

证券研究报告

## AI系列深度报告（四）

# 大模型篇：大模型发展迈入爆发期，开启AI新纪元

计算机行业 强于大市（维持）

证券分析师

闫磊 投资咨询资格编号：S1060517070006

黄韦涵 投资咨询资格编号：S1060523070003

研究助理

王佳一 一般从业资格编号：S1060123070023

2024年8月15日

请务必阅读正文后免责条款

平安证券

## 核心摘要

- **技术：大模型发展呈现“规模定律”，Transformer为技术基座。**大模型（LLM）发展普遍呈现“规模定律”特征，即：模型的性能与模型的规模、数据集大小和训练用的计算量之间存在幂律关系。当前主流大模型普遍是基于Transformer模型进行设计的，Transformer由Encoder（编码器）和Decoder（解码器）两类组件构成，而OpenAI的GPT是Transformer演化树中Decoder-only架构的代表。我们纵观GPT系列模型的发展历程，从GPT-1到GPT-3，参数量从1.1亿大幅提升至1746亿，GPT-4非官方估计达到万亿参数（根据Semianalysis消息，GPT-4包含1.8万亿参数），实现性能的突破。GPT在众多大模型角逐中能够取得瞩目，技术角度上，主要源自其可拓展的训练架构与学习范式以及对于数据质量与数据规模的重视。然而，我们认为GPT的成功也并非全部源自技术性因素，OpenAI能够从早期众多的技术路线中识别到并坚定执行这条路线，这需要大模型团队足够的技术前瞻和定力。
- **市场：全球大模型竞争白热化，国产大模型能力对标GPT-3.5Turbo。**全球大模型竞争中，OpenAI、Anthropic、谷歌三大厂商为第一梯队，2024年以来三家大模型能力呈现互相追赶态势。开源大模型厂商中，Meta AI（Llama）、欧洲Mistral AI等厂商的大模型性能保持前列，图像生成大模型以Midjourney、Stable Diffusion、OpenAI的DALL·E为代表，视频生成以Runway的Gen、Pika和OpenAI的Sora为代表。2023年底国产大模型迈入爆发期，从参与者来看，国内AI大模型厂商大致可以分为互联网/科技公司、AI公司、学术/科研机构、行业专家品牌四类。从模型能力来看，根据SuperCLUE上半年最新的评测结果，国内绝大部分闭源模型已超过GPT-3.5Turbo，其基准上表现最好的国产大模型为阿里云的开源模型Qwen2-72B，超过众多国内外闭源模型，与GPT-4o仅差4分。
- **变现：API同质化、订阅实现难，Agent与MaaS探索破局之路。**OpenAI绝大多数收入来自API调用和ChatGPT订阅，但它的成功并不易复制，值得注意的是，一方面，API是较为普遍的变现方式，但由于大模型性能趋向同质化，全球API价格呈现下降趋势，另一方面，相较API，实现订阅的难度更高，由于AI-first应用目前在用户留存度和粘性上稍显落后，当前在AI产品市场还未出现真正意义上的Killer App。我们尝试总结了两条大模型变现的破局之路：1）C端：AI Agent是通往AGI与Killer App的钥匙。国内以互联网/科技公司、大模型厂商、企业服务SaaS类厂商为代表的越来越多的企业参与进AI Agent市场，产品形态既包括面向企业和开发者的Agent构建平台/框架，也包括服务于各个垂直行业的专业Agent。2）B端：面对企业级需求的快速增长，目前，微软云Azure、阿里云、华为云、腾讯云、百度云、京东云等云服务提供商都已经推出了MaaS（Model as a Service），帮助企业降低模型使用门槛。
- **算力：大模型发展催生海量算力需求，预计带来千亿美元市场规模。**大模型技术与应用发展催生海量算力需求，根据Jaime Sevilla等人的研究，2010-2022年在深度学习兴起背景下，机器学习训练算力增长了100亿倍。持续增长的大模型与AI产品研发需求同时也推升了科技巨头的资本支出，2024年2季度微软、Meta、谷歌、亚马逊四家公司合计资本支出为571亿美元，同比增长66%，环比1季度增长22%。在此背景下，大模型的持续迭代升级将为AI芯片及服务器发展提供强劲动力，我们对不同参数规模的大模型在训练与推理（问答场景）两个阶段的算力需求进行了测算，假设未来有100家大模型实现持续经营，最终测算得到AI服务器的市场规模为2301亿美元。如果再考虑问答场景之外的推理场景，AI服务器的市场规模将更加巨大。根据IDC数据，2023年的全球AI服务器市场规模是211亿美元，相比2023年全球AI服务器市场规模，大模型的持续迭代升级将为AI服务器市场带来广阔的市场空间。
- **投资建议：**当前，全球范围内大模型领域的竞争依然白热化，我国大模型厂商持续迭代升级算法能力，2023年底国产大模型市场迈入爆发期，根据SuperCLUE上半年最新的评测结果，国内绝大部分闭源模型已超过GPT-3.5Turbo，将有望加速国产大模型在各场景的应用落地。同时，大模型的持续迭代升级将为AI芯片及服务器发展提供强劲动力，根据我们测算，假设未来有100家大模型实现持续经营，在训练与推理（问答场景）两个阶段，大模型将为AI服务器带来约2301亿美元的市场空间。我们继续看好AI主题的投资机会，标的方面：1）算力方面，推荐浪潮信息、中科曙光、紫光股份、神州数码、海光信息、龙芯中科，建议关注寒武纪、景嘉微、软通动力；2）算法方面，推荐科大讯飞；3）应用场景方面，强烈推荐中科创达、恒生电子、盛视科技，推荐金山办公，建议关注万兴科技、同花顺、彩讯股份；4）网络安全方面，强烈推荐启明星辰。
- **风险提示：**1) AI算力供应链风险上升。2) 大模型产品的应用落地低于预期。3) 国产大模型算法发展可能不及预期。



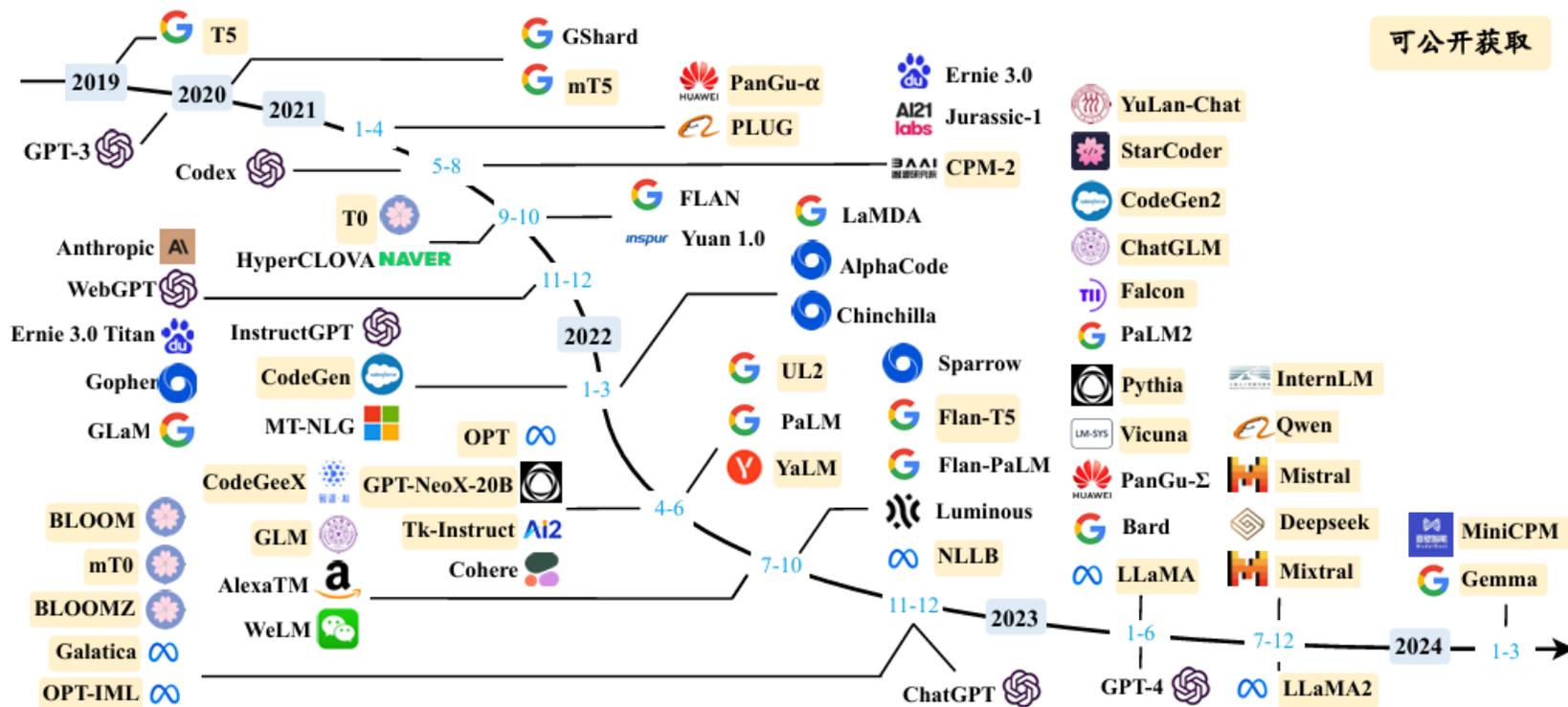
## 目录CONTENTS

- 技术：大模型发展呈现“规模定律”，Transformer为技术基座
- 市场：全球大模型竞争白热化，国产大模型能力对标GPT-3.5Turbo
- 变现：API同质化、订阅实现难，Agent与MaaS探索破局之路
- 算力：大模型发展催生海量算力需求，预计带来千亿美元市场规模
- 投资建议及风险提示

