，进行缺失值处理、异常值处理、噪声处理、重复数据处理、数据格式转换等。	20%-30%	减少错误和不准确数据对模型的干扰，提高模型准确性和可靠性。	目前数据清洗仍以手动为主，但在某些数据较为标准化的场景中（如日志数据、网络流量分析），一般可以通过编写自动化的脚本或者使用一些现成的工具来实现，以去除无效或者重复的数据；随着人工智能快速发展，智能化程度的提升，数据清洗全面自动化也是有可能的。
数据存储	将机器学习算法需要用到的数据保存到磁盘或内存中，以便后续的训练、测试和预测。		数据分为训练集（约60%）、验证集（约20%）、测试集（约20%）；需要选择合适的数据格式存储，不同格式会影响读取速度、空间占比等；大规模数据集需要进行分割后存储。	需要选择合适的数据格式存储，不同格式会影响读取速度、空间占比等；大规模数据集需要进行分割后存储。

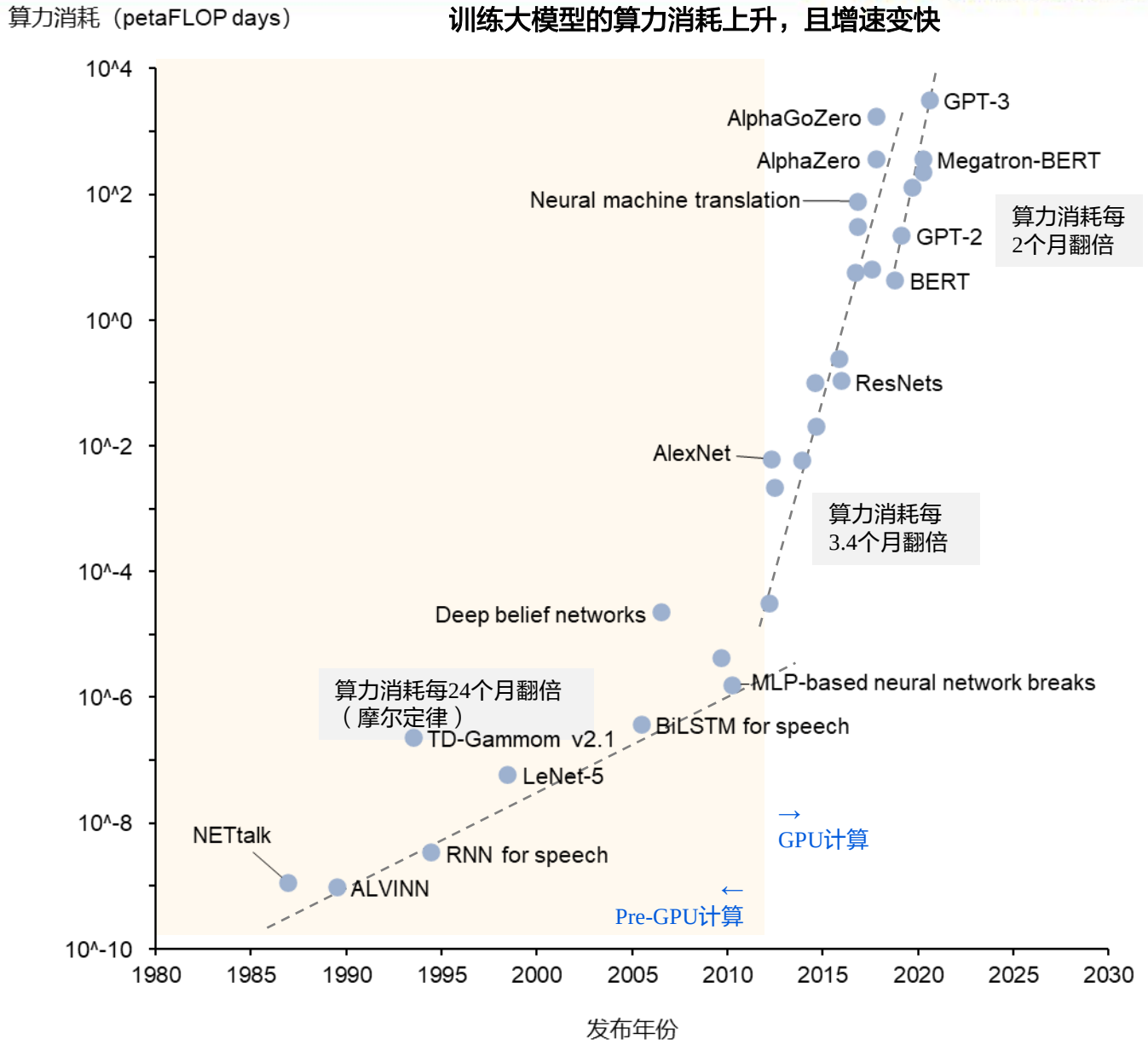
资料来源：专家调研，Appen，Defined.ai，Mindflow，PhoenixNAP，EasyAI，国海证券研究所

算力：随着模型参数量的提升，算力需求显著增加

大模型的参数量整体呈上升趋势



训练大模型的算力消耗上升，且增速变快



AIGC：生产力的革命

类型	任务	应用
文本生成	• 交互文本：闲聊机器人、文本交互游戏； • 非交互文本：结构化/非结构化、辅助性写作。	ChatGPT、Writesonic、Conversion..ai、Snazzy Al、Copysmith、Copy.ai、彩云小梦等。
音频生成	语音克隆、文本生成特定语音、音乐生成等。	Deepmusic、AIVA、Landr、IBM Watson Music、Magenta、网易-有灵智能创作平台等。
图像生成	图像编辑/修复、风格转化、图像生成（AI绘画）等。	GLIDE、DiscoDiffusion、 Big Sleep、StarryAl、VOMBO Dream、百度文心ERNIE-VLG等。
视频生成	视频编辑（AI换脸、特效、删除特定主体、跟踪剪辑等）、自动剪辑等。	Gliacloud、Pencil、VideoGPT、百度智能视频合成平台VidPress、慧川智能等。
3D生成	目前主要是基于图像、文本生成3D建模；AR、VR；3D打印等。	DreamFusion、GET3D、3DiM等。
数字人生成	视频生成、实时交互	腾讯、网易、影谱科技、硅基智能、倒映有声等。
游戏生成	• 元素生成：游戏场景、剧情、NPC生成； • 策略生成：对战策略等。	rct AI、超参数、腾讯AI Lab、网易伏羲等。
代码生成	代码补全、自动注释、根据上下文/注释自动生成代码等。	Codex、Tabnine、CodeT5、Polycoder、Cogram等。
跨模态生成	目前主要是文本生成图像、视频，根据图像视频生成文本等；未来将有更多跨模态应用。	ChatGPT4、百度文心、阿里M6等。

算法

生成式对抗网络（GAN）

- 2014年提出，由**生成器网络（Generator）**和**判别器网络（Discriminator）**组成，相互博弈、对抗，不断提高生成样本真实性和判别器准确性。
- **优点**：生成样本质量高，无需大量数据标注，适用于多种数据类型，可用于数据增强。
- **缺点**：训练不稳定、容易崩溃，生成样本难控制，需要大量计算资源，容易过拟合。

多模态预训练模型

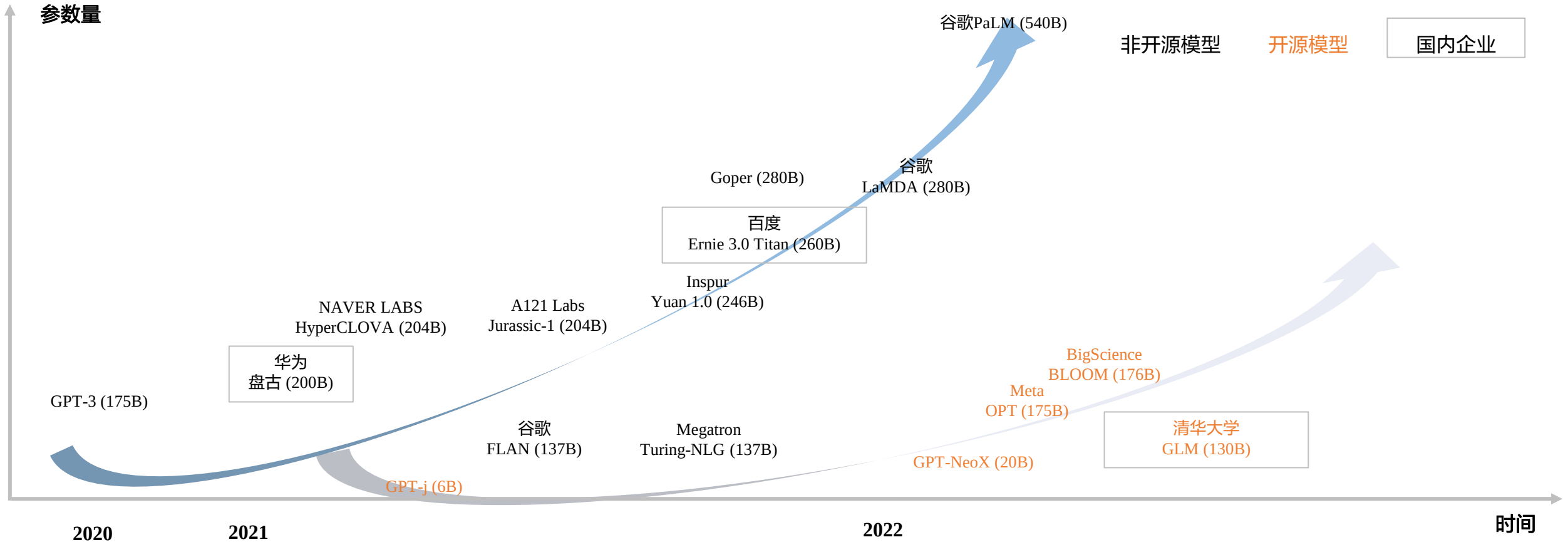
- 2019年提出，多模态数据预训练，实现多种模态数据的联合表示。
- **优点**：泛化能力、数据利用率、模型鲁棒性和可迁移性高。
- **缺点**：数据、算力需求大，特定任务需调参。

扩散模型（Diffusion）

- 2021年提出，相较于GAN，是图像生成领域的一大进步，不需要明确地计算数据的先验概率分布。通过“扩散”来执行隐空间中的推断。
- **优点**：更加灵活建模，样本多样性、可控性更高，训练过程简单、可扩展。
- **缺点**：数据、算力需求大，过程复杂，模型鲁棒性较低。

资料来源：量子位，wuhu动画人空间，数据派THU，Forbes，《Generative Adversarial Networks》Goodfellow等，《Using Pre-Training Can Improve Model Robustness and Uncertainty》Hendrycks等，《Diffusion Models Beat GANs on Image Synthesis》Dhariwal等，国海证券研究所

AIGC模型：参数量持续提升、开源模型逐渐丰富



2010-2015
小模型阶段

小模型（small models）占主导地位，小模型擅于分析任务，可以用于交货时间预测、欺诈分类等工作。但是表达能力不够，无法生成人类级别的写作。

2015-2021
规模竞赛阶段

Google Research的一篇里程碑式的论文提出了一种新的神经网络架构，可以生成高质量的语言模型，需要的训练时间更少，而且可以相对容易地针对特定领域进行定制。

2022之后
更好、更快和更便宜

算力变得更便宜，新技术降低了训练和运行所需的成本。开发人员的访问权限在某些情况下扩展到开源。

NLP算法：迎来突破，但算力、数据需求过高等问题待解决

国海证券 SEALAND SECURITIES

循环神经网络（RNN）

1990s兴起，可处理任一长度输入序列，同时具有记忆功能；但容易出现梯度消失或梯度爆炸。

长短期记忆网络（LSTM）

1997年提出，RNN的变体，控制信息的流动，并引入记忆单元，解决了梯度消失问题，可处理更长序列。

门控循环单元（GRU）

2014年提出，LSTM的简化和改进，相对LSTM处理长序列能力略逊一筹，但参数较少，训练速度更快。

- 词性标注
- 机器翻译
- 文本生成
- 文本预测

卷积神经网络（CNN）

1998年正式提出，具备参数共享和平移不变性，因此可以有效处理具有局部相关性的数据，可以处理文本分类和匹配任务；但不擅长捕捉序列中的长期依赖关系，需要结合CNN或自注意力机制。

- 文本分类
- 对话系统
- 问答系统
- 情感分析

异质图神经网络（HGNN）

近年来逐步发展，基于GNN，引入异质性注意力机制，捕捉不同类型的节点和边之间的关系，可以处理具有异质性的文本数据。

- 文本分类
- 情感分析
- 问答任务

自注意力机制（Self-Attention）

2017年提出，相比传统的RNN、CNN，可以处理变长的序列输入，不需要将序列进行填充或截断；可以自适应地学习不同位置的重要性，从而更好地捕捉文本中的重要信息；可以并行计算，因此可以加速模型的训练和推断过程。

- 语言建模
- 文本预测
- 文本分类
- 句子相似度计算
- 机器翻译

Transformer架构

- 2017年由谷歌提出；
- 自注意力机制+多头注意力机制+前馈神经网络；
- 适用NLP、CV、ARS领域；
- 可并行计算，更长的依赖关系，更好的泛化能力，较少的参数。

