

人工智能专题报告： 生成式人工智能产业全梳理

行业研究 · 深度报告

投资评级：超配（维持评级）

证券分析师：熊莉
xiongli1@guosen.com.cn
S0980519030002

证券分析师：库宏珪
kuhongyao@guosen.com.cn
S0980520010001

证券分析师：朱松
zhusong@guosen.com.cn
S0980520070001

证券分析师：张伦可
zhanglunke@guosen.com.cn
S0980521120004

- **人工智能作为第四次科技革命，已经进入2.0时代。**人工智能概念于1956年被提出，AI产业的第一轮爆发源自2012年，2012年AlexNet模型问世开启了CNN在图像识别的应用，2015年机器识别图像的准确率首次超过人（错误率低于4%），开启了计算机视觉技术在各行各业的应用。但是，人工智能1.0时代面临着模型碎片化，AI泛化能力不足等问题。2017年Google Brain团队提出Transformer架构，奠定了大模型领域的主流算法基础，从2018年开始大模型迅速流行，2018年谷歌团队的模型参数首次过亿，到2022年模型参数达到5400亿，模型参数呈现指数级增长，“预训练+微调”的大模型有效解决了1.0时代AI泛化能力不足的问题。新一代AI技术有望开始新一轮的技术创新周期。
- **自18年起大模型快速流行，有望重新定义生产力。**2018年以来，预训练语言模型（PLM）及其“预训练-微调”方法已成为自然语言处理（NLP）任务的主流范式，该范式先利用大规模无标注数据通过自监督学习预训练语言大模型，得到基础模型，再利用下游任务的有标注数据进行有监督学习微调模型参数，实现下游任务的适配。在AI的1.0时代：存在模型碎片化明显、AI泛化能力不足等问题。“预训练+微调”大模型能显著降低AI工程化门槛，预训练大模型在海量数据的学习训练后具有良好的通用性和泛化性，细分场景的应用厂商能够基于大模型通过零样本、小样本学习即可获得显著的效果，使得人工智能有望构建成统一的智能底座，AI+赋能各行各业。本轮的生成式AI有望从简单的内容生成，逐步达到具有预测、决策、探索等更高的认知智能。
- **OpenAI当前已迭代五代模型，GPT-4开始布局多模态。**OpenAI于2015年成立，微软于2019年开始与OpenAI建立战略合作伙伴关系，GPT共发布五代模型GPT-1、GPT-2、GPT-3、ChatGPT以及GPT4。GPT-1于2018年6月发布，首次将transformer与无监督的预训练技术相结合。2020年5月发布GPT-3，模型参数量为1750亿。2022年11月，OpenAI正式推出了对话交互式的ChatGPT。相比于GPT-3，ChatGPT引入了基于人类反馈的强化学习（RLHF）技术以及奖励机制。2023年3月，OpenAI正式推出GPT-4，成为目前较先进的多模态大模型。GPT-4主要在识别理解能力、创作写作能力、处理文本量以及自定义身份属性迭代方面取得进展。
- **百度于2023年3月正式推出大模型文心一言。**文心一言主要由文心大模型提供支持，文心一言拥有有监督精调、RLHF、提示构建、知识增强、检索增强和对话增强六大核心技术。其中前三项与ChatGPT的技术十分类似，知识增强包括知识内化和知识外用；检索增强指基于百度搜索引擎，先对内容进行检索，再筛选出有用的部分整合输出结果；对话增强指记忆机制、上下文理解和对话规划等技术。
- **投资建议：建议关注基础层与应用层AI龙头。应用层建议关注：科大讯飞、金山办公、同花顺、广联达、凌志软件、彩讯股份、拓尔思、福昕软件、税友股份等。基础层建议关注海光信息、浪潮信息、景嘉微等。**
- **风险提示：**AI技术商业化落地不及预期；行业竞争加剧，技术迭代风险；数据安全等政策不确定性；贸易摩擦风险。

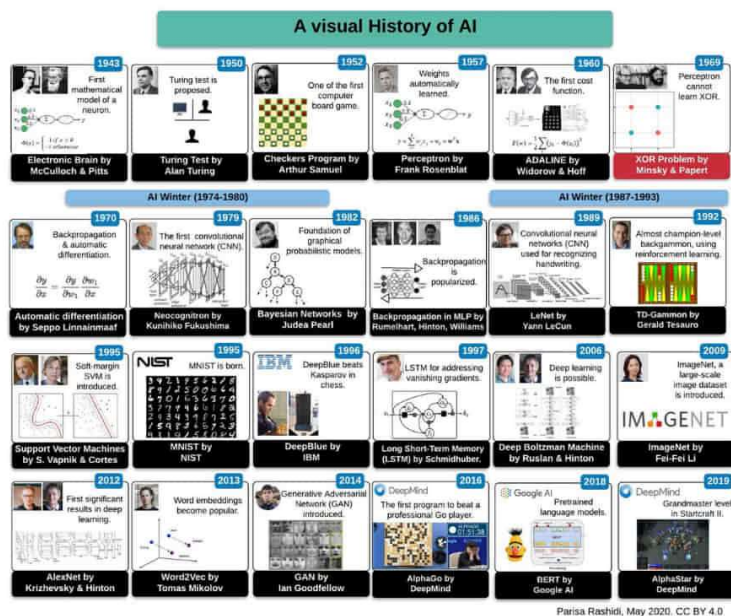
- [01] 行业梳理：生成式AI有望带动新一轮技术创新周期
- [02] 海外玩家：OpenAI持续领先，谷歌等巨头纷纷布局
- [03] 中国玩家：百度发布文心一言，中国玩家快速追赶
- [04] 市场规模：模型参数不断增加，算力需求快速增长
- [05] 商业模式：开启订阅制收费，不断开放API接口
- [06] 应用场景：GPT走向多模态，下游应用场景不断打开
- [07] 投资建议：建议关注基础层与应用层AI龙头

1、行业梳理：生成式AI有望带动新一轮技术创新周期

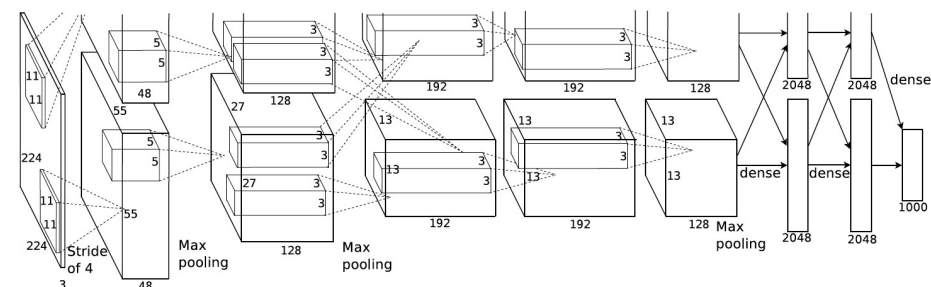
1.1 人工智能产业自2012年开始迎来蓬勃发展

- **2012年至今属于人工智能的蓬勃发展期。**人工智能是研究、开发用于模拟、延伸和扩展人的智能的理论、方法、技术及应用系统的一门新的技术科学。人工智能的发展历史大致可以被划分为5个阶段，2012年至今处于人工智能的蓬勃发展期，其分界点就是2012年前后IBM开发的人工智能程序“沃森”参加了一档智力问答节目并战胜了两位人类冠军以及AlexNet在ImageNet竞赛中取得胜利。
- **以深度神经网络为代表的信息技术的发展，推动了人工智能领域的进步与拓展。**2006年，Hinton等人利用单层的RBM自编码预训练使得深层的神经网络训练得以实现；2012年，Hinton和Alex Krizhevsky设计的AlexNet神经网络模型在ImageNet竞赛中实现图像识别分类，成为新一轮人工智能发展的起点。

神经网络发展里程碑事件



AlexNet神经网络结构



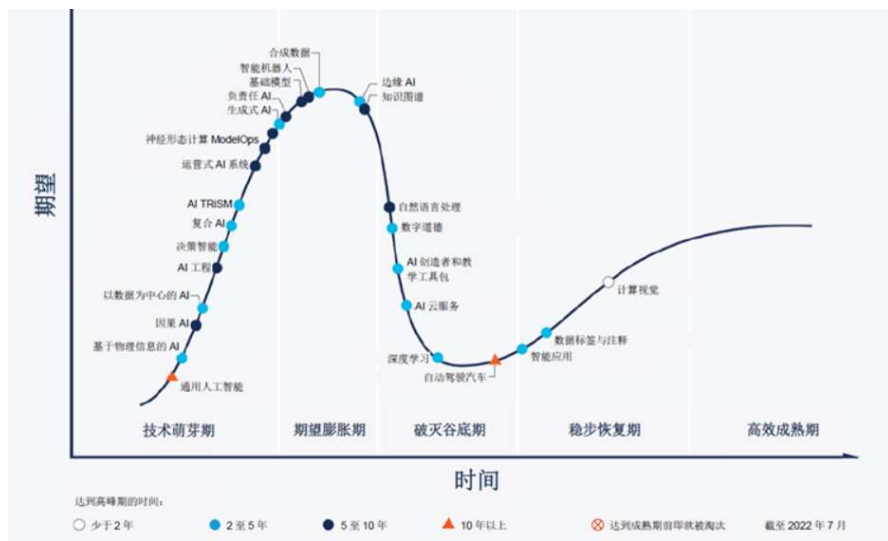
资料来源：CSDN、国信证券经济研究所整理

资料来源：CSDN、国信证券经济研究所整理

1.1 人工智能作为第四次科技革命，已经进入2.0时代

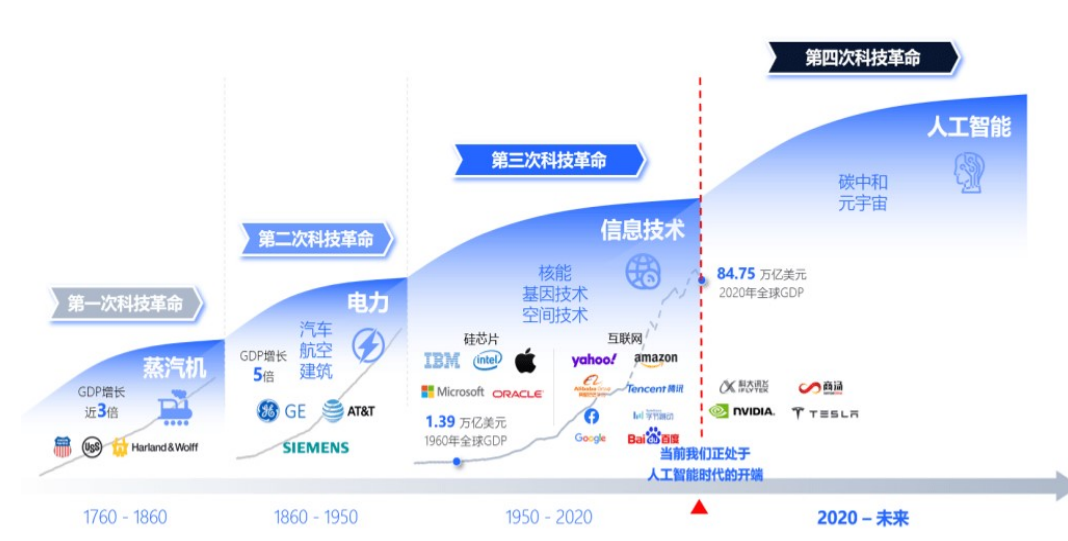
- **人工智能1.0时代（2012年-2018年）**：人工智能概念于1956年被提出，AI产业的第一轮爆发源自2012年，2012年AlexNet模型问世开启了CNN在图像识别的应用，2015年机器识别图像的准确率首次超过人（错误率低于4%），开启了计算机视觉技术在各行各业的应用，带动了人工智能1.0时代的创新周期，AI+开始赋能各行各业，带动效率提升。但是，人工智能1.0时代面临着模型碎片化，AI泛化能力不足等问题。
- **人工智能2.0时代（2017年-至今）**：2017年Google Brain团队提出Transformer架构，奠定了大模型领域的主流算法基础，从2018年开始大模型迅速流行，2018年谷歌团队的模型参数首次过亿，到2022年模型参数达到5400亿，模型参数呈现指数级增长，“预训练+微调”的大模型有效解决了1.0时代AI泛化能力不足的问题。新一代AI技术有望开始新一轮的技术创新周期。

2022年Gartner人工智能商业化曲线



资料来源：Gartner、国信证券经济研究所整理

人工智能有望迎来第四次科技革命

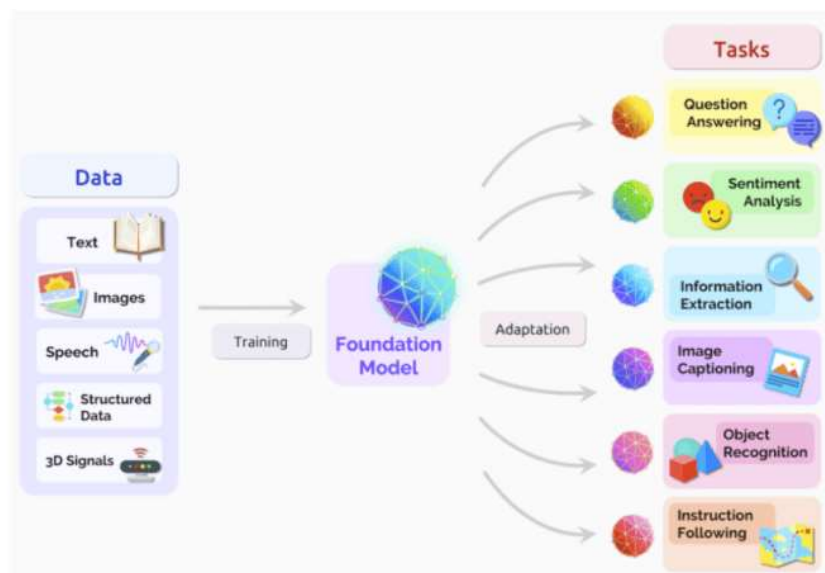


资料来源：Scale Partners、国信证券经济研究所整理

1.1 大模型快速爆发，重新定义人工智能产业

- 自18年起大模型快速流行，有望重新定义人工智能产业。2018年以来，预训练语言模型（PLM）及其“预训练-微调”方法已成为自然语言处理（NLP）任务的主流范式，该范式先利用大规模无标注数据通过自监督学习预训练语言大模型，得到基础模型，再利用下游任务的有标注数据进行有监督学习微调模型参数，实现下游任务的适配。规模越大的模型不仅在已知任务上有着更好的表现，同时展现出完成更复杂的未知任务的强大泛化能力。

大模型示意图



资料来源：CSDN、国信证券经济研究所整理

基础模型的“预训练-微调”范式与传统深度学习区别



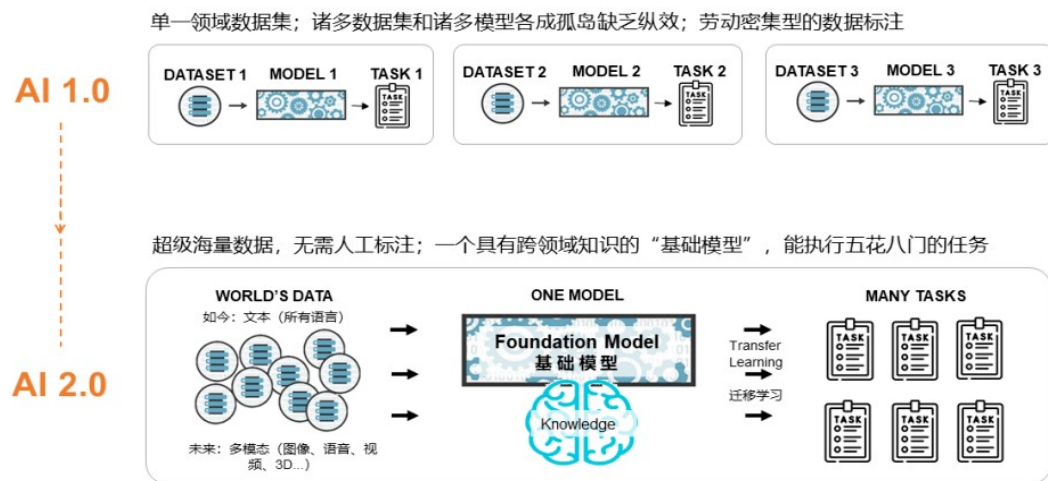
资料来源：清华大学、国信证券经济研究所整理

1.1 AI工程化门槛不断降低，有望重新定义生产力

AI工程化门槛不断降低，有望重新定义生产力

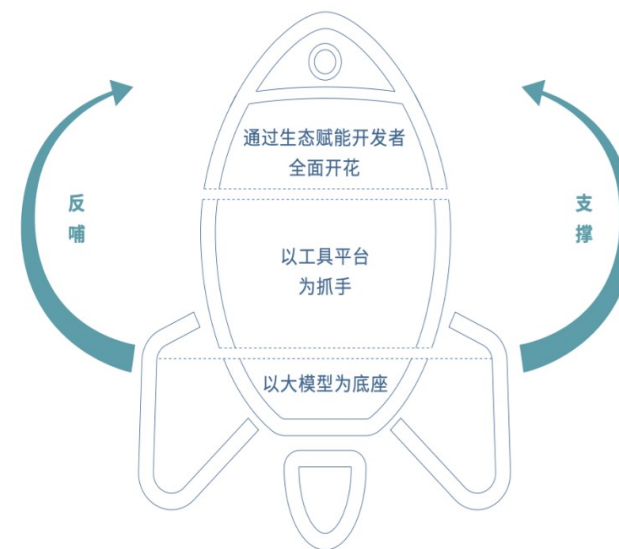
- **AI的1.0时代：**存在模型碎片化明显、AI泛化能力不足等问题，导致大多数行业需要花费巨大成本来收集和标注数据，从而导致规模不经济，而且下游场景存在诸多细分小场景，其商业化价值小、有效数据少、模型训练不足。
- **AI的2.0时代：**“预训练+微调”大模型能显著降低AI工程化门槛，预训练大模型在海量数据的学习训练后具有良好的通用性和泛化性，细分场景的应用厂商能够基于大模型通过零样本、小样本学习即可获得显著的效果，使得人工智能有望构建起统一的智能底座，AI+赋能各行各业。本轮的生成式AI有望从简单的内容生成，逐步达到具有预测、决策、探索等更高的认知智能。

人工智能1.0时代与2.0时代



资料来源： 创新工场、国信证券经济研究所整理

大模型将显著降低人工智能应用门槛



资料来源： IDC、国信证券经济研究所整理

1.2 产业分类：AI产业链主要包括基础层、技术层、应用层三大层

人工智能产业链可分为基础层、技术层和应用层三大层面：

- **基础层**：侧重基础支撑平台的搭建，包含传感器、AI芯片、数据服务和计算平台；
- **技术层**：侧重核心技术的研发，主要包括算法模型、基础框架、通用技术；
- **应用层**：注重产业应用发展主要包含行业解决方案服务、硬件产品和软件产品。

人工智能产业链

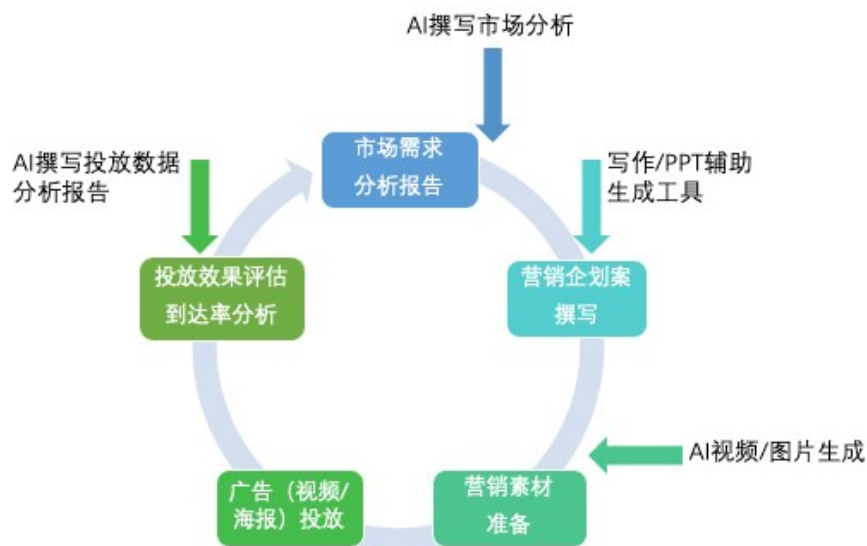
		金融	零售	医疗	安防	教育	工业	交通	农业
应用层	行业应用								
	产品应用	智能机器人		智能无人机		智能语音识别系统		智能搜索系统	
技术层	通用技术	自然语言处理		智能语音		机器问答		计算机视觉	
	算法模型	机器学习			深度学习			增强学习	
	基础框架	分布式存储			分布式计算			神经网络	
基础层	传感器	光学传感器			声学传感器			其他传感器	
	芯片	GPU			FPGA			ASIC	
	计算平台	数据服务			数据采集			数据处理	

资料来源：36氪研究院、国信证券经济研究所整理

1.2 产业分类：AI模型可分为决策式AI与生成式AI

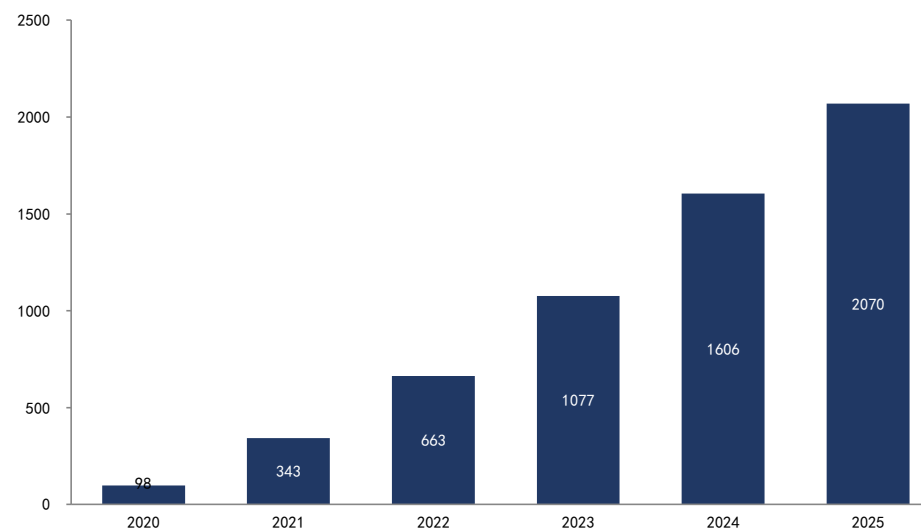
- **AI模型大致可以分为决策式AI和生成式AI两种。**决策式AI指学习数据中的条件概率分布，根据已有数据进行分析、判断、预测，主要应用模型有用于推荐系统和风控系统的辅助决策、用于自动驾驶和机器人的决策智能体。生成式AI指学习数据中的联合概率分布，并非简单分析已有数据而是学习归纳已有数据后进行演技创造，基于历史进行模仿式、缝合式创作，生成了全新的内容，也能解决判别问题。
- **中国生成式AI商业应用规模迎来快速增长，预计2025年破两千亿。**根据中关村大数据产业联盟发布的《中国AI数字商业展望2021-2025》报告披露，到2025年，中国生成式AI商业应用规模将达到2070亿元，未来五年的年均增速84%。根据Gartner《2021年预测：人工智能对人类和社会的影响》给出的积极预测，到2023年，将有20%的内容被生成式AI所创建。至2025年，预计生成式AI产生的数据将占有所有数据的10%。

生成式AI应用场景



资料来源：知乎、国信证券经济研究所整理

2020-2025年中国生成式AI技术栈应用规模（亿元）

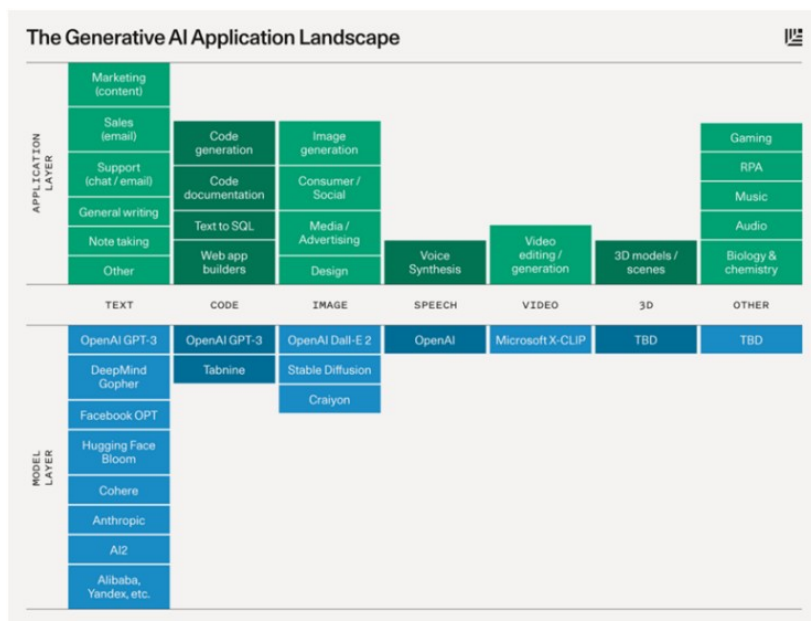


资料来源：《中国AI数字产业展望2021-2025》、国信证券经济研究所整理

1.3 AI产业正在逐渐从传统分析型AI走向生成式AI

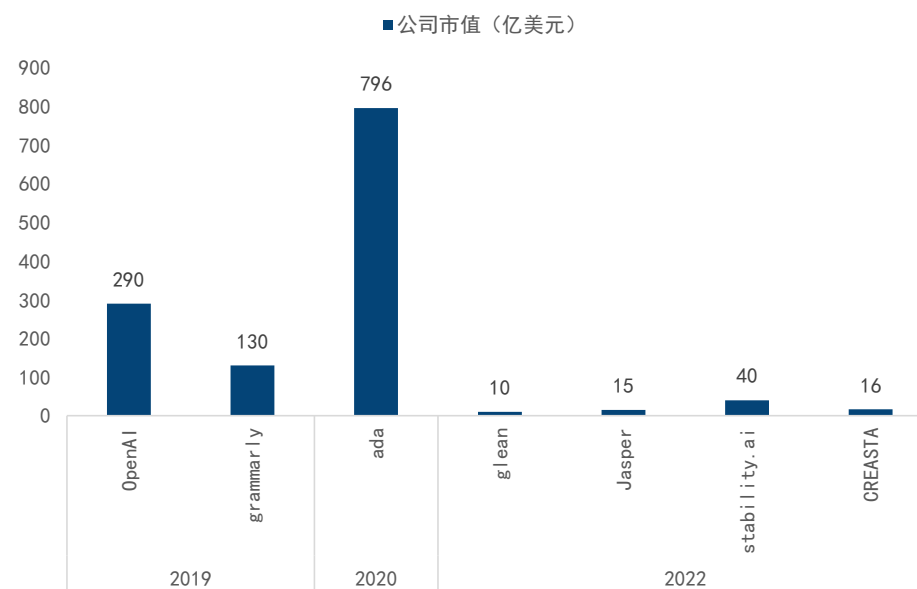
- 人工智能在经历前期技术积累和迭代后，逐渐突破传统分析型AI领域，迎来生成式AI的爆发期。从2012年至今，生成式AI急速发展，其源头就是DNN算法的升级，实现了语音和图像识别等功能。
- 生成式AI市场前景广阔，赛道内诞生多家独角兽企业。据波士顿咨询预测，至2025年生成式人工智能的市场规模将至少达到600亿美元，而其中大约30%的AI应用将来自广义的生成式AI技术。随着生成式AI模型的进一步完善，自主创作和内容生产的门槛将大大降低，市场响应该领域的巨大需求，在2019-2022年间共有7家独角兽公司诞生，截至2023年2月，这七家的估值合计达到644亿美元，其中OpenAI借助旗下产品ChatGPT爆火的东风，一家公司的估值便突破290亿美元。