# 1.1 大模型“大力出奇迹”的背后：Scaling Law

大规模语言模型（Large Language Models, LLM）泛指具有超大规模参数或者经过超大规模数据训练所得到的语言模型。与传统语言模型相比，大语言模型的构建过程涉及到更为复杂的训练方法，进而展现出了强大的自然语言理解能力和复杂任务求解能力。

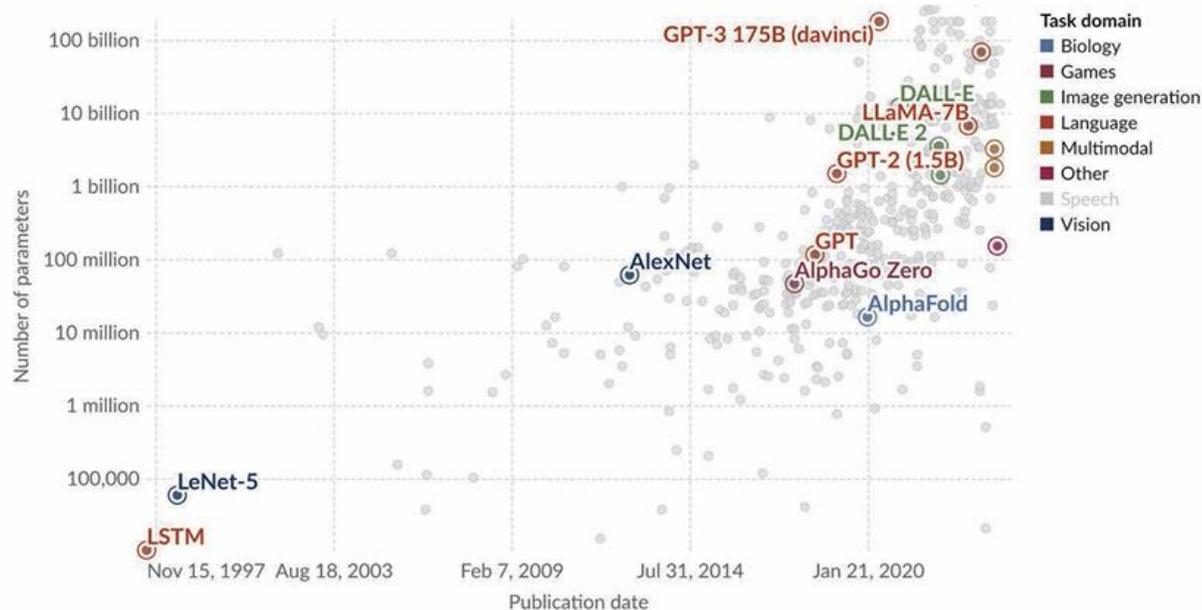
LLM发展时间线



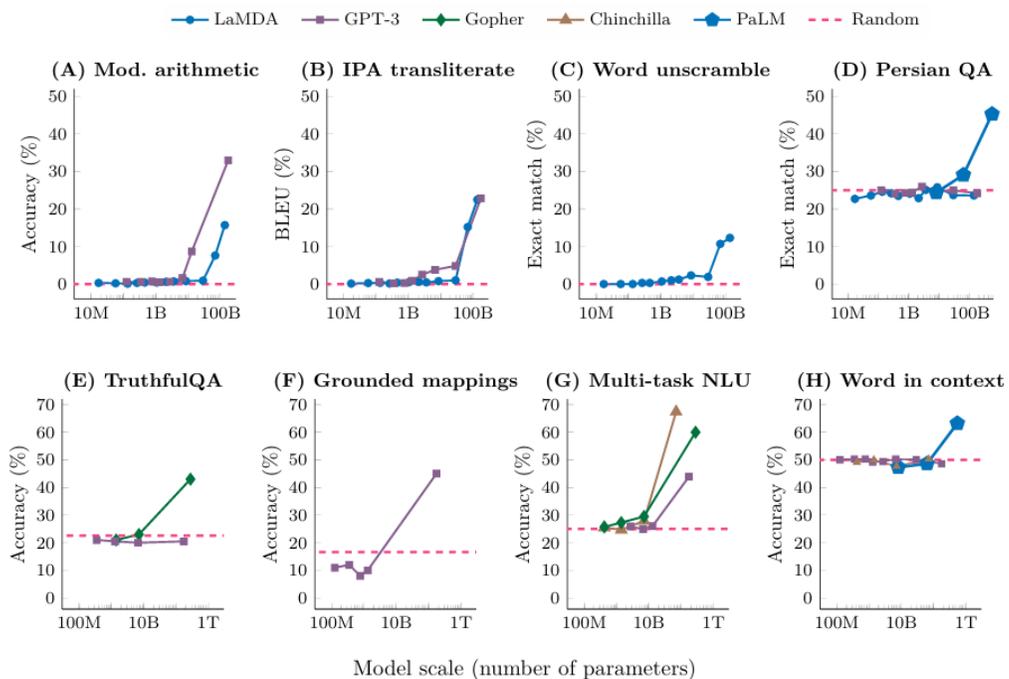
# 1.1 大模型“大力出奇迹”的背后：Scaling Law

大模型发展普遍呈现“规模定律”（Scaling Law）特征，即：模型的性能与模型的规模、数据集大小和训练用的计算量之间存在幂律关系，性能会随着这三个因素的指数增加而线性提高。大模型的参数规模远大于传统深度学习模型，传统模型参数量通常在数万至数亿之间，大模型的参数量则至少在亿级，并已发展到过万亿级的规模。如OpenAI的GPT-1到GPT-3，参数量从1.1亿大幅提升至1746亿，GPT-4非官方估计达到万亿参数（根据Semianalysis消息，GPT-4包含1.8万亿参数），实现性能的突破。根据Google论文，这种大模型具有但小模型不具有的能力通常被称为“涌现能力”（Emergent Abilities）。

## 主流大模型参数量



## 涌现能力：当模型扩展到一定规模时，模型的特定任务性能突然出现显著跃升的趋势

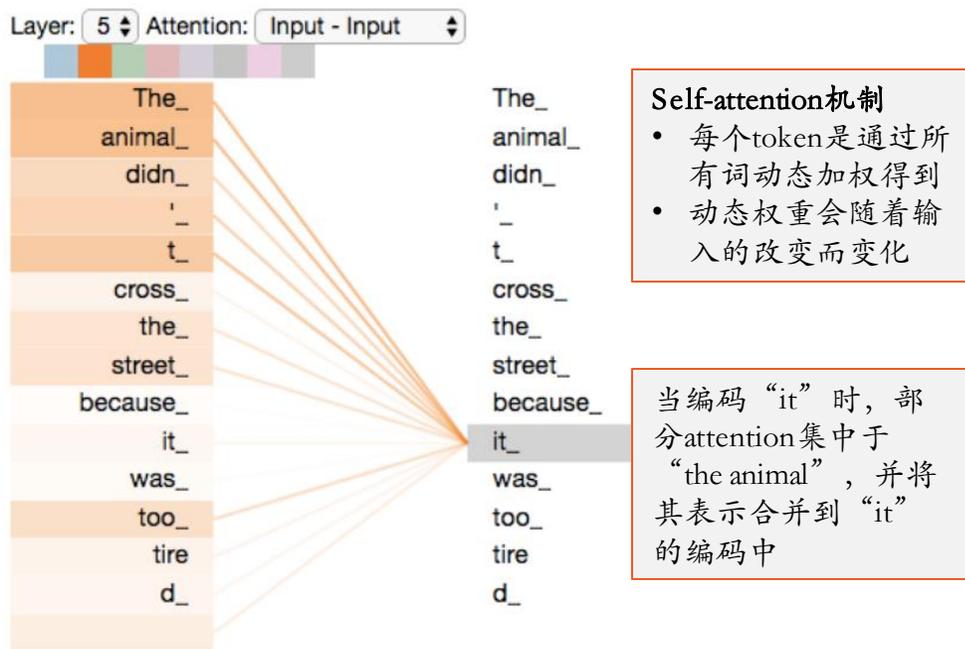


# 1.2 Transformer是LLM基座，核心优势在于Self-attention机制

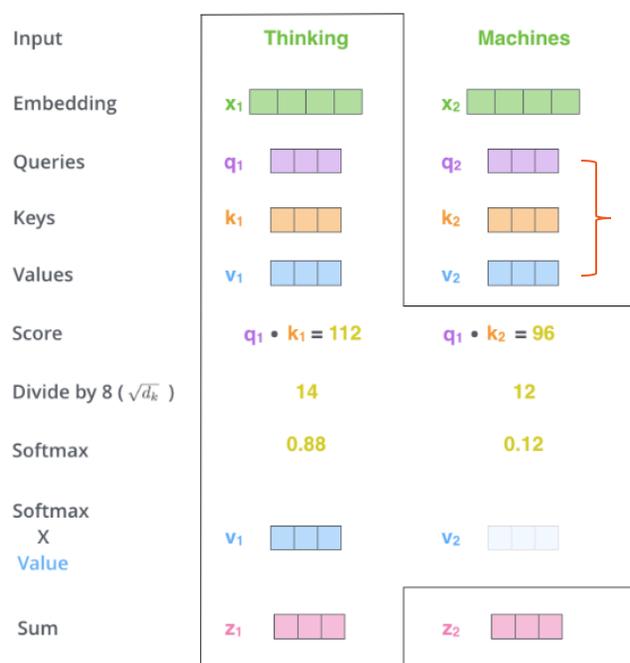
当前主流大模型普遍是基于Transformer模型进行设计的。Transformer模型在Google团队2017年论文《Attention Is All You Need》中被首次提出，Transformer的核心优势在于具有独特的自注意力（Self-attention）机制，能够直接建模任意距离的词元之间的交互关系，解决了循环神经网络（RNN）、卷积神经网络（CNN）等传统神经网络存在的长序列依赖问题。