模型	发布时间	发布者	特点
GPT	2018	OpenAI	单向自回归方式来预训练模型，可以生成连贯的文本，但可能存在信息丢失的问题。
BERT	2018	谷歌	双向训练架构，从而可以适应各种下游任务，但需要更多的文本数据和训练资源。
RoBERTa	2019	Meta	BERT的改进，去掉了下一句预测任务，更大规模的数据集和动态掩码，较BERT提升模型鲁棒性和泛化能力，但训练和推理的计算成本更高，训练时间更长，训练数据要更多。
XLNet	2019	CMU、谷歌	BERT的改进，自回归+自编码训练，较BERT具有更好的建模能力、更强泛化能力，但需要更多的训练数据和更高的计算成本。
T5	2019	谷歌	通用型的文本生成模型，适用各种NLP任务，但需要大量计算资源和时间，在某些任务上的性能略逊于特定领域模型。
Switch Transformer	2021	谷歌	1.6万亿参数（2021.1），目前参数量最大的NLP模型；基于T5模型，采用创新的简化稀疏路由机制，相较传统自回归模型，在效率、可扩展性和生成质量等都具备较大优势，但需要更大的模型和更多的训练数据。

目前主要问题

- 模型参数量大，计算和存储成本高；
- 所需数据量较大；
- 模型复杂，可解释性问题；
- 语言多样性问题；
- 目前主要是无监督学习，因此对于文本处理和理解还不够智能；
- 数据隐私、版权等问题。

发展性

- 多模态融合，多语言处理，真正的理解和推理、更多的应用场景，更小、更高效的模型。

资料来源：EasyAI，机器之心，《Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling》Junyoung Chung等，《Gradient-Based Learning Applied to Document Recognition》Yann LeCun等，《Heterogeneous Graph Neural Network》Chuxu Zhang等，《Attention Is All You Need》Ashish Vaswani等，OpenAI官网，《Improving Language Understanding by Generative Pre-Training》Alec Radford等，《BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding》Jacob Devlin等，《RoBERTa: A Robustly Optimized BERT Pretraining Approach》Yinhan Liu等，《XLNet: Generalized Autoregressive Pretraining for Language Understanding》Zhilin Yang等，《Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer》Colin Raffel等，《Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity》William Fedus等，国海证券研究所

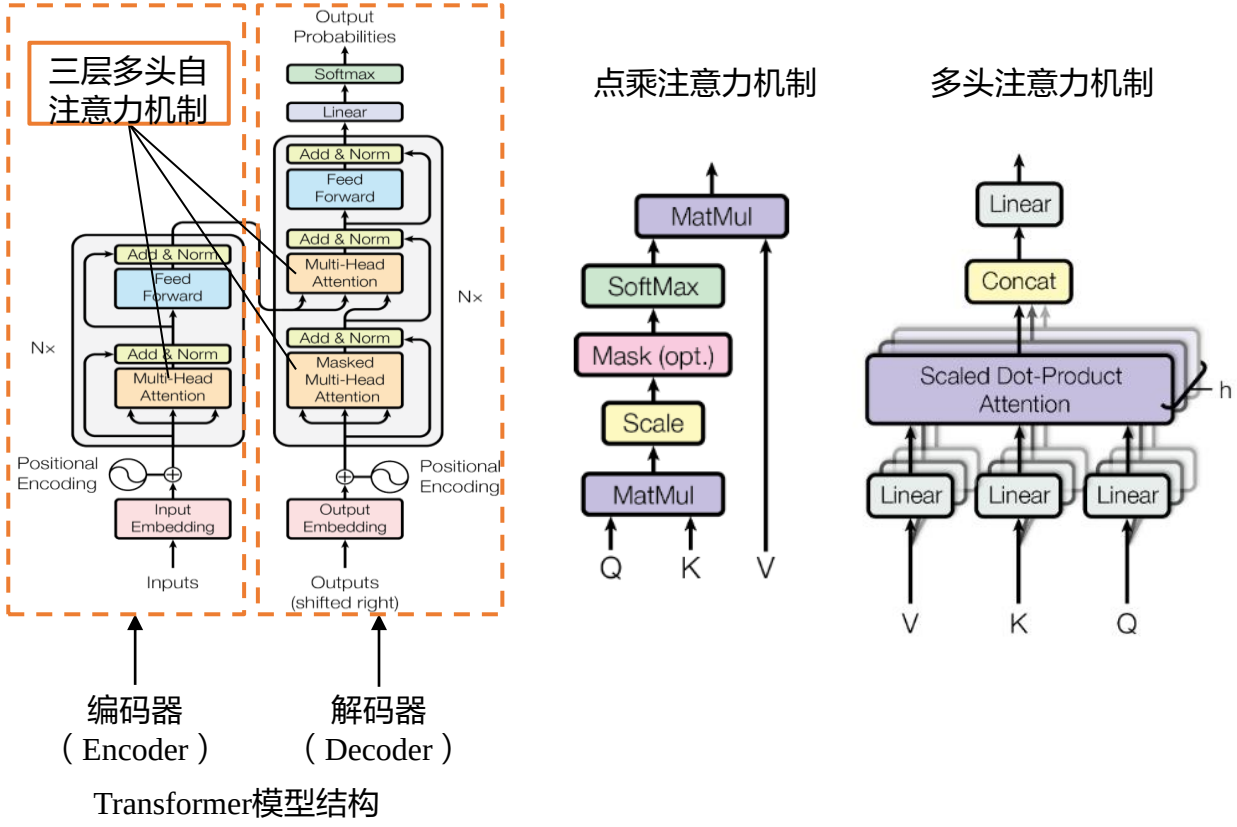
NLP算法：Transformer开辟NLP新路径，架构优化促成衍生模型

Transformer模型——特征提取器

2017年6月，Google发布论文《Attention is all you need》，提出了解决seq2seq（sequence to sequence）问题的Transformer模型。该模型引入自注意力机制（Self-Attention）代替了长短期记忆网络模型（LSTM），抛弃了之前在Encoder-Decoder模式下必须结合CNN或RNN的传统模式。

Self-Attention（自注意力机制）：

- 例：翻译The animal didn't cross the street because it was too tired.
- 以前的模型在处理该句子时，无法像人类一样根据上下文判断it指代animal还是street，而Self-Attention机制的引入使得模型不仅能够关注当前位置的词，而且能够关注句子中其他位置的词，从而在翻译时关联it和animal，提高翻译质量。

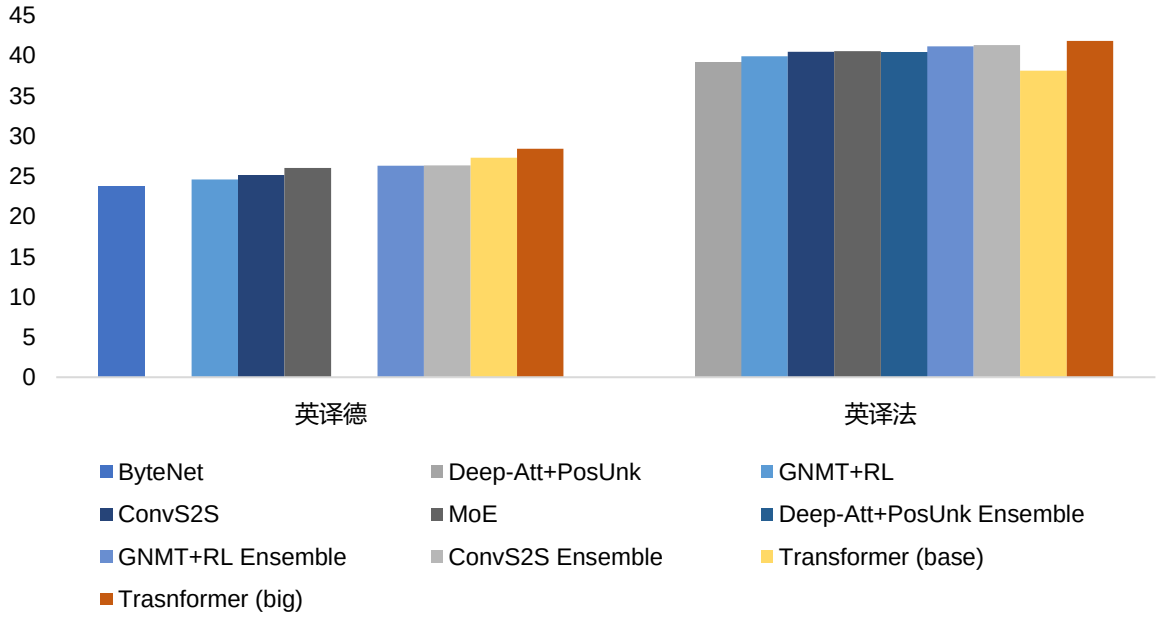


Transformer模型结构

Transformer架构不断优化，对比神经网络模型优势显著：

- Transformer的衍生模型仍在不断优化。研究者们正通过引入知识图谱及知识库、增加特定任务等方式不断优化Transformer的架构。
- 国际机器翻译评价指标采用BLEU（Bilingual Evaluation Understudy）的测试方法。在英译德测试中，Transformer Base/Big及三款Transformer的衍生模型的BLEU值显著高于两款基于RNN和CNN的模型，体现出Transformer模型优异的机器翻译能力；在英译法测试中，亦有三款Transformer模型的BLEU值显著高于RNN、CNN模型。

Transformer与早期模型的机器翻译BLEU值对比



资料来源：《Attention is all you need》Vaswani等，国海证券研究所

ChatGPT：史上用户数增长最快，源于算法的突破、高质量的数据库

国海证券 SEALAND SECURITIES

2015-11-11	2018-6	2019-2-14	2020-5-28	2022-1-27	2022-11-30	2023.3
Open AI宣布成立	**GPT-1发布**	**GPT-2发布**	**GPT-3发布**	**InstructGPT发布**	**ChatGPT发布**	**预计推出ChatGPT4**
• 初期为非盈利AI研究公司性质； • 启动资金**10亿美元**。	• **参数量1.17亿**；BooksCorpus数据集作为语料库，Tokens为**1.3B**； • 结合无监督学习及有监督的微调。	• **参数量15亿**；Tokens为**15B**； • 学习在无明确监督情况下执行多种任务。	• **参数量1750亿**；Tokens为**499B**； • 结合少样本学习及无监督学习。	• **参数量13亿**； • 运用RHLF，利用奖励模型训练学习模型； • 在遵循指令及输出内容等性能方面优于GPT-3。	• 基于**GPT-3.5**预训练模型； • 截至2023年1月末，**活跃用户超过1亿**，成为**史上用户增长最快的应用**。	• 预计为**多模态大模型**（语音、图像、视频）； • 新必应已集成ChatGPT4。

ChatGPT发布后市场反应热烈

表象：杰出的用户体验

- **ChatGPT功能**：回答后续问题、承认错误、质疑不正确的要求以及拒绝不适当的请求。
- **理解用户输入信息意图**，回答内容完整有逻辑、有条理，重点清晰；
- 真正做到**多轮沟通**，对上下文有理解和记忆，对话能力更强。

主因：算法的突破

- **LLM（large language model）**：当模型规模超过某个阈值之后，对于通用任务的效果会显著提升；
- **无监督学习（Unsupervised pre-training）**：又称可预测学习，该学习方式使得ChatGPT在无人工标注数据的条件训练，数据更多、数据成本更低，模型泛化能力更强；
- **CoT(Chain of Thought)思维链**：该算法使得模型生成推理路径，并在敏感话题方面避免了无法回答的问题；
- **RLHF(Reinforcement Learning From Human Feedback)人类反馈强化学习**：ChatGPT能够凭借强化学习的方式不断优化人类反馈的语言模型。

背后：算力、数据、人才、资金的投入

- **算力、资金的投入**：GPT3.5训练阶段总算力消耗约3640PF-days，约使用10000个GPU+285000个CPU，OpenAI耗费10亿美元租用Azure；截止202301，每个月,ChatGPT预计花费公司1200万美元；
- **数据**：ChatGPT1训练数据来自公开的电子书；ChatGPT2训练数据来自Reddit； ChatGPT2训练数据来自82.2%预处理的CommonCrawl语料库、13.5%的线上图书（GPT1数据集及Bibliotik,）、3.8% Reddit 。
- **OpenAI创始人、技术团队**：OpenAI创始人均为科技人才，现有375名正式员工，OpenAI一年人员支出高达8931万美元，人均约为23.8万美元/年/人。