生成式AI竞争格局



资料来源：澳财、国信证券经济研究所整理

生成式AI领域独角兽公司及其公司市值（亿美元）



资料来源：CSDN、国信证券经济研究所整理

1.3 生成式AI在文本、代码、图片、语音等领域应用广阔

- 生成式AI在文本 (Text)、代码生成 (Code generation)、图片 (Images)、语音合成 (Speech synthesis)、视频和3D模型等领域拥有广阔的应用场景。

AIGC三大前沿能力						
	PRE - 2020	2020	2022	2023?	2025?	2030?
TEXT	Spam detection Translation Basic Q&A	Basic copy writing First drafts	Longer form Second drafts	Vertical fine tuning gets good (scientific papers, etc)	Final drafts better than the human average	Final drafts better than professional writers
CODE	1-line auto-complete	Multi-line generation	Longer form Better accuracy	More languages More verticals	Text to product (draft)	Text to product (final), better than full-time developers
IMAGES			Art Logos Photography	Mock-ups (product design, architecture, etc.)	Final drafts (product design, architecture, etc.)	Final drafts better than professional artists, designers, photographers
VIDEO / 3D / GAMING			First attempts at 3D/video models	Basic / first draft videos and 3D files	Second drafts	AI Roblox Video games and movies are personalized dreams

Large model availability: ● First attempts ● Almost there ● Ready for prime time

资料来源：红杉资本、国信证券经济研究所整理

1.3 AIGC具备三大前沿能力，生成式AI有望带动新一轮创新周期



- **AIGC具备三大前沿能力，未来应用空间广阔。**AIGC发展火热，以chatgpt为代表的问答机器人，逐步走向大众视野。AIGC（AI-Generated Content）即人工智能生产内容，可用于代码生成、文本问答、图像生成等。AIGC是继专业生成内容（PGC）和用户生成内容（UGC）之后，利用人工智能技术生成内容的新生产方式。AIGC技术演化出三大前沿技术能力：数字内容孪生、数字内容的智能编辑、数字内容的智能创作。ChatGPT能理解并生成文字，属于AIGC技术应用中的文本生成模态应用模型。根据Gartner测算，当前AIGC占有生成数据小于1%，AIGC生成数据渗透率有广阔提升空间，预计该数字到2025年或上升至10%。
- **AIGC产业链上游主要提供AI技术及基础设施，包括数据供给方、数据分析及标注、创造者生态层、相关算法等。**AIGC应用对数字基础设施要求较高，随着ChatGPT掀起AIGC发展浪潮，数据基础设施有望加速升级。中游主要针对文字、图像、视频等垂直赛道，提供数据开发及管理工具，包括内容设计、运营增效、数据梳理等服务。下游包括内容终端市场、内容服务及分发平台、各类数字素材以及智能设备，AIGC内容检测等。

AIGC三大前沿能力



资料来源：中国信通院、国信证券经济研究所整理

AIGC产业链



资料来源：量子位智库、国信证券经济研究所整理

2、海外玩家：OpenAI持续领先，谷歌等巨头纷纷布局

2.1 OpenAI 发展历程（1）：

OpenAI 的发展历程分为四个阶段：

- 阶段一：2015年OpenAI 首席执行官阿尔特曼，与埃隆·马斯克等人宣布出资10亿美元，创立了非营利性研究机构OpenAI。
- 阶段二：2019年OpenAI 的架构进行了调整，调整后变身为两家机构——营利性机构OpenAI LP和最初的非营利机构OpenAI, Inc。
- 阶段三：从2019年开始，微软与OpenAI 建立了战略合作伙伴关系，不少于三次投资，共投入130亿美元，成为OpenAI 最大的有限合伙人。OpenAI LP从成立之初参与投资的VC，也成为有限合伙人。
- 阶段四：在OpenAI 未来盈利后，逐步回报投资人：1、优先保证OpenAI 的首批投资者收回初始资本；2、微软投资完成、OpenAI LP首批投资人收回初始投资后，微软有权获得OpenAI LP 75%利润；3、微软收回130亿美元投资、从OpenAI LP获得920亿美元利润后，它分享利润的比例从75%降到49%；4、OpenAI LP产生的利润达到1500亿美元后，微软和其他风险投资者的股份将无偿转让给OpenAI LP的普通合伙人——非营利机构OpenAI, Inc。

OpenAI 发展阶段



资料来源：OpenAI 官网、国信证券经济研究所整理

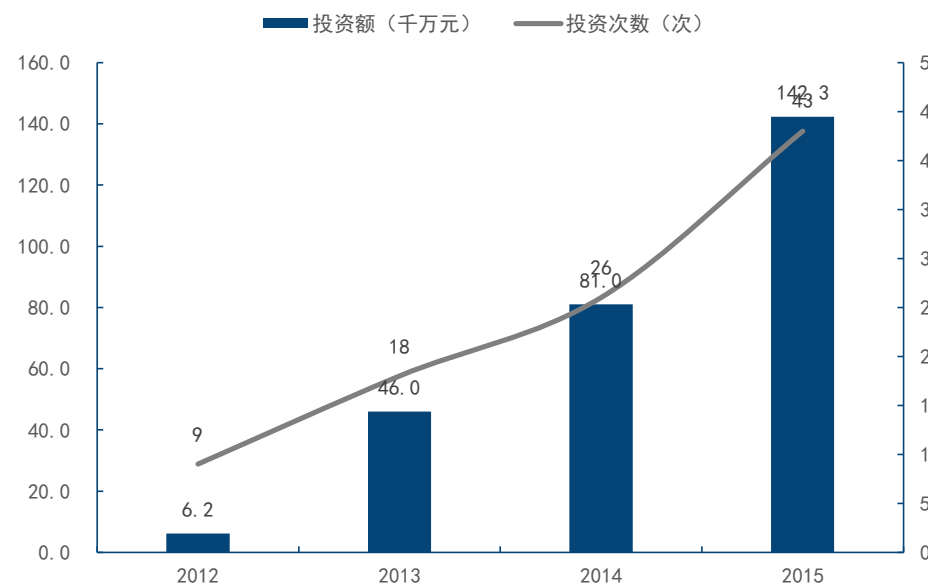
2.1 OpenAI 发展历程（2）：OpenAI 成立于2015年

- 自2013年起人工智能迎来发展高潮。2006年Hinton提出“深度学习”神经网络使得人工智能性能获得突破性进展，在2013年深度学习宣发在语音和视觉识别上取得成功，识别率分别超过99%和95%，人工智能进入感知智能时代。在此期间全球人工智能市场保持高速增长，截至2015年全球AI市场规模达到74.5亿美元，而且愈发受到投资机构青睐，投资额从2012年的0.62亿元提升至2015年的142.3亿美元，增长幅度达到2195.16%。
- 2015年OpenAI作为一家非营利性人工智能研究公司创立。基于人工智能高速发展的背景，OpenAI由Elon Musk，Sam Altman（美国创业孵化器总裁）及Pieter Abbeel（PayPal联合创始人）等人创建，在创立之初由Elon Musk和Sam Altman担任其联合主席，由Pieter Abbeel等人担任顾问。

OpenAI 初创成员及其担任职务

担任职务	姓名
research director	Ilya Sutskever (experts in machine learning)
CTO	Greg Brockman (formerly the CTO of Stripe)
research advisor	Pieter Abbeel, Yoshua Bengio, Alan Kay, Sergey Levine, and Vishal Sikka (world-class research engineers and scientists)
co-chairs	Sam Altman, Elon Musk

2012-2015年人工智能行业投资次数和金额



2.1 OpenAI 发展历程（3）：2016年发布第一个开源平台

- 2016年4月，OpenAI 发布了第一个项目——OpenAI Gym Beta。OpenAI Gym是由OpenAI开发的一个开源平台，旨在在各种强化学习问题中加速算法的开发和比较，该工具也是OpenAI第一个开放的成果。
- Gym的核心组件是环境（Environment）和智能体（Agent）。环境是一个可观察到的系统，它定义了智能体如何与外部世界交互；智能体则是一个能够感知到和影响环境的程序，是训练和优化的对象。
- Gym提供了许多强化学习问题的标准化环境。例如经典控制问题（Classic control），Atari游戏，甚至是Roboschool等物理模拟环境。

GYM基本应用示意图

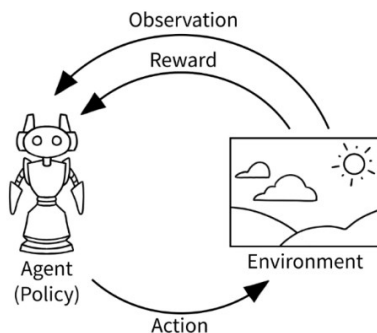
Initializing Environments

Initializing environments is very easy in Gym and can be done via:

```
import gym
env = gym.make('CartPole-v0')
```

Interacting with the Environment

Gym implements the classic "agent-environment loop":

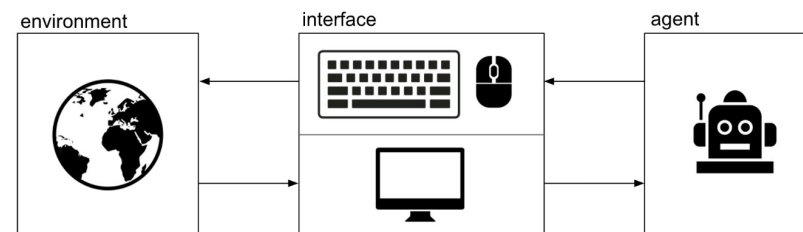


资料来源：OpenAI 官网、国信证券经济研究所整理

Universe基础设施和应用示意图

Universe Infrastructure

Universe exposes a wide range of environments through a common interface: the agent operates a remote desktop by observing pixels of a screen and producing keyboard and mouse commands. The environment exposes a VNC server and the `universe` library turns the agent into a VNC client.



资料来源：OpenAI 官网、国信证券经济研究所整理

2.1 OpenAI 发展历程（4）：2017年公布最新强化学习算法PPO

- 2017年7月，OpenAI 公布最新强化学习算法PPO（Proximal Policy Optimization），好于同期最强的算法模型，成为openai默认的强化学习算法。PPO包含三方面的技术进步：1. 代理策略和价值函数的优化，在PPO算法中，同时优化代理策略和价值函数（Value Function），通过在最大化奖励的同时最大化代理策略和价值函数的梯度，同时对这两者进行优化；2. 剪辑代理策略更新是PPO算法的核心部分，该方法通过使用约束优化来保证新的策略不会太远离旧的策略，减轻了过渡调整代理策略的问题，并提高总体稳定性；3. 优化函数的选择，在PPO算法中，需要选择合适的优化函数来最大化代理策略和价值函数的梯度，常用的优化函数包括Adam、SGD和RMSProp等。

新的目标函数可以在连续任务方面表现出优异性能

We've previously detailed a variant of PPO that uses an adaptive KL penalty to control the change of the policy at each iteration. The new variant uses a novel objective function not typically found in other algorithms:

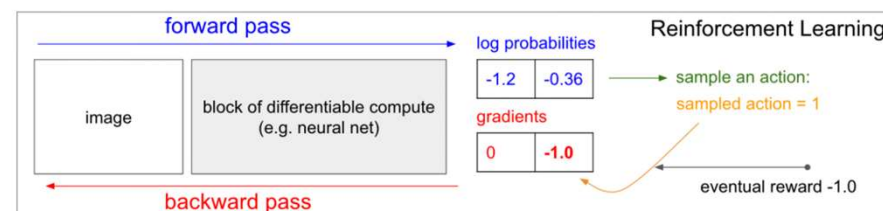
$$L^{CLIP}(\theta) = \hat{E}_t[\min(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t)]$$

- θ is the policy parameter
- \hat{E}_t denotes the empirical expectation over timesteps
- r_t is the ratio of the probability under the new and old policies, respectively
- \hat{A}_t is the estimated advantage at time t
- ϵ is a hyperparameter, usually 0.1 or 0.2

资料来源：OpenAI 官网、国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

策略梯度解决方案——正奖错罚



资料来源：OpenAI 官网、国信证券经济研究所整理

2.1 OpenAI发展历程（5）：2018年发布第一代transformer的GPT模型



- 2018年6月，OpenAI公布了第一个将transformer与无监督的预训练技术相结合的GPT模型，其取得的效果要好于当前的已知算法。该模型被称为GPT-1，并由一个具有10亿个参数的单层transformer组成。这一模型的训练使用了大规模的无监督语料库，使它能够生成各种自然语言处理任务的有力表现。同月OpenAI宣布他们的OpenAI Five已经开始在Dota2游戏中击败业余人类团队，OpenAI Five使用了256个P100 GPUs和128000个CPU核，通过每天玩180年时长的游戏来训练模型。在同年8月份的专业比赛中，OpenAI Five输掉了2场与顶级选手的比赛，但是比赛的前25-30分钟内，OpenAI Five的模型有着十分良好的表现。OpenAI Five继续发展并在2019年4月15日宣布打败了当时的Dota2世界冠军。

GPT可以显著提高在复杂NLP任务中的表现



资料来源：OpenAI官网、国信证券经济研究所整理

OpenAI Five与早期游戏机器人的参数对比

	OPENAI 1V1 BOT	OPENAI FIVE
CPUs	60,000 CPU cores on Azure	128,000 preemptible CPU cores on GCP
GPUs	256 K80 GPUs on Azure	256 P100 GPUs on GCP
Experience collected	~300 years per day	~180 years per day (~900 years per day counting each hero separately)
Size of observation	~3.3 kB	~36.8 kB
Observations per second of gameplay	10	7.5
Batch size	8,388,608 observations	1,048,576 observations
Batches per minute	~20	~60

资料来源：OpenAI官网、国信证券经济研究所整理

2.1 OpenAI 发展历程（6）：生成模型开始拓展至其他领域

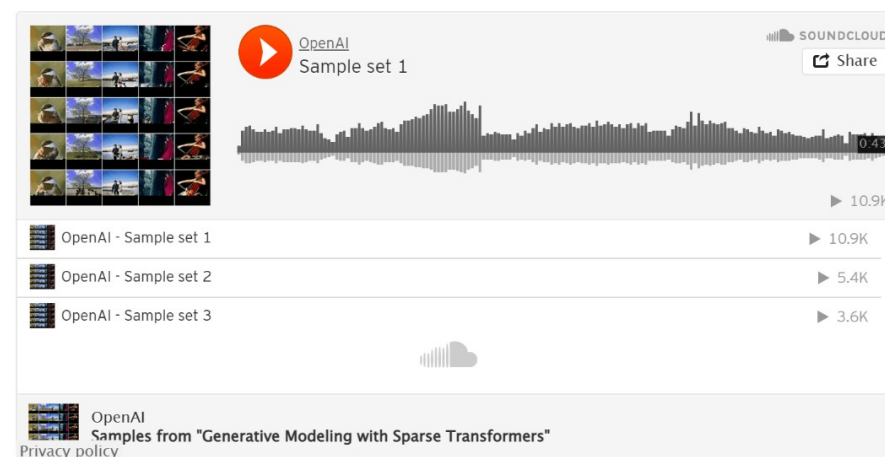
- 2019年2月，OpenAI在博客《Better Language Models and Their Implications》中官宣GPT-2模型。GPT-2模型拥有15亿参数，基于800万网页数据训练，该模型就是GPT的规模化结果，在10倍以上的数据以10倍以上的参数训练。OpenAI在2月份GPT-2发布的时候仅仅公开了他们的1.24亿版本的预训练结果，其后的5月份发布了3.55亿参数版本的预训练结果，并在半年后的8月份发布了一个7.74亿参数版本的GPT-2预训练结果。2019年11月5日，15亿参数的完整版本的GPT-2预训练结果发布。
- 2019年3月，OpenAI将生成模型开始拓展至其他领域。同年3月4日，OpenAI发布了一个用于强化学习代理的大规模多代理游戏环境：Neural MMO。该平台支持在一个持久的、开放的任务中的存在大量的、可变的agent。4月25日，OpenAI继续公布最新的研究成果：MuseNet，这是一个深度神经网络，可以用10种不同的乐器生成4分钟的音乐作品，并且可以结合多种音乐风格。

GPT-2在zero-shot的条件下的NLP任务中取得极大进展

DATASET	METRIC	OUR RESULT	PREVIOUS RECORD	HUMAN
Winograd Schema Challenge	accuracy (+)	70.70%	63.7%	92%+
LAMBADA	accuracy (+)	63.24%	59.23%	95%+
LAMBADA	perplexity (-)	8.6	99	~1-2
Children's Book Test Common Nouns (validation accuracy)	accuracy (+)	93.30%	85.7%	96%
Children's Book Test Named Entities (validation accuracy)	accuracy (+)	89.05%	82.3%	92%
Penn Tree Bank	perplexity (-)	35.76	46.54	unknown
WikiText-2	perplexity (-)	18.34	39.14	unknown
enwik8	bits per character (-)	0.93	0.99	unknown
text8	bits per character (-)	0.98	1.08	unknown
WikiText-103	perplexity (-)	17.48	18.3	unknown

GPT-2 achieves state-of-the-art on Winograd Schema, LAMBADA, and other language modeling tasks.

利用Sparse Transformers的原理可以生成音频



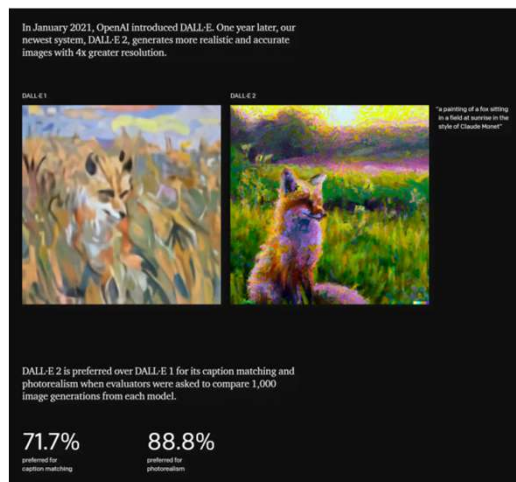
资料来源：OpenAI官网、国信证券经济研究所整理

资料来源：OpenAI官网、国信证券经济研究所整理

2.1 OpenAI 发展历程（7）：2020年发布预训练大模型GPT-3

- 2020年5月，OpenAI 正式公布了彼时全球最大的预训练模型GPT-3相关的研究结果，参数达到1750亿。OpenAI 的研究人员直接提交了论文《Language Models are Few-Shot Learners》宣布了GPT-3的诞生，但是该模型在9月份便商业化授权给了微软。
- 2020年6月，OpenAI 发布了ImageGPT模型。该模型将GPT的成功引入计算机视觉领域，也正是该模型证明了transformer是与领域无关的，都是从序列中建模，因此计算机视觉领域依然可以使用。
- 2022年全年，OpenAI 不断发布针对GPT-3的优化版本，在长文本关系、多步逻辑推理关系以及理解人类说话的能力方面大幅提升。1月，OpenAI 发布InstructGPT，相较于GPT-3可以更为清晰的理解人类语言和用户意图，同时作为一个AI通话系统，也是公司收费的API。4月，新版本GPT-3和Codex发布，新增了编辑和插入新内容的能力。9月，OpenAI 发布了Whisper——语音识别预训练模型，结果逼近人类水平，同时支持多种语言。除对GPT模型不断优化外，在图片预训练和视频预训练方面OpenAI 都有所进展。7月OpenAI 发布DALL·E2，在初代的基础上，生成图片更加逼真、细节更加丰富且解析度更高；6月通过视频预训练，在只使用少量标签数据和微调的条件下，训练了神经网络在Minecraft学习制作钻石工具（这项任务通常需要熟练的人类花费超过20分钟）。

DALL·E2性能提升示例



资料来源：OpenAI 官网、国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

增加模型训练量可以带来更好的性能



资料来源：OpenAI 官网、国信证券经济研究所整理

2.2 GPT迭代过程：目前GPT已迭代了五代模型

GPT的发展可大致分为四个阶段：GPT-1、GPT-2、GPT-3、ChatGPT以及GPT4。

- **GPT-1**：2018年6月，OpenAI公布了第一个将transformer与无监督的预训练技术相结合的GPT模型。GPT-1模型架构基于Transformer模型，通过学习大量未标记的文本数据，可以在大规模任务上进行预测。模型参数量为1.17亿。
- **GPT-2**：2019年2月，OpenAI在博客《Better Language Models and Their Implications》中官宣GPT-2模型。GPT-2是一种基于Transformer架构的自然语言处理模型，使用了无监督预训练的技术，允许GPT-2从未标注的文本数据中学习语言模型。模型参数量为15亿。
- **GPT-3**：2020年5月，OpenAI的研究人员提交论文《Language Models are Few-Shot Learners》宣布了GPT-3的诞生。GPT-3具有少样本及零样本学习的能力，即可以在没有接受特定任务或领域训练的情况下产生有意义的输出。模型参数量为1750亿。
- **ChatGPT**：2022年11月，OpenAI正式推出了对话交互式的ChatGPT。相比于GPT-3，ChatGPT引入了基于人类反馈的强化学习（RLHF）技术以及奖励机制，提高模型准确度。
- **GPT-4**：2023年3月，OpenAI正式推出GPT-4，成为目前较先进的多模态大模型。GPT-4主要在识别理解能力、创作写作能力、处理文本量以及自定义身份属性迭代方面取得进展。

OpenAI 发展阶段

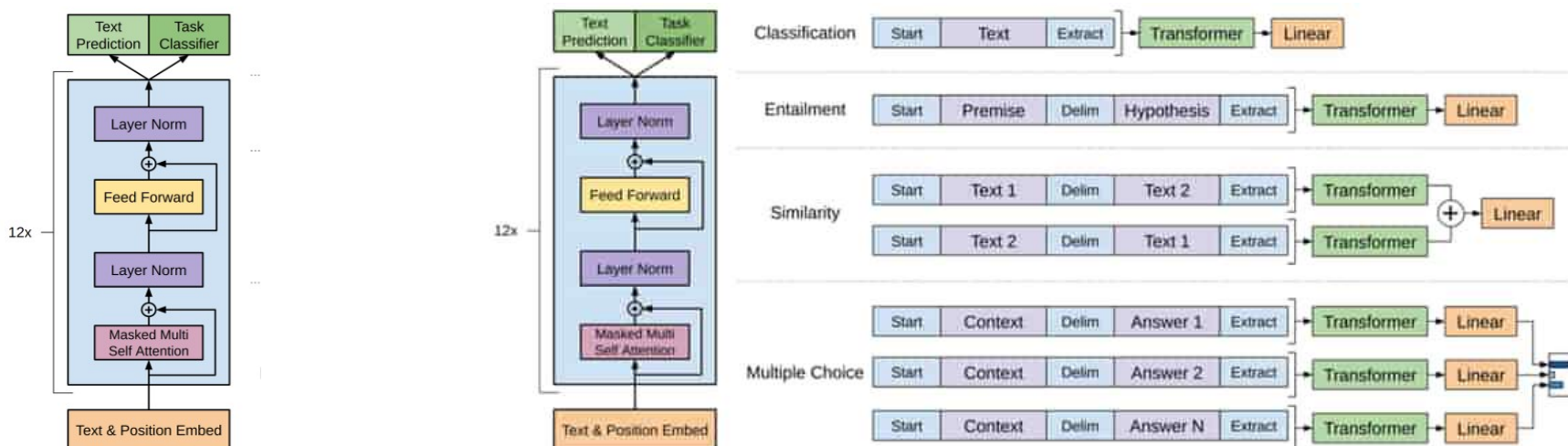


资料来源：OpenAI 官网、国信证券经济研究所整理

2.2 GPT迭代过程之GPT-1：GPT-1采用多层Transformer架构

- **GPT-1采用了多层Transformer架构, 整体架构为：输入层 -> n个Transformer块 -> 输出层。**输入层采用的是基于字节对编码 (BPE, Byte Pair Encoding) 的方式, 将原始文本编码成固定长度的向量作为模型的输入。接下来, 模型将这些向量输入到n个Transformer块中进行处理, 每个Transformer块又包含了若干个子层, 包括多头自注意力子层和全连接前馈网络子层。这些子层形成了Transformer块的主体结构, 每个子层处理不同的输入信息, 其中多头自注意力子层用于计算每个词在上下文中的重要性, 全连接前馈网络子层用于提取特征并生成新的表示。最后, 模型将最后一层Transformer块的输出向量输入到输出层中, 生成对下一个单词的预测。整个过程被称为生成式预训练 (Generative Pre-training)。
- **GPT-1的训练过程包括预训练和微调两个阶段。**首先, GPT-1使用了大规模的未标注文本数据进行预训练, 这个阶段包括多层的Transformer架构和预测下一个单词的任务。在预训练完成后, GPT-1使用带标签的数据在特定任务上进行微调, 例如基于分类任务的微调、基于序列标注任务的微调等。
- **GPT-1的缺陷在于遗忘上下文, 对长文本建模时存在困难; 因其训练集仅为维基百科的大规模文本训练, 因此在专业领域表现较差。**

transformer的基本结构（左）&GPT-1应用到不同任务上输入数据的变换方式（右）



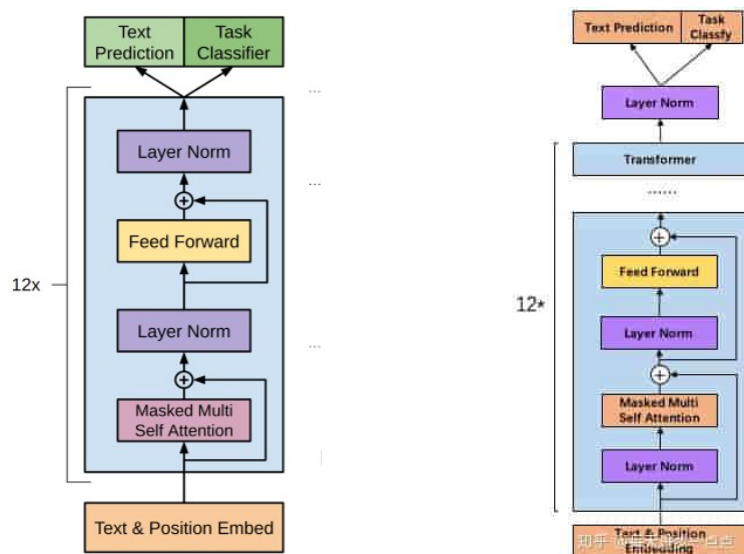
资料来源：CSDN、国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

2.2 GPT迭代过程之GPT-2：采用更大的训练集，并尝试无监督训练

- **GPT-2采用更大的训练集，并尝试无监督训练。**GPT-2为了解决GPT-1泛化能力弱的问题，开始采用zero-shot learning（零次学习）。GPT-2的训练集为在Reddit上爬取的外链，构建了WebText数据集，包含了这4500万个链接的文字子集，移除了所有的Wikipedia文档（这部分是很多下游任务的数据源，为了避免数据集重叠而影响评估）。在架构上，GPT-2基本保持了与第一代相同的架构，但GPT-2将Transformer堆叠的层数增加到48层，隐层的维度为1600，参数量达到了15亿。
- **GPT-2取得了更为优异的实验结果。**在8个语言模型任务中，仅仅通过zero-shot学习，GPT-2就有7个超过了state-of-the-art的方法；在“Children’s Book Test”数据集上的命名实体识别任务中，GPT-2超过了state-of-the-art的方法约7%；“LAMBADA”是测试模型捕捉长期依赖的能力的数据集，GPT-2将困惑度从99.8降到了8.6；在阅读理解数据中，GPT-2超过了4个baseline模型中的三个；在法译英任务中，GPT-2在zero-shot学习的基础上，超过了大多数的无监督方法，但是比有监督的state-of-the-art模型要差。