相较于RNN，Transformer具有两个显著的优势。1) 处理长序列数据：RNN受限于循环结构，难以处理长序列数据。Self-attention机制能够同时处理序列中的所有位置，捕捉全局依赖关系，从而更准确地理解、表示文本含义。2) 实现并行化计算：RNN作为时序结构，需要依次处理序列中的每个元素，计算速度受到较大限制，而Transformer则可以一次性处理整个序列，大大提高了计算效率。

## Transformer通过Self-attention将“it”与“animal”联系起来



## Attention的计算



### 矩阵计算方法:

$$\text{softmax} \left( \frac{Q \times K^T}{\sqrt{d_k}} \right) \cdot V = Z$$

1、将输入序列x的每个元素（词或字）映射到一个向量表示。这些向量表示称为查询（Query）、键（Key）和值（Value）

2、分别计算 $q_1$ 与 $k_1$ 、 $k_2$ 的点积。点积结果表示两个元素之间的相关性。

3、使用softmax函数对点积结果进行归一化。

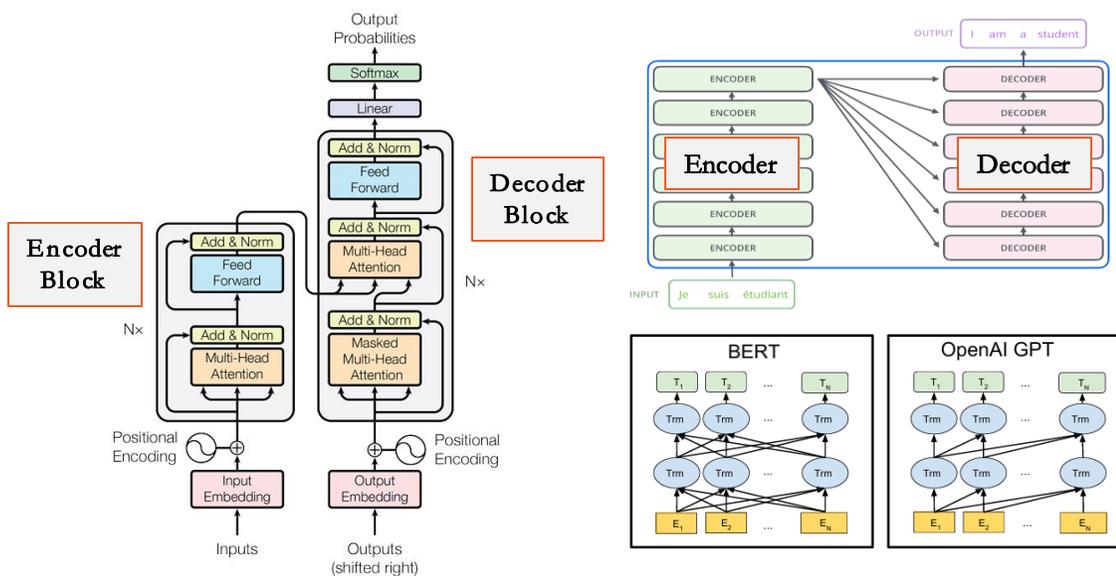
4、将归一化后的权重与对应的值v向量相乘，并对所有元素求和。

# 1.2 Transformer 组件：Encoder 重理解，Decoder 重生成

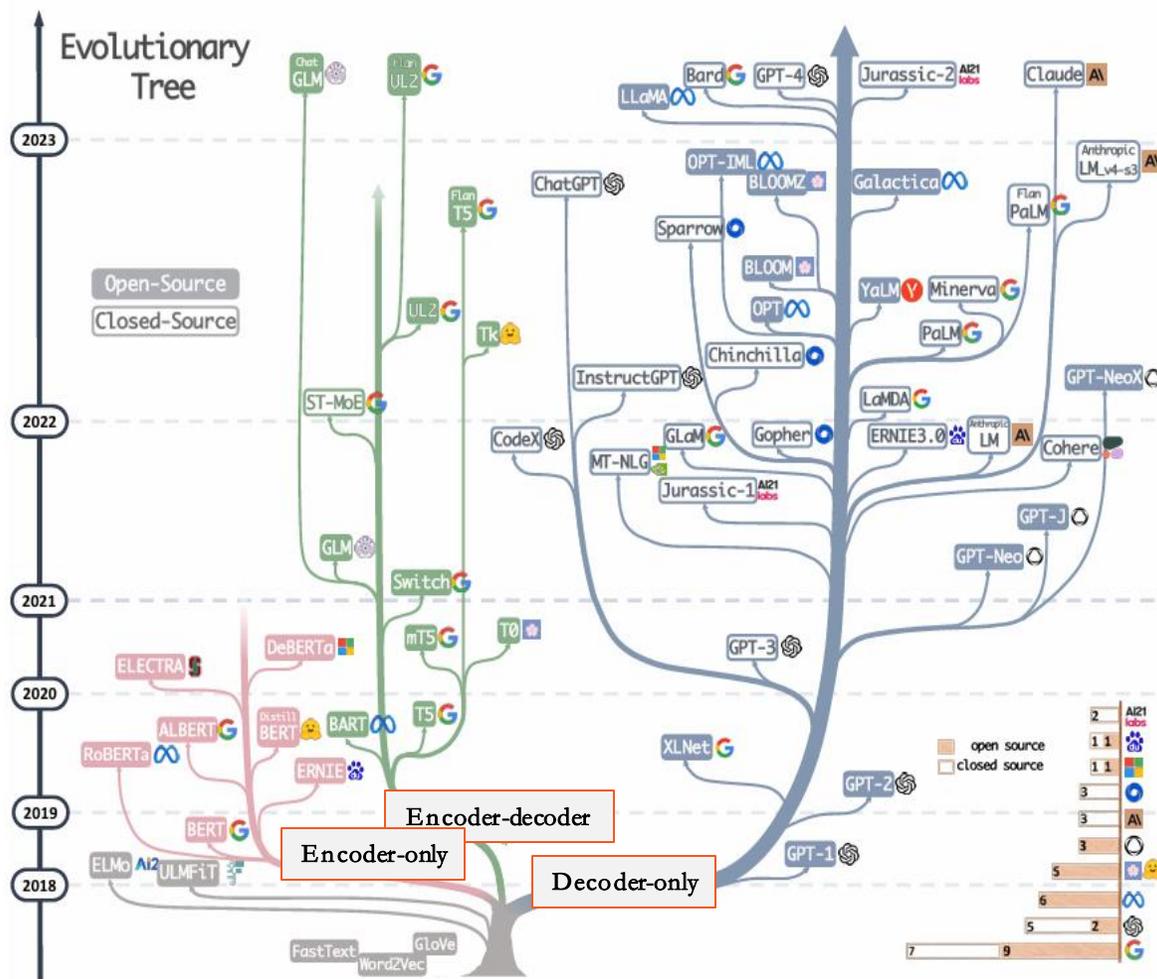
Transformer 由两类组件构成：Encoder（编码器）和 Decoder（解码器）。通常，Encoder 结构擅长从文本中提取信息以执行分类、回归等任务，而 Decoder 结构则专用于生成文本。

实际上，两类组件可以独立使用，当前主流大模型中，诞生了以 BERT 为代表的 Encoder-only 架构、以 T5 为代表的 Encoder-decoder 架构、以 GPT 为代表的 Decoder-only 架构的大规模预训练语言模型。

Transformer 模型网络架构



主流大模型网络架构演化树

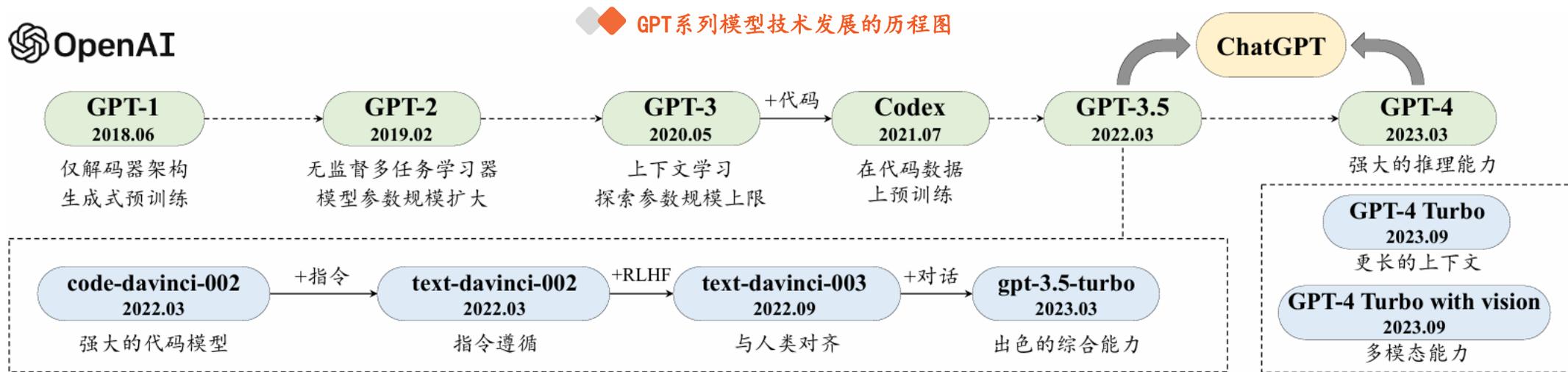


## 1.3 GPT系列模型技术发展历程回顾

从技术角度来看，结合中国人民大学《大语言模型》的观点，GPT在众多大模型角逐中能够取得瞩目，有以下几点值得注意——

- 1) **可拓展的训练架构与学习范式**：当谷歌2017年推出基于注意力机制的Transformer模型后，OpenAI团队能够迅速洞察到其作为大规模可扩展训练的理想架构的潜在优越性，最终将Transformer扩展到百亿、千亿甚至万亿参数规模，并且将预训练任务统一为通用学习范式。
- 2) **对于数据质量与数据规模的重视**：高质量数据、超大规模数据成为GPT成功的关键基础，比如，OpenAI将人类生成的对话数据和高质量的标注数据用于训练ChatGPT，使得ChatGPT在与人机对话测试中展现出了优秀能力。