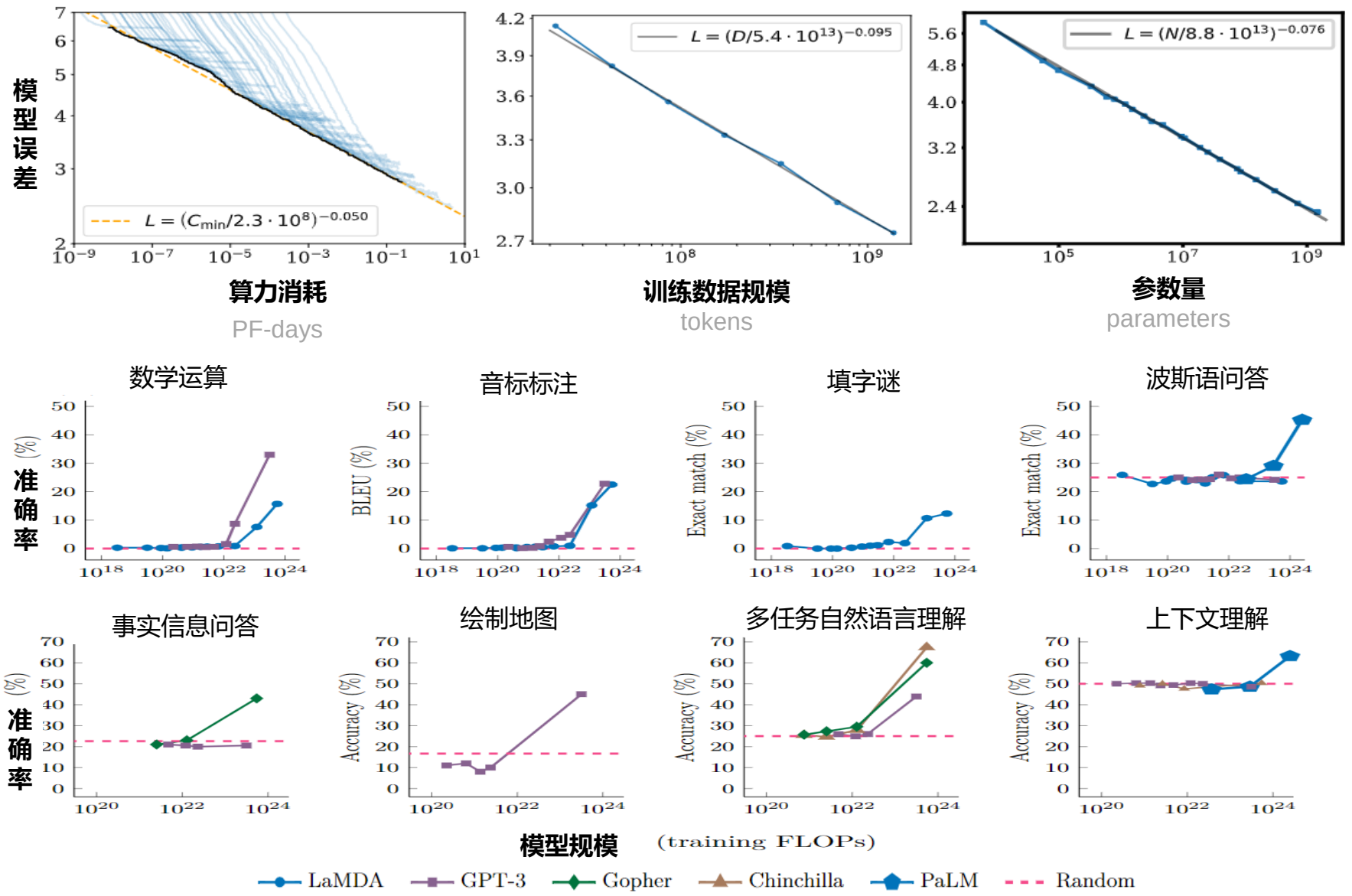
ChatGPT-算法：当模型规模达到某个阈值时，模型出现涌现能力

2020年1月，OpenAI发表论文《Scaling Laws for Neural Language Models》，探讨模型效果与模型规模之间的关系。

结论是：模型的表现与模型的规模之间服从Power Law，即随着模型规模指数级上升，模型性能实现线性增长。

而在**2022年8月**，Google发表论文《Emergent Abilities of Large Language Models》，重新探讨了模型效果与模型规模之间的关系。

结论是：当模型规模达到某个阈值时，模型对某些问题的处理性能突然呈现快速增长。作者将这种现象称为Emergent Abilities，即涌现能力。



资料来源：《Scaling Laws for Neural Language Models》Kaplan等，《Emergent Abilities of Large Language Models》Wei等，国海证券研究所

ChatGPT-算法：采用RLHF学习机制，效果优于GPT-3的无监督学习

GPT-3采用无监督学习机制，优点在于无需人工进行数据标注，可以节省模型训练成本，模型泛化能力更强。

而ChatGPT采用RLHF学习机制，即人工反馈的强化学习，属于强化学习。不同于传统的相比于传统的有监督学习机制，ChatGPT无需提前对训练数据进行标注，而只需要对人工对模型输出的结果进行评分，从而可以节省人力。虽然相比于GPT-3，ChatGPT需要消耗一定的人力，但是模型结果会更加符合人类偏好。

有监督学习vs无监督学习

			数据标注成本	模型精度
完全监督学习	所有数据都有标签	模型	高	高
半监督学习	含标签数据 无标签数据	模型	中	中
无监督学习	无标签的数据	模型	低	低

RLHF学习机制

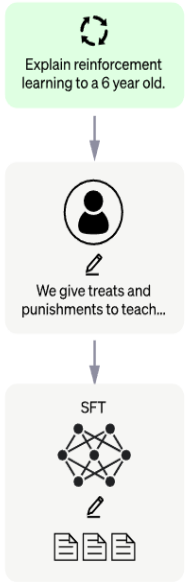
阶段1

收集演示数据并训练

随机从信息库抽取指令

专业的标注者对制定的提示给出高质量回答

专业人员用标注数据来调优GPT-3.5



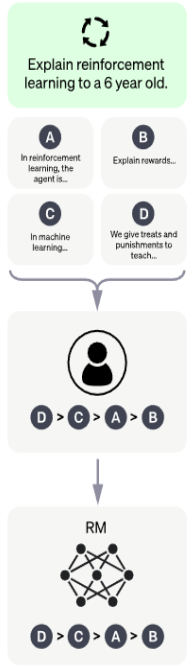
阶段2

通过人工标注训练数据来训练回报模型

标注一批模型产出及提示

标注人员根据多种标准对许多答案从优到差进行排序

利用以上排序结果来训练回报模型



阶段3

使用PPO强化学习法优化回报模型 - 根据RM评分结果更新预训练模型的参数

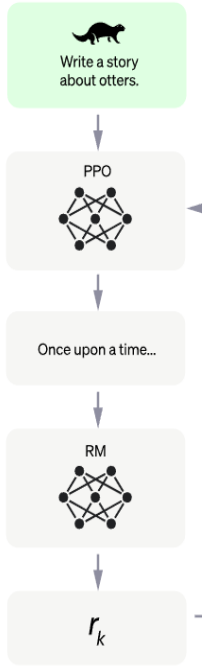
从用户提交的指令/问题中随机抽取一批新的命令

由监督模型初始化PPO模型的参数

PPO模型生成回答

用回报模型计算前一阶段训练好的模型给出的回答，得到分数

回报分数/策略梯度可以更新PPO模型参数

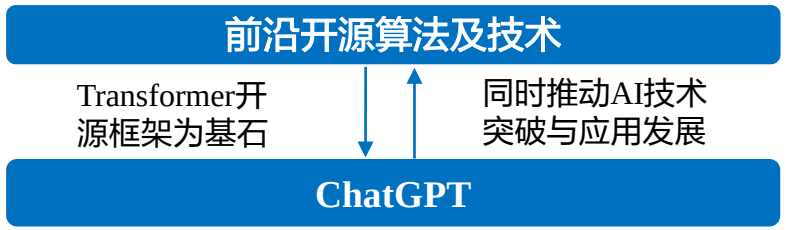


ChatGPT-反思：站在巨人的肩膀之上，开源开放期待更多可能和变革

国海证券 SEALAND SECURITIES

为什么ChatGPT脱颖而出?

- **ChatGPT成功的原因之一：站在巨人（Google、微软等）肩膀上。**



- **ChatGPT推出的意义**不仅仅是**技术上的突破，**更多的是预示着，**人工智能将越来越多的走出实验室，拥抱人类。**

- 只要人工智能技术领域，学术与产业界一直保持**开源、开放的生态**，人工智能还会迎来更多的技术突破，竞争格局变革的更多可能。
- 而行业的奠基者，也一定因其长久的技术积累存在较大的先发优势，持续的投入也会迎来更多的突破和可能。

反思

还有什么问题需要解决?

- **“黑匣子”般的理解和思考过程**

ChatGPT采用未标注数据训练而来，它的回答看起来相对以往更加聪明，但其理解和思考的过程更像是“黑匣子”，在某些领域或环节，并没有像人一样去理解，而是“不懂装懂”的感觉，给予的回答并没有那么靠谱。

- **底层算法的局限性**

存在信息丢失问题；模型参数量庞大，结构复杂，计算资源和训练数据需求较大。

- **在专业领域表现可能不如小模型**

在垂直领域上，ChatGPT在没有足够的标注数据训练的背景下，面对新数据时，性能可能弱于专业领域小模型；此外较小的模型微调更加容易，推理成本也更低。

- **训练数据存在时差，ChatGPT认知与真实世界不同频**

ChatGPT目前的训练数据局限在2021年之前，而模型推理时无法访问外部知识，超出其训练数据的相关问题无法进行回答，所以如何缩短其知识与真实世界的时差，也是决定其智能性的关键。

- **种族歧视、偏见、价值观、伦理等问题**

ChatGPT在回答问题时假设提问者的问题是价值观正确的，对于一些违背伦理的输入可能会给出价值观有问题的输出。

- **商业模式还需要进一步探索**

目前ChatGPT商业模式包括三种：直接向C端用户进行套餐收费（ChatGPT已推出Plus版本，定价每月20美元，可获得高峰时段免排队、快速响应、稳定多轮回答等服务）；向企业、第三方开发者提供API接口，如目前GPT-3采取按量收费的方式；与微软自身的必应、Azure等产品相结合，带动业务增长。

资料来源：Aimultiple，OpenAI官网，《Improving Language Understanding by Generative Pre-Training》Alec Radford等，Redian新闻，《Nature》，微软官网，国海证券研究所

三、应用篇：技术突破实现应用创新，已在多领域落地

AIGC何时突破工业红线：重点关注数据、算法的突破和商业模式的发展

国海证券 SEALAND SECURITIES

	大模型：通用型、任务型、行业级	小模型：专业领域，细分行业
参与方	**大模型技术巨头+第三方服务商** 巨头：微软、谷歌、meta、百度、阿里、华为、腾讯等； 第三方服务商：SaaS厂商、其他技术厂商等。	**AI企业** 商汤、科大讯飞、旷视、云从、依图、虹软、格灵深瞳、拓尔思等。
解决问题	**数据是瓶颈：数据增强、迁移学习、数据合成、数据要素市场实现数据共享、数据反哺加速商业化飞轮。** 1）数据获取：大模型所需数据量较大，而现实世界缺乏大量且优质数据； 2）数据存储、传输、管理：海量数据训练，读取和处理速度非常关键。	专业领域、长尾场景数据较少。
	算力是支撑：短期-国内云厂商等均早有囤货布局；长期-硬件进步、算法优化、并行计算、量子计算。 大模型往往需要大量计算资源，且模型参数仍在快速膨胀；但AI芯片全球短缺，英伟达A100、H100被禁止向中国供货。	**商业价值闭环：技术突破、AI企业深耕垂直细分行业（know-how、先发优势）、规模效应+飞轮效应双轮驱动。**
	人才是关键：“挖角”、企业高校合作。 美国人工智能一直领先，国内顶尖技术人才从数量、质量都存在较大差距，AI领域（尤其是CV）优秀的华人很多，但更多的在谷歌、微软、Meta等企业；北京的微软亚洲研究院的人才输出几乎撑起中国AI“半壁江山”。	
	商业价值闭环：技术进步、国家支持、巨头推动、生态建设、市场化教育。 • 海外软件生态成熟，企业/个人用户付费意愿更高；国内市场无论是生态和市场都存在较大差距。 • 国内外目前商业模式、付费逻辑尚未跑通。	技术成本（前期训练成本、数据成本、人才成本，后期使用的推理成本），与带来的增量或给企业实现降本增效相比，还不足以驱动企业投入AI。
	价值观、伦理、政治风险等：从技术层面让AI更可控，不要发展的那么快。	