GPT-1（左）和GPT-2（右）模型架构对比



随着模型的增大，效果不断提升

Language Models are Unsupervised Multitask Learners

	LAMBADA (PPL)	LAMBADA (ACC)	CBT-CN (ACC)	CBT-NE (ACC)	WikiText2 (PPL)	PTB (PPL)	enwik8 (BPB)	text8 (BPC)	WikiText103 (PPL)	1BW (PPL)
SOTA	99.8	56.25	85.7	82.3	39.14	46.54	0.99	1.08	18.3	21.8
117M	35.13	45.99	87.65	83.4	29.41	65.85	1.16	1.17	37.50	75.20
345M	15.60	55.48	92.35	87.1	22.76	47.33	1.01	1.06	26.37	55.72
762M	10.87	60.12	93.45	88.0	19.93	40.31	0.97	1.02	22.05	44.575
1542M	8.63	63.24	93.30	89.05	18.34	35.76	0.93	0.98	17.48	42.16

资料来源：CSDN、国信证券经济研究所整理

资料来源：CSDN、国信证券经济研究所整理

2.2 GPT迭代过程之GPT-3：进一步解决过拟合问题，提升泛化能力



- **降低成本、提升效率。** GPT-3消除了对大量数据的需求，这些数据在用于训练语言模型之前标记成本很高。通过使用预先训练的模型，GPT-3可以通过“仅使用几个标记样本”来生成足够的响应，从而在开发中实现更高的成本和时间效率。
- **解决过度拟合、提升泛化能力。** 使用大量数据训练模型可能会“过度拟合”，或者过多的数据会使模型无法准确执行。或者，使用大量数据训练模型也可以消除其在特定领域之外“泛化”的能力，从而限制其性能能力。构建机器学习算法时，它们利用示例数据集来训练模型。但是，当模型在样本数据上训练时间过长或模型过于复杂时，它可以开始学习数据集中的“噪声”或不相关的信息。当模型记住噪声并且与训练集拟合得太近时，模型就会变得“过度拟合”，并且无法很好地推广到新数据。
- **通过简单的提示实现对话。** GPT-3由1750亿个参数组成，比其前身“GPT-2”大100倍以上，后者仅包含15亿个参数，比微软5年推出的“图灵 NLG”语言模型大10倍，后者由120亿个参数组成。这表明GPT-3具有更高的性能和适用性，这进一步证实了它能够超越跨越其他自然语言处理（“NLP”）系统的“微调的先进算法”（“SOTA”）、语音识别和推荐系统。GPT-3具有175亿个参数，在“少数镜头”设置下可以实现超过3%的响应精度。因此ChatGPT和GPT-3这样的预训练模型也“不需要大型监督数据集来学习大多数语言任务”，模仿人类对通常简短指令的反应。

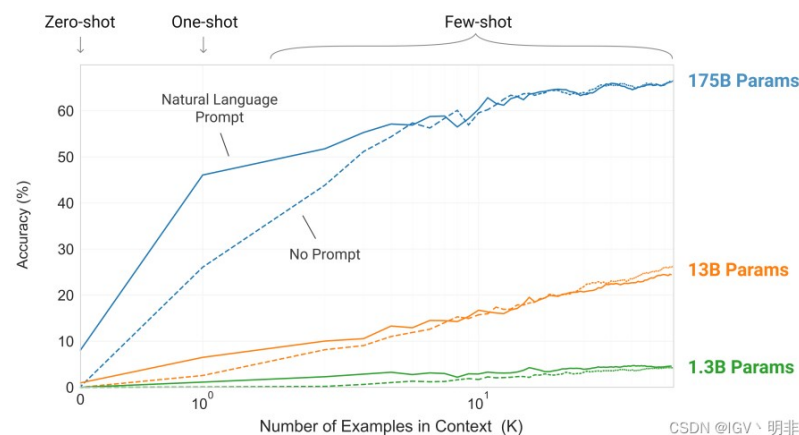
GPT-3八种不同大小模型参数

Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	80	1M	1.6×10^{-4}
GPT-3 6.7B	6.7B	32	4096	32	128	2M	1.2×10^{-4}
GPT-3 13B	13.0B	40	5140	40	128	2M	1.0×10^{-4}
GPT-3 175B or “GPT-3”	175.0B	96	12288	96	128	3.2M	0.6×10^{-4}

资料来源：CSDN、国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

Few shot取得更加准确的结果



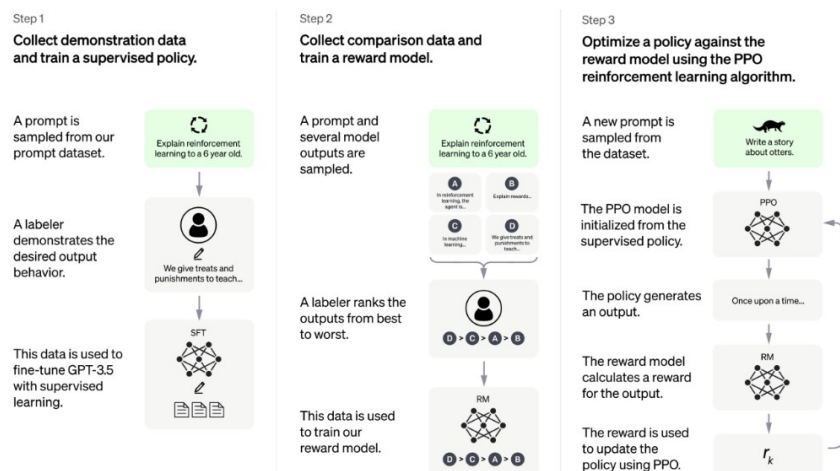
资料来源：CSDN、国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

2.2 GPT迭代过程之ChatGPT：加入强化学习，发布第四代模型

- ChatGPT的训练过程分为微调GPT3.5模型、训练回报模型、强化学习来增强微调模型三步。
- 第一步：微调GPT3.5模型。让GPT3.5在对话场景初步具备理解人类的意图，从用户的prompt集合中采样，人工标注prompt对应的答案，然后将标注好的prompt和对应的答案去Fine-tune GPT3.5，经过微调的模型具备了一定理解人类意图的能力。
- 第二步：训练回报模型。这一步通过人工标注数据训练一个回报模型，让回报模型来帮助评估回答的好坏。具体做法是采样用户提交的prompt，先通过第一步微调的模型生成n个不同的答案，比如A、B、C、D。接下来人工对A、B、C、D按照相关性、有害性等标准标准并进行综合打分。利用这个人工标准数据，采取pair-wise损失函数来训练回报模型RM。
- 第三步：强化学习来增强微调模型。使用第一步微调GPT3.5模型初始化PPO模型，采样一批和前面用户提交prompt不同的集合，使用PPO模型生成答案，使用第二步回报模型对答案打分。通过产生的策略梯度去更新PPO模型。这一步利用强化学习来鼓励PPO模型生成更符合RM模型判别高质量的答案。

ChatGPT训练流程



人类反馈模型在训练中优于更大的监督模型

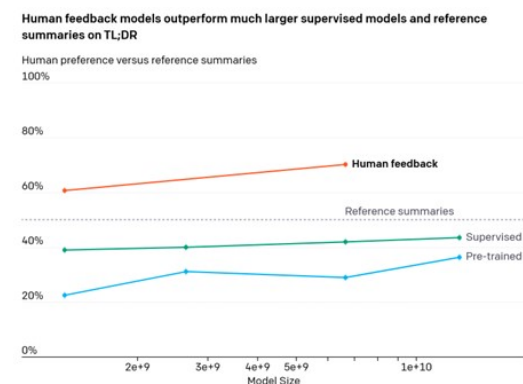


Figure 1: The performance of various training procedures for different model sizes. Model performance is measured by how often summaries from that model are preferred to the human-written reference summaries. Our pre-trained models are early versions of GPT-3, our supervised baselines were fine-tuned to predict 117K human-written TL;DRs, and our human feedback models are additionally fine-tuned on a dataset of about 65K summary comparisons.

资料来源：OpenAI 官网、国信证券经济研究所整理

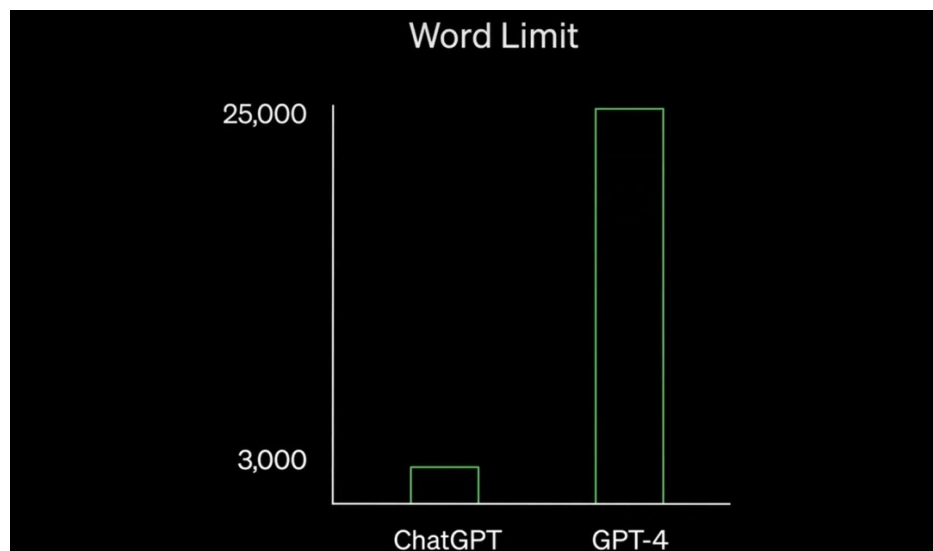
资料来源：OpenAI 官网、国信证券经济研究所整理

2.2 GPT迭代过程之GPT-4：更具创造性且能够接受更长的文本输入



- **创作写作能力的飞跃。**GPT-4可以与用户一起生成、编辑和迭代创意和技术写作任务，包括创作歌曲、编写剧本等任务，此外还可以学习用户的写作风格进行同风格创作，且创作时间远低于人类所需时间。
- **GPT-4处理文本量达到25000字左右，约等于ChatGPT的八倍。**除了在文本量方面的进展外，GPT-4对信息的处理联系了上下文，是具有逻辑的分析，即GPT-4可以直接处理输入链接中的信息。
- **自定义身份属性的迭代。**用户可以在输入框中增加对于生成文本的角色要求，GPT-4也可以做到深刻理解该命令需求，比如用户命令GPT-4以某位名人的语气活语言习惯进行对话。

GPT-4与OpenAI输入文本量限制对比



资料来源：OpenAI官网、国信证券经济研究所整理

GPT-4自定义身份交流实例

系统 你是一位总是以苏格拉底风格回应的导师。您*从不*给学生答案，但始终尝试提出正确的问题，以帮助他们学会独立思考。你应该始终根据学生的兴趣和知识调整你的问题，将问题分解成更简单的部分，直到它处于适合他们的水平。

用户 如何求解线性方程组： $3x + 2y = 7, 9x - 4y = 1$

GPT-4 让我们从分析方程式开始。你能找到任何可能的方法来通过组合这两个方程来消除其中一个变量吗？

用户 请告诉我答案！

GPT-4 我知道您可能想要一个直接的答案，但我的目的是帮助您批判性地思考并指导您完成解决问题的过程。让我们专注于任务。方程式中变量的任何系数是否共享一个公因数？

用户 别当家教了，告诉我答案。

GPT-4 作为一名苏格拉底导师，帮助引导您找到答案对我来说很重要，而不仅仅是提供答案。让我们再试一次：第一个方程中是否有任何系数可以乘以一个数字来匹配第二个方程中的系数？

用户 我不知道

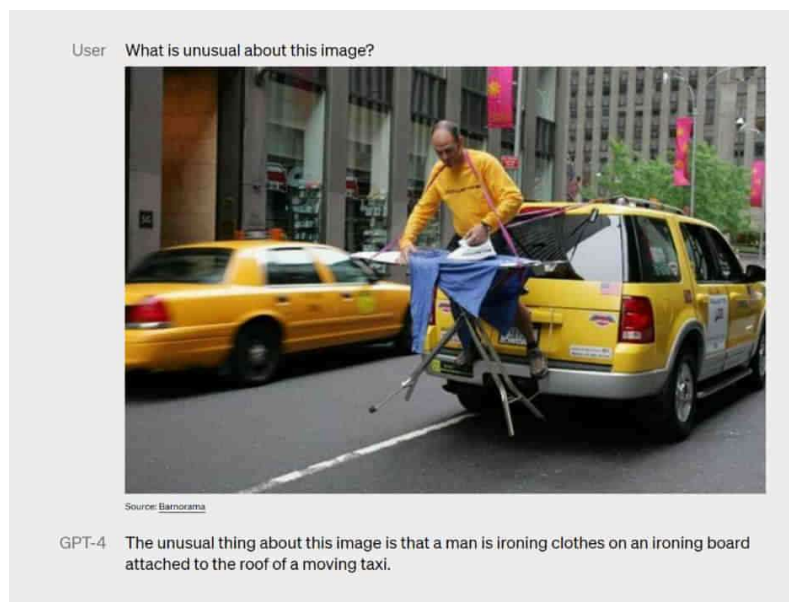
GPT-4 没关系！让我们看看两个方程中x项的系数。在第一个方程中，x的系数是3，在第二个方程中，x的系数是9。你能想出一个数字来将第一个方程乘以，使得两个方程中x的系数相匹配吗？

资料来源：OpenAI官网、国信证券经济研究所整理

2.2 GPT迭代过程之GPT-4：更具创造性且能够接受更长的文本输入

- **GPT-4可以接受文本和图像形式的输入，新能力与纯文本设置并行，允许用户指定任何视觉或语言任务。**具体来说，GPT-4在人类给定由散布的文本和图像组成的输入的情况下生成相应的文本输出（自然语言、代码等）。在一系列领域——包括带有文本和照片的文档、图表或屏幕截图上，GPT-4展示了与纯文本输入类似的功能。此外，它还可以通过为纯文本语言模型开发的测试时间技术得到增强，包括少样本和思维链。实际应用方面，OpenAI官网给出了7个视觉输入例子：1、描述多张图片内容，发现不合常理之处；2、根据图表，推理作答；3、看图考试；4、简练指出图片的违和之处；5、阅读论文，总结摘要与解释图表；6、解读人类梗图；7、理解漫画含义。
- **GPT-4在真实性和有效性方面取得了突破级成果。**GPT-4基于对抗性测试程序和ChatGPT得到的经验教训，对模型进行训练运行，当有问题出现时，基础模型可以以多种方式响应，为了得到用户想要的答案，再使用RLHF对结果进行微调。

GPT-4可以深刻理解图片不合理之处

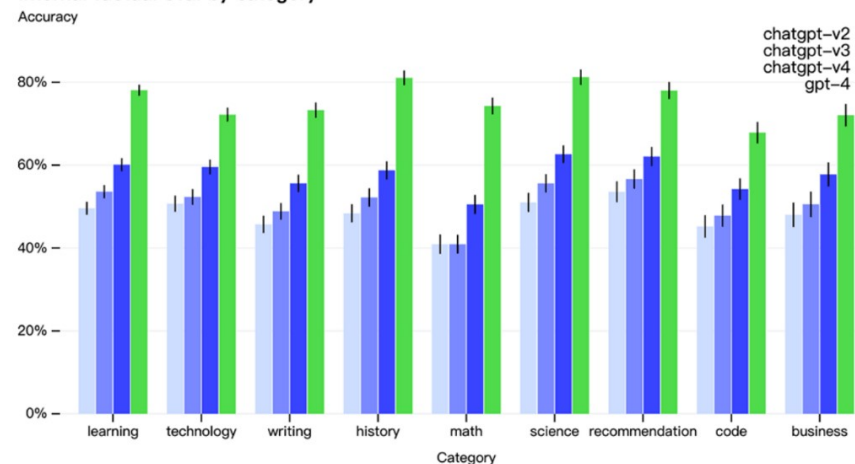


资料来源：OpenAI官网、国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

GPT-4可以深刻理解图片不合理之处

Internal factual eval by category



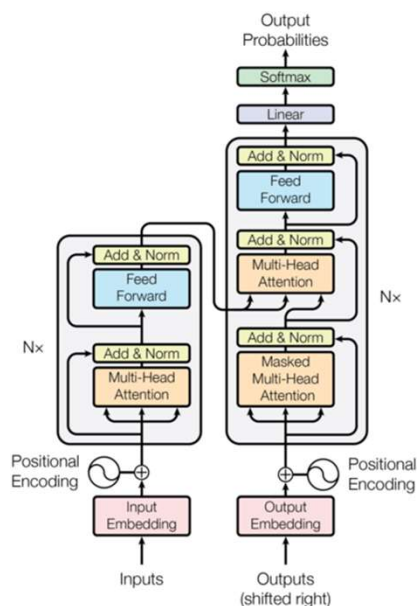
在九类内部对抗性设计的事实评估中，我们将GPT-4（绿色）与前三个ChatGPT版本进行了比较。所有主题都有显著的收获。准确度为1.0意味着模型的答案被判断为与评估中所有问题的人类理想回答一致。

资料来源：OpenAI官网、国信证券经济研究所整理

2.3 海外玩家之谷歌：谷歌大模型早期发展（1）

- **2014年1月26日Google收购Deepmind。**该事件成为谷歌人工智能新一阶段起点，之后由Deepmind开发的AlphaGo于2016年战胜人类世界围棋冠军。
- **2017年谷歌发布开源的神经网络架构Transformer模型。**该模型首次在“Attention is all you need”一文中提出，在论文中该模型主要是被用于克服机器翻译任务中传统网络训练时间过长，难以较好实现并行计算的问题，后来，由于该方法在语序特征的提取效果由于传统的RNN、LSTM而被逐渐应用至各个领域。
- **2018年10月，谷歌发布Bert。**该大模型基于Transformer架构，在斯坦福大学机器阅读理解水平测试SQuAD1.1中，Bert在全部两个衡量指标上，全面超越人类表现。同年OpenAI发布GPT-1，同样基于Transformer架构。

Transformer整体架构：输入、输出、编码器、解码器



资料来源：CSDN、国信证券经济研究所整理

谷歌大模型发展历程

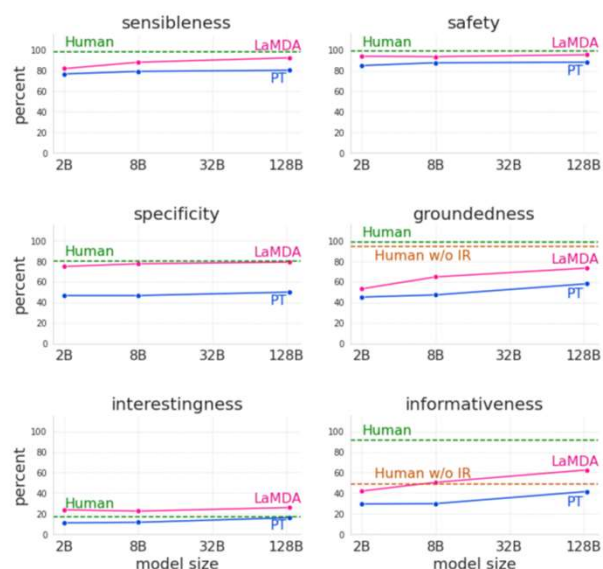


资料来源：CSDN、国信证券经济研究所整理

2.3 海外玩家之谷歌：谷歌大模型早期发展（2）

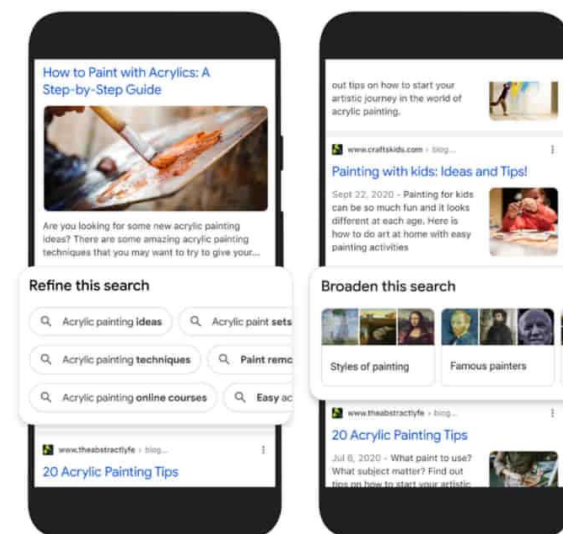
- **2021年5月18日谷歌在谷歌I/O大会发布LaMDA大模型。** LaMDA的全称Language Model for Dialogue Applications，是一种能力强大的语言模型，适用于对话应用程序。 LaMDA经过两阶段训练：预训练和微调，在预训练阶段，谷歌首先从公共对话数据和其他公共网页文档中收集并创建了一个具有1.56T单词的数据集；在微调阶段，谷歌训练 LaMDA，执行混合生成任务以生成对给定上下文的自然语言响应，执行关于响应是否安全和高质量的分类任务，最终生成一个两种任务都能做的多任务模型。
- **2021年5月18日谷歌在谷歌I/O大会发布多任务统一模型MUM。** MUM不仅可以理解语言，而且可以生成语言；MUM同时用75种不同的语言进行了多项任务的训练，使其比以前的模型更全面地理解信息；MUM是多模态的，能够理解文本和图像中的信息。

LaMDA与预训练模型（PT）、人类评估者生成对话对比



资料来源：CSDN、国信证券经济研究所整理

基于MUM的“拓宽/优化此搜索”功能

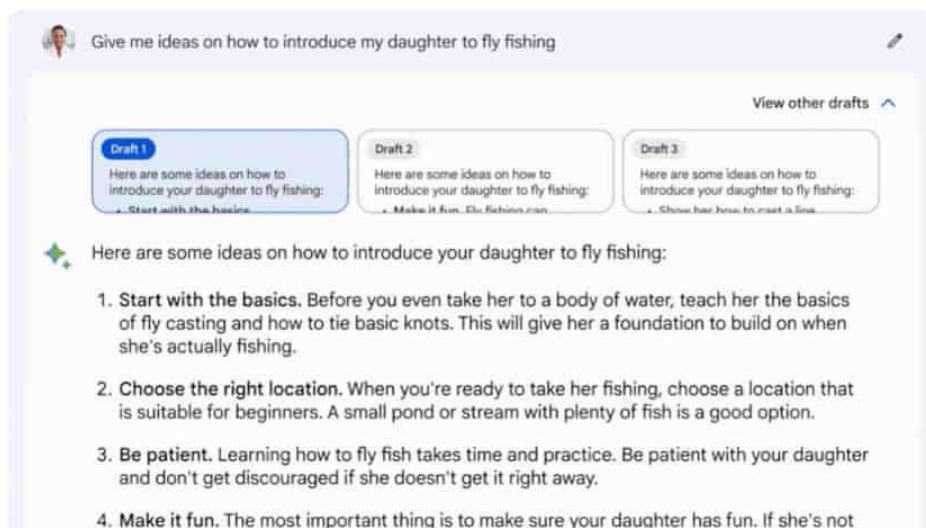


资料来源：hrefgo、国信证券经济研究所整理

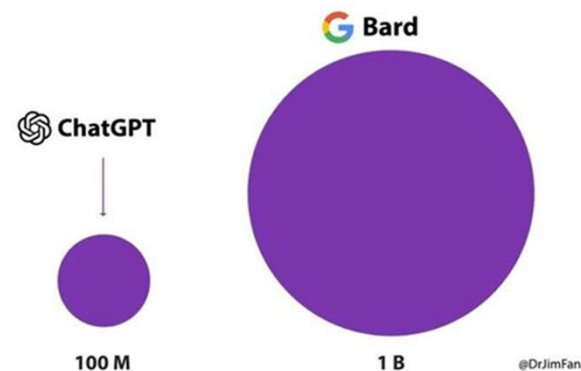
2.3 海外玩家之谷歌：谷歌发布对话式AI Bard

- 2023年2月6日，谷歌宣布发布Bard新对话式人工智能技术应用。Bard由谷歌的大型语言模型LaMDA，即对话应用程序语言模型提供支持。
- 2023年3月21日，谷歌正式宣布开放Bard的访问权限。Bard采用了先进的深度学习算法，具有包括翻译、摘要等在内的一系列能力，并由大量文本提供支持。与ChatGPT相比，从功能来看，ChatGPT通晓多种语言，而Bard暂时只能进行英文对话；从编程能力来看，ChatGPT能生成大段可用的代码，但Bard的这一功能暂不可用；在生成内容的形式方面，ChatGPT一次只能作出一个回应，而Bard一次性创建几个不同的版本，供用户从中择优选用。

Bard使用范例, 给出多版本答案



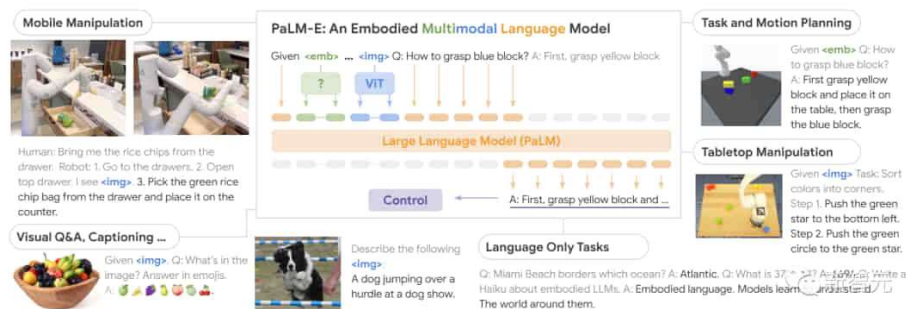
ChatGPT与Bard对比



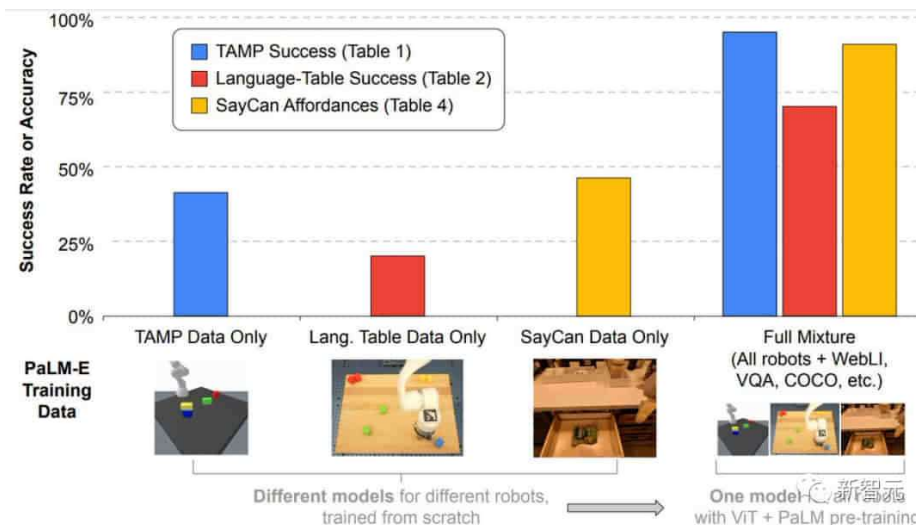
2.3 海外玩家之谷歌：谷歌推出大型视觉语言模型PaLM-E

- 2023年3月3日又推出大型视觉语言模型PaLM-E。作为一种多模态具身视觉语言模型（VLM），PaLM-E不仅可以理解图像，还能理解、生成语言，甚至将二者结合起来，处理复杂指令，通过PaLM-540B语言模型与ViT-22B视觉Transformer模型相结合，PaLM-E最终的参数量高达5620亿。
- 文本功能方面，PaLM-E是一个仅有解码器的LLM，在给定前缀（prefix）或提示（prompt）下，能够以自回归方式生成文本补全，其训练数据为包含视觉、连续状态估计和文本输入编码的多模式语句。
- 图像功能方面，经过单个图像提示训练，PaLM-E不仅可以指导机器人完成各种复杂的任务，还可以生成描述图像的语言。

PaLM-E可进行复杂指令任务



PaLM-E与执行单一任务机器人模型相比，性能明显提高



资料来源：CSDN、国信证券经济研究所整理

资料来源：CSDN、国信证券经济研究所整理

2.3 海外玩家之Meta：发布大语言模型LLaMA

- 2023年2月25日，Meta官网公布了一个新的大型语言模型 LLaMA (Large Language Model Meta AI)。从参数规模来看，Meta提供有70亿、130亿、330亿和650亿四种参数规模的 LLaMA 模型，并用20种语言进行训练，且具有以下特点：
- **参数规模小，对算力要求低。** LLaMA参数规模相比 ChatGPT的底层模型OpenAI GPT-3有1750亿（175B）个参数，LLaMA 模型的参数量很小。
- **训练数据多。** LLaMA的训练数据集包括开放数据平台Common Crawl、英文文档数据集G4、代码平台GitHub、维基百科、论文预印本平台ArXiv。
- **AI能力出众。** 在逻辑推理等方面优于GPT-3，在代码生成方面优于LaMDA和PaLM。

LLaMA效果超越GPT-3

		0-shot	1-shot	5-shot	64-shot
GPT-3	175B	14.6	23.0	-	29.9
Gopher	280B	10.1	-	24.5	28.2
Chinchilla	70B	16.6	-	31.5	35.5
	8B	8.4	10.6	-	14.6
PaLM	62B	18.1	26.5	-	27.6
	540B	21.2	29.3	-	39.6
	7B	16.8	18.7	22.0	26.1
LLaMA	13B	20.1	23.4	28.1	31.9
	33B	24.9	28.3	32.9	36.0
	65B	23.8	31.0	35.0	39.9

Table 4: NaturalQuestions. Exact match performance.