然而，我们认为GPT的成功并不是全部源自技术性因素，Transformer、RLHF算法等等关键技术都并非OpenAI首创，并且也在被其他研究团队广泛使用，我们认为OpenAI能够从早期众多的技术路线中识别到并且坚定地去执行这条路线，这来自OpenAI团队足够的技术前瞻和策略定力。比如，OpenAI在早期GPT-2的论文中，就深入讨论了基于大规模文本预训练的通用任务学习范式；再比如GPT-3将参数规模极限拓展到175B，实际上，OpenAI两篇关于Scaling Law的论文都是在2020年发表的，这说明在前期已经进行了比较充分的实验探索。



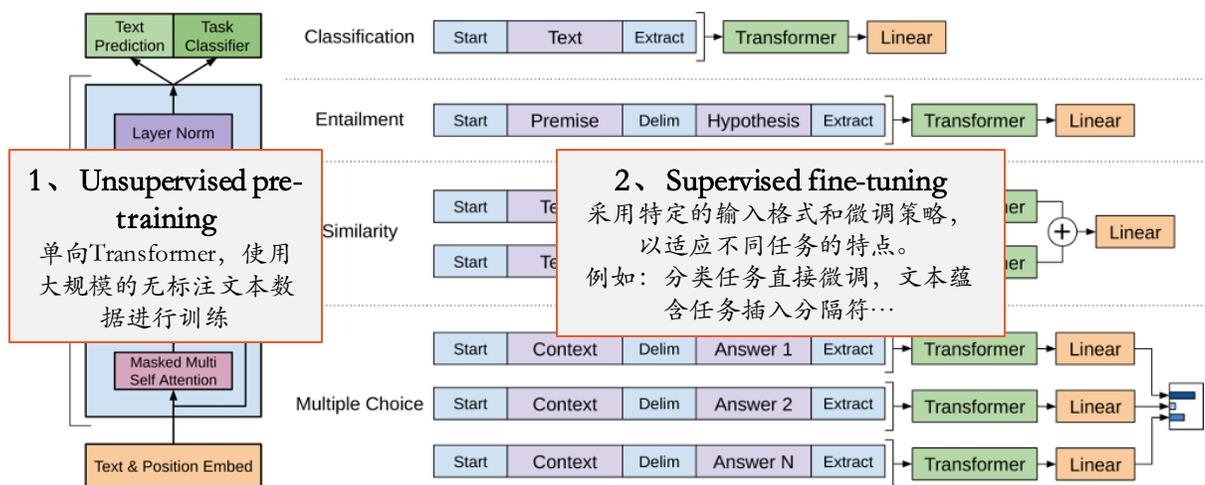
## 1.3 GPT 1~3: 采用Transformer架构, 参数规模拓展至千亿级

**GPT系列模型的技术演变 (GPT 1~3)**: 2017年, Google推出Transformer后, OpenAI迅速着手并于2018年推出的第一个GPT模型, 全称为Generative Pre-trained Transformer。GPT-1基于生成式、Decoder-only的Transformer架构开发, 由于参数规模相对较小, 模型尚缺乏通用任务求解能力, GPT-1采用Pre-training(预训练)+Fine-tuning(微调)的两阶段范式, 先通过单向Transformer预训练一个通用的模型, 然后在特定子任务上进行微调。

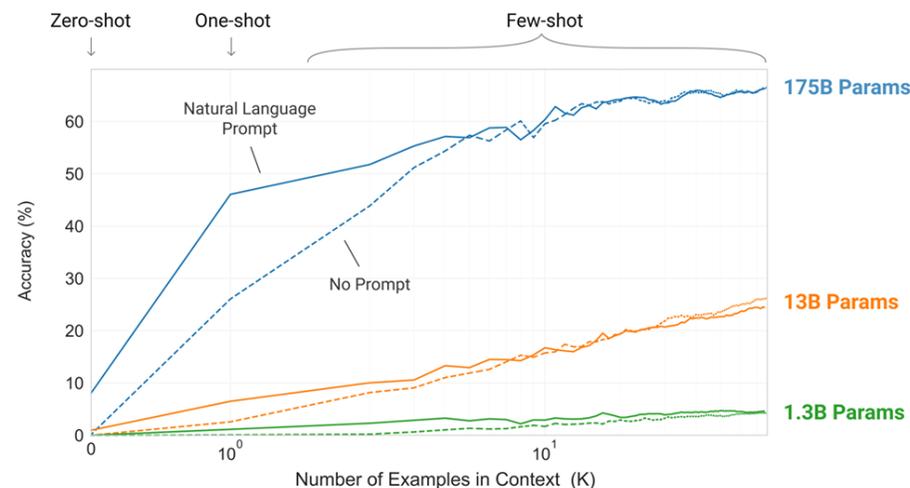
GPT-2沿用GPT-1类似架构, 并将参数规模扩大至1.5B, 并使用大规模网页数据集WebText进行预训练。与GPT-1不同的点在于, GPT-2旨在探索通过扩大模型参数规模来提升模型性能, 并且试图使用无监督预训练的语言模型来解决各种下游任务。

OpenAI经过较为充分的实验探索, 2020年, GPT-3将模型参数扩展到了175B, 较GPT-2提升100余倍, 验证了将神经网络扩展到超大规模可以带来大幅的模型性能提升。同时, GPT-3论文正式提出“上下文学习”, 建立了以提示学习方法为基础技术路线的任务求解范式。

### GPT-1的两阶段范式



### GPT-3证明参数扩展和增加提示学习的正效应



# 1.3 ChatGPT: 一举成为现象级应用，引入RLHF算法改进训练数据

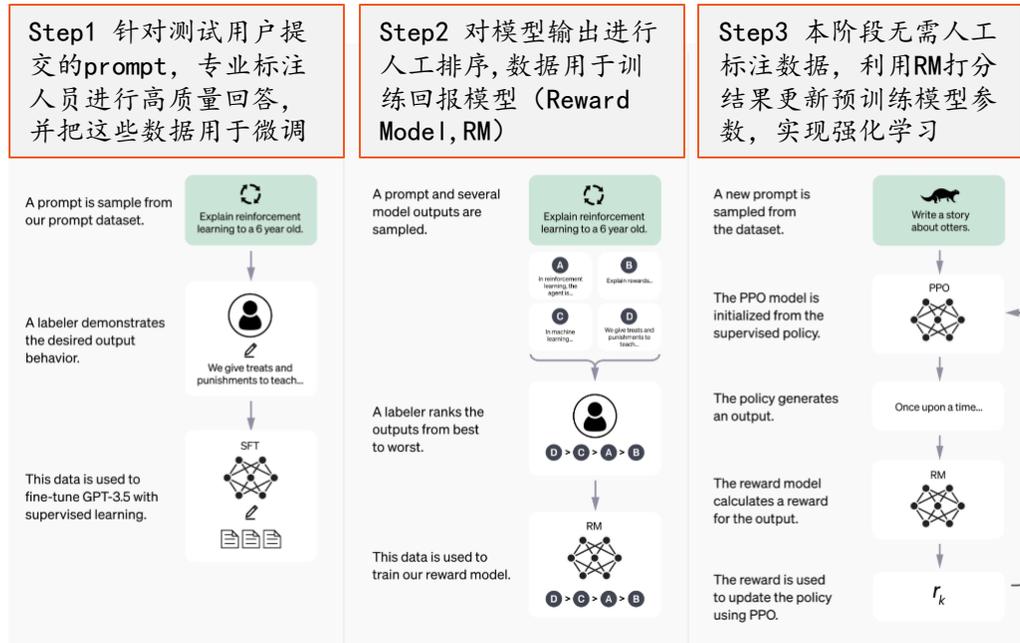
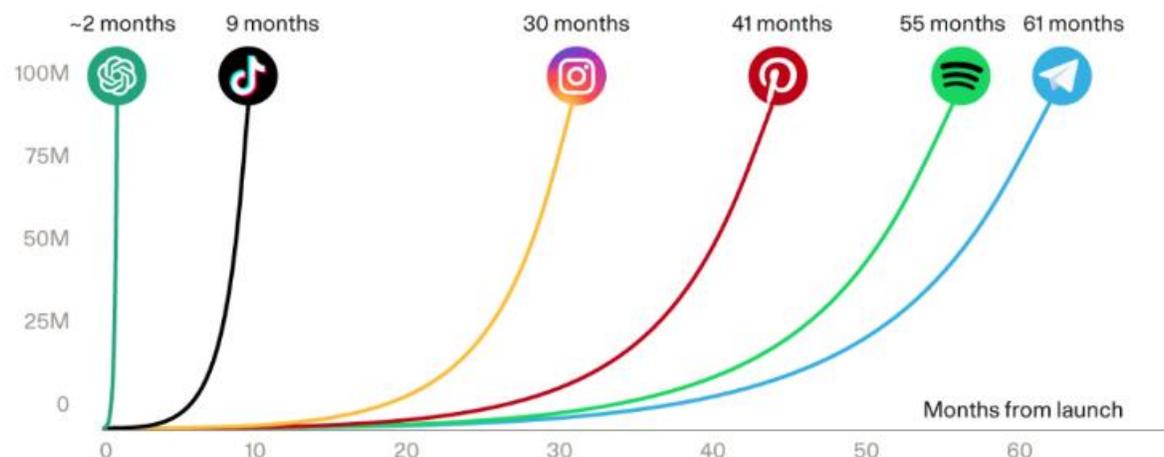
GPT系列模型的技术演变 (ChatGPT)：在GPT-3的基础上，OpenAI又通过代码训练、人类对齐、工具使用等技术对于模型性能不断升级，推出了GPT-3.5系列模型。2022年11月，ChatGPT正式上线，实现以对话形式解决多种任务，使得用户能够通过网络API体验到语言模型的强大功能。ChatGPT 仅用5天时间注册用户达到100万，约2个月注册用户达到1亿，成为AIGC领域的现象级应用。