AIGC应用：已在影视、传媒领域规模应用

领域		应用方式	AI生成的应用	AI生成的优势	产品/案例
影视领域	剧本生成	智能创作	文本生成	通过对海量剧本数据进行分析归纳，并按照预设风格快速生产剧本，创作者再进行筛选和二次加工。	开阔创作思路，缩短创作周期	
		智能改编		通过AI语义分析，将小说中的描述语言重新理解、拆解、组合，重组成包含重要场景、对白、动作等视听语言的剧本格式文本。	大幅缩减时间、人力成本。	
	角色场景	人脸合成	图片处理 视频处理	AI换脸：替换劣迹艺人、实现演员角色年龄的跨越、高难度动作合成等。面部调整：通过 AI 深度视频合成技术精准调整演员的面部特征，让演员的口型和不同语种的配音或字幕相匹配。	减少由于演员自身局限对影视作品宣传与发行的影响	百度大脑AI融合技术平均处理时长仅需数百毫秒，一键上传人脸照片即可体验融合效果。
		声音合成	声音处理	配音演员无法继续参与语音收录工作的情况下，利用AI算法学习还原配音演员声音。		百度推出语音合成技术 Meitron。
		场景合成	模型/场景 视频处理	通过人工智能合成虚拟物理场景，将无法实拍或成本过高的场景生成出来。利用数字建模+实时抠像技术，将演员动作与虚拟场景进行融合，最终生成视频。	拓宽影视作品想象力的边界，更优质的视觉效果和听觉体验	
	后期制作	修复还原	视频处理	对影视图像进行修复、还原，提升影像资料的清晰度，保障影视作品的画面质量。	提高修复效率。	百度与电影频道合作的“智感超清联合项目”，共同打造智感超清解决方案。
		生成预告	视频生成	机器学习自动生成影视预告片，人工进行最终调整。	节约时间成本。	预告片制作周期从一个月左右缩减到24小时。
		3D转制	模型/场景 视频处理	将影视内容自动从2D 向 3D 自动转制。	提升3D转制效率。	聚力维度推出的人工智能 3D 内容自动制作平台“峥嵘”支持对影视作品进行维度转换，将院线级 3D 转制效率提升 1000 多倍。
传媒领域	选题策划	热点捕捉	文本分析	应用AI大数据处理与分析技术，从已有的数据库中挖掘受众的阅读内容和阅读重点，总结这一阶段公众所关注的事件和话题。	提高信息精确度、选题精准度	
		热点预测	模型建立	通过AI科学地建立模型，对之后的热点进行预测。	提高时效性、选题精准度。	
	信息采编	信息收集	文本分析	AI大数据分析技术从全网海量信息中自动抓取相关素材，进行智能分类、聚类。	提高收集筛选效率，降低人力成本。	
		信息处理	语音识别 文本生成	借助AI语音识别技术实现录音语音转写和实时翻译，直接形成文字稿。	提高时效性，降低人力成本。	冬奥期间，科大讯飞智能录音笔跨语种转写功能助力记者2分钟快速出稿。
	内容制作	稿件写作	文本生成	基于AI算法自动生产新闻稿，将部分工作自动化，帮助媒体生产内容更快速准确且智能。	提高稿件质量和时效性，降低人力成本。	Wordsmith写稿机器人（美联社）、快笔小新（新华社）、Dreamwriter（腾讯财经）、写稿机器人“小南”（南方都市报&凯迪网）
		视频制作	视频生成	通过AI多媒体信息识别能力，对视频素材进行人像、文字、语音识别，迅速剪辑和生成字幕；实现画面人物的动态追踪、去除视频的拍摄抖动、多方位修复视频画质。	提高效率，降低人工成本，将几小时的工作量缩减至几分钟	
	播报传递	AI主播	语音合成 动画合成	通过实时语音及人物动画合成，让虚拟主持人根据输入的文本内容播报新闻，确保其表情、唇动与音频一致；除了常规式主持播报，虚拟主持人也开始陆续支持多语种和手语播报。	效率高、零出错、全天候	AI主播“小融”（人民日报）、AI记者“i思”（新华社）、AI记者“小聪”（浙江卫视）
		新闻分发	视频处理	使用横屏速转竖屏，视频拆条、视频集锦等AI智能工具，适应各平台分发要求。	提高传播效率。	

AIGC应用：已在电商、C端娱乐规模应用

领域		应用方式	AI生成的应用	AI生成的优势	产品/案例
电商领域	产品展示 3D外观	模型/场景	基于不同角度的商品图像，借助视觉生成算法自动化生成商品的3D几何模型和纹理。在分钟级时间内完成商品的3D拍摄和生成，精度可达毫米级。	全方位展示商品主题外观，大幅降低用户选品沟通时间，提升用户体验感，快速成交。	阿里巴巴的每平每屋业务就利用AIGC技术，实现线上“商品放家中”的模拟展示效果。
	产品展示 在线试穿	模型/场景	AI算法生成的3D商品模型还可用于在线试穿，高度还原商品或服务试用的体验感。	高度还原商品或服务试用的体验感，3D购物转化率平均值70%，较行业平均水平提升9倍，同比正常引导成交客单价提升超200%，退货率明显降低。	优衣库虚拟试衣、 阿迪达斯虚拟试鞋、周大福虚拟试珠宝、Gucci 虚拟试戴手表和眼镜、 宜家虚拟家具搭配、保时捷虚拟试驾等。
	人机交互 直播带货	语音合成 动画合成	基于视觉、语音、文本生成技术打造虚拟主播，为观众提供24小时不间断的货品推荐介绍。	填补真人主播直播间隙，24小时不间断连播；加速店铺、品牌年轻化进程，拉近与新消费人群距离；人设更稳定可控，不怕“塌房”。	欧莱雅、飞利浦、卡姿兰、完美日记等品牌均已推出自己的品牌虚拟主播。
	人机交互 客服咨询	信息处理 文字生成	AI虚拟客服，智能问答、自动回复、及时解决消费者问题	全年无休、无接待上限、相较人工节约80%左右成本。	京东言犀2.0每天可提供1000万次的智能服务、每月200万小时通话语音。
	广告营销 素材生成	图像生成 文字生成	AI自动生成营销文案、宣传图片。	提高个性化和针对性。	阿里巴巴的AI设计师“鲁班”可以为商家生成广告素材。
	虚拟货场 场景搭建	模型/场景	通过从二维图像中重建场景的三维几何结构，快速、低成本、大批量生成可交互的三维购物环境。	提升沉浸感和消费体验。	阿里巴巴的虚拟现实计划“Buy+”， 提供 360°虚拟的购物现场开放购物体验。
C端娱乐	趣味图像视频生成 AI换脸应用	图像处理 视频处理	生成AI换脸图像或视频。用途多为二创、社交平台分享。	较大满足用户猎奇的需求， 成为破圈利器。	FaceAPP、ZAO、Avatarify 等图像视频合成应用一经推出，就立刻病毒式在网络上引发热潮，登上 App Store 免费下载榜首位。
	趣味图像视频生成 照片动漫化	图像处理 视频处理	提供原始照片、输入关键词和参数、选择风格和模板，即可自动生成动漫化图片。	操作简单、满足用户猎奇需求。	可以将照片转为动漫风格的产品如Befunky、ToonApp、Colorinch、SocialBook。
	语音合成 智能配音	语音合成	为用户生产内容添加智能配音。只需输入文字内容、选择人声类型即可生成配音，进一步降低视频制作门槛。	降低制作门槛，比普通机械配音更真实有感情。	可以智能配音的产品如Syntesys、Murf、Listnr、Lovo、Play.ht。
	语音合成 变声功能	语音合成	支持用户体验大叔、萝莉等多种不同声线。	增加互动娱乐性，丰富用户体验。	Voicemod可以把用户的声音变成摩根弗里曼、变成飞行员、变成航天员等八种不同角色。
	趣味文字内容生成	文字生成	要用于为小说续写、人物世界观构建提供灵感。	降低写作门槛，为小说作者提供便利。速度快、操作简便。	彩云小梦AI续写功能提供三条不同的故事走向供用户选择。用户可以与自己创建的人物开启文字和语音互动，AI根据人物设定自动生成聊天内容
	虚拟数字人	图像生成 模型/场景	AI构建拟真人虚拟形象及可定义的虚拟形象等“数字化身”，可通过虚拟形象进行交互，如试穿服装、虚拟社交等。数字人内容生产处于起步阶段，随着短视频的崛起，创作者普遍对高效生产存在诉求，而数字人结合文本/音频驱动，可快速实现短视频内容生产，市场潜力较大。	降低虚拟形象定制门槛，丰富用户体验。	字节旗下的抖音，推出名为“沸寂（pheagee）”的业务，其平台定位是“数字时尚创意平台”。此外还有AI Studios、DeepBrain AI、Character Creator等产品。

AIGC应用：已在游戏领域逐步应用

国海证券 SEALAND SECURITIES

领域			应用方式	AI生成的应用	AI生成的优势	产品/案例
AIGC+游戏	游戏内	NPC生成	动作生成	模型/场景	大量的运动数据集上学习人体动作，生成行走、跑步、跳跃等动作。	降低动作生成的工作量，提高开发效率。	网易AI Lab研究人员设计语音文本匹配的全身动作序列。腾讯游戏光子S工作室《和平精英》团队携手腾讯AI Lab、腾讯游戏 CROS GVoice（腾讯游戏语音）团队，基于深度学习在语音编解码器上的不断突破，将 AI Codec 应用于《和平精英》游戏中，在行业内率先实现 AI Codec 更低码率更高质量的语音编码，由此成为首个将 AI Codec 技术全面应用于游戏语音领域的产品。
			表情生成	模型/场景	整合文字转语音研究，根据语音同步影响嘴型、表情等面部变化。	降低表情生产的工作量，提高开发效率。	
			外型生成	模型/场景	借助AI生成面容、服饰、声音和性格特征。	降低生产成本，提升生成角色规模和效率，增加角色个性化程度，千人千面。	
		环境生成	地图生成	模型/场景	采用AI算法生成较大的开放世界环境。	降低生产成本，提升地图规模，增加地图丰富程度。	
		逻辑生成	关卡生成	模型/场景 文本生成	根据游戏数据库和玩家输入的实时信息，自动生成新的关卡。	提高关卡生成效率，降低开发成本。	腾讯AI Lab对自研的2D小游戏可实现几小时内生成1000+关卡。
			剧情生成	文本生成	Rct AI能够根据玩家实时输入信息，动态地生成NPC交互反应，从而构建几乎无限且不重复的剧情，增强用户体验并延长生命周期。	提高剧情动态性、交互性、多样性，丰富用户体验。	
			对战模式	模型/场景	AI模拟游戏对战模式，与玩家进行游戏对战。	简单高效，为玩家提供更好的游戏体验。	
			平衡性测试	数据模拟	AI模拟玩家在某一套数值体系下的游戏体验，提出优化策略，为玩家带来更加平衡的游戏交互体验	增加游戏测试效率，降低测试成本。之前需要邀请人类玩家试玩1-2个月，现在由AI直接在内部完成。	腾讯“绝悟”AI通过强化学习的方法来模仿王者荣耀真实玩家，包括发育、运营、协作等指标类别，以及每分钟手速、技能释放频率、命中率、击杀数等具体参数，让AI更接近正式服玩家真实表现，将测试的总体准确性提升到95%。
	游戏周边	对战训练	AI陪练	数据模拟	身兼队友与老师，AI与玩家在真实对战环境中交流协作，并在过程中向玩家传授职业级的策略与操作技术。	帮助玩家迅速熟悉游戏玩法。	王者荣耀在引入王者绝悟教学后，玩家单局游戏主动沟通的次数明显提升，提高了PVE玩法的可玩性。
			模拟对手	数据模拟	AI通过模仿职业选手，掌握他们的典型个人风格，玩家感觉像在与真实的职业选手对抗。	增加游戏可玩性，丰富玩家体验。	基于拟人化AI研究，腾讯AI Lab与《穿越火线》手机版合作打造了「明星玩法」——挑战职业选手。AI通过模仿职业选手，掌握他们的典型个人风格，玩家则感觉像在与真实的职业选手对抗。该玩法上线后大受欢迎，对局数量较平时平均数提升了3-4倍。
		游戏衍生	集锦生成	视频处理	根据比赛视频实现自动剪辑、自动配乐、自动配文案。	集锦生成速度提高，生成规模增大。	
			赛事解说	语音合成	根据比赛情况实现自动解说。	节省人力成本，实现24小时不间断解说。	

AIGC应用：在金融、计算机、教育、工业、医疗等专业领域还在持续拓展

国海证券 SEALAND SECURITIES

领域		AI应用模式		AI价值	相关机构/企业/产品
金融领域	金融资讯	文本分析 文本生成 视频生成	基于算法自动编写咨讯，将部分采编工作自动化。	提高咨询生成速度，提高热点捕捉能力，增加资讯的时效性。	
	虚拟客服	文本生成 语音生成	为金融客户提供虚拟客服咨询。	降低人力时间成本，为客户提供个性化服务。	
	数据报告	数据分析 图像生成	以工作量优势辅助分析师抓取数据、进行数据分析、初步的报告生成。	提高数据分析的时效性、全面性、准确性，减少人工成本，提高分析效率。	
计算机领域	应用开发	算法设计 代码生成	AI智能设计程序或算法，辅助人工进行应用开发。	减少人工成本，提高应用开发效率。	
	代码编写	代码生成	AI智能编写代码。	降低代码错误，提高代码运行效率，降低人工成本。	GitHub Copilot Replit 的 Ghostwriter 为人工智能驱动的编程助手，可以对代码提出建议，辅助编程。 OpenAI大模型+演进算法训练AI写代码并修改代码。
	硬件设计	模型构建	辅助进行芯片设计等计算机硬件设计。	降低人工成本，提高设计效率。	2021年，谷歌在Nature上发文章表示AI能在6个小时内生成芯片设计图，而且比人类设计得更好。
教育领域	2D教材转3D	模型/场景 文本生成 语音生成	将2D教材转写为3D教材， 丰富教材内容和教学内容。	提高学生学习兴趣，使教学内容更具体直观。	
	合成虚拟教师	模型/场景 语音生成	合成虚拟教师，具有与传统教师相似的教学能力与学生进行互动。可个性化定制。	提高教学趣味性，降低教学成本。	
工业领域	辅助 CAD 设计	智能设计	将工程设计中重复的、耗时的和低层次的任务自动化。	缩短工程设计周期	SketchGraphs 能够协助建筑师、工程师等用户使用 AutoCAD 和 SOLIDWORKS 设计 2D 和 3D 原型。
医疗领域	个性化康复	模型/场景 语音合成	为失声者合成语言音频、为残疾者合成肢体投影、心理疾病患者合成无攻击感的医护陪伴。	为患者提供个性化康复服务。	聆心智能是一个为心理健康提供数字化诊断和治疗服务的平台。
	智能诊断	图像分析 文本生成	辅助医生进行初步诊断，并且智能生成诊断报告。	提高医疗效率，减轻医生工作量，帮助精确评估和患者病情及时对比。	百度生物计算平台“螺旋桨PaddleHelix”提供了一整套开源工具集和计算平台，支持构建针对新药研发、疫苗设计、精准医疗场景的技术方案。