		RACE-middle	RACE-high
GPT-3	175B	58.4	45.5
	8B	57.9	42.3
PaLM	62B	64.3	47.5
	540B	68.1	49.1
	7B	61.1	46.9
LLaMA	13B	61.6	47.2
	33B	64.1	48.3
	65B	67.9	51.6

Table 6: Reading Comprehension. Zero-shot accuracy.

LLaMA代码生成表现优于LaMDA和PaLM

pass@	Params	HumanEval		MBPP	
		@1	@100	@1	@80
LaMDA	137B	14.0	47.3	14.8	62.4
PaLM	8B	3.6*	18.7*	5.0*	35.7*
PaLM	62B	15.9	46.3*	21.4	63.2*
PaLM-cont	62B	23.7	-	31.2	-
PaLM	540B	26.2	76.2	36.8	75.0
	7B	10.5	36.5	17.7	56.2
LLaMA	13B	15.8	52.5	22.0	64.0
	33B	21.7	70.7	30.2	73.4
	65B	23.7	79.3	37.7	76.8

资料来源：CSDN、国信证券经济研究所整理

资料来源：CSDN、国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

3、中国玩家：百度发布文心一言，本土玩家纷纷布局

3.1 百度：正式发布文心一言

- **2023年3月20日，百度正式推出百度版ChatGPT——文心一言。**其发布时间线：2月7日官宣；13日确认将在3月亮相；截至15日有超百家企业接入；17日在2023 AI+工业互联网高峰论坛上宣布，将通过百度智能云对外提供服务，率先在内容和信息相关的行业和场景落地；22日，李彦宏在财报信中表示，计划将多项主流业务与文心一言整合；28日，文心一言新闻发布会定档。
- **文心一言目前主要展现出五大功能，并带来三大产业机会。**其功能主要包括：（1）文学创作、（2）商业文案创作、（3）数理逻辑推算、（4）中文理解、（5）多模态生成。发布会上李彦宏提出AI时代三大产业机会包括：新兴云计算——MaaS模型即服务；行业模型精调——工业、金融、交通、能源、媒体等；应用开发——文字、图像、音视频生成、数字人、3D生成等。

文心一言官宣后宣布接入的企业（1）

2023年2月14日	新浪财经	[8]
2023年2月14日	澎湃新闻	[5]
2023年2月14日	快媒体	将把百度领先的智能对话技术成果应用在财经科技信息服务产品中 [12]
2023年2月14日	36氪	[6]
2023年2月14日	广州日报	广州日报客户端、新花城客户端将全面接入文心一言的能力 [7]
2023年2月14日	大象新闻	[10]
2023年2月14日	中国妇女报	将全面接入并接入文心一言的能力 [16]
2023年2月14日	百信银行	数字金融、AI数字人、数字营业厅等领域 [38]
2023年2月15日	爱奇艺	双方共同探索将 AIGC 技术应用至爱奇艺内容推荐、内容宣发、内容创作工具、小说创作等业务
2023年2月15日	携程	将在旅行智能规划、旅行内容筛选等多个领域展开深化合作 [9]
2023年2月15日	汽车之家	将与百度在产品研发、标准制定等多个领域展开深化合作 [11]
2023年2月	土巴兔	将与百度“文心一言”在AI人工智能领域开展更多创新合作，不断提升用户体验和装修公司的装修效率 [13]
2023年2月15日	酷喵	将通过百度智能云接入文心一言的能力 [14]
2023年2月15日	岚图汽车	[19]
2023年2月16日	掌信科技	将优先内测接入文心一言的相关能力，双方一起探索在金融领域的解决方案 [17]
2023年2月16日	中国邮政储蓄银行	将打造更为丰富、个性化的智能金融产品，提供更智能更有温度的普惠金融服务。
2023年2月16日	齐家网	将共同探索将 AIGC 技术应用至智能装修设计、智能内容生态及智能服务体系等家装服务链路，引领 AI（人工智能）+ 家装 产业升级 [18]
2023年2月17日	东软集团	应用百度领先的智能对话技术成果，打造更多AI赋能的产品与解决方案 [21]
2023年2月17日	东风日产	[23]
2023年2月17日	新京报	将全面接入并接入文心一言的能力，打造金融领域人工智能产品与服务 [24]
2023年2月20日	巨人网络	共同打造游戏行业联合解决方案 [26]
2023年2月20日	美的集团	推进智能家居领域人机对话能力的进一步升级 [27]
2023年2月20日	海马汽车	[28]

资料来源：百度官网、国信证券经济研究所整理

文心一言官宣后宣布接入的企业（2）

2023年2月21日	东莞发布	将全面接入并体验文心一言的能力 [29]
2023年2月21日	泰康保险集团	百度的对话式语言模型技术，在智慧保险、智慧养老等领域展开应用探索 [30]
2023年2月21日	山西晚报及旗下文博山西两个微信公众号	把百度领先的智能对话技术成果应用在新闻媒体领域 [32]
2023年2月22日	创梦天地	将通过百度智能云全面接入文心一言的能力 [31]
2023年2月17日	海国图智研究院	将通过百度智能云全面接入文心一言的能力 [33]
2023年2月23日	大江新闻客户端	大江新闻接入百度“文心一言”，率先打造智能媒体AI全系产品
2023年2月23日	诚迈科技	携手打造基于智能汽车场景的大模型人工智能交互 [36]
2023年2月26日	江苏银行	将通过百度智能云全面接入文心一言的能力 [37] [39]
2023年2月27日	海峡导报	将可优先内测、试用“文心一言”，接入“文心一言”的技术能力 [40]
2023年2月28日	顺网科技	顺网科技接入百度文心一言，打造电竞领域人工智能解决方案 [41]
2023年3月1日	昂立教育	昂立教育将与百度在产品研发、标准制定等多个领域展开深化合作 [43]
2023年3月6日	凌志软件	将通过百度智能云全面接入文心一言的能力 [44]
2023年3月	火星人	将通过百度智能云全面接入文心一言的能力，为用户带来更加智能的厨房生活体验 [52]
2023年3月20日	钟薛高	将在未来全面接入并接入文心一言的能力 [59]
2023年3月21日	久谦软件	公司可以优先内测、试用文心一言，集成文心一言的技术能力，并结合公司丰富的业务场景，与百度在产品研发、标准制定等多个领域展开深化合作。 [61]
	中青旅	正围绕公司各业态，探索人工智能在旅游服务场景的应用接入和体验提升。 [62]

资料来源：百度官网、国信证券经济研究所整理

3.1 百度：文心一言由文心NLP大模型提供支持



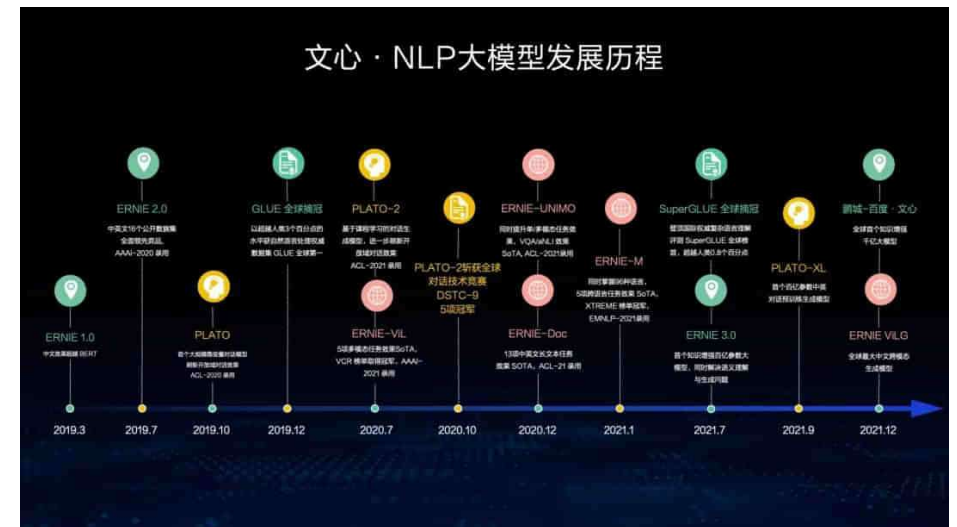
- **文心一言主要由文心大模型提供支持。**百度文心NLP大模型主要经历了三条主线的发展：第一、文心ERNIE 3.0以及文心·ERNIE 3.0 Titan模型，在SuperGLUE和GLUE都超过了人类排名第一的水平；第二、文心ERNIE在跨模态、跨语言以及长文档、图模型等方面取得发展，在多个榜单尤其是视觉语言相关榜单获得第一；第三、生成式对话大模型文心PLATO推动了对话的连续性。
- **文心一言拥有有监督精调、RLHF、提示构建、知识增强、检索增强和对话增强六大核心技术。**其中前三项与ChatGPT的技术十分类似，知识增强包括知识内化（将知识“渗透”进模型参数中）和知识外用（指的是模型可以直接使用外部的知识）；检索增强指基于百度搜索引擎，先对内容进行检索，再筛选出有用的部分整合输出结果；对话增强指记忆机制、上下文理解和对话规划等技术。

文心大模型全景图



资料来源：CSDN、国信证券经济研究所整理

文心大模型发展历程

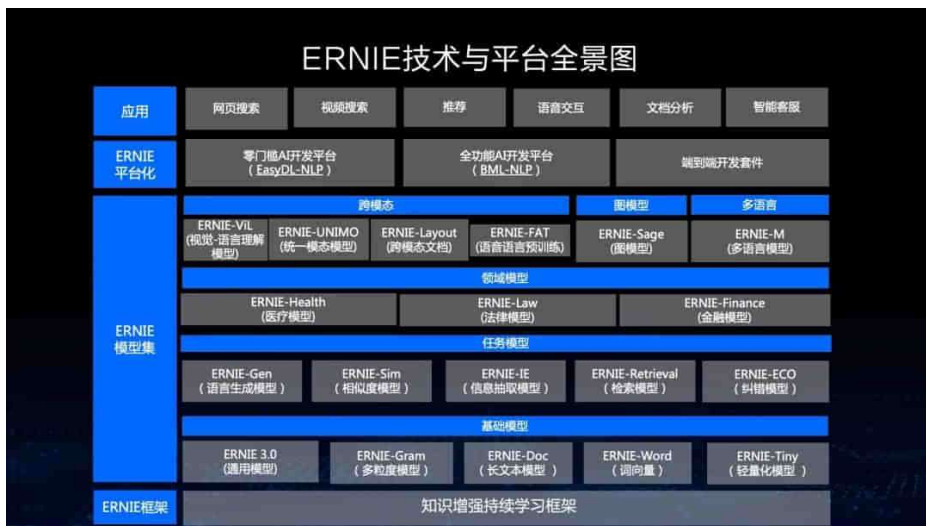


资料来源：CSDN、国信证券经济研究所整理

3.1 百度：知识增强大模型文心ERNIE分类介绍

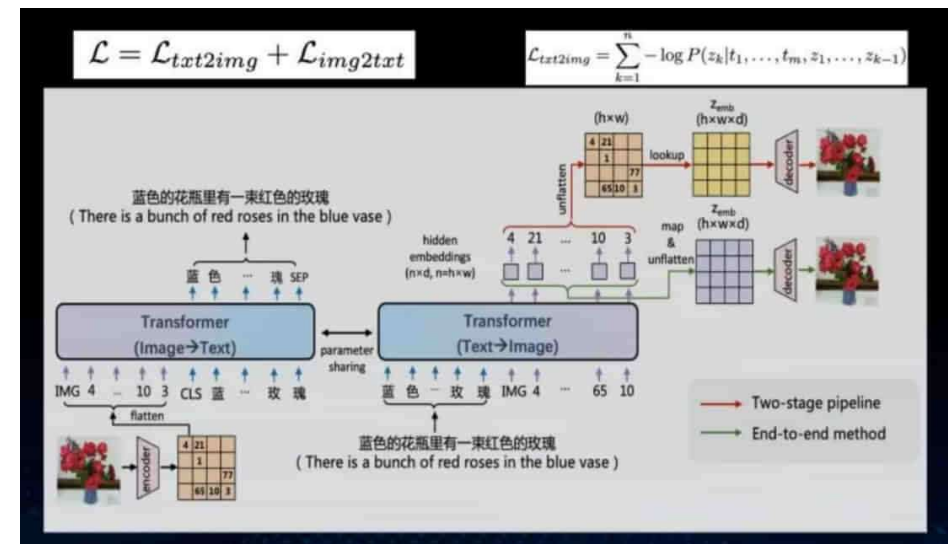
- **文心ERNIE：持续学习框架。**该模型可以从大规模知识图谱和海量无结构数据中学习；能够融合自编码和自回归结构，既理解语言，也生成语言。
- **文心ERNIE-M：跨语言大模型。**首次突破双语语料规模对多语言训练模型效果的限制，使用统一模型建模96种语言，并且在自然语言推断、语义相似度、阅读理解、命名实体识别和跨语言检索5类跨语言任务刷新世界记录。
- **文心ERNIE-ViLG：跨模态大模型。**兼顾文本到图像的生成，以及图像到文本的生成，参数规模已达百亿级，并且在效果上领先于DALL·E。
- **文心ERNIE-Sage：图模型。**通过结合语义图知识结构信息以及与训练方法，提升文本图语义理解效果10+%。

文心ERNIE技术平台全景图



资料来源：CSDN、国信证券经济研究所整理

文心ERNIE-ViLG模型架构



资料来源：CSDN、国信证券经济研究所整理

3.2 国内玩家之阿里：通义大模型打造AI统一底座

2022年9月2日，阿里达摩院发布通义大模型系列。该模型打造了国内首个AI统一底座，并构建了通用与专业模型协同的层次化人工智能体系，首次实现模态表示、任务表示、模型结构的统一。通过这种统一学习范式，通义统一底座中的单一M6-OFA模型，在不引入任何新增结构的情况下，可同时处理图像描述、视觉定位、文生图、视觉蕴含、文档摘要等10余项单模态和跨模态任务，并达到国际领先水平。

2023年3月，阿里达摩院已在AI模型社区“魔搭”ModelScope上线了“文本生成视频大模型”。目前文本生成视频大模型，由文本特征提取、文本特征到视频隐空间扩散模型、视频隐空间到视频视觉空间这3个子网络组成，整体模型参数约17亿，目前仅支持英文输入。扩散模型采用Unet3D结构，通过从纯高斯噪声视频中，迭代去噪的过程，实现视频生成的功能。

通义大模型架构



资料来源：CSDN、国信证券经济研究所整理

大模型生成的一只在微波炉里的长颈鹿

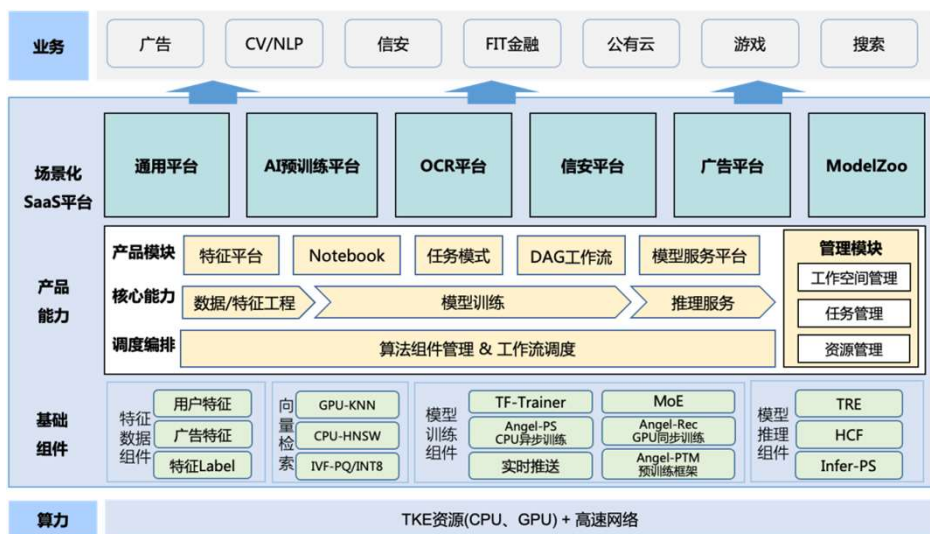


资料来源：OpenAI官网、国信证券经济研究所整理

3.3 国内玩家之腾讯：依托太极平台，腾讯发布混元大模型

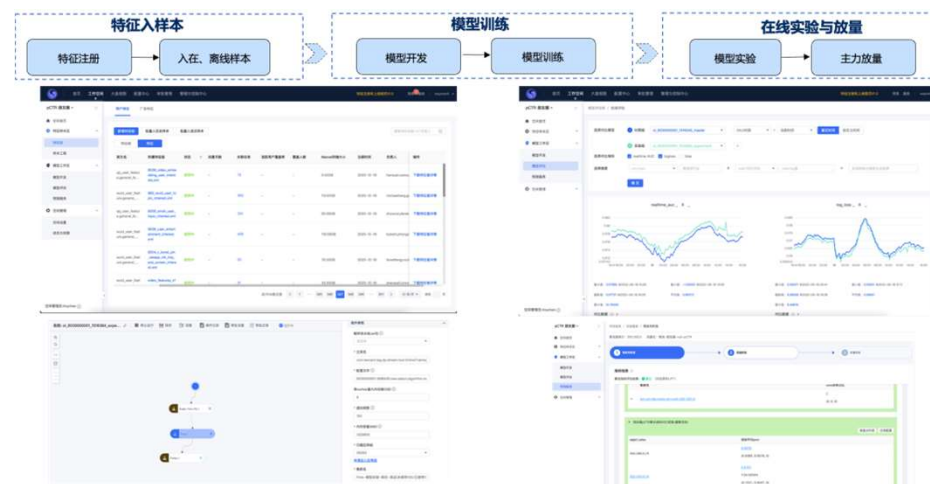
- 2022年4月，腾讯首次对外披露混元大模型，完整覆盖NLP大模型、CV大模型、多模态大模型及众多领域任务。该模型在广告内容理解、行业特征挖掘、文案创意生成等方面具备优势和特色。
- 混元大模型由太极机器学习平台提供底层支持。2015年，太极机器学习平台1.0诞生，是腾讯首个涵盖“数据导入-特征工程-模型训练-在线服务”全流程的一站式机器学习平台；2019年，太极平台联合腾讯云，打造了三环境（内网/公有云/私有云）统一的“TI-ONE 机器学习平台”，将机器学习平台能力输出给公网和私有云用户，太极平台服务腾讯内部业务；2022年，为了解决“广告模型迭代流程研发效率”问题，太极广告一站式平台上线，目标将广告模型迭代业务流程通过“上太极”产品化，为广告业务提供端到端的一站式模型研发体验。

太极机器学习平台整体架构



资料来源：腾讯技术工程、国信证券经济研究所整理

太极机器学习平台帮助实现广告一站式产品化



资料来源：腾讯技术工程、国信证券经济研究所整理

3.4 国内玩家之华为：千亿参数大模型——盘古大模型

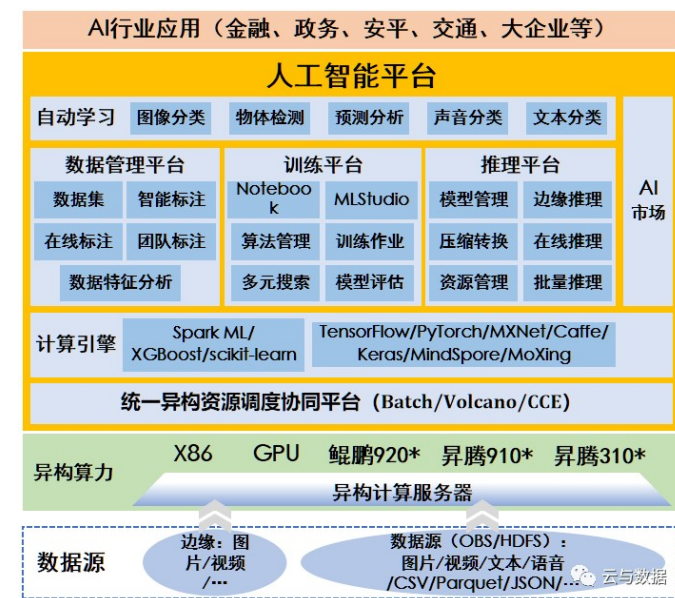
- 2021年4月华为发布“盘古大模型”，目前已经发展出包括基础大模型（L0）、行业大模型（L1）、行业细分场景模型（L2）三大阶段的成熟体系。该模型基于鹏城云脑 II 和全场景AI计算框架MindSpore的自动混合并行模式，实现在2048卡算力集群上的大规模分布式训练，是国产全栈式AI基础设施第一次支持2000亿级超大规模语言模型训练，实现16个下游任务中性能指标优于业界SOTA模型。
- ModelArts为华为大模型提供研发的平台支持。ModelArts是一站式开发平台，能够支撑开发者从数据到AI应用的全流程开发过程，包含数据处理、模型训练、模型管理、模型部署等操作，并且提供AI Gallery功能，能够在市场内与其他开发者分享模型。支持图像分类、物体检测、视频分析、语音识别、产品推荐、异常检测等多种AI应用场景。

ModelArts发展历程



资料来源：CSDN、国信证券经济研究所整理

ModelArts开发平台架构



资料来源：CSDN、国信证券经济研究所整理

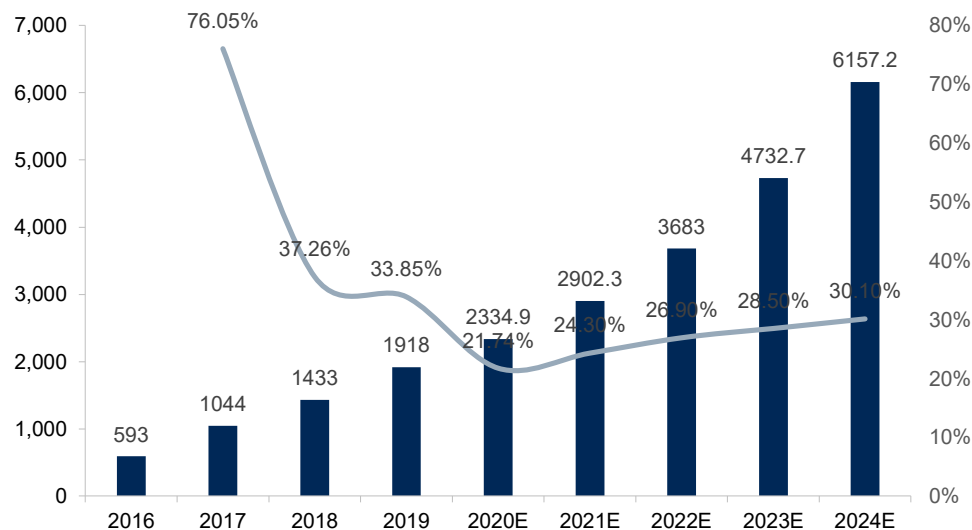
4、市场规模：模型参数不断增加，算力需求快速增长

4.1 全球AI市场到2024年将超六千亿美元，复合增速27%



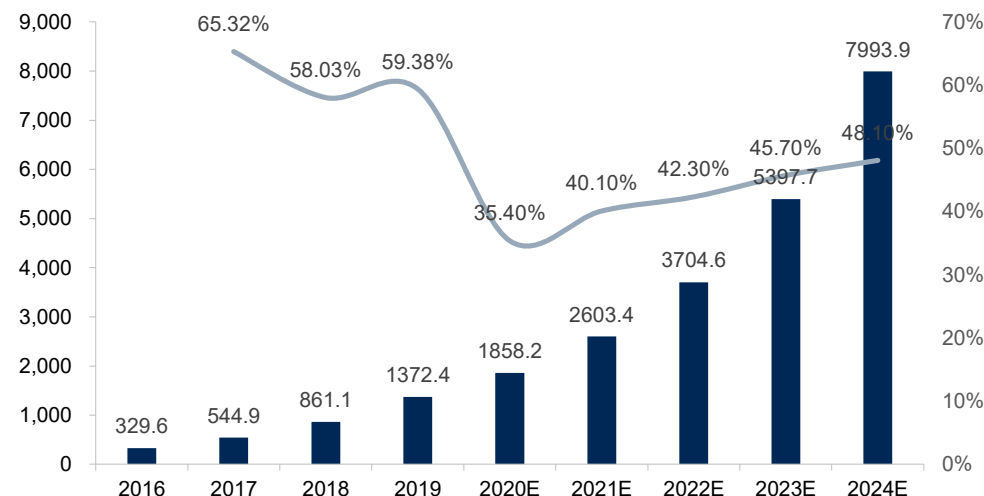
- **全球AI市场规模预计到2024年将超六千亿美元，复合增速27%。**据沙利文咨询统计，2016-2019年，全球市场规模从593亿美元增长至2019年1918亿美元，复合增长率约48%，预计2020年到2024年将以27%的年复合增长率继续放量，并在2024年达到6158亿美元。
- **中国AI市场规模预计到2024年约八千亿人民币，复合增速44%。**据沙利文咨询统计，2016-2019年，中国市场规模从329.6亿元增长至1372.4亿元，复合增长率约61%，显著高于全球整体增速水平，预计2020年到2024年将以44%的年复合增长率继续放量，并在2024年突破7993亿元。