OpenAI在官网文章中介绍，ChatGPT主要是沿用了2022年1月推出的InstructGPT。InstructGPT的核心技术是基于人类反馈的强化学习算法，即RLHF算法 (Reinforcement Learning from Human Feedback)，旨在改进模型与人类对齐的能力。具体实现上，人类标注人员扮演用户和代理进行对话，产生对话样本并对回复进行排名打分，将更好的结果反馈给模型，让模型从两种反馈模式——人类评价奖励和环境奖励中学习策略，对模型进行持续迭代式微调。

◆ ChatGPT达到1亿注册用户仅用约2个月

◆ RLHF算法的三个阶段

## Path to 100 Million Users (stylized)



# 1.3 GPT-4系列：能力跃升，增加多模态能力，最新版4o突破性价比

GPT系列模型的技术演变（GPT-4~GPT-4o）：继ChatGPT后，OpenAI于2023年3月发布GPT-4，它首次将GPT系列模型的输入由单一文本模态扩展到了图文双模态。GPT-4在解决复杂任务方面的能力显著强于GPT-3.5，在一系列面向人类的考试中都获得了非常优异的结果。

基于GPT-4，OpenAI在2023年9月进一步发布了GPT-4V，重点关注GPT-4视觉能力的安全部署。GPT-4V在多种应用场景中表现出了强大的视觉能力与综合任务解决能力。2023年11月，OpenAI在开发者大会上发布GPT-4 Turbo，引入了一系列技术升级，如：将模型内部知识库更新至2023年4月，将上下文长度提升至128K，价格更便宜，引入若干新的功能（如函数调用、可重复输出等）。

今年5月14日，OpenAI春季发布会，发布了新版旗舰模型GPT-4o。GPT-4o将文本、音频和视觉集成到一个模型中，提供更快响应时间、更好的推理能力以及在非英语语言中的更佳表现，不仅在传统文本能力上与GPT-4 Turbo性能相当，还在API方面更快速，价格便宜50%。与GPT-4 Turbo相比，GPT-4o速度提高了2倍，限制速率提高了5倍，目前的上下文窗口为128k，模型知识截止日期为2023年10月。

## GPT-4允许图片输入

Example of GPT-4 visual input:

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



Source: <https://www.reddit.com/r/hmm/comments/sbabi6/hmm/>

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

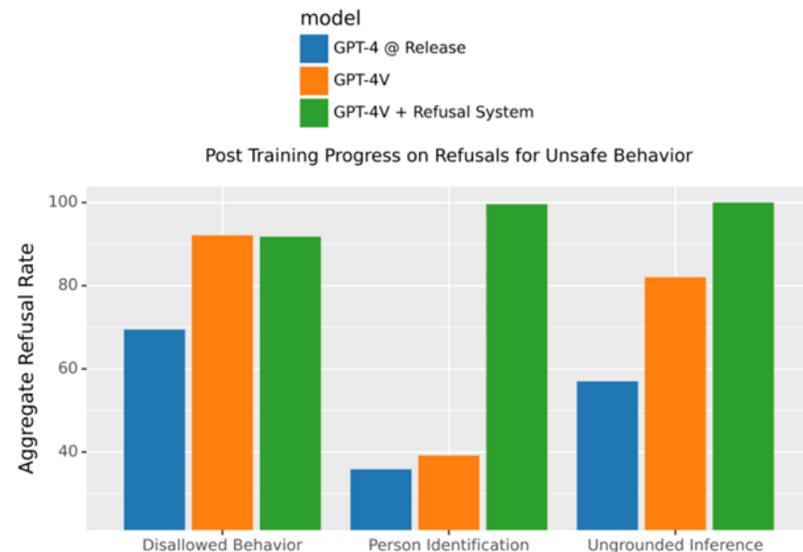
Panel 1: A smartphone with a VGA connector (a large, blue, 15-pin connector typically used for computer monitors) plugged into its charging port.

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

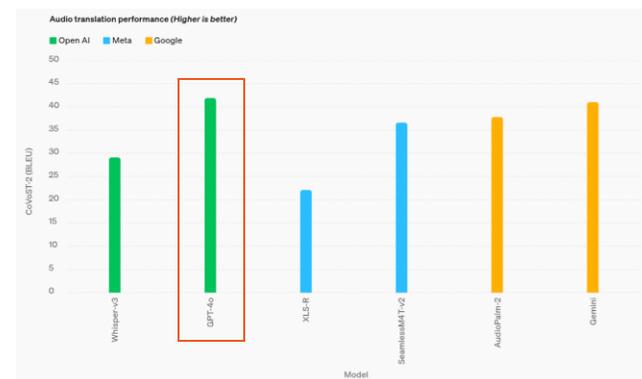
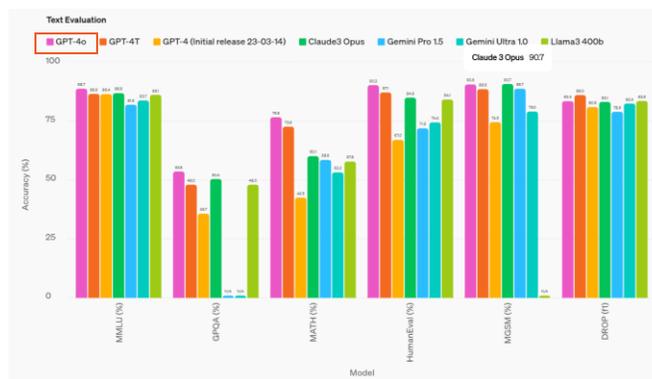
Panel 3: A close-up of the VGA connector with a small Lightning connector (used for charging iPhones and other Apple devices) at the end.

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

## GPT-4V采取与视觉输入相关的风险评估措施



## GPT-4o文本能力与GPT-4 Turbo相当，音频翻译表现在同业突出





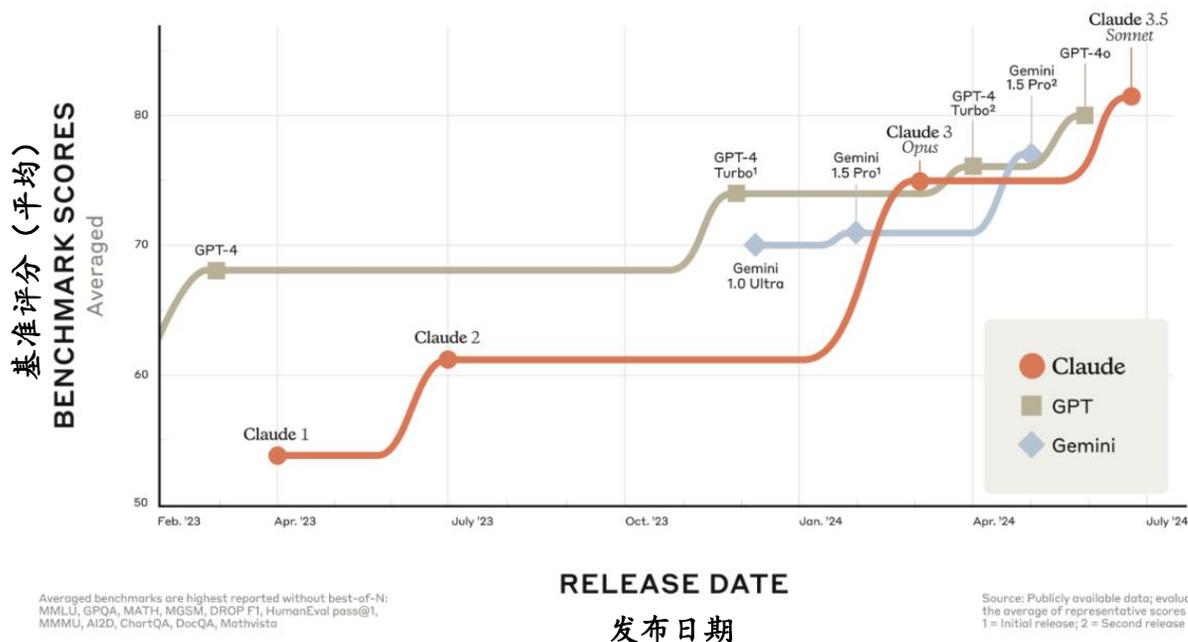
## 目录CONTENTS

- 技术：大模型发展呈现“规模定律”，Transformer为技术基座
- 市场：全球大模型竞争白热化，国产大模型能力对标GPT-3.5Turbo
- 变现：API同质化、订阅实现难，Agent与MaaS探索破局之路
- 算力：大模型发展催生海量算力需求，预计带来千亿美元市场规模
- 投资建议及风险提示

## 2.1 海外大模型：通用大模型竞争白热化，闭源LLM三足鼎立

全球大模型竞争中，OpenAI、Anthropic、谷歌三大厂商为第一梯队，OpenAI先发推出GPT-4，在2023年基本稳定在行业龙头地位，而Anthropic凭借Claude、谷歌凭借Gemini后发，可以看到，2024年以来，三家大模型能力呈现互相追赶态势。开源大模型厂商中，Meta AI (Llama)、欧洲Mistral AI (Mistral)、Google (Gemma) 等厂商的大模型性能保持前列。此外，伴随Sora推出以及Pika的出圈，图像、视频生成领域的超预期进展获得极大关注，全球图像生成大模型以Midjourney、Stable Diffusion、OpenAI的DALL·E为代表，视频生成以Runway的Gen、Pika和OpenAI的Sora为代表。

三家大模型厂商追赶模型能力



海外大模型厂商梯队及模型最新版本

闭源 LLM	团队	OpenAI	ANTHROPIC	Google DeepMind
	模型	GPT-4o	Claude 3.5 Sonnet	GEMINI 1.5
开源 LLM	团队	Meta AI	MISTRAL AI	Google DeepMind
	模型	Llama 3	Mistral NeMo	Gemma 2
图像生成	团队	Midjourney	stability ai	OpenAI
	模型	MidJourney V6	Stable Diffusion 3	DALL-E 3
视频生成	团队	runway	Pika Labs	OpenAI
	模型	Gen-3 Alpha	Pika	Sora