资料来源：量子位，Gartner公司，百度AI，GitHub，statmodeling，NVIDIA，paddlepaddle，国海证券研究所

AIGC应用：在法律、农业、设计等专业领域还在持续拓展

领域		AI应用模式		AI价值	相关机构/企业/产品
法律领域	法条检索	文本分析 内容检索	依据案情特点快速检索相关法律条款，辅助梳理法律适用条件。	方便快捷，节约时间，减少人工成本，提高准确性。	2023年哥伦比亚法院在裁判中使用了ChatGPT中的文本生成功能来增加说理依据，在裁判文书中，ChatGPT给出了具体的法律条款、适用情形、立法目的以及宪法法院判例等内容，能够有效提升诉讼案件的处理。
	案情分析	文本分析	案件事实加以分析，并给出在不同情形下应当考虑的法律认定因素，对人工起到辅助判断作用。		
	法律文书撰写	文字生成	AI按格式和要求生成法律文书。		
	辅助司法裁判	分析比对	提供既有法律资料比对，尤其是法律条文和司法判决，可以实现裁判文书的辅助生成、案件信息的自动回填等功能。		
虚拟生命	AI伴侣	空间感知 语音合成	提供家用陪伴功能，可以提供沟通交流、实体互动、智能管家等功能。可按客户需求定制外形、性格等特点。	为客户提供更智能的情感、陪伴体验，满足客户个性化需求。	艾伦人工智能实验室为家用机器人提出了一些模型，如AI2THOR 和 ManipulaTHOR，可以让他们感知房间环境。
	互动型NPC	语音合成 文本合成	虚拟空间（如游戏）中的可互动NPC，提供沟通交流、虚拟互动、情感陪伴等体验。	丰富玩家游戏体验。	
农业/食品领域	辅助质检	图像分析 文本生成	AI辅助农产品质检，并生成质检报告。	提高质检效率、准确度，减少人工成本。	百事薯片工厂应用机器视觉系统，相机为传感器，用以计算土豆大小、重量和数量，判断土豆是否质量合格。
	配货建议	数据分析 文本生成	AI根据历史数据分析，生成农产品配货比例建议。	提高农产品销售效率，低门槛。	可口可乐试将60台自动售货机全链路数据喂给AI算法，让其生成配货建议，其交易量增加15%，进补货次数反而少了18%。
艺术/设计领域	3D模型生成	模型构建	AI根据设计需求，生成3D设计模型。	降低设计门槛，提高个性化，减少人工成本，提高设计效率。	英伟达Magic3D AI可实现一句话生成3D模型。
	AI艺术创作	图像生成 视频生成 文字生成	AI辅助艺术创作，或为艺术创作提供灵感。	提高创作效率，提供多样艺术形式。	

资料来源：法学学术前沿，量子位，国海证券研究所

四、企业布局：科技巨头全面布局，中下游厂商百花齐放

厂商布局：互联网大厂全面布局，中小厂商主要发力中下游环节

国海证券 SEALAND SECURITIES

AIGC产业链图谱

上游

■ 云计算
（000977.SZ）浪潮信息
（9988.HK）阿里
（9888.HK）百度集团
（0700.HK）腾讯
（未上市）华为

■ 芯片
（300474.SZ）景嘉微
（9888.HK）百度集团
（NVDA.O）英伟达
（9988.HK）阿里巴巴
（688256.SH）寒武纪
（002405.SZ）四维图新
（688981.SH）中芯国际
（未上市）地平线

■ IDC
（300738.SZ）奥飞数据
（603019.SH）中科曙光
（9698.HK）万国数据
（CD.US）秦淮数据

■ 光模块
（300308.SZ）中际旭创
（300502.SZ）新易盛
（220081.SZ）光迅科技

■ 服务器液冷
（600756.SH）浪潮信息
（300017.SZ）网宿科技
（000938.SZ）紫光股份

■ 数据供给方
（688787.SH）海天瑞声

中游

■ 多模态
（9888.HK）百度
（9988.HK）阿里巴巴
（0700.HK）腾讯控股
（300612.SZ）宣亚国际
（300418.SZ）昆仑万维
（603466.SH）风语筑
（688327.SH）云从科技
（2121.HK）创新奇智
（MSFT.O）微软
（GOOGL.O）谷歌
（NVDA.O）英伟达
（META.O）Meta
（未上市）珍岛
（未上市）中科闻歌
（未上市）澜舟科技

■ 策略生成
（未上市）rct AI
（未上市）超参数科技

■ NLP
（9988.HK）阿里巴巴
（002230.SZ）科大讯飞
（9888.HK）百度集团
（002230.SZ）科大讯飞
（688111.SH）金山办公
（300058.SZ）蓝色光标
（002292.SZ）奥飞娱乐
（学术机构）清华大学

■ 3D生成
（未上市）聚力维度

■ 代码生成
（MSFT.O）微软
（学术机构）清华大学
（学术机构）中国科学技术大学（学术机构）哈尔滨工业大学

■ 虚拟人
（300229.SZ）托尔思
（002467.SZ）二六三
（688088.SH）虹软科技
（002362.SZ）汉王科技
（300113.SZ）顺网科技
（未上市）小冰公司
（未上市）倒映有声
（未上市）相芯科技
（未上市）心识宇宙

■ 视频生成
（688039.SH）当虹科技
（0020.HK）商汤
（未上市）迈吉客
（未上市）影谱科技

下游

■ 电商
（300785.SZ）值得买

■ 传媒
（301270.SZ）汉仪股份
（300364.SZ）中文在线
（000681.SZ）视觉中国
（300781.SZ）因赛集团
（300624.SZ）万兴科技

■ 营销
（301052.SZ）果麦文化
（002803.SZ）吉宏股份
（301171.SZ）易点天下

■ 教育
（300081.SZ）恒信东方

■ 虚拟人
（300182.SZ）捷成股份
（002354.SZ）天娱数科

■ 游戏
（002624.SZ）完美世界
（0700.HK）腾讯控股
（300459.SZ）汤姆猫

■ 政务
（300075.SZ）数字政通
（002530.SZ）金财互联

■ C端应用
（MSFT.O）微软
（GOOGL.O）谷歌
（未上市）写作猫
（未上市）写作狐
（未上市）盗梦师
（未上市）诗云科技
（未上市）ZMO.ai
（未上市）影谱科技
（未上市）帝视科技
（未上市）不咕剪辑

产业链各环节发展趋势

类型	代表机构	上游		中游		下游		竞争优势
		算力	数据	大模型	小模型	行业合作	内部赋能	
互联网大厂（全面布局）	百度	百度云 昆仑芯片	百度各产品数据 行业合作伙伴数据	文心大模型	包括在文心大模型中的各类行业模型	与B端企业有广泛合作	百度搜索 百度各类产品的内容推荐	先发优势 具有较多的行业数据和专业知识数据
	阿里	阿里云 平头哥芯片	淘宝、天猫电商数据 阿里云B端数据	阿里M6大模型	-	合作较多	电商搜索 阿里云和企业服务	在大模型研发上具有资金和人才优势
	腾讯	腾讯云	微信用户数据 腾讯视频、新闻数据 腾讯游戏数据	混元大模型	腾讯游戏AI	合作较少	腾讯游戏AI NPC 微信等产品的自媒体创作、内容推荐	具有较多的用户数据和娱乐内容数据
	华为	华为云 海思芯片	手机用户数据	盘古大模型	盘古大模型中的各类行业模型	合作较少	较少	深耕上游和中游 赋能下游厂商
	谷歌	谷歌云	搜索数据 谷歌学术 Youtube数据	Imagen、ExTS、PaLM等	-	合作较少	Bard+Google	AI赋能搜索业务，同时快速积累新用户
	微软	Azure云	Office用户数据 Bing搜索数据	LayoutLM、DiT以及OpenAI旗下的大模型	-	较多企业接入chatGPT接口	chatGPT+Bing chatGPT+Office	AI赋能搜索和办公业务，同时快速积累新用户
学术机构（中游为主）	清华大学 中国科学技术大学 哈尔滨工业大学等	主要通过外购	互联网公开数据	√	√	合作方向主要为学术研究	-	政府支持 人才储备
中小厂商（中下游为主）	中游小模型厂商	主要通过外购	垂直行业数据	-	垂直行业模型	√	√	行业know-how 积累行业数据
	下游应用厂商	主要通过外购	垂直行业数据	-	-	√	√	客户粘性 用户粘性
产业链核心竞争要素		规模效应 政府补助 前期研发投入	数据规模 数据质量 数据获取成本	资金能力 技术能力 人才储备	行业Know-how 行业数据	先发优势 行业know-how	内部用户规模和业务数据积累；业务和AI技术结合的可行性	
产业链未来发展方向		头部效应↑ 边际成本↓	通用类数据集中于大厂，而垂直行业数据分散	头部效应↑	百花齐放	通用型内容生成集中于大厂，而垂直行业解决方案百花齐放	大厂对外提供服务的同时内部赋能，小厂采取外购的方式更加经济	

资料来源： 澜舟科技，AspenCore，海思，前瞻产业研究院，IDC，机器之心，国海证券研究所

AIGC相关标的——上游企业

领域	代码	公司	相关业务
光模块	300308.SZ	中际旭创	2022年公司800G产品、相干光模块产品等已实现小批量出货，保持了产品领先性。
	300502.SZ	新易盛	公司一直致力于高性能光模块的研发、生产和销售，为数据中心客户提供100G、200G、400G和800G高速光模块产品；为电信设备商客户提供5G前传、中传和回传光模块、以及应用于城域网、骨干网和核心网传输的光模块产品；为智能电网和安防监控网络服务商提供光模块解决方案。
	002281.SZ	光迅科技	公司主要产品有光电子器件、 模块和子系统产品， 按应用领域可分为传输类、 接入类、 数据通信类。其中传输类产品中的传输收发模块包括100G/400G 等速率、10km/40km/80km/120km 等距离的光模块产品。
IDC	600756.SH	浪潮信息	公司建成了亚洲最大的液冷数据中心研发生产基地，构筑了从研发、测试、生产、交付的全链条液冷智造能力，年交付能力超过 10 万台，为不同类型的数据中心提供液冷系列产品、改造和部署方案，整体 PUE 降至 1.1 以下。
	300017.SZ	网宿科技	作为IT基础平台服务提供商，公司一直围绕信息技术基础设施平台进行能力建设及业务开拓，在CDN、IDC等成熟业务的基础上，正在进行向“云安全”与“边缘计算方向的革新；并积极拓展私有云/混合云、MSP、数据中心液冷解决方案等新业务。
	000938.SZ	紫光股份	根据 IDC 数据， 2022 年上半年公司多项产品市场占有率持续领先，在中国以太网交换机、企业网交换机、数据中心交换机市场，分别以 36.9%、 37.9%、 37.8%的市场份额排名第一。
	300738.SZ	奥飞数据	公司是专业的数据中心业务运营商和通信综合运营企业，在华南地区自建了多个数据中心，并在全国各地运营着众多高标准数据中心，具备覆盖全国的服务能力。
	603019.SH	中科曙光	公司AI计算服务主要应用在数据中心端，为各个需要人工智能技术支撑的领域提供服务。
	9698.HK	万国数据	中国领先的数据中心基础设施和服务提供商之一，数据中心覆盖中国国内重点核心城市。
	CD.US	秦淮数据	创立于2015年，总部位于北京，是全球首家以泛亚太新兴市场作为业务发展核心区域的超大规模数据中心解决方案运营商，同时也是专注信息技术产业生态基础设施规划、投资、设计、建造和运营的综合服务提供商。
芯片	NVDA.O	英伟达	2020-2021年推出GANverse3D，能够生成可自定义并生成动画的3D图形。2022年发布了Magic3D，一个可以从文字描述中生成3D模型的AI模型。
	300474.SZ	景嘉微	公司在图形显控领域拥有图形显控模块、图形处理芯片、加固显示器、加固存储和加固计算机等五类产品，其中图形显控模块是公司最为核心的产品。
	688041.SH	海光信息	旗下有AI芯片海光8000系列
	INTC.US	英特尔	旗下有AI芯片 Nervana NNP-I1000 和 Nervana NNP-T1000
	9988.HK	阿里巴巴	旗下有平头哥（芯片研发）
	9888.HK	百度	百度昆仑芯片
	未上市	燧原科技	旗下有AI芯片邃思1.0和邃思2.0
	未上市	壁仞科技	旗下有AI芯片BR100
	未上市	天数智芯	旗下有AI芯片天垓100
	未上市	华为	旗下有AI芯片昇腾310和昇腾910