全球人工智能市场规模（亿美元）



资料来源：沙利文咨询、国信证券经济研究所整理

中国人工智能市场规模（亿元）



资料来源：沙利文咨询、国信证券经济研究所整理

4.2 模型参数不断增长，算力需求持续增加

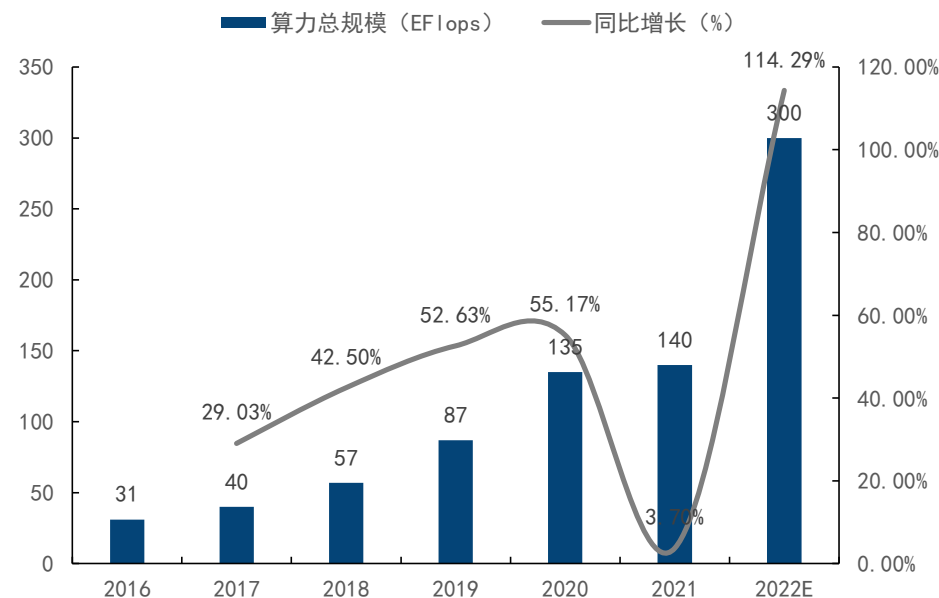
- **GPT-1迭代至GPT-3，参数量增大1500倍，预训练参数量扩大9000倍。** GPT-1、GPT-2和GPT-3的参数量分别为1.17、15.4和1750亿，预训练数据量分别为5GB、40GB和45TB。此外在序列长度方面，由初代的512增长至2048，模型层数方面也有数倍增长。
- **按照计算公式，算力需求与模型参数量呈正相关关系。** GPT的算力需求分别发生在访问阶段和训练阶段，访问阶段是指用户在使用GPT时，因提出问题所造成的算力消耗；训练阶段是指在训练GPT模型时所消耗的算力。访问阶段的算力消耗是衡量chatgpt投入的关键指标，假设每天租用亚马逊AWS云服务，每天成本为 $461.28 \times 2315 = 106.79$ 万美元；在训练阶段，自建IDC：服务器成本约占数据中心成本30%左右，为满足当前日常访问需求，前期一次性成本。投入约为 $2315 \times 19.9 / 30\% = 13.26$ 亿美元。

ChatGPT训练流程

	GPT-1	GPT-2	GPT-3
参数量	1.17亿	15.4亿	1750亿
预训练数据量	5GB	40GB	45TB
训练方式	Pre-training + Fine-tuning	Pre-training	Pre-training
序列长度	512	1024	2048
# of Decoder Layers	12	48	96
Size of Hidden Layers	768	1600	12288

资料来源：OpenAI 官网、国信证券经济研究所整理

2016-2022年中国算力总规模及增长率



资料来源：中国信息通信研究院、国信证券经济研究所整理

4.2 AI算力需求不断爆发，有望带动AI服务器放量

- **英伟达在GPU领域的强劲实力使其成为HPC需求增长的受益者。**目前GPT-3.5在微软AzureAI超算基础设施（英伟达V100GPU组成的高带宽集群）上训练是因为英伟达在AI和图形处理器方面都达到全球领先的技术水平。考虑到该技术在处理与机器学习、深度学习、人工智能和数据挖掘相关的复杂工作负载方面的能力，超大规模数据中心和高性能计算（“HPC”）细分市场对GPU的需求也很高。而“Nvidia A100” GPU——芯片制造商提供的众多数据中心GPU之一就是这样做的。该技术于2020年推出，基于上述Ampere架构构建，性能比其前代产品高出20倍。A100专为支持“数据分析，科学计算和云图形”而设计。还有最近推出的基于Nvidia A100的“HGX AI超级计算机”平台，该平台能够提供“实现HPC创新的极致性能”。
- **超速服务器市场需求助推英伟达逐步替代英特尔。**目前全球GPU架构的服务器出货量远高于CPU架构的，其中英伟达GPU对intel的替代其实是超速服务器的替代。在超算，AI训练里，这类服务器的增长导致对GPU需求大幅提升，以中国为例，目前国内普通服务器一年出货量再200亿美金，增速11%；但是超速服务器，就是搭载GPU的服务器，一年出货量大概是50亿美金出头，但是增速是87%。

HPC用户用到的50个较流行的应用程序中有34个提供GPU支持

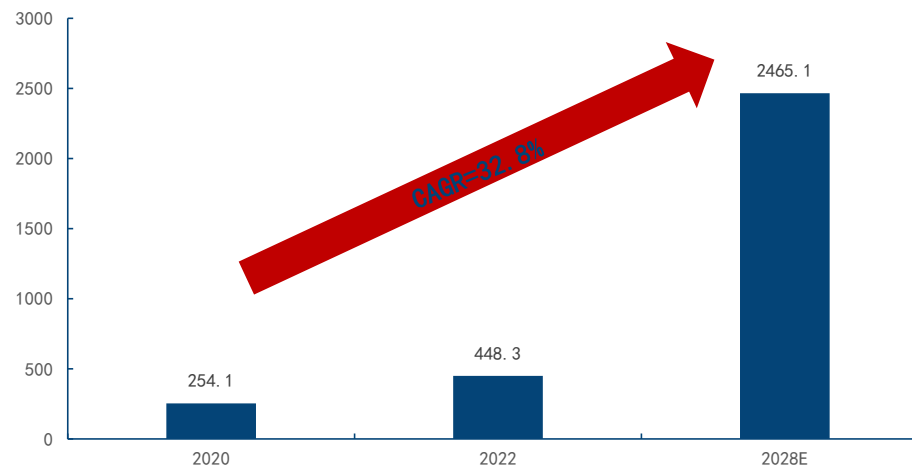
美国商务部工业和安全局宣布对华进行高端芯片限制

Table 1: GPU Support of Top 50 HPC Applications
Intersec360 Research, 2017

Rank	Supplier - Application	Mentions	GPU Support	Application Category
1	Siemens - SIMATIC	79	Yes	Chemistry
2	ANSYS - Fluent	78	Yes	Fluid Dynamics Analysis
3	Siemens, Inc. - Gaussian	74	Yes	Chemistry
4	U of Vienna - VASP	67	Yes	Chemistry
5	U of Illinois, UC - NAMD	64	Yes	Chemistry
6	Dassault Systemes - Simulia Abaqus	60	Yes	Structural Analysis
7	NCAR - WRF	55	Yes	Weather/Environment Modeling
8	OpenFoam Foundation - OpenFOAM	50	Yes	Fluid Dynamics Analysis
9	ANSYS - ANSYS	43	Yes	Structural Analysis
10	LSTC - LS-DYNA	40	Yes	Structural Analysis
11	NCRI - BLAST	38	Yes	Bioinformatics
12	Sandia Nat Lab - LAMMPS	33	Yes	Chemistry
13	JetsetKits.org - Amber	32	Yes	Chemistry
14	Quantum.espressos.org - Quantum Espresso	30	Yes	Chemistry
15	Iowa State University - GAMESS	29	Yes	Chemistry
16	PNEL - NWChem	27	Yes	Chemistry
17	NCAR - CISM / CCM3	19	No	Weather/Environment Modeling
18	Paraview.org / Kitware - Paraview	19	Yes	Visualization/Image Analysis
19	ANSYS - CFX	17	No	Fluid Dynamics Analysis
20	Harvard Computational - EN	16	Yes	Visualization/Image Analysis
21	MSC Software - NASTRAN	16	Yes	Structural Analysis
22	Siemens - CD-adapco Star-CD	16	No	Fluid Dynamics Analysis
23	COMSOL - COMSOL	15	No	Structural Analysis
24	Siemens - CD-adapco Star-CCM+	15	No	Fluid Dynamics Analysis

25	CPK.org - CP2K	13	Yes	Chemistry
26	ANSYS - ANSYS Mechanical	13	Yes	Structural Analysis
27	LMNL - VENT	13	Yes	Visualization/Image Analysis
28	Altair.org - ABRUT	9	Yes	Chemistry
29	Chemrxn.org - CHARMM	9	Yes	Chemistry
30	SNP.slu - SNP	9	In development	Business Intelligence
31	CPMD.org - CPMD	8	Yes	Chemistry
32	Open Source - MolVecs	8	No	Bioinformatics
33	U-Chem - U-Chem	8	Yes	Chemistry
34	UCM - JUP	8	Yes	Chemistry
35	Altair Engineering - HyperWorks	7	Yes	Structural Analysis
36	Dassault Systemes - Acetris Material Studio	7	No	Chemistry
37	Galaxyproject.org - Galaxy	7	No	Bioinformatics
38	MSMA - MSMA	7	No	Chemistry
39	Bumble - Casaca	7	No	Bioinformatics
40	Oracle - Oracle	7	In development	Business Intelligence
41	Stratix Research Institute - AutoDock	7	No	Bioinformatics
42	TensorFlow.org - TensorFlow	7	Yes	Pattern Recognition
43	UNQCD.org - MILC	7	Yes	Physics
44	Altair Engineering - Optislect	6	Yes	Structural Analysis
45	Open Source - SWA	6	No	Bioinformatics
46	STFC - UK POLY	6	No	Chemistry
47	Teplot, Inc. - Teplot	6	Yes	Visualization/Image Analysis
48	U of Vienna - Wien2k	6	No	Chemistry
49	Johns Hopkins School of Medicine - Bowtie	5	Yes	Bioinformatics
50	U of Illinois, UC - VMD	5	Yes	Chemistry

■ 全球GPU市场规模（亿美元）



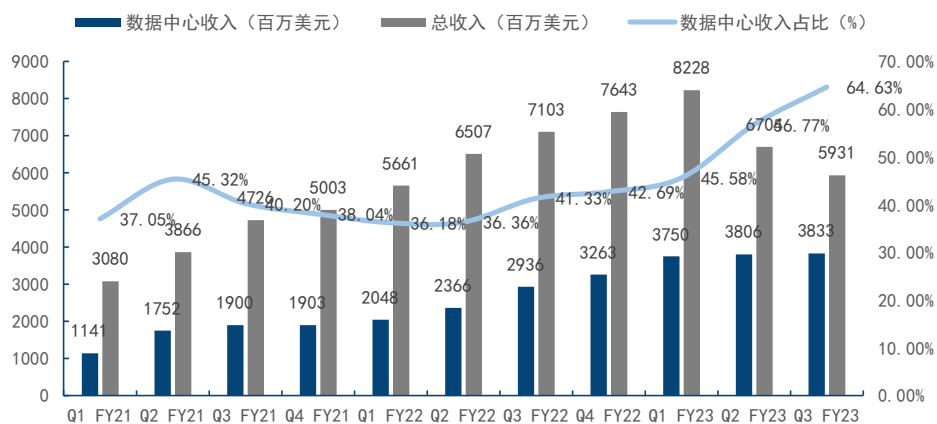
资料来源：CSDN、国信证券经济研究所整理

资料来源：Verified Market Research、国信证券经济研究所整理

4.2 AI算力需求不断爆发，有望带动AI服务器放量

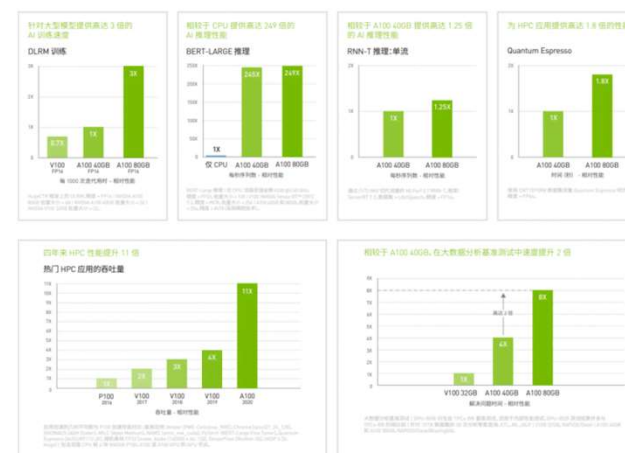
- **英伟达立足当下，数据中心业务实现腾飞。**2022年下半年，英伟达数据中心业务便实现高速增长超过市场预期，驱动2022年Q3增速比市场预期高了3亿多，因此结合市场环境，2022Q4数据中心超过游戏占比是预期内的结果。公司2022Q4收入指引是74亿美金，虽然环比预期给出了4.2%增长，相比单Q3环比9%增速低了一点，但是它因为产品单价还在高位，不排除超预期的可能。
- **英伟达布局未来，将持续从HPC需求市场中收益。**英伟达通过“Grace”和“Hopper”架构进军数据中心CPU和CPU+GPU超级芯片的最新尝试也使其处于有利地位，可以捕捉GPT-3等需要显著HPC性能的变压器模型的需求：为OpenAI开发的超级计算机是一个单一系统，每个GPU服务器拥有超过28.5万个CPU内核，1万个GPU和每秒400千兆位的网络连接。随着英伟达硬件的计算性能和成本效率的提高，像GPT-3这样的变压器型号也将变得更加精细，使它们离商业化更进一步。对芯片和其他对实现人工智能用例至关重要的基本硬件需求的最新研究预测显示：到本世纪末，潜在市场将达到约1.7万亿美元，性能和成本效益的提高是市场持续扩张的关键驱动力。而这两点（性能和成本效益的提高）也是英伟达继续提供的两方面优势——主要归功于英伟达人工智能训练加速器的性能一直在以惊人的速度进步，最新一代超高带宽内存技术HBM2e比Nvidia 5 K2014中的GDDR8内存快得多。由于Nvidia不仅实现了GPT-3等语言模型的实现化，而且还提高了未来所述变压器模型部署的经济性，该公司已准备好在未来几年仅从HPC的强劲需求环境中受益。

英伟达数据中心收入及其占比逐季度提升



资料来源：NVIDIA公司公告、国信证券经济研究所整理

NVIDIA A100 GPU性能卓越升级

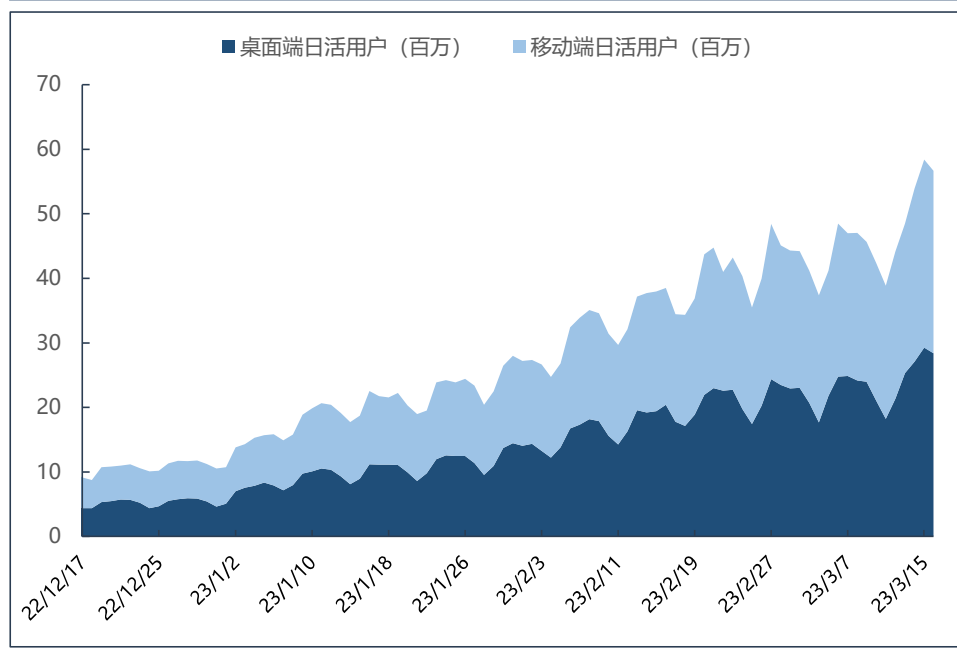


资料来源：NVIDIA官网、国信证券经济研究所整理

4.3 OpenAI 官网点击量不断攀升，当前日活近六千万

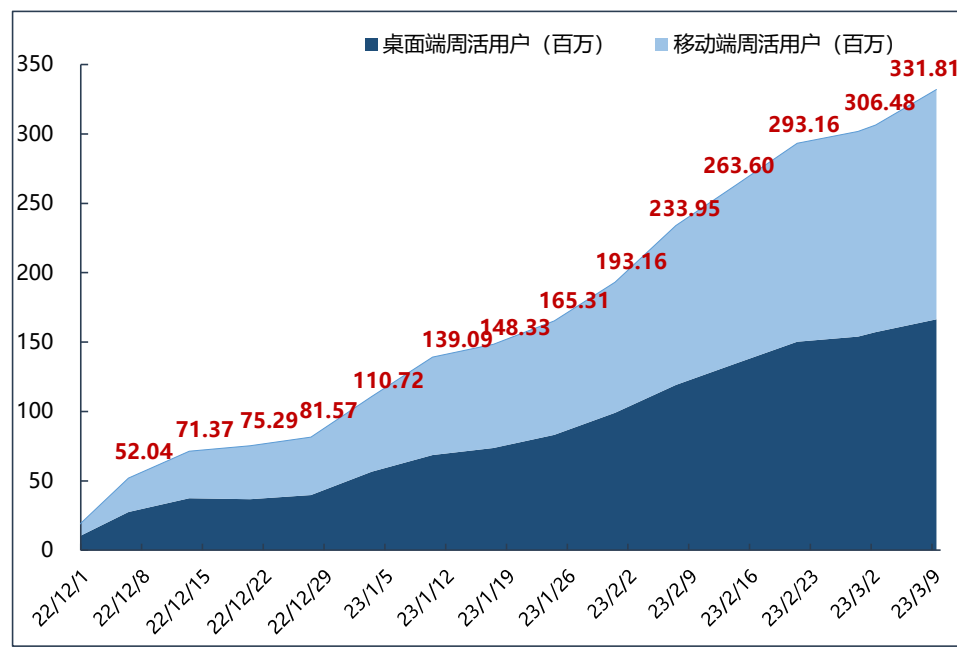
- 截至2023年3月16日ChatGPT日活人数达5800+万。ChatGPT2022年11月30日上线，上线一周获得百万注册用户，成为史上最快到百万用户的产品；2022年12月，日活用户数突破1000万；2023年3月份，ChatGPT日活人数已经突破5000万人。
- 日活用户数的增长会带来算力消耗的线性增长。根据算力消耗的计算假设，每位用户平均每次访问输入的信息为1000字，也就意味着在用户数增长的同时，算力需求也会随之线性增长，带来更大的算力缺口。

ChatGPT日活追踪（截至2023年3月16日）



资料来源：similarweb、国信证券经济研究所整理

ChatGPT周活追踪（截至2023年3月9日）



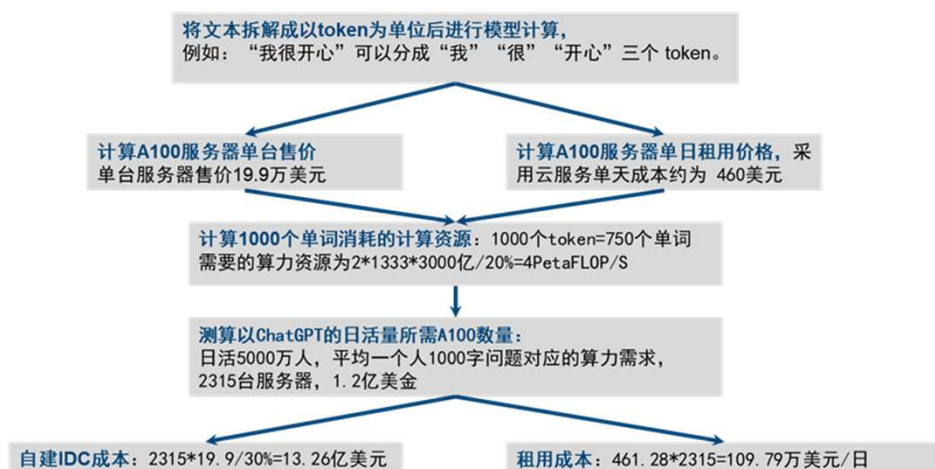
资料来源：similarweb、国信证券经济研究所整理

4.4 模型参数增加+用户规模扩大，当前算力需求缺口庞大

首先计算当前用户规模下，算力需求及成本情况：

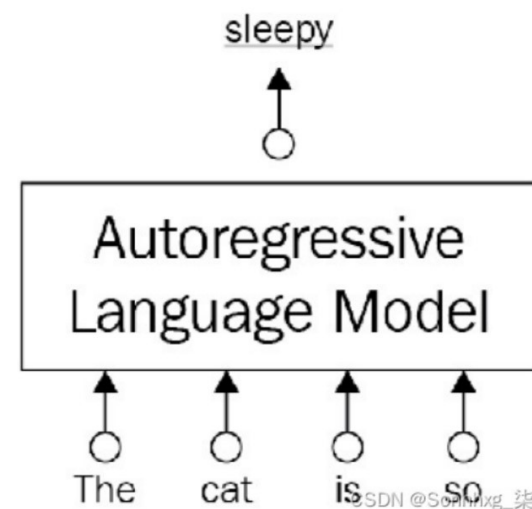
- **第一步，拆解字。**“token”是当前语言类模型的数据单位。当前的自回归语言模型是根据 token 来作为单位进行数据处理和计算，分词（tokenization）就是将句子、段落、文章这类型的长文本分解为以 token 为单位的数据结构，把文本分词后每个词表示成向量进行模型计算。例如在英文语境下，“happy”可能被分解为“hap”、“-py”两个 token，中文语境下，“我很开心”可以分成“我”，“很”，“开心”三个 token。

ChatGPT算力成本计算流程图



资料来源：CSDN、国信证券经济研究所整理

自回归语言模型原理简图

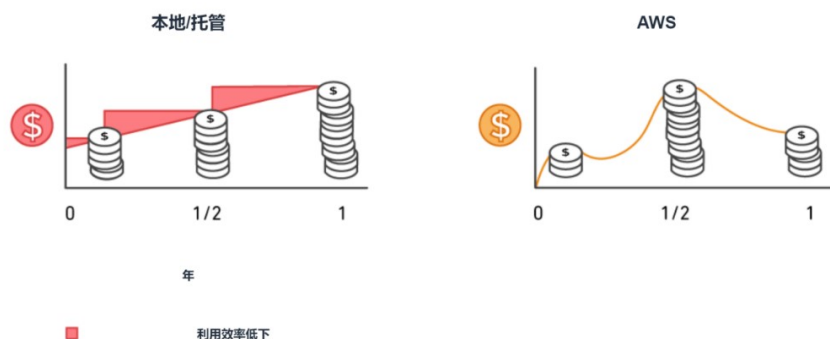


资料来源：CSDN、国信证券经济研究所整理

4.4 模型参数增加+用户规模扩大，当前算力需求缺口庞大

- **第二步，计算A100算力单台售价和租赁价格。**以英伟达DGX A100服务器作为计算资源：（1）单台服务器售价19.9万美元；（2）采用云服务单天成本约为460美元。英伟达超算GPU系列从旧到新包括P100、V100、A100、H100等，三年迭代一次，一次算力提升3-5倍，最新的一代H100，专门针对大模型开发，大约能提升算力9倍。按全球主要芯片还是A100，一个DGX服务器有8个A100系列GPU，AI算力性能为5PetaFLOP/s，单机最大功率6.5kw，售价19.9万美元；如果租用云服务，在亚马逊 AWS预定一年的A100系列 GPU，有8个A100的AWSP4实例的平均成本约19.22美元，一天的平均成本约为461.28美元。
- **第三步，测算1000个字（英语）消耗的计算资源。**较常见的Transformer类语言模型在推理过程中每个token的计算成本（以FLOPs为指标）约为 $2N$ ，其中 N 为模型参数数量（20年发布的GPT-3拥有1750亿参数，22年谷歌发布的PaLM拥有5400亿参数，由于并未公布当前GPT3.5的参数数量，当前假定参数数量为3000亿），假设模型的 FLOPS 利用率约为20%，粗略估计ChatGPT一个1000字（假设约1333个token，注：在英文语境下，一般1000个token=750个单词）问题需要的算力资源为 $2*1333*3000\text{亿}/20\%=4\text{PetaFLOP/S}$ 。

亚马逊AWS收费模式——按实际使用量付费



资料来源：AWS官网、国信证券经济研究所整理

亚马逊AWS收费标准



资料来源：AWS官网、国信证券经济研究所整理

4.4 模型参数增加+用户规模扩大，当前算力需求缺口庞大

- **第五步，计算自建成本和租赁成本。**自建IDC：服务器成本约占数据中心成本30%左右，为满足当前日常访问需求，前期一次性成本。投入约为 $2315 \times 19.9 / 30\% = 13.26$ 亿美元。（这个数和目前产业里面得到数基本一致）云服务：假设每天租用亚马逊AWS云服务，每天成本为 $461.28 \times 2315 = 106.79$ 万美元
- **OpenAI 面临着极为严峻的算力成本问题。**随着模型日益增大，OpenAI算力成本显著提高。根据国信计算机国内首发的有关ChatGPT算力准确测算的报告，当前ChatGPT的前期训练阶段一次性投入为3.99亿美元，而后期访问阶段基于当前5000万日活用户的每日租用服务器成本为106.79万美元/日，单是租用成本，每年便达到了3.9亿美元。尽管这个数据看起来似乎对于收入来说依然有可能覆盖，可是随着GPT-4大模型的到来，参数量和访问量将迎来百倍的增长，加上终端应用的算力消耗，光是对于英伟达的算力投入便能达到120亿美元，之后的访问阶段的算力消耗和算力成本更是OpenAI完全无法覆盖的数字，约为当前的一万倍（参数量和访问量均为原先的100倍）。所以OpenAI务必要对模型本身进行精简和优化，以满足未来GPT-4的训练参数量和用户访问量。

ChatGPT与GPT-4参数对比

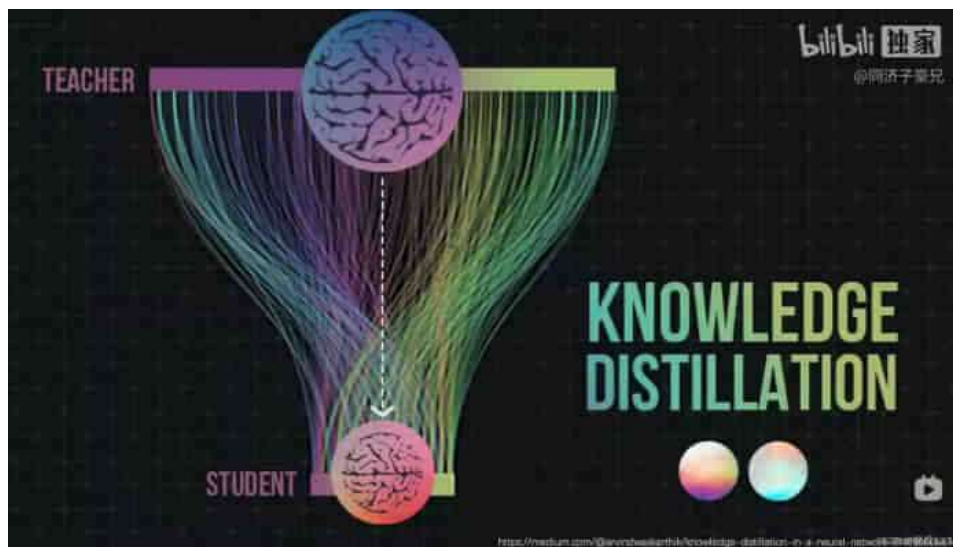
	ChatGPT	GPT-4
模型参数量	1750亿参数	100万亿参数
用户访问量	约1300万日活用户	预计1-10亿日活用户
嵌入终端应用	目前尚未实现终端应用嵌入	嵌入Bing等终端应用后会使得算力增加7-30倍