## 2.1 OpenAI: 上半年重磅发布Sora, GPT-4o取得性能与实用性双突破

OpenAI发布文生视频大模型Sora, 在全球视频大模型领域取得里程碑式进展。今年2月15日, OpenAI在官网正式发布Sora, 根据OpenAI官网介绍, Sora可以在保持视觉质量和遵循用户的文本提示的情况下, 生成长达1分钟的视频, 遥遥领先于以往的视频生成时长。

GPT-4o实现性能与实用性双突破, 有望加速大模型应用落地。5月14日, OpenAI在春季发布会上推出GPT-4o, 并表示将免费提供给所有用户使用。GPT-4o可接受文本、音频和图像的任意组合作为输入、输出, 在英语文本和代码方面的性能可对标GPT-4 Turbo, 同时在API中也更快且便宜50%。根据OpenAI官网信息, 在GPT-4o之前, 使用语音模式与ChatGPT对话, GPT-3.5/GPT-4的平均延迟分别为2.8/5.4秒。而GPT-4o可以在短至232毫秒的时间内响应音频输入, 平均时长为320毫秒, 与人类在一次谈话中的响应时间相似。

7月18日, OpenAI正式推出了GPT-4o mini, 将取代ChatGPT中的旧模型GPT-3.5 Turbo, 向ChatGPT的免费用户、ChatGPT Plus和团队订阅用户开放。OpenAI表示, GPT-4o mini的成本为每百万输入标记(token) 15美分和每百万输出标记60美分, 比GPT-3.5 Turbo便宜超过60%。

◆ Sora根据文字指令生成的视频画面



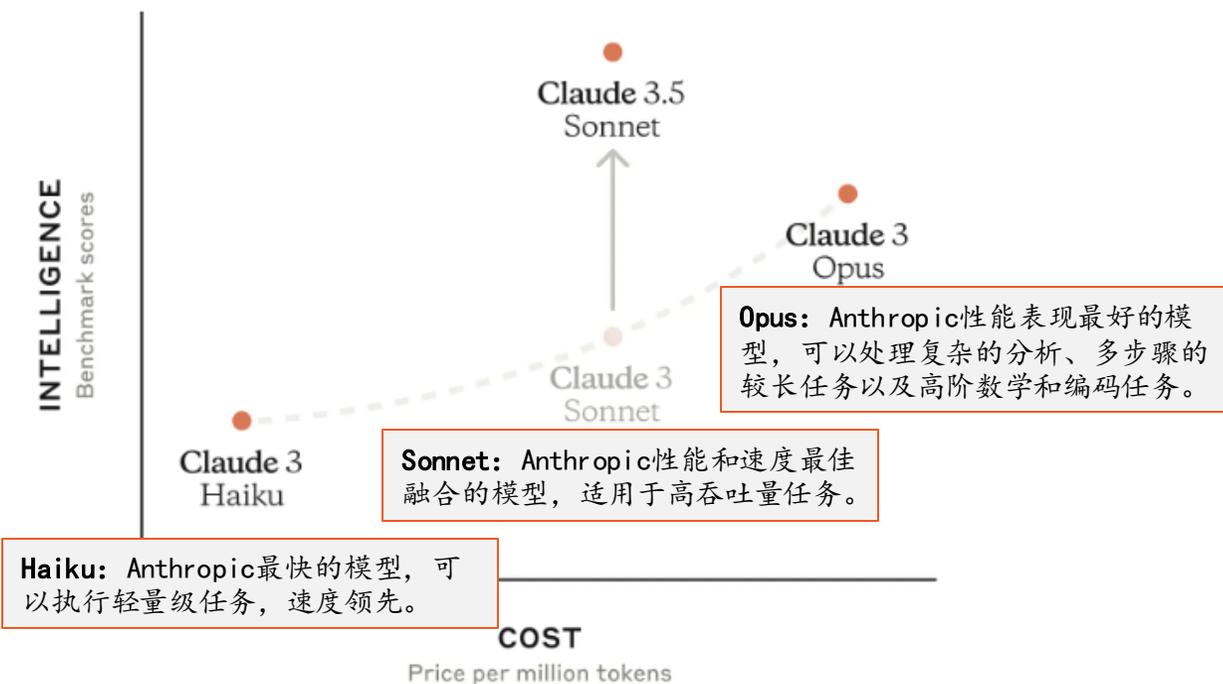
◆ OpenAI 各代GPT比较

GPT各代	发布时间	训练数据	模型参数	序列长度 (Token)	输入	输出
GPT-1	2018年6月	约5GB	1.17亿	512	文本	文本
GPT-2	2019年2月	40GB	15亿	1024	文本	文本
GPT-3	2020年5月	45TB	1746亿	2048	文本	文本
GPT-3.5	2022年11月	百TB级别	-	4096/16385	文本	文本
GPT-4	2023年3月	百TB级别	万亿级别	8192/32768	文本、图像	文本、图像
GPT-4 Turbo	2023年11月	百TB级别	万亿级别	128000	多模态	多模态
GPT-4o	2024年5月	百TB级别	万亿级别	128000	多模态	多模态

## 2.1 Anthropic: 推出Claude 3家族, Claude 3.5 Sonnet性能追赶GPT-4o

Claude 3.5 Sonnet在绝大多数基准评估中都超越了竞品大模型。今年3月4日, Anthropic发布Claude 3系列模型, 包含Opus、Sonnet和Haiku, Opus代表Anthropic最高级、最智能的模型, Sonnet代表中等级别的模型, 在性能和成本效益之间取得平衡, Haiku代表入门级别或最基础的快速模型, 其中, Claude 3 Opus为Claude 3系列模型的最强版本, 根据官方资料其性能全面超过了GPT-4。值得注意的是, Claude 3全系列模型具有与其他领先模型同等的复杂视觉功能, 可以处理各种视觉格式, 包括照片、图表、图形和技术图表。6月21日, Anthropic发布了全新大模型 Claude 3.5 Sonnet, 号称为“迄今为止最智能的模型”。据Anthropic介绍, Claude 3.5 Sonnet在绝大多数基准评估中都超越了竞品大模型和自家前代最强 Claude 3 Opus, 与此同时, 运行速度、成本与自家前代 Claude 3 Sonnet相当。

### ◆ Claude 3家族Haiku、Sonnet、Opus三类模型定位



### ◆ Claude 3.5 Sonnet多项基准评分优于GPT-4o

	Claude 3.5 Sonnet	Claude 3 Opus	GPT-4o	Gemini 1.5 Pro	Llama-400b (early snapshot)
Graduate level reasoning <i>GPQA, Diamond</i>	59.4%* 0-shot CoT	50.4% 0-shot CoT	53.6% 0-shot CoT	—	—
Undergraduate level knowledge <i>MMLU</i>	88.7%** 5-shot	86.8% 5-shot	—	85.9% 5-shot	86.1% 5-shot
	88.3% 0-shot CoT	85.7% 0-shot CoT	88.7% 0-shot CoT	—	—
Code <i>HumanEval</i>	92.0% 0-shot	84.9% 0-shot	90.2% 0-shot	84.1% 0-shot	84.1% 0-shot
Multilingual math <i>MGSM</i>	91.6% 0-shot CoT	90.7% 0-shot CoT	90.5% 0-shot CoT	87.5% 8-shot	—
Reasoning over text <i>DROP, F1 score</i>	87.1 3-shot	83.1 3-shot	83.4 3-shot	74.9 Variable shots	83.5 3-shot Pre-trained model
Mixed evaluations <i>BIG-Bench-Hard</i>	93.1% 3-shot CoT	86.8% 3-shot CoT	—	89.2% 3-shot CoT	85.3% 3-shot CoT Pre-trained model
Math problem-solving <i>MATH</i>	71.1% 0-shot CoT	60.1% 0-shot CoT	76.6% 0-shot CoT	67.7% 4-shot	57.8% 4-shot CoT
Grade school math <i>GSM8K</i>	96.4% 0-shot CoT	95.0% 0-shot CoT	—	90.8% 11-shot	94.1% 8-shot CoT

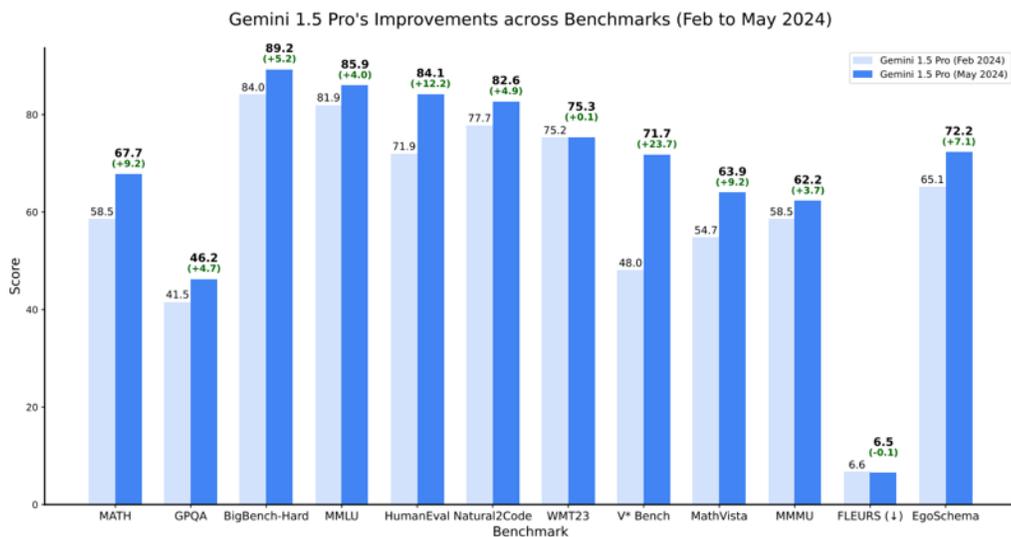
## 2.1 谷歌：全面升级Gemini系列，发布视频生成模型Veo

**谷歌大模型全线升级。**继2023年12月谷歌推出其规模最大、功能最强的多模态大模型Gemini系列之后，今年2月15日，谷歌发布Gemini 1.5，在版本中首个登场的多模态通用模型Gemini 1.5 Pro，将稳定处理上下文的上限扩大至100万tokens。5月14日，在谷歌IO开发者大会上，谷歌大模型迎来多项升级，基础大模型方面，Gemini 1.5 Pro将上下文窗口长度从之前的100万tokens进一步扩展至200万tokens，谷歌首席执行官Pichai称这是目前市场处理上下文长度规模最大的基础大模型。