AIGC相关标的——上游企业

领域	代码	公司	相关业务
超算中心	9988.HK	阿里巴巴	2022年8月阿里云启动两个超算中心：一是张北超级智算中心，总建设规模为12 EFLOPS（每秒1200亿亿次浮点运算）AI算力，将超过谷歌（9 EFLOPS）和特斯拉（1.8 EFLOPS）的智算中心。二是乌兰察布超级智算中心，建设规模为3 EFLOPS（每秒300亿亿次浮点运算）AI算力，位于“东数西算”内蒙古枢纽节点。两座超级智算中心皆是以“飞天智算平台”为技术底座，在规模和效率上实现双向突破，将为AI大模型训练、自动驾驶、空间地理等人工智能探索应用提供强大的智能算力服务。
	0700.HK	腾讯控股	2022年9月，腾讯长三角人工智能先进计算中心及生态产业园投入使用。该计算中心建成后，服务器数量将达到80万台，算力是目前世界排名第一的超算中心的10倍，届时将成为全国单体规模最大、达到世界领先水平的数据中心。将承担各类人工智能、即时通信、图像处理、科学计算等任务，以强大的数据处理能力为全社会提供云计算服务。腾讯长三角人工智能先进计算中心及生态产业园项目是上海市重大建设项目，将引入腾讯的科恩、优图、微翎三大实验室。除了腾讯三大实验室外，还有常山北明、东华软件、T3出行、灵雀云等50余家腾讯生态链企业将入驻生态产业园项目。
	9888.HK	百度	2019年10月，百度在保定，同时自建两个超大型云计算中心，分别为徐水智能云计算中心和百度定兴智能云计算中心。2020年8月20日，百度在保定自建的超大型数据中心宣布开放，承载36万台AI服务器。
	002230.SZ	科大讯飞	讯飞于2009年开始算力基础设施建设，目前已建成4城7中心深度学习计算平台，讯飞的算力不仅完全满足AI算法模型训练，及面向开放平台数百万开发者和其他行业伙伴提供相关AI服务的需求。
	0020.HK	商汤	2022年1月商汤科技宣布，商汤科技人工智能计算中心启动运营。商汤AIDC是一座开放、大规模、低碳的先进计算基础设施，是SenseCore商汤AI大装置的重要算力基座，其设计的峰值算力高达3740 Petaflops（1 Petaflop等于每秒1千万亿次浮点运算。
	601728.SH	中国电信	2022年7月，中国电信安徽智算中心正式启动，该中心累计投资将达100亿元人民币，建设16000个高密度机架，支持算力规模可达到2.2EFLOPS，使安徽省的整体算力规模翻番，将成为华东区域具有重要影响力的超大型数据中心之一。
	600941.SH	中国移动	中国移动围绕智算建设运营和AI产业生态培育两大目标，构建全栈智能信息服务体系，助力国内AI产业长远发展。一是制定新型智算技术体系，构建智算基础设施底座。二是以网强算形成智能算力集群，提升算力服务效率。三是是联合产业打造“芯合”算力原生平台，以软件为牵引构建新生态。
	600050.SH	中国联通	2022年11月，国家超级计算西安中心与中国联通陕西公司算网融合实验室揭牌仪式在西安举行。双方将在算网一体化、高性能计算仿真应用、AI创新应用、算力交易、数据要素交易、算力调度、算力编排等技术进行应用研究，共同打造联通高性能计算专区，为特殊行业及高校、科研院所提供联通云高性能计算仿真云。
	未上市	旷视科技	2019年，旷视在芜湖投资建设的AI超算中心投入使用，该超算中心包括超算中心CPU+GPU机房和具有自主知识产权的深度学习平台。其中，超算中心机房分为2个计算集群和1个存储集群，可提供6000+计算核心和6.7PB的集群存储空间，具备城市级数据计算处理能力。
	未上市	华为	2021年5月华为武汉人工智能计算中心投运，目标算力规模 200 PFLOPS FP16（扩容后）；2021年9月华为西安未来人工智能计算中心投运，目标算力规模300 PFLOPS FP16；2021年10月华为许昌中原人工智能计算中心投运，目标算力规模100 PFLOPS FP16.
渲染引擎	未上市	光线云	公司已布局AIGC渲染引擎，可以生成不同模态的渲染数据内容，能够降低内容开发成本，提升内容创作效率。光线云实时渲染引擎RAYSENGINE采用端云协同架构，既解决了纯端架构大量依赖端侧算力，设备性能需求高的痛点，又解决了纯云架构云服务成本高、网络带宽影响卡顿的问题。
液冷	002837.SZ	英维克	液冷技术作为针对高热密度等解决方案，公司规划较早，技术的积累包括基础材料、器件、端到端的系统、甚至售后运维等方面，而且对技术演进的路线也有充分的准备。就数据中心、储能、电力电子散热、新能源车热管理等不同的应用场景，公司始终遵循包括液冷、风冷、电子散热等不同技术综合的最优原则在寻求解决方案。
	872808.NQ	曙光数创	公司主营业务收入主要来自于数据中心液冷基础设施产品，公司核心技术应用于公司的浸没相变液冷数据中心基础设施产品、冷板液冷数据中心基础设施产品中。

AIGC相关标的——中游企业

领域	代码	公司	相关业务
NLP	002292.SZ	奥飞娱乐	参股公司光年无限有自己的AI对话机器人产品-图灵机器人开放平台，开发者可自行快速接入并创建个性化机器人，包含聊天机器人、智能客服等，目前累计注册企业开发者超过150万。
	002230.SZ	科大讯飞	科大讯飞主导承建了认知智能全国重点实验室，多年来始终保持关键核心技术处于世界前沿水平，同时已面向认知智能领域陆续开源了6大类、超过40个通用领域的系列中文预训练语言模型，成为业界最广泛流行的中文预训练模型系列之一。 公司已经在当前核心技术、产业场景、行业数据等深厚积累的基础上，于2022年12月份进一步启动生成式预训练大模型任务攻关，科大讯飞AI学习机将成为该项技术率先落地的产品，将于2023年5月6日进行产品级发布，该技术突破将在AI学习机的中英文作文辅导、中英文口语学习等方面带来显著提升。 科大讯飞在2023年会持续升级该系列技术，并陆续应用于公司在教育、医疗、汽车、消费者等多个行业赛道的既有产品。
	688111.SH	金山办公	金山黑马校对v30拥有 79 个涵盖各领域的大规模专业词库，不仅能校对大部分中文错别字，还可校对大部分多字、少字、多余标点、成对标点等计算机录入的常见错误，OCR 识别和语音转文字后产生的难以注意的异体字、异形词等也能高效识别。
虚拟人	300058.SZ	蓝色光标	公司的AIGC布局不仅包括“分身有数”，“蓝标智播”等AI产品，也涵盖“销博特”等多人协同创作平台。撰稿机器人“妙笔”是公司于2018年推出的14款智能营销产品之一，目前是公司的参股公司。同时，“妙笔”还在与人民网和北大合作，通过自然语言处理和知识图谱等AI技术，支持政府和企业的智慧党建工作。AI营销平台销博特（XiaoBote）能够一键自动化生成策划案、品牌分析报告、消费者洞察等内容，已累计注册用户超60000家，覆盖10多个行业，400多个品类。
	002467.SZ	二六三	公司制作的虚拟数字人，融合NLP、动作捕捉、知识图谱等AI（人工智能）技术，能主动地、智能地与真人交流，可在各类活动中担任虚拟主持人、AI讲师、AI客服等工作
	0020.HK	商汤	SenseMARS Avatar是商汤推出的虚拟人解决方案，针对直播、短视频、智能相机、虚拟社交、虚拟会议等场景提供行业解决方案。主要通过卡通风格Avatar虚拟数字人，增强用户虚实互动，提升用户趣味体验，减少用户个人隐私信息暴露。
	002657.SZ	中科金财	中科金财已布局WEB3.0内容制作引擎、多模态超写实数智人、全媒体智能客服（数字人智能大脑解决方案）、RPA机器人等解决方案中运用了多模态输入/输出、视觉合成、知识处理等人机交互技术。
	300229.SZ	拓尔思	2022年世界杯期间，公司将利用自研互联网大数据资讯平台，对世界杯相关的热点和话题进行大数据分析和研判，通过AIGC的内容自动创作和虚拟数字人进行联合，开展“大数据虚拟人”的机会，将前期代替人的“灵魂”应用匹配新的“形象”科技，面向金融、政府、安全等行业的提供更丰富的专业服务场景。
	688088.SH	虹软科技	2022年11月初进博会上，虹软通过AI建模技术与3D建模技术提供两种不同模式的数字人。该数字人可将其融入现实空间中并在现实空间中完成既定动作或模仿人的各类肢体动作和表情进行对话与互动。
	002362.SZ	汉王科技	公司在三个条线已有相关的技术储备及初步产品：一是在专业的行业领域，提供数字员工，其利用AI图像文本识别技术及RPA技术替代人工做一些需要高人力、高重复性的工作。针对政法、民生等领域的送达机器人等产品也已有落地试点。二是在金融行业，双录智能风控机器人、流水稽核机器人及智能财报读取机器人等虚拟机器人已经在银行、金融服务公司等落地应用。三是公司在动力机械方面的技术积累，目前公司仿生扑翼飞行器已经推出多种形态并商业化推广销售。
	300113.SZ	顺网科技	公司面向电竞酒店打造了数字人“晓竞”，融合了音频、文本、图像等多模态信息。公司是浙江大数据重点研究院之一，在大数据、AI、区块链、GPU云领域都有储备。
	未上市	小冰公司	作为“AI being”派虚拟人。小冰的产品始终是人+交互+内容。具体包括虚拟人（夏语冰等somebody instance、虚拟男友等nobody instance和国家队人工智能裁判与教练系统观君等在垂直场景中工作的虚拟人类）、音频生成（主攻超级语言及歌声，在线歌曲生成平台与歌手 歌声合成软件X studio）、视觉创造（毕业作品集《或然世界》、为国家纺织品开发中心、万事利等数百家机构提供了图案和纹样设计）、文本创造（2017年即推出小冰诗集）、虚拟社交、Game AI（Xiaoice Game Studio）等。
	未上市	倒映有声	倒映有声将其虚拟人的高自然度归结于神经渲染（Neural Rendering）、TTSA(基于文本和语音合成实时生成音频和视频)、ETTS(富情感语音合成)、Digital Twin。通过神经渲染技术快速构建AI数字分身，通过语音+图像生成技术，生成和驱动数字分身的唇形、表情、动作、肢体姿态，创造表情自然、动作流畅、语音充满情感的高拟真度数字分身IP。2021年3月倒映有声和中央广播电视总台音频客户端「云听」签署战略合作协议。
	未上市	相芯科技	相芯科技专注于计算机图形学和人工智能技术的深度融合， 推动XR技术创新和产业应用，自主研发的“虚拟数字人引擎”和“超写实数字物平台”已在逾千家国内外企业得到规模化应用。
	未上市	心识宇宙	心识宇宙旗下的产品MindOS在2022年11月初发布了内测版本，面向少数B端客户试点。这是一个AI角色生成引擎，通过简单的填写配置、拖拽上传，就能完成一个具备专业知识、记忆和人格的AI角色，大大提升应用交互的体验。仅2023年1月，MindOS就获得了百万元订单。