资料来源：CSDN、国信证券经济研究所整理

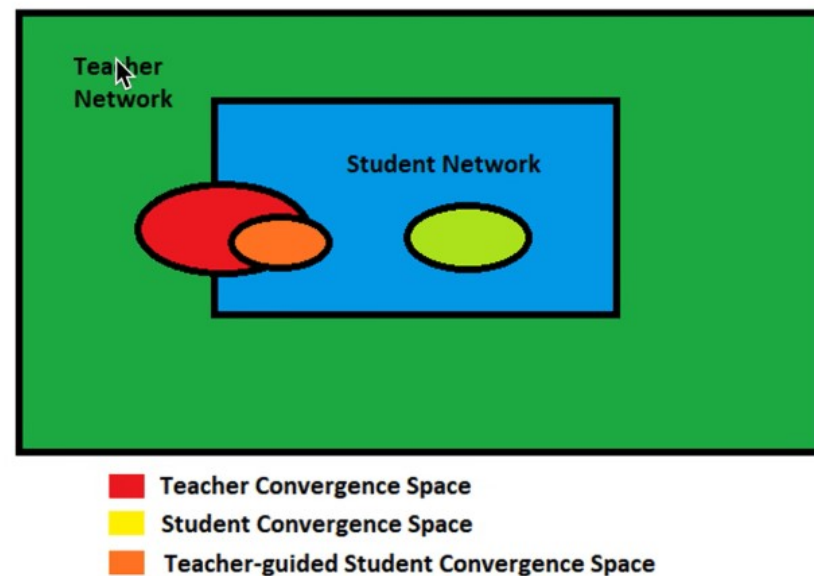
4.5 蒸馏算法等算法可以有望降低算力成本

- 知识蒸馏等算法可以解决模型臃肿和算力要求过高等问题。**蒸馏算法是指将知识丰富但是臃肿的“教师网络”经过精准转换将特定领域的知识传授给“学生网络”以实现网络结构的轻量化。如下图所示，知识蒸馏背后的原理是：绿色是教师网络求解空间，蓝色是学生网络求解空间。红色为教师网络的答案空间，浅绿色为学生网络的答案空间，橙色是在知识蒸馏的情况下得到的答案空间也是最优解，如果不加以引导（无监督训练），最后找到浅绿色的答案，而加入教师网络后，教师会给予学生指导，让学生网络得到更为准确的答案，所以知识蒸馏会得到更加精简且效果更好的模型。

教师网络与学生网络的关系



教师网络可以指导学生网络得到更为精准的答案



资料来源：哔哩哔哩、国信证券经济研究所整理

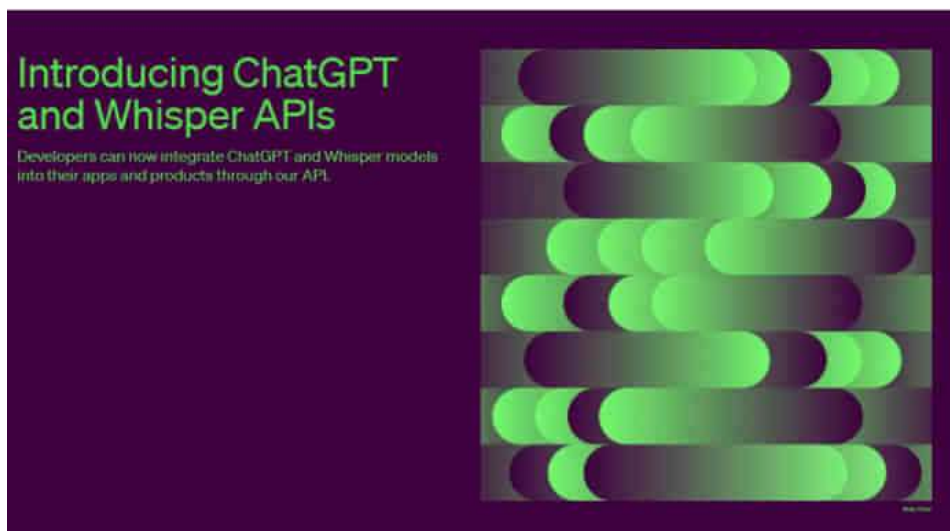
资料来源：CSDN、国信证券经济研究所整理

5、商业模式：开启订阅制收费，不断开放API接口

5.1 开启SaaS订阅服务，商业化变现不断打开

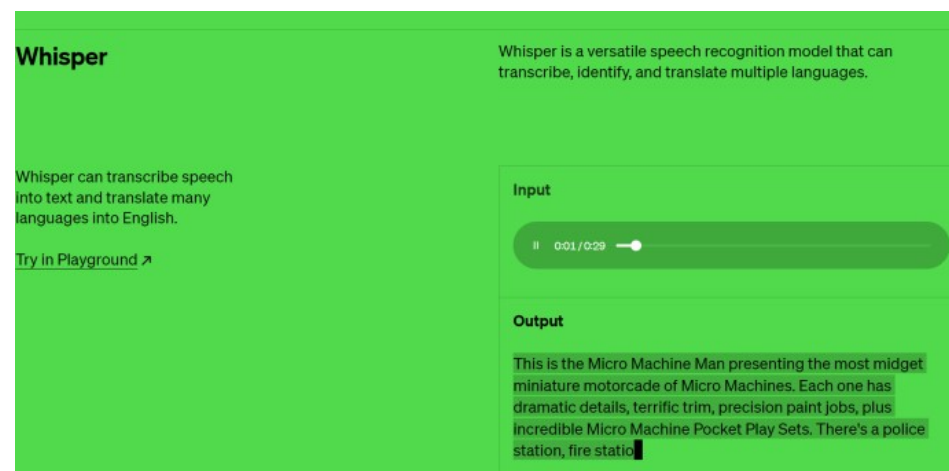
- 北京3月2日，OpenAI 宣布以收费形式对外部公司提供API 接入ChatGPT和Whisper模型的服务。API（Application Programming Interface）即应用程序接口，可以通过调用被外部使用，意味着ChatGPT和Whisper可以接入外部公司的各类软件。新发布的ChatGPT模型和之前提供的ChatGPT产品所使用的模型均为GPT-3.5，但是其收费标准比之前的产品便宜90%，为\$0.002/1k tokens。Whisper是2022年9月OpenAI推出的语音转文本模型，现在通过API推出V2版本，并以\$0.006/min进行收费。根据OpenAI官网介绍，截至3月2日，共有5家公司接入ChatGPT API，分别是Snapchat母公司Snap Inc、全球学习平台Quizlet、日用百货配送公司Instacart、电子商务软件开发商Shopify以及教育软件公司Speak。

OpenAI宣布开发ChatGPT以及Whisper模型API



资料来源：OpenAI官网、国信证券经济研究所整理

Whisper网页端界面



资料来源：OpenAI官网、国信证券经济研究所整理

5.2 订阅、API许可费和微软的深度合作是OpenAI主要收入渠道

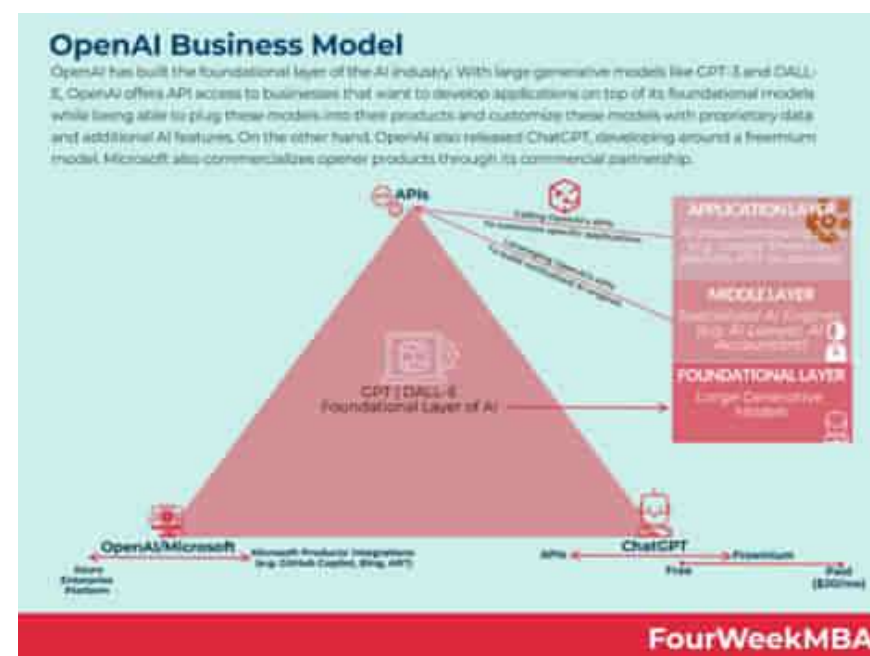
- 订阅、API许可费和微软的深度合作产生的商业化收入是OpenAI目前主要的收入渠道。在订阅渠道，ChatGPT已经推出付费版本ChatGPT plus每月收费20美元，用于提升软件服务质量以及后续开发成本；在API许可费方面，OpenAI刚刚开放API接口，主要希望吸引B端用户，一方面由于C端用户的问题杂乱且无序，所以其算力成本要高过B端用户9倍，OpenAI此举既可以吸引更多的B端用户迅速抢占市场份额以应对生成式AI激烈的竞争环境，另一方面还可以寻求降低算力成本的途径；与微软的深度合作收入更多属于营业外收入，自2019年开始，微软合计投入超过130亿美元，用于OpenAI的研发工作，以占领更多的云计算和搜索引擎市场。

2020年OpenAI LP财务数据

Total Revenue	\$3,481,784	
Total Functional Expenses	\$12,990,543	
Net income	-\$9,508,759	
Notable sources of revenue		
		Percent of total revenue
Contributions	\$2,661,461	76.4%
Program services	\$0	
Investment income	\$96,633	2.8%
Bond proceeds	\$0	
Royalties	\$0	
Rental property income	\$515,000	14.8%
Net fundraising	\$0	
Sales of assets	\$208,690	6.0%
Net inventory sales	\$0	
Other revenue	\$0	
Notable expenses		
		Percent of total expenses
Executive compensation	\$250,000	1.9%
Professional fundraising fees	\$0	
Other salaries and wages	\$363,861	2.8%
Other		
Total Assets	\$21,376,567	
Total Liabilities	\$113,129	
Net Assets	\$21,263,438	

资料来源：FourWeekMBA、国信证券经济研究所整理

OpenAI商业模式示意图



资料来源：FourWeekMBA、国信证券经济研究所整理

5.2 订阅渠道商业模式分析

- 2023年2月2日，OpenAI宣布，推出其人工智能聊天机器人ChatGPT的付费订阅版本，同时继续提供免费访问。该公司宣布，新的订阅服务月费为20美元，订阅者可以提前获得聊天机器人的新功能和改进，以及更快的响应时间。
- 高月活为OpenAI带来极大潜在收入空间。ChatGPT在发布后两个月时间内月活人数便突破1亿，若按照最低消费标准和10%的付费比例进行计算，则单是ChatGPT一项便会给OpenAI带来超过24亿美元的潜在年收入。

OpenAI宣布推出ChatGPT的付费订阅版本



资料来源：OpenAI官网、国信证券经济研究所整理

OpenAI旗下产品及其功能介绍

AI应用产品	产品介绍
GPT-3	是一种深度学习模型，该模型经过训练后可以预测序列中的下一个标记，并能生成流畅逼真的文本，覆盖文本分析和摘要、文案撰写、代码生成等应用场景
DALL·E 2	可以根据文本描述生成对应的逼真图像，还可以根据文本提示修改图像内容，以及根据给定的图像拓展生成多种主题风格一致的“变体”
Whisper	是一个自动语音识别系统。它使用从网络上收集的68万小时的多语言和多任务监督数据进行训练。使用庞大且多样化的数据集可以提高对口音、背景噪音和技术语言的鲁棒性。此外，它还支持多种语言的转录，以及将这些语言翻译成英语。
Codex	是一种基于GPT-3的代码生成助手，有助于将自然语言指令转换为十几种流行的编码语言。Codex专注于生成可执行代码，使其成为开发人员的绝佳助手。它已集成到Visual Studio中无缝集成到开发周期中
ChatGPT	定位聊天机器人，能根据用户的提示，在对话中生成类似人类的文本响应，完成回答问题、提供信息或参与对话等各种任务

资料来源：OpenAI官网、国信证券经济研究所整理

5.2 API许可渠道商业模式分析

- 开启API后，OpenAI将以0.002美元的价格提供1000个token，比之前版本便宜90%。成本下降后有助于推动ChatGPT与B端C端应用加速结合，实现产品加速落地，例如结合新版Bing以及嵌入Office相关应用。
- 基于GPT模型收取费用已有成功案例。例如AI创作公司Jasper其商业模式是以类SAAS服务的形式进行收费，大致分为初级、高级和订制三个模式。2021年、2022年营收分别为4000万美元、7500万美元，2022年10月，Jasper获1.25亿美元A轮融资，估值达15亿美元。此外AI辅助编程工具Copilot在2022年6月开始收费后首月边拥有40万订阅人数，用户付费率达1/3，远超行业均值。

OpenAI 各类模型收费标准

模型类型	具体分类	使用价格	模型类型	具体分类	训练价格	使用价格
嵌入模型	Ada	\$0.0004/1k tokens	微调模型	Ada	\$0.0004/1k tokens	\$0.0016/1k tokens
	1024×1024	\$0.0020/image		Babbage	\$0.0006/1k tokens	\$0.0024/1k tokens
图像模型	512×512	\$0.0018/image		Curie	\$0.0030/1k tokens	\$0.0120/1k tokens
	256×256	\$0.0016/image		Davinci	\$0.0300/1k tokens	\$0.1200/1k tokens
语音模型	Whisper	\$0.006/min				
聊天模型	gpt3.5-turbo	\$0.002/1k tokens				

资料来源：OpenAI官网、国信证券经济研究所整理

首批接入ChatGPT的企业及其应用

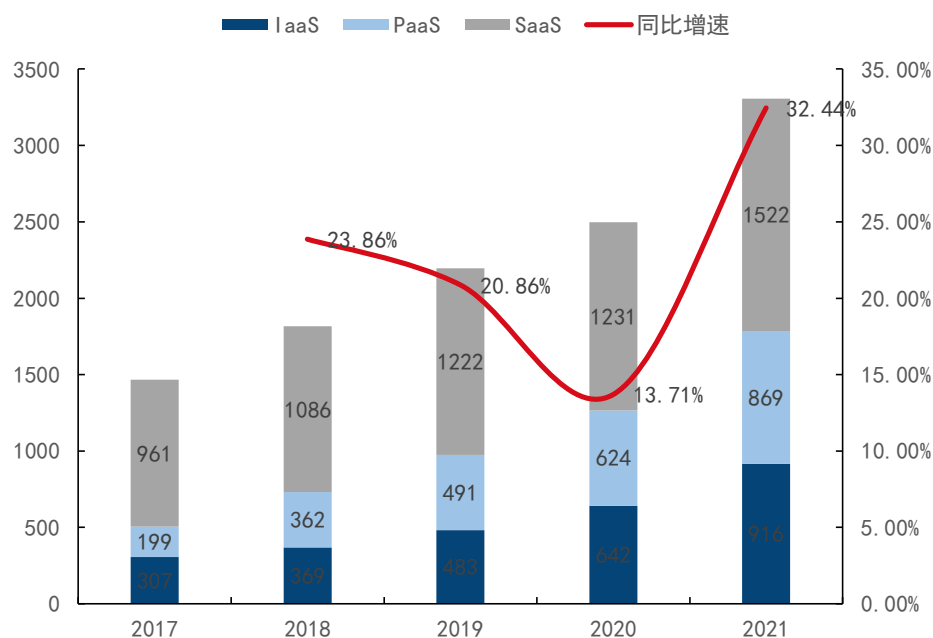
合作公司	具体应用
Snap Inc	Snapchat拥有7.5亿月活用户，ChatGPT API可为Snapchat+提供AI服务，主要可为用户创建一个友好且支持自定义聊天的机器人。
Quizlet	Quizlet是一个有超过6000万学生的全球学习平台，在过去三年中一直与OpenAI合作，包括词汇学习和实践测试。随着ChatGPT API的推出，Quizlet将推出Q-Chat，这是一款完全自适应的AI导师，它让学生根据通过有趣的聊天体验提供的相关学习材料提出个性化定制问题。
Instacart	Instacart正在扩充自己的应用程序，这使得ChatGPT与Instacart自己的AI和75000多家零售合作伙伴门店的产品数据互联互通，以给出客户开放式的购物想法和用餐建议。
Shop	Shop是一个拥有1亿用户的购物平台，通过使用ChatGPT可以帮助用户进行更为精准的个性化推荐以及更加高效的寻找到想要购买的商品。
Speak	Speak是韩国增长最快的英语应用程序，由AI驱动。该程序已经在使用Whisper API为一款新的配套产品提供开放式对话练习和高准确度的反馈，并推广至世界其他地区。

资料来源：OpenAI官网、国信证券经济研究所整理

5.2 与微软深度合作商业模式分析

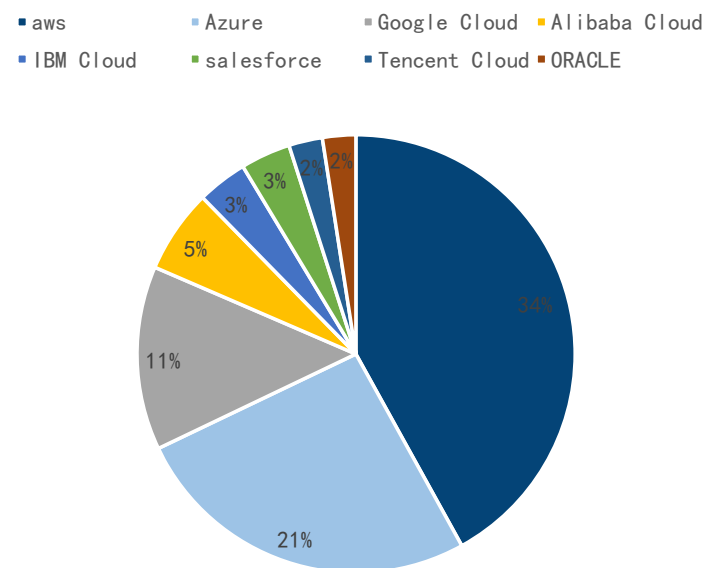
- **2021年全球云计算市场规模已突破3000亿美元。**2021年全球云计算市场规模达到3307亿美元，同比增长32.44%，市场空间广阔，保持稳定增长。按照业务划分基础服务，平台服务以及软件服务分别占27.70%、26.28%和46.02%。
- **OpenAI与微软的深度合作主要瞄准云计算和搜索引擎市场。**微软投资OpenAI主要还是针对其主要竞争对手进行布局：一方面加强Azure云计算领域的竞争力来对抗亚马逊，另一方面，根据StatCounter，2022年6月全球搜索引擎市场份额谷歌占据91.88%，微软有望利用GPT模型集成到公司旗下的搜索引擎必应，打破谷歌在搜索方面的垄断地位。

全球云计算市场规模及其增速（亿美元，%）



资料来源：信通院、国信证券经济研究所整理

2022Q3云计算行业公司市场份额占比 (%)



资料来源：澳财、国信证券经济研究所整理


6、应用场景：GPT走向多模态，下游应用场景不断打开

6.1 GPT的多模态化筑基AI多元应用

- **多模态指的是多种模态的信息，包括：文本、图像、视频、音频等。**多模态研究的就是这些不同类型的数据的融合的问题，例如通过NLP的预训练模型，可以得到文本的嵌入表示；再结合图像和视觉领域的预训练模型，可以得到图像的嵌入表示。
- **OpenAI 宣称GPT-4可以接受图像和文本输入，是目前较先进的多模态大模型。**ChatGPT仅可以接受文本信息的输入，而GPT-4在接收到文本和图片的融合信息后，模型可以输出纯文本信息，包括但不限于自然语言以及代码，且具备与接受纯文本信息后相同的输出能力。除此之外，GPT-4更加具有创造性，可以生成、编辑并与用户一起迭代创造性和技术性的写作任务，例如写歌、写剧本或学习用户的写作风格；能够处理超过25000个单词的文本，允许使用长格式内容创建、扩展对话以及文档搜索和分析等用例。

GPT-4可以准确识别出图片所包含信息

User: What is funny about this image? Describe it panel by panel.



Source: hmmm (Reddit)

GPT-4: The image shows a package for a "Lightning Cable" adapter with three panels.

Panel 1: A smartphone with a VGA connector (a large, blue, 15-pin connector typically used for computer monitors) plugged into its charging port.

Panel 2: The package for the "Lightning Cable" adapter with a picture of a VGA connector on it.

Panel 3: A close-up of the VGA connector with a small Lightning connector (used for charging iPhones and other Apple devices) at the end.

The humor in this image comes from the absurdity of plugging a large, outdated VGA connector into a small, modern smartphone charging port.

资料来源：OpenAI官网、国信证券经济研究所整理

GPT-4允许以更丰富的方式长内容创建

Input

[Text from: <https://en.wikipedia.org/wiki/Rihanna>]

Describe her Super Bowl performance and what made it special.

Output

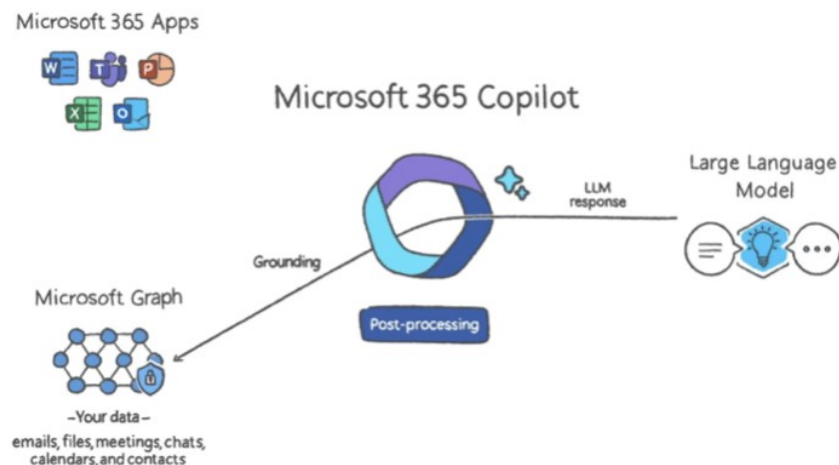
Rihanna's Super Bowl LVII halftime show performance was special for a number of reasons. For one, it marked her first live performance in over five years, and it was her first time headlining the Super Bowl halftime show. Additionally, she made history by becoming the first person to perform at the Super Bowl while pregnant. The show itself was critically acclaimed, with many praising her high-energy performance and impressive vocals. It also garnered a massive viewership, becoming the second most-watched halftime show in history.

资料来源：OpenAI官网、国信证券经济研究所整理

6.2 B端应用——办公软件之Copilot

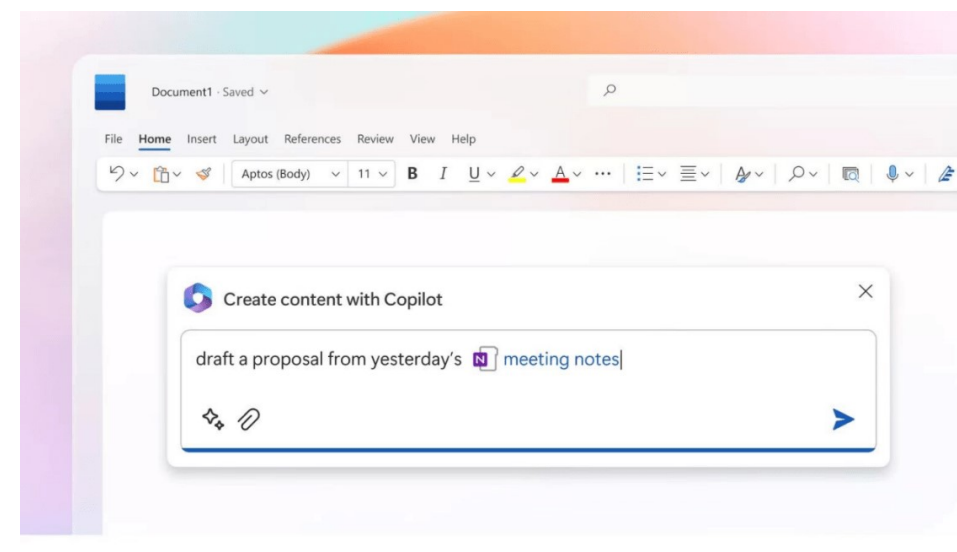
- 2023年3月16日，微软发布了AI服务Microsoft 365 Copilot。微软将其嵌入Word、PowerPoint、Excel 等Office 办公软件中，同时将GPT-4模型集成至产品当中，并结合其业务数据。目前微软正在与约20家企业用户测试新产品功能。
- Copilot主要通过两种方式集成到Microsoft 365中。一、直接被内置于Word、PowerPoint、Excel 等Office 办公软件中，以聊天机器人的形式呈现在办公软件的侧边栏上；二、通过最新产品Business Chat使用，该软件被应用于大型语言模型、Microsoft 365应用以及用户的日历、电子邮件、聊天、文档、会议和联系人，只需用户对其使用自然语言命令，便可根据会议、电子邮件和聊天记录等生成状态更新。

微软Copilot运行结构



资料来源：哔哩哔哩、国信证券经济研究所整理

Copilot呈现在侧边栏

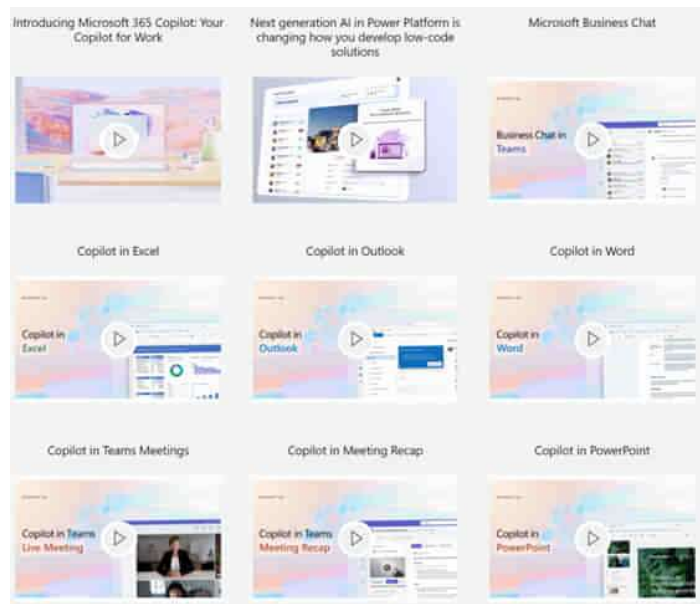


资料来源：新浪财经、国信证券经济研究所整理

6.2 B端应用——办公软件之Copilot

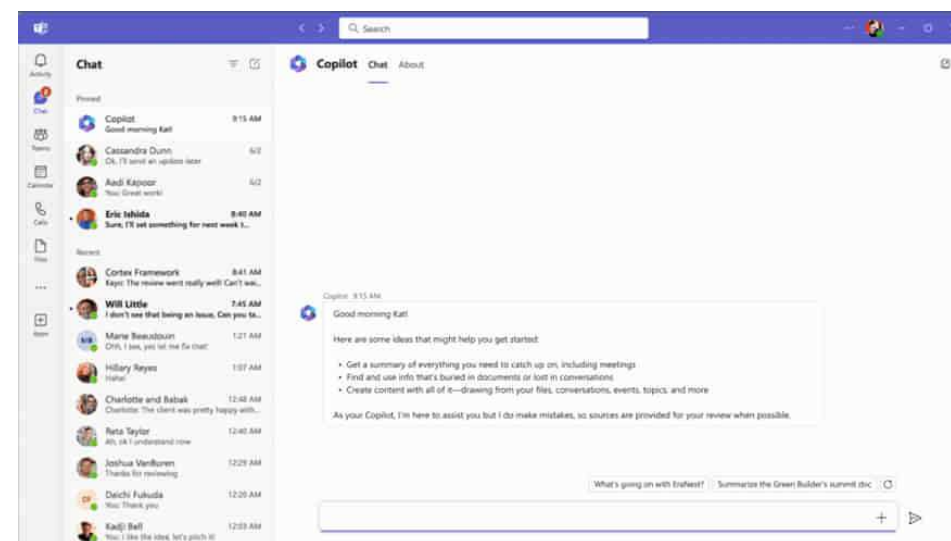
- **Copilot in Word:** 帮助用户提升创作、编辑等效率。
- **Copilot in PowerPoint:** 能够将自然语言命令转化为演示文稿。
- **Copilot in Excel:** 可创建趋势判断等专业形式数据可视化工作。
- **Copilot in Outlook:** 帮助用户整合并管理邮箱。
- **Copilot in Teams:** 能够帮助用户从对话上下文中提供实时摘要和待办事项，提高会议效率。
- **Copilot in Power Platform:** 通过在Power Apps和Power Virtual Agents中引入新功能，帮助开发人员利用低代码工具加速和简化开发。
- **Business Chat:** 汇集了来自文档、演示文稿、电子邮件、日历、笔记和联系人的数据，能够帮助用户总结聊天内容、撰写电子邮件、查找关键日期，以及根据其他项目文件制定计划。