**多模态大模型方面，**谷歌发布了视频生成模型Veo以及文生图模型Imagen3，Veo模型对标Sora，能够根据文字、图片和视频的提示生成长度超过1分钟、分辨率最高1080p的视频；Imagen3文生图模型是Imagen系列的升级版，从细节拟真度来看对标Midjourney v6。

◆ 5月，Gemini 1.5 Pro性能相比2月份有明显提升

◆ 5月，谷歌发布视频生成模型Veo，推出文生图Imagen最新产品Imagen 3



## 2.1 Meta: 推出最强开源大模型Llama3.1, 405B模型能力赶超领先模型

今年4月18日, Meta推出强大的开源人工智能模型Llama 3, 发布包括8B和70B参数的两个版本, 作为Llama2的重大升级。Meta称, 目前正在开发的最大的模型是400B参数, 将在未来几个月内推出, 英伟达科学家Jim Fan认为, Llama3 400B将成为一个分水岭, 即社区将获得开源重量级的GPT-4模型, 它将改变许多研究工作和草根创业公司的计算方式。

7月23日, Llama 3.1正式发布, 405B版本在150多个基准测试集上的表现追平或超越现有领先的基础模型, 包括GPT-4、GPT-4o和Claude 3.5 Sonnet。除了与闭源模型相比具有显著更好的成本/性能比之外, 405B模型的开放性将使其成为微调和蒸馏更小模型的优质选择。此外, Meta也推出了8B和70B模型的升级版本, 能力与同等参数下的顶尖模型基本持平, 与具有相似参数数量的开闭源模型构成强竞争力。

### Llama 3.1 405B在多项基准上超越现有领先基础模型

Category Benchmark	Llama 3.1 405B	Nemotron 4 340B Instruct	GPT-4 (0-shot)	GPT-4 Omni	Claude 3.5 Sonnet
General					
MMLU (0-shot, CoT)	88.6	78.7 (non-CoT)	85.4	88.7	88.3
MMLU PRO (5-shot, CoT)	73.3	62.7	64.8	74.0	77.0
IFEval	88.6	85.1	84.3	85.6	88.0
Code					
HumanEval (0-shot)	89.0	73.2	86.6	90.2	92.0
MBPP EvalPlus (base) (0-shot)	88.6	72.8	83.6	87.8	90.5
Math					
GSM8K (8-shot, CoT)	96.8	92.3 (0-shot)	94.2	96.1	96.4 (0-shot)
MATH (0-shot, CoT)	73.8	41.1	64.5	76.6	71.1
Reasoning					
ARC Challenge (0-shot)	96.9	94.6	96.4	96.7	96.7
GPQA (0-shot, CoT)	51.1	-	41.4	53.6	59.4
Tool use					
BFCL	88.5	86.5	88.3	80.5	90.2
Nexus	58.7	-	50.3	56.1	45.7
Long context					
ZeroSCROLLS/QUALITY	95.2	-	95.2	90.5	90.5
InfiniteBench/En.MC	83.4	-	72.1	82.5	-
NIH/Multi-needle	98.1	-	100.0	100.0	90.8
Multilingual					
Multilingual MGSM (0-shot)	91.6	-	85.9	90.5	91.6

### Llama 3.1 8B和70B的较小模型性能优于对标开闭源模型

Category Benchmark	Llama 3.1 8B	Gemma 2 9B IT	Mistral 7B Instruct	Llama 3.1 70B	Mixtral 8x22B Instruct	GPT 3.5 Turbo
General						
MMLU (0-shot, CoT)	73.0	72.3 (5-shot, non-CoT)	60.5	86.0	79.9	69.8
MMLU PRO (5-shot, CoT)	48.3	-	36.9	66.4	56.3	49.2
IFEval	80.4	73.6	57.6	87.5	72.7	69.9
Code						
HumanEval (0-shot)	72.6	54.3	40.2	80.5	75.6	68.0
MBPP EvalPlus (base) (0-shot)	72.8	71.7	49.5	86.0	78.6	82.0
Math						
GSM8K (8-shot, CoT)	84.5	76.7	53.2	95.1	88.2	81.6
MATH (0-shot, CoT)	51.9	44.3	13.0	68.0	54.1	43.1
Reasoning						
ARC Challenge (0-shot)	83.4	87.6	74.2	94.8	88.7	83.7
GPQA (0-shot, CoT)	32.8	-	28.8	46.7	33.3	30.8
Tool use						
BFCL	76.1	-	60.4	84.8	-	85.9
Nexus	38.5	30.0	24.7	56.7	48.5	37.2
Long context						
ZeroSCROLLS/QUALITY	81.0	-	-	90.5	-	-
InfiniteBench/En.MC	65.1	-	-	78.2	-	-
NIH/Multi-needle	98.8	-	-	97.5	-	-
Multilingual						
Multilingual MGSM (0-shot)	68.9	53.2	29.9	86.9	71.1	51.4

## 2.2 国产大模型：迈入爆发期，模型能力追赶GPT-4 Turbo

自2022年11月底ChatGPT发布以来，AI大模型在全球范围内掀起了有史以来规模最大的人工智能浪潮，国内学术和产业界也在抓紧追赶突破。SuperCLUE将国内大模型发展大致分为三个阶段，1) 准备期：2022年11月ChatGPT发布后，国内产学研迅速形成大模型共识。2) 成长期：2023年初，国内大模型数量和质量开始逐渐增长。3) 爆发期：2023年底至今，各行各业开源闭源大模型层出不穷，形成百模大战的竞争态势。

### 国内AI大模型2023-2024年关键进展



## 2.2 国产大模型：迈入爆发期，模型能力追赶GPT-4 Turbo

从参与者来看，目前我国AI大模型厂商大致可以分为四类，1) 互联网/科技公司：以百度、阿里、腾讯、字节、快手、华为等为代表。2) AI公司：以智谱AI、昆仑万维、科大讯飞、商汤科技为代表的专注于AI研发与应用的科技公司。3) 学术、科研机构：包括清华、北大、复旦、中科院等国内一流高校，以及智源研究院、IDEA研究院等科研机构。4) 行业专家品牌：以月之暗面（Moonshot AI）、百川智能、MiniMax等由AI专家创业成立的公司为代表。

国内大模型厂商分为四大流派

国内大模型全景图



资料来源：各公司官网，SuperCLUE，平安证券研究所

## 2.2 国产大模型：迈入爆发期，模型能力追赶GPT-4 Turbo

从模型能力来看，根据SuperCLUE最新的评测结果，以开源Llama-3-70B、Llama-3-8B的模型能力为分界线，国内大模型可以形成三大梯队。

可以看到，国内大模型上半年发展尤为迅速，

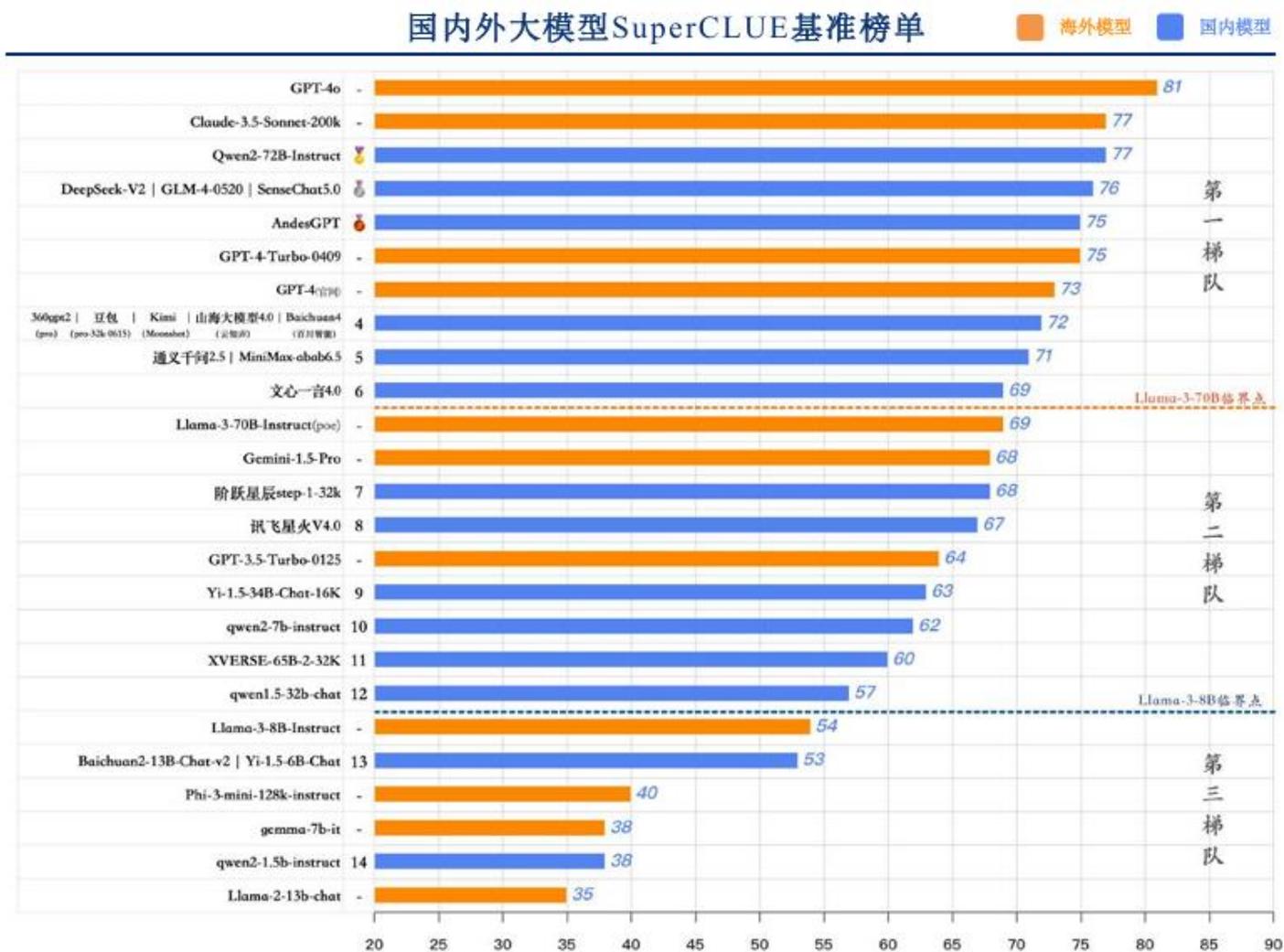
1) 开源模型Qwen2-72B在SuperCLUE基准中表现非常出色，超过众多国内外闭源模型，与Claude-3.5持平，与GPT-4o仅差4分。