AIGC相关标的——中游企业

领域	代码	公司	相关业务
多模态	9888.HK	百度	在百度万象大会上已发布多项AIGC的技术和产品，包括“创作者AI助理团”和“百度APP数字人计划”。百度百家号携手澎湃新闻、新京报等数十家权威媒体成立“AIGC媒体联盟”。“创作者AI助理团”通过文心大模型、文心一格、图文转视频等技术，为创作者提供“AI文案助理”、“AI插画助理”、“AI视频制作助理”。“百度APP数字人计划”将数字人技术与图文转视频、TTS语音合成技术结合，为媒体及创作者定制真人孪生数字人。
	9988.HK	阿里巴巴	达摩院牵头推出魔搭社区 ModelScope，社区首批上架超 300 个模型，其中中文模型超过 100 个，覆盖了视觉、语音、自然语言处理、多模态等 AI 主要领域，覆盖主流任务超过 60 个，均全面开源并开放使用。
	0700.HK	腾讯控股	腾讯“混元”AI大模型在MSR-VTT，MSVD，LSMDC，DiDeMo和ActivityNet 五大跨模态视频检索数据集榜单中先后取得第一名的成绩，实现了该领域的大满贯。特别是在MSR-VTT榜单上，“混元”AI大模型将文字-视频检索精度提高到55%，领先第二名1.7%，位居行业第一。
	9618.HK	京东集团	京东AI研究院研究领域包括文字识别、人脸与人体识别、图像及视频理解、自然语言处理、智能创作、语音技术、内容审核、商品理解，一方面为内部的京东零售和京东物流赋能，同时对外提供行业解决方案（包括智能供应链、智能运营、智能营销、智能零售、智慧城市、智能硬件）。
	未上市	字节跳动	字节AI Lab成立于2016年，研究领域覆盖自然语言处理、数据挖掘、计算机视觉、机器学习、计算机图形&增强现实、系统&网络、安全&隐私、语音与音频。AI Lab 专注于人工智能领域的前沿技术研究，涵盖了计算机视觉、语音 & 音频处理、 NLP、CV、Speech、音乐、机器学习等多技术研究领域，同时致力于将研究成果落地，为公司现有的产品和业务提供核心技术支持和服务。
	300612.SZ	宣亚国际	公司自主研发的“巨浪技术平台”依托AI、大数据、云计算等技术为客户在文字、图片、视频等内容生产的采集、加工和审核环节提供智能化、自动化支持，为客户提供多样化、场景化的营销素材，提升客户的内容营销效率。
	300418.SZ	昆仑万维	公司AIGC方向发布的天工巧绘、天工乐府、天工妙笔、以及天工智码具备AI生成图像、音乐、文本、以及代码的能力，其中天工巧绘开发面向C端用户的小程序。2022年12月，昆仑万维发布了「昆仑天工」AIGC全系列算法与模型，并宣布模型开源。
	603466.SH	风语筑	公司已结合AIGC技术在文生文、文生图、文生音视频等领域进行场景应用，还将强化在3D建模和虚拟空间生成等领域的定向训练和模型优化。
	688327.SH	云从科技	2020年开始，已经陆续在NLP、OCR、机器视觉、语音等多个领域开展预训练大模型的实践，这不仅进一步提升了公司各项核心算法的性能效果，同时也大幅提升了公司的算法生产效率，并已经在城市治理、金融、智能制造等行业应用中体现价值。
	2121.HK	创新奇智	创新奇智依托于自研的人工智能平台，研发出面向制造领域的AIGC产品AInnoGC。AInnoGC具有样本生成、产线布局生成、对话式AI回答等能力。
	MSFT.O	微软	微软正在迅速推进OpenAI的工具商业化，计划将包括ChatGPT、DALL-E等AI工具整合进微软旗下的所有产品中，并将其作为平台供其他企业使用。
	NVDA.O	英伟达	英伟达提出了 SPADE、MUNIT 等多个图像及视频合成模型。并开源了PyTorch 库「Imaginaire」，共包含 9 种英伟达开发的图像及视频合成方法。
	META.O	Meta	2022年推出了统一多模态的子自监督学习框架Data2Vec。
	GOOGL.O	谷歌	2019年10月，Google发布了统一的模型框架T5，基于编码解码器的T5（BERT只有编码，GPT只用解码），最大模型110亿参数并开放。
	未上市	珍岛	珍岛的营销云（T云）可以基于AI赋能营销推广的应用场景与大数据营销获客分析，为企业提供基于AI的“营销综合测评—— 一键智能建站——商机智能发布与推广——口碑营销——广告获客再营销——图文视频等营销素材智能生成——短视频获客——流量询盘转化”等企业智能营销全链路管理。
	未上市	中科闻歌	旗下主要有三款产品。①天湖数据智算平台：基于跨模态深度语义理解、社会计算与因果推理、决策推演的数据智能与决策智能技术平台；面向领域全链条，软件工程低代码、敏捷开发环境和业务交付平台。②闻海全球开源数据平台：国内最大的多语言、跨模态媒体数据库，提供多语种、跨模态、全渠道信息专业检索和深度分析云服务，赋能政企客户实现人机共融智能决策。③红旗融媒体平台：面向国家媒体融合发展重大战略部署，以大数据和人工智能为驱动的新一代媒体融合协同发展平台。
	未上市	澜舟科技	公司旗下的AIGC产品包括：①文学辅助写作，为文学创作提供文学实体渲染、自定义模板生成、关键词扩写等可控文本生成API；2）营销文案写作，面向营销类文案场景，提供基于商品关键词生成、基于商品信息的广告标题生成，以及营销类文本续写等能力；③熊猫小说家，一个能够让用户快速和朋友一起创作小说的小程序；4）论文助写，论文风格提示工具，帮助用户快速找到论文写作灵感并提供写作风格提示；5）文图生成，通过文字描述生成图像/视频内容。

AIGC相关标的——中游企业

领域	代码	公司	相关业务
3D生成	未上市	聚力维度	聚力维度前身为十二维度（北京）科技有限公司，创始于2012年，专注计算机视觉和人工智能核心技术研发，并深入影视领域，参与《机械战警》、《警察故事》、《狼图腾》等多部一线影视制作。由科幻成真实验室开启的人工智能影视变革第一步：人工智能立体设计师——峥嵘，能自动将2D视频转换成3D视频。
视频生成	688039.SH	当虹科技	公司AIGC相关技术在媒体演播室、智能媒资平台等众多场景上的应用已经较为成熟，可为电视台、新媒体、互联网、泛媒体行业等客户提供通过AI技术进行短视频、长视频的生产/加工等解决方案。包括AI智能剪辑；用AI生成短视频，简化了流程提高了效率，保证了视频上线的时效性。
	0020.HK	商汤	商汤科技全新推出SenseNeo商汤智广“一站式”广告营销平台，集内容创作、媒体投放、效果分析于一体，可一键生成创意短视频，囊括脚本生成、背景替换、横竖屏转换、配字幕等覆盖短视频广告生产的多种服务，将传统由人工主导的短视频拍摄和制作模式，转变为AI辅助的自动化生成模式。
	未上市	迈吉客	迈吉客科技的内容生产工具尝试将影视动画CG技术消费化，把动画内容的生产门槛降低，将图文内容快速生成可视化内容。透过在混合现实仿真融合、表情/动作捕捉与实时自然交互等方面的技术创新和应用场景探索，Magics工具赋予IP生命，赋能其活灵活现的“演绎”内容，实现高频零门槛、千人千面、视觉化互动的内容生产，帮助品牌品牌立体化呈现、媒体形式升级、营销内容高效生产。
	未上市	影谱科技	在视频生成相关领域支持结构化视觉分析、影像自动合成技术（将视频短片、图片、音轨等按照规定效果批量化自动拼接）、智能视频编辑（基于视频中多模态信息的特征融合进行学习，按照氛围、情绪等高级语义限定，对满足条件片段进行检测并合成）、视频内容生产（对视频中的镜头、元素和场景采用不同的生成方式，同时对组件的组合方式进行学习，实现视频的自动化生产）、行为动作分析、场景信息恢复、跨模态转换。
音频生成	9999.HK	网易	2022年1月，网易正式推出首个AI音乐创作平台——网易天音，利用AI技术来帮助创作者快速制作音乐。同时，普通用户也能在“网易天音”微信小程序中，定制春节拜年歌曲，最快仅需10秒钟就能完成歌曲制作并分享。
	1698.HK	腾讯音乐	2022年12月，腾讯音乐天琴实验室正式推出AI音乐视觉生成技术MUSE(Music Envision),能依据用户选择的歌曲文意自动画出相应歌词海报或生成歌词动效视频等视觉内容,实现了行业首创的规模化音乐海报绘制技术,给用户创作UGC分享内容时提供较大便利,更加简单、高效,助力用户多元化音乐体验大大提升。
	未上市	DeepMusic（灵动音科技）	产品包括针对视频生成配乐的配乐猫、支持非音乐专业人员创作的口袋音乐、可AI生成歌词的 LYRICA、AI作曲软件LAZYCOMPOSER。目前已与国内多家音乐平台厂商达成合作。其音乐标注团队已形成了全球最精确的话语歌曲音乐信息库。
策略生成	未上市	rct AI	通过简单设计并调整不同的参数，rct AI的混沌球（Chaos Box）算法可以在游戏中大规模地轻松生成具有智能意识的虚拟角色。他们的行为和对话不会重复，皆为动态生成。在游戏场景中，部署具有不同性格的智能 NPC，通过对话、行为等动态交互，增加玩家的游戏时长，同时提供新的变现途径。具体包括性格化NPC、对抗式AI、互动式AI、大规模智能 NPC 部署、智能留存及智能运营策略等。目前，rct AI已凭借核心技术Chaos Box帮助了10余家企业，完成包括对战游戏、虚拟人铸造等多种类型的项目，与世界范围内 20+ 游戏厂商建立了深入合作，触达超过 2 亿用户。
	未上市	超参数科技	超参数科技提供的AI bot支持玩家陪玩（3D生存游戏AI猎户座α）、多人团队竞技（球球大作战）、非完美信息博弈AI（斗地主、德扑、麻将等）等。自有游戏AI平台“Delta”采用全新的“AI+游戏”研发管线，为开发侧和体验侧两端带来范式创新。目前，超参数AI Bot已在多款千万日活的产品中上线，每日在线数峰值将近百万，业内率先实现在3D FPS领域的大规模商业化落地。

AIGC相关标的——下游企业

领域	代码	公司	AIGC相关业务
电商	300785.SZ	值得买	自2017年开始尝试通过算法和机器的方式生产内容，即MGC（Machine-Generated Content）。2021年，机器贡献的内容占比为18.97%。
传媒	301270.SZ	汉仪股份	公司使用AI算法，根据300-500个字提取其风格特征，就可以生成与样字风格一致的全套字库，从而提高字库开发效率。
	300364.SZ	中文在线	公司已推出AI绘画功能和AI文字辅助创作功能，其中AI文字辅助创作功能已上线，该功能已向公司旗下17K文学平台全部作者开放。公司深度结合作者的真实写作场景，作者在使用AIGC功能时，通过针对不同的描写场景填写关键词和辅助短语，即可生成对应的文字内容描写，提高写作效率。
	000681.SZ	视觉中国	公司与百度旗下的AI作画平台文心一格将在创作者赋能和版权保护等方面展开多项合作。另外，公司入驻腾讯会议应用市场，为腾讯会议用户提供包括插画、摄影图片、动态图片以及AIGC-人工智能生成图片在内的各类虚拟背景图片。公司旗下元视觉平台有发行AI生成的数字艺术品，相关业务尚处于初期阶段。
	300781.SZ	因赛集团	公司开发GPT相关技术，应用于公众号标题和内容文案的自动生成与游戏社群互动工具。 公司的AI创意生成及管理平台“因赛引擎INSIGHTengine”已应用于公司的主要行业客户以及NFT（数字藏品）的平面及视频内容生成。
	300624.SZ	万兴科技	司旗下AI绘画产品万兴爱画目前已提供AI文字绘画、AI以图绘图、AI简笔画三种AI创作模式；万兴喵影及Wondershare Filmora也已集成文身图的AIGC能力；此外文档创意产品亿图脑图也已率先布局办公绘图AIGC应用，开启AIGC功能内测。
营销	301052.SZ	果麦文化	公司通过采集互联网大数据精选文章、本地文件导入转化为自己的内容库，有机训练：段落、词句、文章、知识。机器通过持续深度学习，可以生成语句通顺、可读性强的优质内容，素材专业、多元实现一键自动成稿，改写后文意相同、内容相似，实现底稿优化转换。达到高效孵化图书营销软文的目的。目前AI相关业务给公司带来的营业收入及利润占比极小，对公司生产经营未产生重大影响。
	002803.SZ	吉宏股份	公司在跨境电商板块和SaaS吉喵云中使用AIGC技术，如智能素材、智能广告、智能投放、智能客服等。公司先后与华为、亚马逊合作，共同完善业务场景的技术。
	301171.SZ	易点天下	公司已经开始使用AIGC技术为客户提供创意素材生成服务，并根据自身业务发展需求，不断探索AIGC在其它业务领域的应用。
教育	300081.SZ	恒信东方	公司的人工智能家庭教育机器人“斯泰同学”是一款公司自主研发、面向2-8岁儿童打造的AI家庭教育机器人，以智能主持人为核心，打造内容+场景+知识点的儿童教育趣味课堂，是AI合家欢平台的家庭接入端。
虚拟人	300182.SZ	捷成股份	公司子公司世优科技数字人已经接入chatGPT，正在通过数字人自身的人设背景等相关数据集，并基于OpenAI来训练这个数字人专有大脑形成个性化模型。截至目前，世优科技已为客户复活的虚拟数字人超500只。
	002354.SZ	天娱数科	公司虚拟数字人已经接入ChatGPT等模型，并已在Tik Tok跨境电商直播、虚拟主播直播互动等场景实现应用。
游戏	002624.SZ	完美世界	公司已将AI相关技术应用于游戏中的智能NPC、场景建模、AI剧情、AI绘图等方面。例如，《梦幻新诛仙》采用智能NPC与IK技术，使得NPC具有丰富的微表情，为玩家提供真实自然的交互体验；公司在研的仙侠题材MMORPG端游《诛仙世界》创新运用了全天候天气智能AI演算技术，实现了对雨、雪、大雾等天气的全局还原和细节处理，天气变化切实融入到玩家的游戏体验中。在研发方面，公司通过AI技术完成场景建模、纹理渲染等；此外公司还在游戏研发过程中使用AI绘图等技术提升游戏研发效率。
	0700.HK	腾讯控股	腾讯AI Lab自2017年开始研发策略协作型AI「绝悟」，应用于MOBA游戏「王者荣耀」，及FPS游戏「使命召唤」与「穿越火线」手机版。
	300459.SZ	汤姆猫	公司已尝试应用ChatGPT模型进行AI语音互动产品功能原型测试，努力将“会说话的汤姆猫”在未来迭代为“会聊天的汤姆猫”，为用户打造更逼真、更人性化的交互体验。目前，公司语音互动产品功能原型测试主要围绕“聊天”“问答”等场景，相关产品能否上线，以及是否能够达到预期，存在一定不确定性。
政务	300075.SZ	数字政通	公司的12345城市热线解决方案，综合应用了NLP语义挖掘、GES产品及服务、AI感知引擎、知识图谱等人工智能技术，实现了热线大数据的自动化、精细化分析，真正发挥了“数据金矿”的价值，使得12345热线数据真正有效赋能城市治理工作。相关业务已在上海、深圳等24个城市落地。目前公司正在研究ChatGPT在行业内的相关落地应用，测试生成文本在一些场景下的学习能力。
	002530.SZ	金财互联	公司全资子公司研发的数字税务员工“税宝”系基于AIGC技术研发的，已上岗服务一年多。
C端应用	MSFT.O	微软	微软正在迅速推进OpenAI的工具商业化，计划将包括ChatGPT、DALL-E等AI工具整合进微软旗下的所有产品中，并将其作为平台供其他企业使用。
	GOOGL.O	谷歌	2022年，开放测试Imagen模型，并发布了AI 写作协助工具 LaMDA Wordcraft、结合 Imagen Video 和 Phenaki 优势的超长连贯性视频生成模型。