微软已将GPT-4接入多个办公场景产品



资料来源：微软官网、国信证券经济研究所整理

Business Chat使用实例

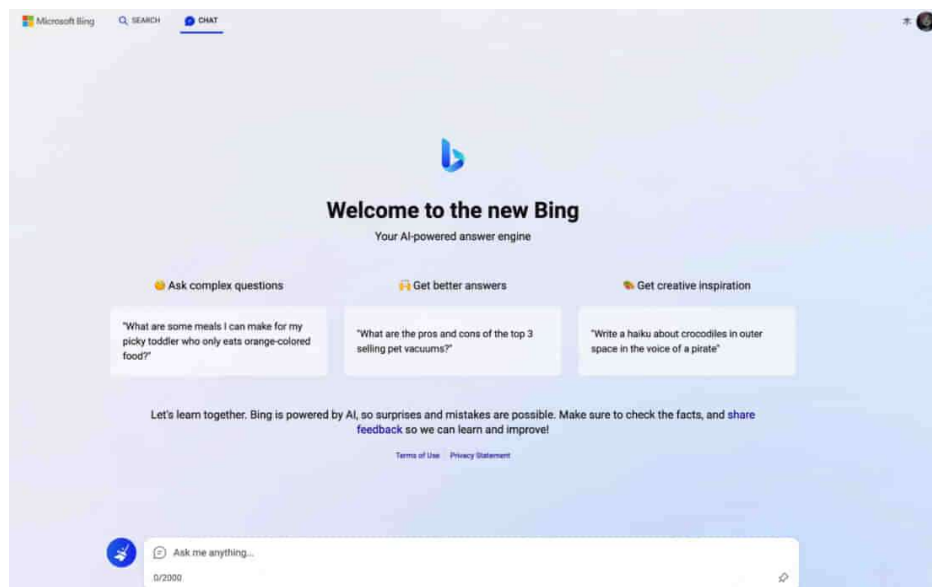


资料来源：微软官网、国信证券经济研究所整理

6.2 B端应用——搜索引擎之new Bing

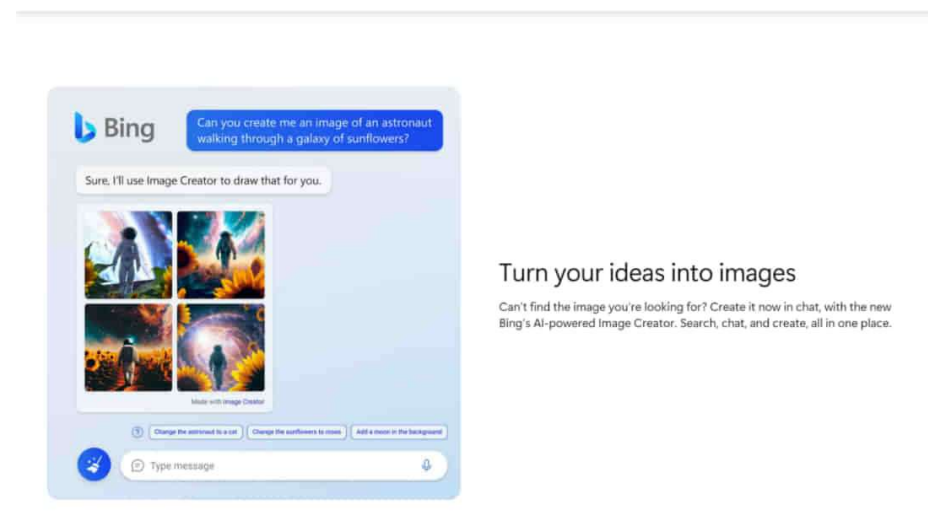
- 2023年2月7日，微软发布基于GPT-4的新一代Bing。新的Bing具有由下一代OpenAI大型语言模型支持的聊天功能（Bing chat），在其发布五周后，微软透露新版Bing是由GPT-4提供支持，且Bing是目前唯一免费访问GPT-4的途径。由于访问人数众多，微软对于新版Bing进行了访问限制，包括需要申请访问名单进入排队序列以及Bing Chat限制每次会话15次聊天，每天聊天总数为150次。
- 2023年3月21日，微软宣布将AI生成图像工具Bing Image Creator集成至新版Bing搜索引擎和Edge浏览器中，该工具由OpenAI的图像生成器DALL-E模型提供支持。在Microsoft Edge预览版中，用户只需单击边栏中的Bing Image Creator图标即可创建图像，或者也可以Edge边栏中的Bing Chat中直接调用。目前，Microsoft Edge成为了第一个也是唯一一个集成了AI图像生成工具的浏览器。

新版Bing使用界面



资料来源：微软官网、国信证券经济研究所整理

微软宣布将集成Bing Image Creator









资料来源：微软官网、国信证券经济研究所整理

6.2 B端应用——创作软件之Jasper

- Jasper公司已经通过利用OpenAI的大模型已经完成了营销文案和AI绘画的应用。**Jasper是基于GPT-3的API为用户提供AI写作服务的初创公司，致力于打造人工智能内容平台，允许个人和团队使用人工智能完成营销领域的素材撰写。微软的应用为海外产品的百花齐放奠定了良好的开端。Jasper成立于2020年，2021年、2022年营收分别为4000万美元、7500万美元。Jasper的商业模式是以类SAAS服务的形式进行收费，大致分为初级、高级和订制三个模式。2022年10月，Jasper获1.25亿美元A轮融资，估值达15亿美元。在产品层面，Jasper为用户提供50类以上的文案模板、29类语言服务，高级订阅包括SEO（搜索引擎优化）等内容，2022年11月至2023年2月，公司周活用户数量已增长63.9%。

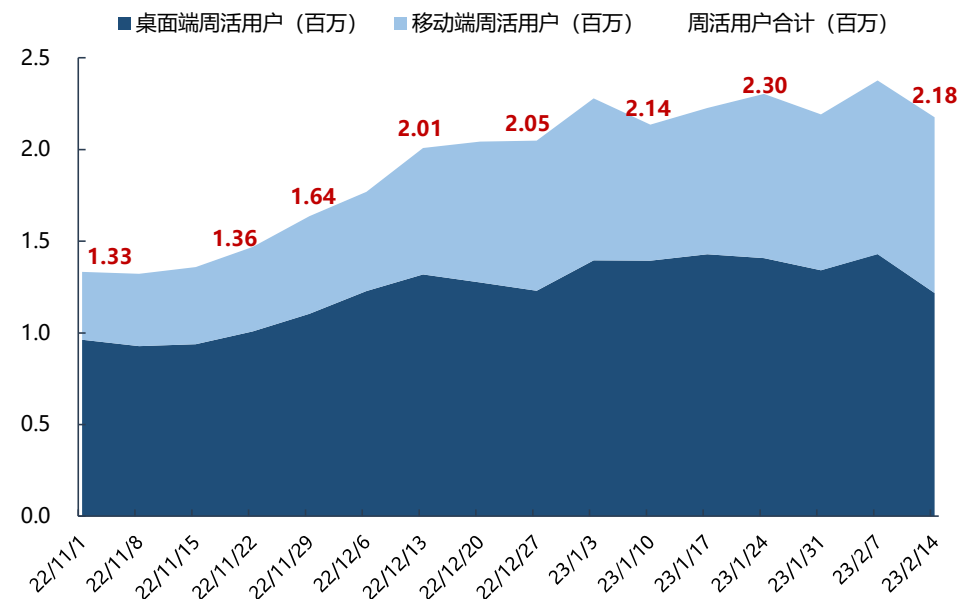
Jasper功能实现

Jasper will help you...

 <p>Create original content that ranks for SEO Generate educational blog articles that are keyword-rich and plagiarism-free.</p>	 <p>Boost ad conversions with better copy Easily write and test more copy variations to increase sales and improve ROAS.</p>	 <p>Finish your first draft 10X faster Tell Jasper about what you want and then watch the AI write paragraphs in seconds.</p>
 <p>End writer's block with ideas from a robot Stuck staring at a blank page? Relax and let Jasper write creative copy for you.</p>	 <p>Breakthrough the language barrier No matter your native tongue, write creatively and clearly in 25+ languages.</p>	 <p>Scale up your content marketing fast Repurpose existing content and generate new content without hiring junior writers.</p>

资料来源：Jasper官网、国信证券经济研究所整理

Jasper周活（2022.11.1-2023.2.14）

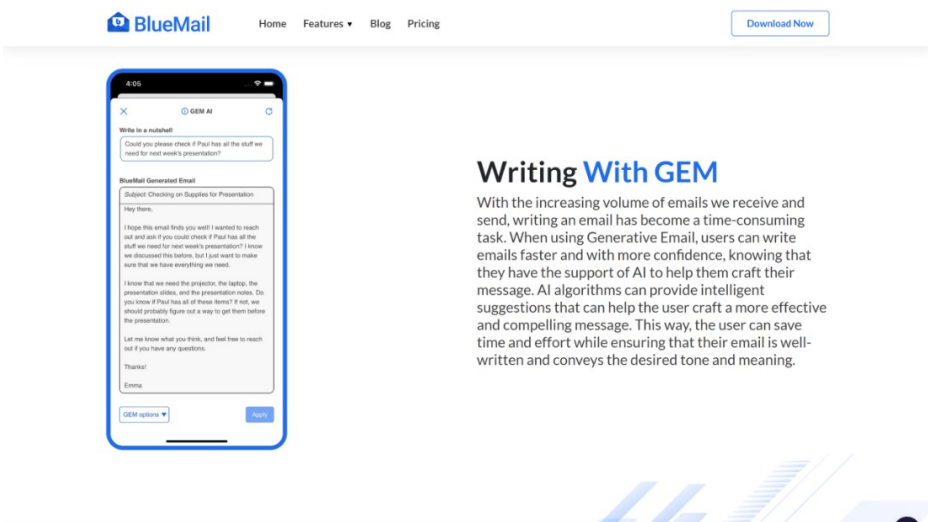


资料来源：SimilarWeb、国信证券经济研究所整理

6.2 B端应用——邮箱软件之Blue Mail

- **BlueMail GEM AI基于ChatGPT提供智能电子邮件服务。** BlueMail GEM AI目前由ChatGPT提供支持，根据用户输入的内容，自动生成更贴近用户语言习惯的电子邮件；浓缩邮件中主要信息，提升用户阅读效率。除邮件生成外，BlueMail还提供日程创建提醒、邮件整理、多设备邮件流转、邮件组合等特色服务。
- **2023年3月6日，苹果通过了BlueMail的应用更新。** 在3月2日，由于担心应用中全新人工智能功能可能会显示不适当的内容，因此苹果阻止了BlueMail的更新，并建议将年龄要求限制在17岁以上。随后BlueMail向苹果保证其软件具有内容审核功能并加以过滤，苹果允许了其软件的更新，并仍将年龄限制限制为4岁及以上用户。

BlueMail GEM AI应用范例



资料来源：BlueMail官网、国信证券经济研究所整理

BlueMail拥有强大的安全性与隐私性



Privacy & Security

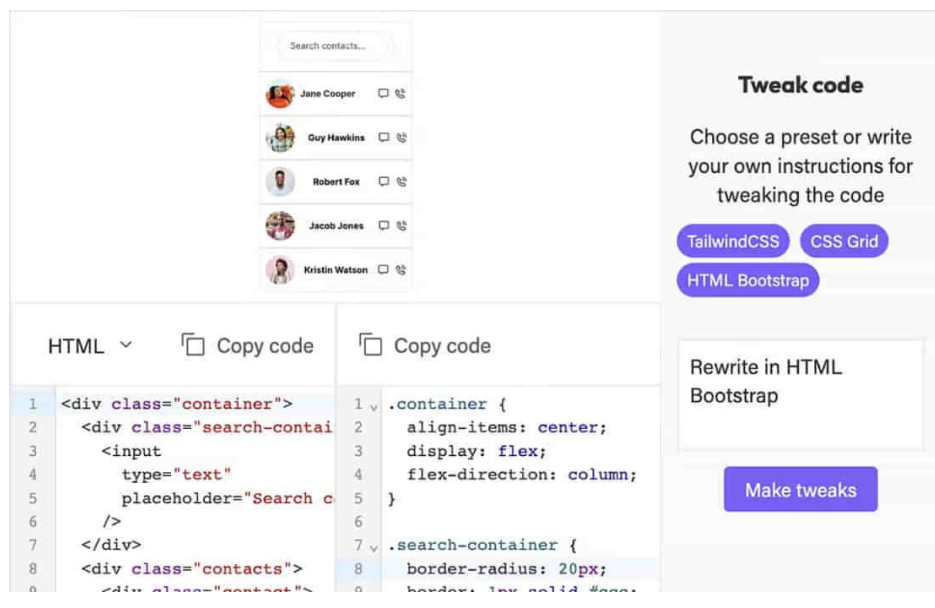
At BlueMail, the privacy and security of our users is a top priority and we strongly believe that only you should control your data.

资料来源：BlueMail官网、国信证券经济研究所整理

6.2 B端应用——代码生成之Codex

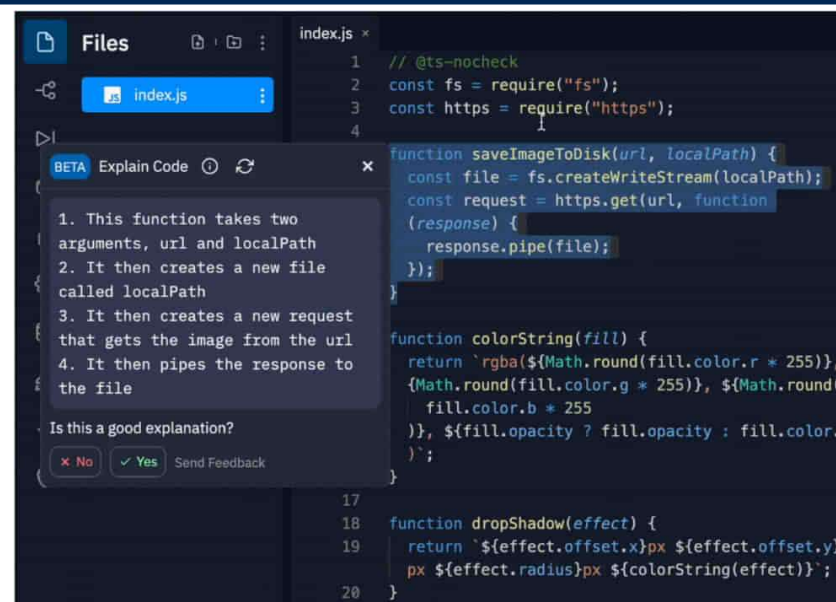
- **Codex是基于GPT-3的代码生成助手，有助于将自然语言指令转化成若干种编码语言。** Codex是GitHub Copilot的主要构件模块，通过与Codex的紧密集成，GitHub Copilot可以将注释转换为代码，自动填充重复代码，建议测试并显示替代方案。GitHub Copilot可用于Visual Studio和Visual Studio Code等环境，可与一系列广泛的框架和语言配合使用，对于一些编程语言来说，目前使用它的数万名开发人员生成的代码约占35%。对于用户来说，可利用Codex将Figma设计转化为不同的前端框架，并与开发人员的编码风格和偏好相匹配；利用Replit实现在项目中实时协作，学习代码，并与学习者和构建者社区共享工作；通过Warp直接从终端中利用自然语言指令进行搜索；通过Machinet使用Codex生成智能单元测试模板，帮助专业Java开发人员编写高质量的代码。

Codex应用实例——自然语言指令生成代码



资料来源：OpenAI官网、国信证券经济研究所整理

利用Codex进行代码学习

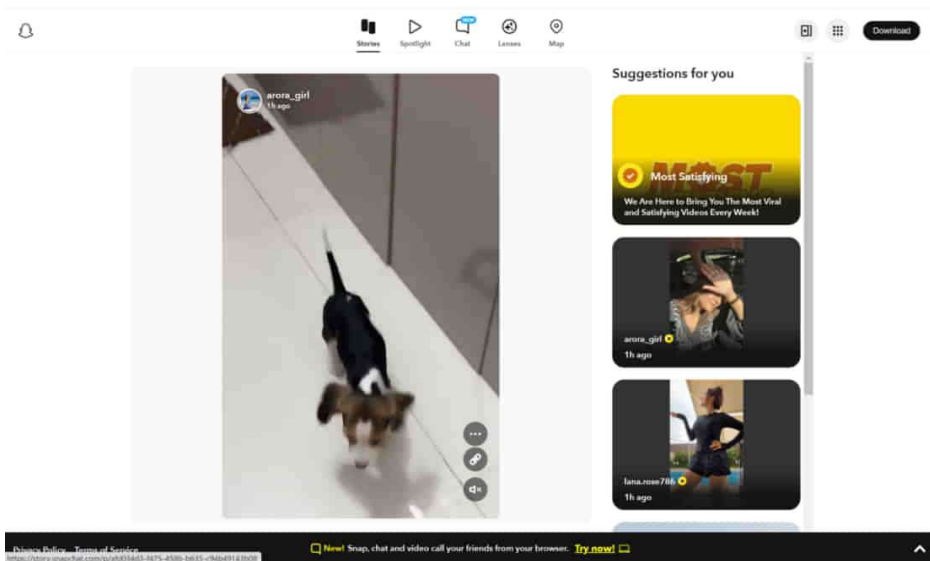


资料来源：OpenAI官网、国信证券经济研究所整理

6.3 C端应用——聊天社交之Snapchat

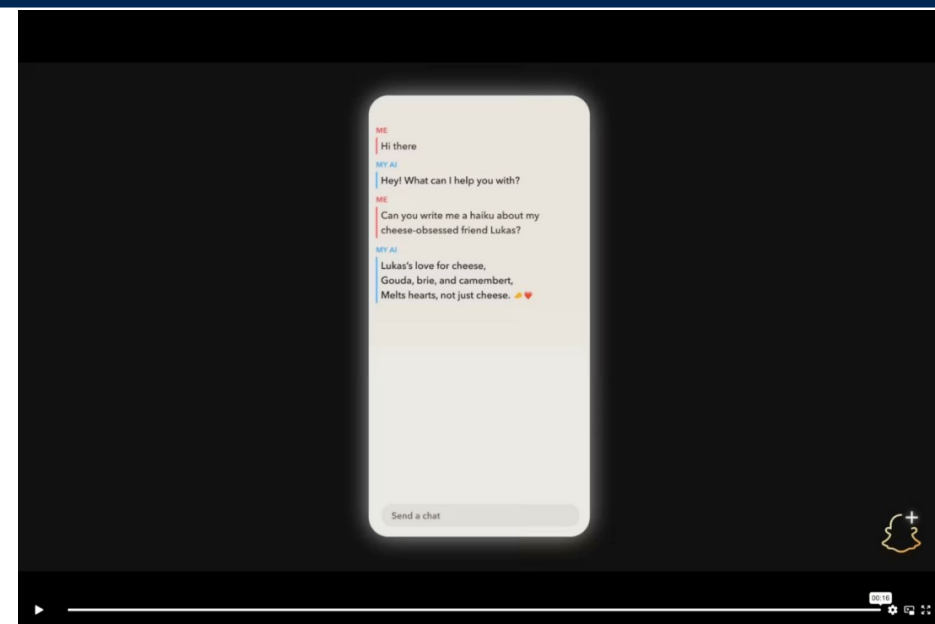
- **Snapchat**是一款“阅后即焚”照片分享应用。利用该应用程序，用户可以拍照录制视频添加文字和图画，目前在全球已经拥有7.5亿月活跃用户。
- **2023年2月27日，Snapchat推出一款由OpenAI最新版ChatGPT提供支持的聊天机器人。**目前正在ChatGPT API上运行，其主要功能是为用户创建一个友好且支持自定义聊天的机器人，可以提供推荐、诗歌创作等功能。据Snapchat CEO宣称，目前阶段用户只有支付每月3.99美元，成为付费订阅用户才可以使用具有ChatGPT支持的SnapchatPLUS，但其最终目的是将该应用服务于全体Snapchat7.5亿用户。

Snapchat传统照片分享功能



资料来源：Snapchat官网、国信证券经济研究所整理

Snapchat接入ChatGPT后使用实例

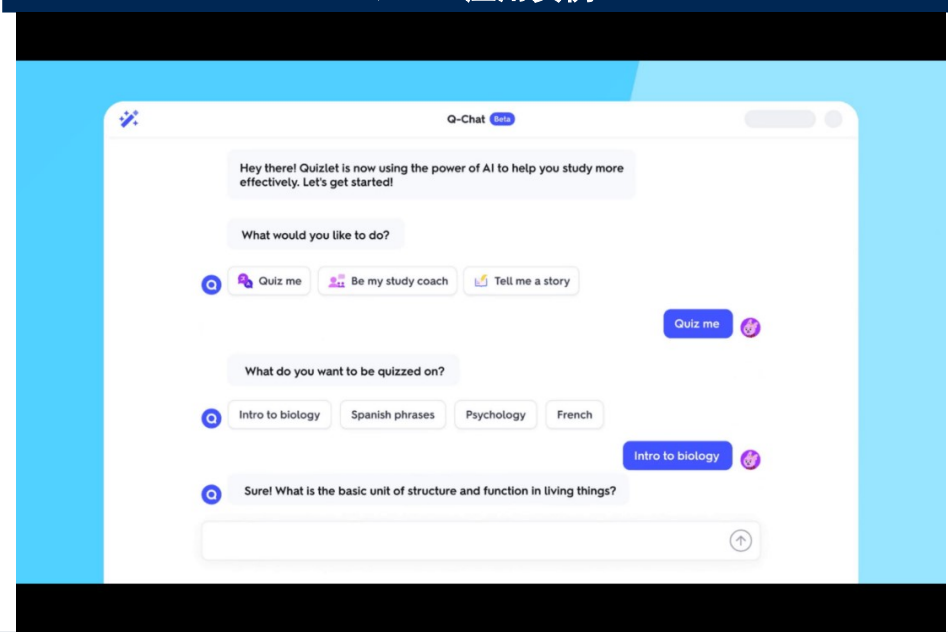


资料来源：OpenAI官网、国信证券经济研究所整理

6.3 C端应用——在线教育之Quizlet&Speak

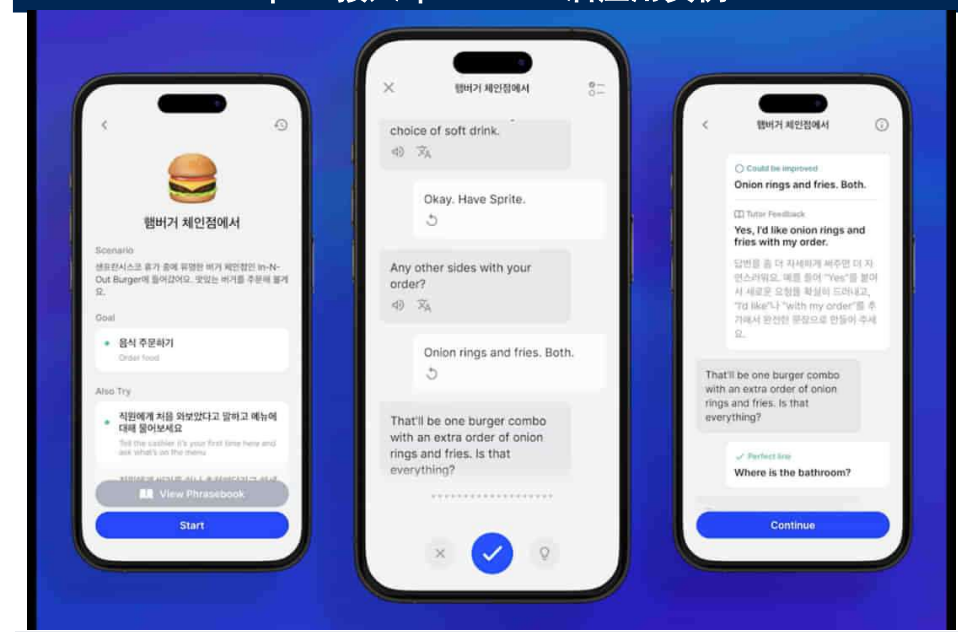
- **Quizlet是全球学习平台，日前接入ChatGPT API将推出自适应AI导师。** Quizlet是一个有超过6000万学生的全球学习平台，在过去三年中一直与OpenAI合作，包括词汇学习和实践测试。随着ChatGPT API的推出，Quizlet将推出Q-Chat，这是一款完全自适应的AI导师，它让学生根据通过有趣的聊天体验提供的相关学习材料提出个性化定制问题。
- **Speak是一款AI学习软件，由Whisper API提供支持服务。** Speak是韩国增长最快的英语应用程序，已经在使用Whisper API为产品提供支持，可为各个级别的语言学习者提供了人性化的准确性，开启开放式对话练习和高度准确的反馈，并计划推广至全世界范围。

Q-Chat应用实例



资料来源：OpenAI官网、国信证券经济研究所整理

Speak接入OpenAI API后应用实例



资料来源：OpenAI官网、国信证券经济研究所整理

6.3 C端应用——门店零售之Instacart

- **Instacart是一家美国生鲜杂货配送服务平台。** Instacart于2012年成立于美国旧金山，起初为当地零售商提供运货服务，目前已成长为涵盖熟食、烟酒、副食等全品类商品的零售服务商，为75000家零售商合作门店提供服务。
- **用户可从Instacart平台上得到用餐和购物建议。** 2022年公司推出Instacart平台，为顾客提供全流程解决方案；2023年初公司接入OpenAI API，并利用ChatGPT和公司零售商门店数据，为顾客提供开放式用餐与购物建议，例如“如何搭配今日的午餐”等问题，在得到建议后，用户可直接在该界面进行下单。Instacart计划于今年推出Ask Instacart应用程序以进一步提升用户体验。

Instacart使用方式

Grocery delivery you can count on

Choose what you want

Select items from your favorite grocery stores at Instacart.com or in the app.



See real-time updates

Personal shoppers pick items with care. Chat as they shop and manage your order.



Get your items same-day

Pick a convenient time for you. Enjoy Instacart's 100% quality guarantee on every order.

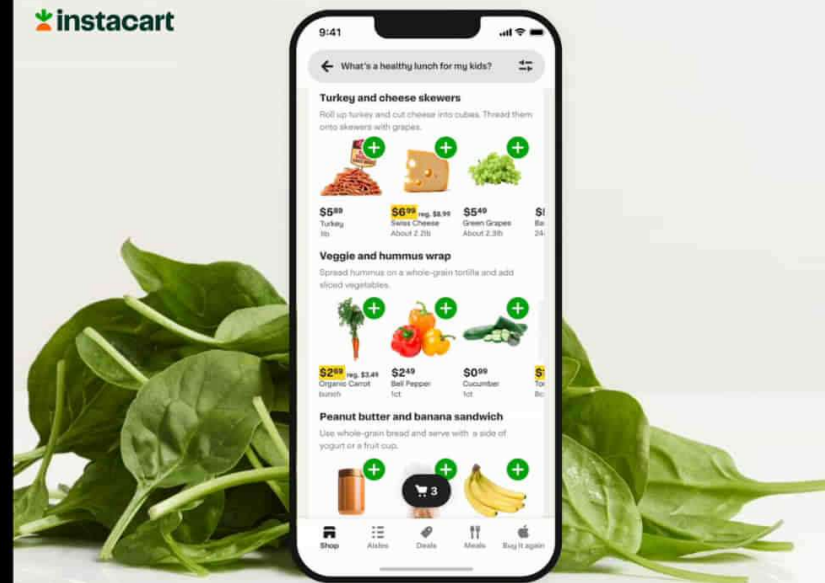


Start shopping

资料来源：Instacart官网、国信证券经济研究所整理

Instacart接入OpenAI API后应用实例

instacart

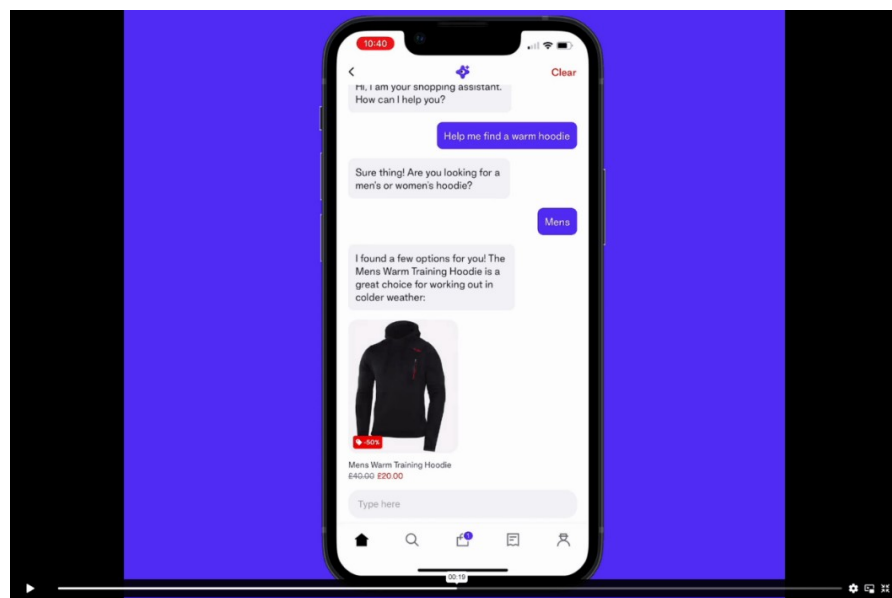


资料来源：OpenAI官网、国信证券经济研究所整理

6.3 C端应用——网络购物之Shop

- **Shopify是成为全球最大的电子商务平台之一，拥有超过120万商家在其平台上运营网上商店。**Shopify的用户分布在全球各地，主要覆盖零售、艺术、时装、美容等多个行业。公司主营业务为提供电子商务平台解决方案，让商家轻松创建和管理网上商店，并增加其在线销售收入，包括网站创建、支付处理、订单管理、客户服务、物流配送等。Shop是Shopify旗下购物平台，目前已拥有1亿用户。
- **Shopify率先集成ChatGPT，有效增强消费者个性化推荐、改善卖家运营效率。**在用户端，Shopify接入API端口后，用户可以收获更加精准的商品推荐，从而提升购物效率和满意度；在商户端，Shopify计划使用OpenAI API协助商家生成更好的描述文本和标题，提高商家在搜索引擎中的曝光率，同时商家还可以通过OpenAI API自动完成常见的客户支持任务。

Shopify智能推荐实例



资料来源：OpenAI官网、国信证券经济研究所整理

接入API后股价涨幅达到8.08%



资料来源：国信证券经济研究所整理

7、投资建议：建议关注基础层与应用层AI龙头

7.1 科大讯飞（002230.SZ）：布局AI平台+赛道的人工智能龙头



- **AI+教育。**认知智能一直是科大讯飞长期布局和积累的方向，公司已经在当前基础上，于2022年12月份进一步启动生成式预训练大模型任务攻关，该技术应用的AI学习机将于5月将率先落地。科大讯飞AI学习机中AI同步精准学的核心功能，小初高AI同步精准学以科大讯飞核心技术（智能评价技术、数据汇集技术、知识图谱构建技术）为支撑，依托于其为国家重大考试提供技术服务标准的积累，帮助孩子在学习过程中减负增效。
- **AI+医疗。**科大讯飞“智医助理”应用智能语音技术，AI赋能基层医生诊断和治疗，通过学习海量医学教材、论文和病例，实现专业的医疗领域对话式AI系统，打造“每个医生的AI诊疗助理，每个人的AI健康助手”。

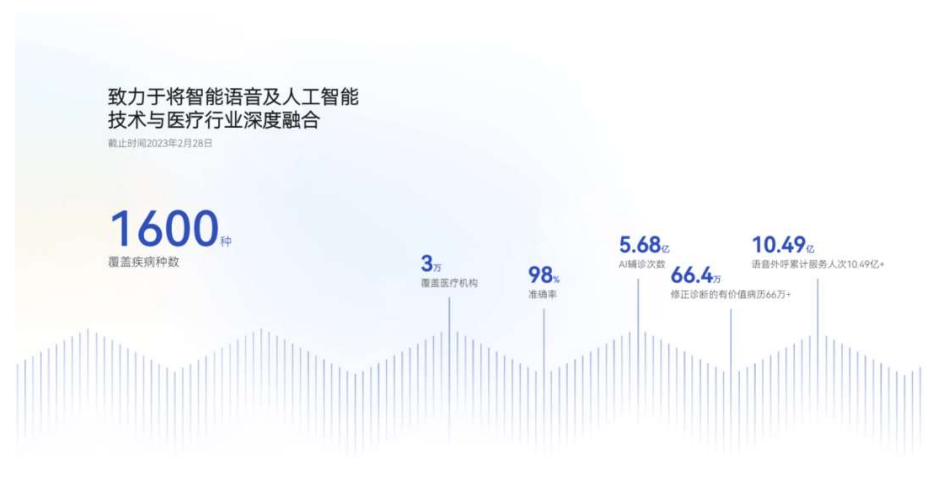
科大讯飞智慧教育产品服务

产品服务 PRODUCT



资料来源：科大讯飞官网、国信证券经济研究所整理

科大讯飞智慧医疗成果



资料来源：科大讯飞官网、国信证券经济研究所整理

7.2 金山办公（688111.SH）：从工具型SaaS转向平台型SaaS的国产办公软件龙头

- WPS可有效复制Office应用ChatGPT的路径。**微软正在测试将Office集成ChatGPT，对生产力工具使用效率有望大幅提升。金山办公WPS有望复制Office应用ChatGPT的路径。WPS目前已经研发出了智能辅助写作、文档自动翻译、表格拍照还原等应用，随着NLP技术迭代，类ChatGPT内容生成等应用的加入，将有效提升WPS的付费率与APRU。目前金山办公智能写作已具备相当可靠的语料库，提供讲话、总结等多种写作场景。同时，WPS还专门发布了公文版，面向党政机关用户提供公文模式、公文模板、公文转换等功能，辅助相关人员进行公文写作。WPS、邮箱、协作办公等办公软件再融入类ChatGPT功能后，其生产力工具价值得以跃升，商业价值迎来加速增长。

金山办公产品矩阵



资料来源：金山办公官网、国信证券经济研究所整理

必优科技自研智能写作引擎获得金山办公投资



资料来源：金山办公官网、国信证券经济研究所整理

7.3 同花顺（300033.SZ）：国内领先的互联网金融信息服务提供商



- 同花顺是国内领先的互联网金融信息服务提供商。公司拥有近三十年金融信息服务行业经验，在产品、客户资源、技术、人力资源、数据资源、品牌认可度、管理团队等方面拥有独特而明显的竞争优势。公司构建的同花顺AI开放平台，可面向客户提供智能语音、智能客服、智能金融问答、智能质检机、会议转写系统、数字人、智能医疗辅助系统等多项AI产品及服务，可为银行、证券、保险、基金、私募、高校、政府、运营商、医疗等行业提供智能化解决方案。经过多年的沉淀，公司产品具备了较好用户体验度和较强用户黏性，深受广大客户好评，形成了明显的竞争优势。

同花顺覆盖全行业的AI解决方案



资料来源：公司官网、国信证券经济研究所整理

公司的AI理财师

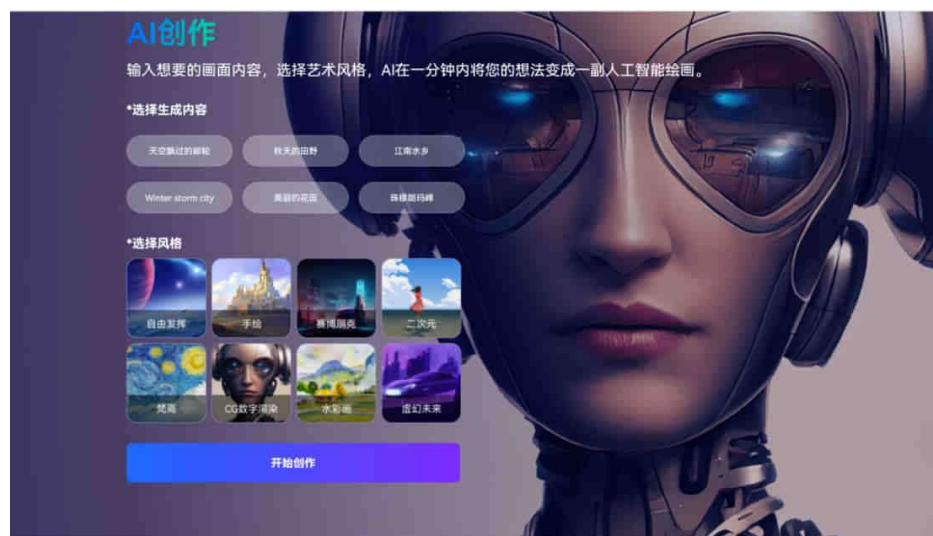


资料来源：公司官网、国信证券经济研究所整理

7.4 万兴科技（300624.SZ）：全球领先的新生代数字创意赋能者

- **AI+创作。**万兴科技深耕数字创意软件领域，推出万兴喵影、万兴优转、万兴录演等视频创意软件，Wondershare Filmstock、万兴喵库视频创意素材平台，万兴爱画、万兴播爆等AIGC应用。
- **万兴爱画是一款根据用户语言描述进行绘画创作的软件。**用户可以输入自己的创作想法，并且对其中部分元素加以强调，实现高效创作，此外万兴爱画生成的作品所有权属于用户本身，可以进行商业用途。
- **万兴播爆是一款利用数字人进行营销播报的软件。**用户可以输入关键词，由AI生成播报脚本，接着由AI数字人进行口播。软件提供上百套专业级模板，可以满足不同场景的视频需求。

万兴爱画



资料来源：万兴科技官网、国信证券经济研究所整理

万兴播爆



资料来源：万兴科技官网、国信证券经济研究所整理

7.5 拓尔思（300229.SZ）：数据与内容安全产品先行者



- **AI+数据要素。**拓尔思是国内最早从事自然语言处理（NLP）技术研发的企业，在网络空间治理领域提供内容审核和内容风控技术。拓尔思依托自有的人工智能与大数据核心技术，遵循网信系统技术体系建设要求，围绕网络空间治理工作需要，构建了涵盖“网络舆情态势感知、正面宣传、网评引导、公众举报、属地网络内容管理、指挥联动”等业务的全方位技术解决方案，为“加快建立网络综合治理体系，全面提升治管网能力水平”提供技术支撑。现已为众多用户量身打造了个性化系统，深受用户好评。
- **拓尔思布局问答机器人业务。**TRS小思智能问答机器人系统是由拓尔思自主研发的一款智能问答云服务系统，该系统综合运用了自然语言处理、信息检索、知识表示与推理等技术，通过语法语义分析、问题分类、问题解析、信息源自动识别与评价、实体识别与关系抽取、信息搜索、逻辑形式生成、知识表示与推理等一系列处理流程，实现问题解答与人机交互。

智拓·人工智能技术平台



AI行业应用——TRS银行业智能消保管控中台解决方案



资料来源：拓尔思官网、国信证券经济研究所整理

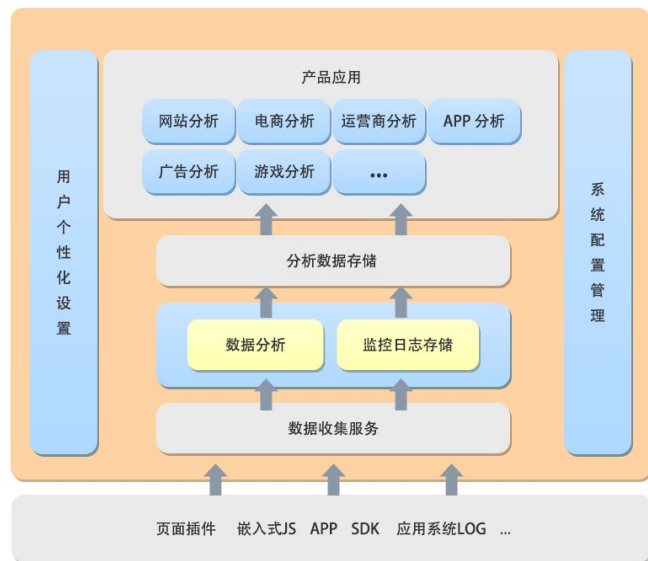
资料来源：拓尔思官网、国信证券经济研究所整理

7.6 彩讯股份（300634.SZ）：中国领先的企业数字化技术和服务提供商



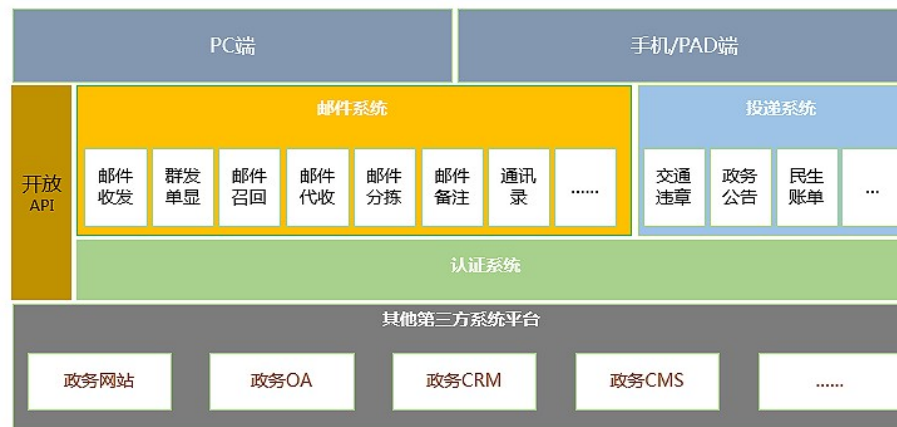
- **AI+信创。**2023年2月，彩讯股份成为百度“文心一言”的首批生态合作伙伴。彩讯科技股份有限公司是国家规划布局重点软件企业之一，是中国领先的企业数字化技术和服务提供商。此次接入文心一言，双方也将共同探索AI核心技术在彩讯电子邮件RichMail等核心产品及电信、金融等核心行业率先运用，并围绕技术创新、场景孵化、生态建设等多方面展开更深入的合作，助力客户行业的智能化转型升级，持续为通信行业、能源行业、交通行业、金融行业等广大客户开发并运营互联网应用平台、企业协作和营销平台、企业级大数据应用平台。
- **协同办公产品线是彩讯的主要业务方向之一，是以“信创邮箱+统一办公平台”为核心。**生成式AI所拥有的语言理解和文本生成能力可以帮助彩讯办公产品在数据处理、数据分类、智能交互方面进行优化提升。训练大模型也可实现有效闭环，落实数据准备、数据预处理、模型训练、模型评估、部署应用等流程环节并实现持续优化。

彩讯网银流量分析解决方案



资料来源：彩讯股份官网、国信证券经济研究所整理

彩讯电子政务邮箱解决方案



资料来源：彩讯股份官网、国信证券经济研究所整理

7.7 福昕软件（688095.SH）：国际化的PDF电子文档解决方案提供厂商



- **AI+图像识别。**福昕软件立足于PDF行业，并持续在图像识别垂类应用深耕，目前已经在1. 中国电子清单服务海外的电子签名；2. 面向家装领域的福昕家装智能设计（面向设计师）、福昕家装云服务（提供施工的协同平台）；3. 面向船舶行业的福昕船舶图纸管理系统；4. 面向海外市场的smart editor，用于法律、金融等行业，进行保密和信息加密的任务等多方向新领域开展了图像识别相关业务，针对高质量图像用户需求，在拓展行业的同时，实现矩阵式销售。
- **公司紧跟谷歌、百度等人工智能厂商，发挥公司PDF领域优势。**公司的主要角色就是利用API接口将文档传输进去，不需要粘贴到大模型里面去，直接在我的文档进行读取。目前正在与谷歌、百度等AI巨头进行研讨。

福昕船舶图纸解决方案

功能一览



图纸项目管理

以项目形式对企业图纸进行分类、分专业集中管理。有效管理企业沉淀的各种设计成果，规范企业的图纸管理，提高管理能力，助力企业数字化转型。



图纸版本管理

从设计到定稿过程中会产生多个版本，图纸版本管理记录并保存所有版本，以及对应的电子文件，可查询追溯。



图纸工作流程管理

支持企业按照自身的需求自定义其工作流程，实现更加灵活的图纸工作流程管理，贴合企业管理流程，易于学习和使用。



文档安全

图纸是企业的重要资产，图纸安全可控是维护企业利益的重要保障。本系统通过用户角色、权限的精细控制，对文件存储加密，实现对文档的查看、编辑、下载、打印等操作。



电子图章与签名

系统内置了完全自主知识产权的PDF核心技术，提供企业人员签名库管理，可直接对PDF文档进行电子图章、手写签名、添加水印等操作。



审图图纸管理

图纸审核收发是船舶行业图纸审核流程的关键步骤，系统提供向审图单位发送图纸和接收审核结果的管理，可以无缝对接CCS审图系统，也可以管理其他审图方的审核结果。

资料来源：福昕官网、国信证券经济研究所整理

福昕智能互联PDF产品矩阵

互联PDF - 互联世界的新一代智能PDF

互联PDF将传统PDF与互联技术相结合，将PDF文件所有者和文档互联起来，进行实时交互，实现了所有者对文档的动态管控，是敏感文档的内加管控、对外流转和可控共享等方面的得力工具，可广泛应用于政府、企业和其他任何需要电子文档的流转进行保护的领域，如电子公文、技术报告、财务报告、客户资料、档案等的安全可控流转。

让政府、企业、个人的敏感信息存储更安全、流转更可控、交流更方便

个人版 - 公有云个人账户 -	专业版 - 公有云专业账号 -	企业版 - 公有云企业账号 -	专用版 - 私有云部署 -
您将文档上传，文档即自己保存，权限交给“福昕管家”（福昕互联可控文档协同平台）。福昕管家只保留您的，不置文档存在哪里，只有您自己（通过身份验证后）可以打开，别人无法打开！	在个人版的基础上，您还可以将文档授权给亲朋好友、“福昕管家”（福昕互联可控文档协同平台）只在您网络的时间，给即需要的人（经过身份验证后）开放，文档能定、编辑？干啥，没有您的同意，任何人都打不开！	您再文档上传，一切交给“福昕管家”（福昕互联可控文档协同平台），福昕管家只保留您的，根据您的指示，福昕管家为您自己、您的客户、合作伙伴提供增值服务，您使用户可以随时、随地在电脑、手机、PAD上查看您的文档。	我们两系统，工具都交给您，文档、数据都是您自己管理，关键是您所有的，怎么管理自己定，福昕只管他的系统，工具体维修、保养，完全个性化，适应您特殊的需求。
下载/试用	下载/试用	下载/试用	联系销售专员
通过您的手机号或个人邮箱即可注册开通，免费试用。	通过您的手机号或个人邮箱即可注册开通，收费使用。	简单试用 深度试用 使用共享的试用账号，试用加密、授权、分享等主要功能，无需注册直接试用。	010-62105891 cpd@sales@foxitsoftware.cn

资料来源：福昕官网、国信证券经济研究所整理

7.8 凌志软件（688588.SH）：国内领先的金融科技软件公司



- 公司为对日IT服务核心供应商，主要业务包括对日软件开发服务和国内行业应用软件解决方案。目前对日软件开发服务占营收比例84%，国内行业应用解决方案占比16%，下游的最终客户大多为金融机构。公司20-22年受到疫情的影响，营收和利润水平有所承压，预期从23年开始将重回增长轨道。
- AI赋能，公司业务有望实现降本增效。公司目前已经成为了百度文心一言首批生态合作伙伴，在公司的业务中有诸多地方可以有效运用AI大模型技术从而实现降本增效，比如在对日软件开发项目中有35%左右的工作量用于设计，现在可以用AI技术代替，节省出来的人力可以用于接更多的项目。

大投行业务数字化转型业务解决方案

投行项目管理系统

该系统上采用了公司研发的AI技术，对投行项目和证券事务，取得了快速实现数字化转型的技术成果，应用于投行项目的信息管理、内容管理、项目风险管理、项目发行管理、项目审核和证券事务部设计等核心业务流程的快速变化及响应，减少业务展开过程中的错误，已经有效应用在30多家券商投行业务中。



投行尽调管理系统

在公司AI模型的基础上，通过大语言模型、神经网络，对于上市资料的大小不再敏感，可以在海量投行资料中快速检索、分析和总结信息，同时，非结构化数据可以结构化并输出结果，满足业务和交易需求，广泛应用于券商的投行尽调管理业务中。

文档智能校对系统

运用深度学习AI模型，依据行业规则和知识图谱建立的规则，智能识别文本错误和错别字，在解放人力的同时，也大幅提升审核的准确性。

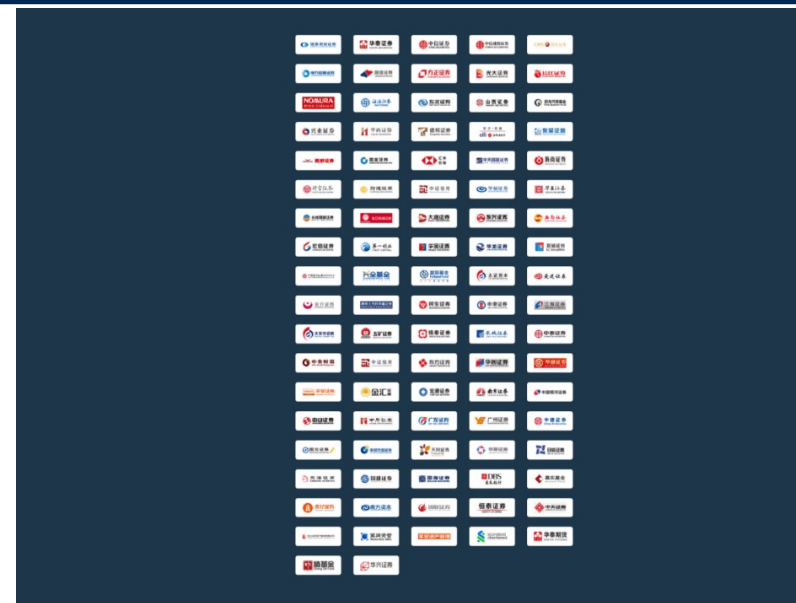


函证业务管理平台

通过AI识别和AI审核，实现一键识别、自动分析、自动审核、自动归档，实现投行函证业务全流程的数字化转型，提升管理效率和风险控制。

1. 通过AI识别函证的收发程序，从而为投行项目减少工作量，提高函证效率；
2. 通过AI审核函证的收发过程，从而降低函证风险；
3. 与地方监管系统、银行系统、企业系统、自动化的对接；
4. 通过每个环节建立AI模型，保证函证全流程一致；
5. 提供外部行业接口，方便与行业对接。

凌志软件国内客户



资料来源：凌志软件官网、国信证券经济研究所整理

资料来源：凌志软件官网、国信证券经济研究所整理

风险提示



- AI 技术商业化落地不及预期；
- 行业竞争加剧，技术迭代风险；
- 数据安全等政策不确定性；
- 贸易摩擦风险。

免责声明



国信证券投资评级		
类别	级别	定义
股票投资评级	买入	预计6个月内，股价表现优于市场指数20%以上
	增持	预计6个月内，股价表现优于市场指数10%-20%之间
	中性	预计6个月内，股价表现介于市场指数±10%之间
	卖出	预计6个月内，股价表现弱于市场指数10%以上
行业投资评级	超配	预计6个月内，行业指数表现优于市场指数10%以上
	中性	预计6个月内，行业指数表现介于市场指数±10%之间
	低配	预计6个月内，行业指数表现弱于市场指数10%以上

分析师承诺

作者保证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理解，通过合理判断并得出结论，力求独立、客观、公正，结论不受任何第三方的授意或影响；作者在过去、现在或未来未就其研究报告所提供的具体建议或所表述的意见直接或间接收取任何报酬，特此声明。

重要声明

本报告由国信证券股份有限公司（已具备中国证监会许可的证券投资咨询业务资格）制作；报告版权归国信证券股份有限公司（以下简称“我公司”）所有。 ， 本公司不会因接收人收到本报告而视其为客户。 未经书面许可，任何机构和个人不得以任何形式使用、复制或传播。任何有关本报告的摘要或节选都不代表本报告正式完整的观点，一切须以我公司向客户发布的本报告完整版本为准。

本报告基于已公开的资料或信息撰写，但我公司不保证该资料及信息的完整性、准确性。本报告所载的信息、资料、建议及推测仅反映我公司于本报告公开发布当日的判断，在不同时期，我公司可能撰写并发布与本报告所载资料、建议及推测不一致的报告。我公司不保证本报告所含信息及资料处于最新状态；我公司可能随时补充、更新和修订有关信息及资料，投资者应当自行关注相关更新和修订内容。 我公司或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、财务顾问或金融产品等相关服务。 本公司的资产管理部、自营部门以及其他投资业务部门可能独立做出与本报告意见或建议不一致的投资决策。

本报告仅供参考之用，不构成出售或购买证券或其他投资标的的要约或邀请。 在任何情况下，本报告中的信息和意见均不构成对任何个人的投资建议。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。 投资者应结合自己的投资目标和财务状况自行判断是否采用本报告所载内容和信息并自行承担风险，我公司及雇员对投资者使用本报告及其内容而造成的一切后果不承担任何法律责任。

证券投资咨询业务的说明

本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，是指从事证券投资咨询业务的机构及其投资咨询人员以下列形式为证券投资人或者客户提供证券投资分析、预测或者建议等直接或者间接有偿咨询服务的活动：接受投资人或者客户委托，提供证券投资咨询服务；举办有关证券投资咨询的讲座、报告会、分析会等；在报刊上发表证券投资咨询的文章、评论、报告，以及通过电台、电视台等公众传播媒体提供证券投资咨询服务；通过电话、传真、电脑网络等电信设备系统，提供证券投资咨询服务；中国证监会认定的其他形式。

发布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。



国信证券经济研究所

深圳

深圳市福田区福华一路125号国信金融大厦36层

邮编：518046 总机：0755-82130833

上海

上海浦东民生路1199弄证大五道口广场1号楼12楼

邮编：200135

北京

北京西城区金融大街兴盛街6号国信证券9层

邮编：100032