五、商业模式：商业化初启，期待产业生态、技术与产品发展完善

商业模式：大模型商业化初启，小模型在部分领域已实现商业价值闭环

国海证券 SEALAND SECURITIES

	大模型			小模型	
	MaaS（Model as a Service）			垂直行业解决方案	
商业模式	1)按调用次数或调用量（Tokens等）收费；	2）按年/月订阅套餐收费；	3）定制服务，特定领域再开发，将大模型和数据库打包，按项目收费。	1）纯软件及平台；	1）一站式解决方案
面向用户	企业、机构、个人		企业、机构	细分行业企业	
毛利率	推理算力成本，毛利率可达80%+。		含再开发项目实施费用。	标准化产品，毛利率可达90%+。	含外购硬件，毛利率30%-70%。
提供商	OpenAI、微软、谷歌、Meta、百度、阿里、华为、腾讯、商汤、科大讯飞、字节、京东等。			科大讯飞、商汤、旷视、云从、依图、虹软、格灵深瞳、云天励飞、拓尔思、海康威视等。	
商业模式	大模型厂商自用，实现增量或降本增效。	云厂商“MaaS+IaaS”打包输出，实现IaaS收入增长和增量服务收入。	替代翻译、美工、原画师、程序员、分析师、设计师等繁琐重复的低端工作。	垂直行业解决方案，包括SDK产品、一站式落地解决方案。	
付费逻辑	谷歌、微软必应搜索引擎，YouTube视频创作等，阿里电商营销产品，腾讯企业微信、腾讯会议相关产品等，字节内容创作等；基于C端用户使用量内部付费。	大模型厂商+SaaS厂商，打造更多可直接面向C端的产品，SaaS厂商根据调用情况付费。	1）企业开发者调用后自用或个人用户自行调用，基于自身需求调用付费；2）为SaaS厂商提供产品付费。	智慧城市、智慧交通、智慧楼宇、智慧园区、智慧医疗、智慧金融、智慧生活、智能制造等多领域均有企业布局，在过去主要是感知、分析、决策式AI，部分存在生成式AI，已有部分行业实现商业价值闭环，主要是传统软件收费逻辑，不同行业略有不同。	
中美差距	差距不大且均有较大需求，甚至国内厂商的产品更加丰富多元。	生态差距较大，美国SaaS厂商面向全球，中国SaaS行业尚在快速发展中。	海外付费意识更高。	中美企业格局略有差异，美国头部效应更为明显，主要由细分行业龙头或者科技巨头提供相关AI驱动的解决方案；中国不局限于科技巨头和行业龙头，还有众多AI企业在众多细分行业、领域布局。	

商业模式：开始商业化尝试，会员制+按次收费为主

国海证券 SEALAND SECURITIES

类型	产品名称	隶属公司	收费方式
AI绘图	文心一格	百度在线网络技术有限公司	免费生成100张图后，9.9元50张图；15.9元100张图；49.9元400张图
	Tiamat	上海退格数字科技有限公司	公测阶段，可免费生成140张图，之后充值方式尚未明确
	6pen art	北京毛线球科技有限公司	最多免费生成100张后，5元10张图；30元100张图，100元400张图，500元2500张图
	盗梦师	西湖心辰科技有限公司	免费生成5张图后，5.5元25张图，24.9元125张图；或会员制：99元/月660张图，299元/月2160张图
	Stable Diffusion	Stability AI	免费生成200张图后，2英镑/200张图
	Disco Diffusion	Google	现阶段免费
	Midjourney	Midjourney	免费生成25张图后，10美元/月生成200张图，或30美元/月无限生成
	Mimic	RADIUS5 Inc.	公测阶段，免费生成30张后，付费生成100张但充值方式尚未明确
	Novel AI	Novel AI	10美元/月最多生成200张图；或25美元/月最多生成2000张图
	DALL-E2	OpenAI	免费生成25张图后，每月免费60张图，15美元生成460张
AI写作	ChatGPT	OpenAI	分为免费和plus付费（20美元/月），Plus会员高并发期无需排队、更快响应、更快尝试新功能等
	Contentnote	珠海横琴容徽信息科技有限公司	免费版每月生成10条，尝鲜版99元/月200条（限一次），初级版459元/月200条，高级版1999元/月1000条
	Effidit	腾讯	现阶段免费
	百度智能创作平台	百度在线网络技术有限公司	智能写作会员198元/月，视频创作会员1698元/月
	Jasper AI	Jasper AI	免费试用5天，40美元/月35k词，82美元/月100k词
	Copysmith	Copysmith	初级版 19美元/月 或 190美元/年 75 credits；专业版 59美元/月 或 590美元/年 100credits
	Rytr	Google Copysmith	免费版每月生成5k词，9美元/月50k词，29美元/月无限生成

资料来源：第一财经，OpenAI，FinPins，容徽，Effidit，百度智能创作平台，国海证券研究所

成本测算-训练成本：总成本持续提升，但同级别参数消耗量将显著下降

表：大模型训练成本中各成本占比概览

成本项	占比
算力成本	**40%-70%**
设备折旧	14%-24.5%
存储	10-17.5%
电费	12%-21%
宽带	4%-7%
数据成本	**20%-35%**
数据收集	6%-10.5%
数据标注	8%-17.5%
数据清洗	4%-7%
数据存储	
人力成本	**10%-25%**

注：参考ChatGPT、百度文心、阿里M6、华为盘古大模型数据

- 随着参数量快速膨胀，算力成本会持续上升；
- 但随着模型压缩、蒸馏等，同参数量级别的模型算力消耗量会显著下降。

- 数据获取：随着应用较快数据反哺，数据获取边际成本将下降；
- 数据标注：有两个方向，一是无监督学习流行、标注自动化提升，数据标注成本下降；而是对于专业领域、图像视频等复杂数据标注需求提升。
- 随着数据量快速膨胀，训练数据集需求越来越大，数据存储成本也将相应提升。

- AI资产复用、自动化程度提升，规模效应，单位人力成本下降。

表：各大模型全局训练成本概览

模型	算力成本占比	数据成本占比	人力成本占比	单次完整训练价格（万美元/次）	全年完整训练次数（次）	全年训练成本（万美元）	已投入金额（万美元）
ChatGPT3	70%	20%	10%	400-1000	1-2	2000左右	4300左右
ChatGPT3.5	60%	25%	15%	400-1000	1-2	不到2000	

风险提示

- **人工智能发展不及预期，AIGC发展不及预期**，人工智能发展历经三次波折，曾因基础硬件软件、投入产出比较低等原因受限，目前AI、AIGC的发展可能仍存在相应的问题和风险；
- **技术发展不及预期**，自ChatGPT大火出圈后，对于人工智能技术的关注度和期望可能过高，技术发展存在不及高预期的风险；
- **商业化拓展不及预期**，目前AI、AIGC商业化模式仍在持续探索和尝试阶段，可能由于付费意愿、实际效果等存在商业化拓展不达预期的风险；
- **行业竞争加剧风险**，AIGC成为新一轮的生产力革命，海内外科技巨头纷纷加速布局，行业竞争加剧；
- **中美科技竞争不确定性风险**，如美国限制对中出口高端芯片，将多个科技企业和相关机构列入“实体清单”等。

研究小组介绍

海外小组介绍

陈梦竹，南开大学本科&硕士，6 年证券从业经验，现任国海证券海外研究团队首席，专注于全球内容&社交互联网、消费互联网、科技互联网板块研究。

尹芮，康奈尔大学硕士，中国人民大学本科，2年证券从业经验，现任国海证券海外互联网分析师，主要覆盖内容&社交互联网方向。

张娟娟，上海财经大学硕士，三年产业工作经验，曾任职于阿里、美团，现任国海证券海外互联网研究助理，主要覆盖消费互联网方向。

陈凯艺，武汉大学硕士，西南财经大学本科，1年证券从业经验，现任国海证券海外研究团队研究助理，主要覆盖科技互联网方向。

罗婉琦，伦敦政治经济学院硕士，现任国海证券海外研究团队研究助理，主要覆盖消费互联网方向。

分析师承诺

陈梦竹，本报告中的分析师均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，客观的出具本报告。本报告清晰准确的反映了分析师本人的研究观点。分析师本人不曾因，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取到任何形式的补偿。

国海证券投资评级标准

行业投资评级

推荐：行业基本面向好，行业指数领先沪深300指数；
中性：行业基本面稳定，行业指数跟随沪深300指数；
回避：行业基本面向淡，行业指数落后沪深300指数。

股票投资评级

买入：相对沪深300 指数涨幅20%以上；
增持：相对沪深300 指数涨幅介于10%～20%之间；
中性：相对沪深300 指数涨幅介于-10%～10%之间；
卖出：相对沪深300 指数跌幅10%以上。

免责声明和风险提示

免责声明

本报告的风险等级定级为R3，仅供符合国海证券股份有限公司（简称“本公司”）投资者适当性管理要求的的客户（简称“客户”）使用。本公司不会因接收人收到本报告而视其为客户。客户及/或投资者应当认识到有关本报告的短信提示、电话推荐等只是研究观点的简要沟通，需以本公司的完整报告为准，本公司接受客户的后续问询。

本公司具有中国证监会许可的证券投资咨询业务资格。本报告中的信息均来源于公开资料及合法获得的相关内部外部报告资料，本公司对这些信息的准确性及完整性不作任何保证，不保证其中的信息已做最新变更，也不保证相关的建议不会发生任何变更。本报告所载的资料、意见及推测仅反映本公司于发布本报告当日的判断，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。报告中的内容和意见仅供参考，在任何情况下，本报告中所表达的意见并不构成对所述证券买卖的出价和征价。本公司及其本公司员工对使用本报告及其内容所引发的任何直接或间接损失概不负责。本公司或关联机构可能会持有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提供投资银行、财务顾问或者金融产品等服务。本公司在知晓范围内依法合规地履行披露义务。

风险提示

市场有风险，投资需谨慎。投资者不应将本报告为作出投资决策的唯一参考因素，亦不应认为本报告可以取代自己的判断。在决定投资前，如有需要，投资者务必向本公司或其他专业人士咨询并谨慎决策。在任何情况下，本报告中的信息或所表述的意见均不构成对任何人的投资建议。投资者务必注意，其据此做出的任何投资决策与本公司、本公司员工或者关联机构无关。

若本公司以外的其他机构（以下简称“该机构”）发送本报告，则由该机构独自为此发送行为负责。通过此途径获得本报告的投资者应自行联系该机构以要求获悉更详细信息。本报告不构成本公司向该机构之客户提供的投资建议。

任何形式的分享证券投资收益或者分担证券投资损失的书面或口头承诺均为无效。本公司、本公司员工或者关联机构亦不为该机构之客户因使用本报告或报告所载内容引起的任何损失承担任何责任。

郑重声明

国海证券 · 研究所 · 海外研究团队

心怀家国，洞悉四海

国海研究上海

上海市黄浦区福佑路8号人保寿险大厦7F

邮编：200010

电话：021-60338252

国海研究深圳

深圳市福田区竹子林四路光大银行大厦28F

邮编：518041

电话：0755-83706353

国海研究北京

北京市海淀区西直门外大街168号腾达大厦25F

邮编：100044

电话：010-88576597