2) 此外，还有4个国内大模型（深度求索DeepSeek-V2、智谱GLM-4、商汤SenseChat5.0、OPPO的AndesGPT）超过GPT-4-Turbo-0409。

3) 除了互联网大厂和AI公司，专家创业团队如Baichuan4、Kimi、MiniMax-abab6.5均有超过70分的表现，位列国内大模型第一梯队。

4) 国内绝大部分闭源模型已超过GPT-3.5 Turbo-0125。

全球大模型SuperCLUE通用能力测评榜单





## 目录CONTENTS

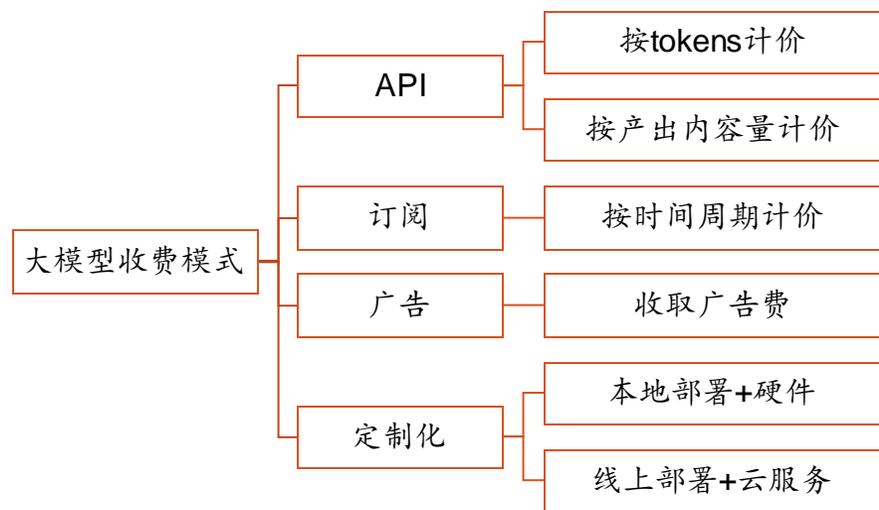
- 技术：大模型发展呈现“规模定律”，Transformer为技术基座
- 市场：全球大模型竞争白热化，国产大模型能力对标GPT-3.5Turbo
- 变现：API同质化、订阅实现难，Agent与MaaS探索破局之路
- 算力：大模型发展催生海量算力需求，预计带来千亿美元市场规模
- 投资建议及风险提示

## 3.1 大模型商业形态多元，B端变现模式更清晰

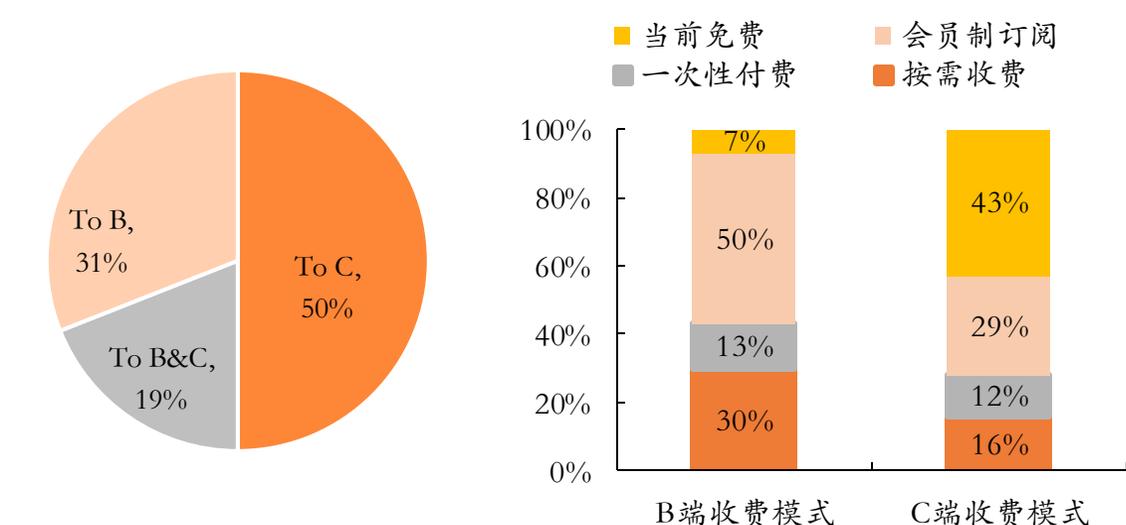
我们认为大模型收费模式可以总结为API、订阅、广告、定制化四种。首先，大模型最常见的商业模式基本遵循软件行业的SaaS（Software as a Service），通用大模型通常会采取API模式，根据tokens/调用次数/产出内容量等计价，大模型形成AI产品后，可以采用订阅制，按月/季/年向用户收取使用费。同时，AI产品若具备一定程度的流量价值，能够吸引商家投放广告，从而收取广告费。此外，服务内容可以不限于大模型本身，针对付费能力强的企业客户，部分厂商会提供软硬件一体的定制化解决方案，我们称之为MaaS（Model as a Service）。

从AI产品商业化程度来看，B端变现模式更加清晰，C端大多数产品仍然以免费为主。根据量子位智库，面向B端的AI产品从通用场景到垂直赛道分布较均匀，收入模式以会员订阅和按需付费为主，商业模式较为清晰，虽然（纯B端）市场占比只有31%，但80%以上的产品均能实现营收。C端AI产品以智能助手以及图像生成类的生产力工具为主，虽然用户量大（纯C端占比50%以上），但近50%的产品当前仍未有明确的收入模式，**以免费为主**。

大模型常见收费模式



AI应用面向用户群体结构与收费模式

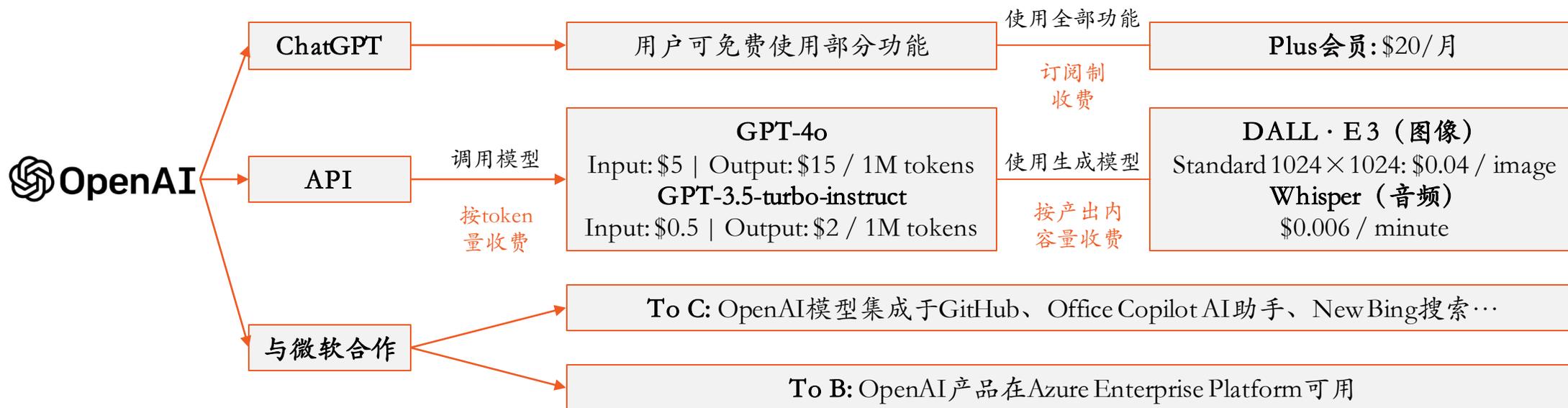


## 3.1 OpenAI以ChatGPT订阅、API为主要营收来源

纵观海内外的大模型厂商，OpenAI确立了最为经典的大模型商业模式，主要包括ChatGPT订阅、API调用、战略合作三种营收方式。1) ChatGPT订阅：OpenAI向C端提供生产力解放工具ChatGPT，并以付费订阅模式变现，针对ChatGPT Plus会员收取每月20美元的订阅费。2) API调用：而对于模型使用灵活性要求更高的用户，提供API服务，基于模型的调用量（tokens）或者产出内容量（如图片张数、时长）收费。3) 战略合作：此外，公司与微软建立了密切合作关系，ToC，OpenAI模型能力嵌入微软的生成式AI工具，如GitHub、Office、Bing等；ToB，微软Azure是OpenAI的独家云服务提供商，Azure全球版企业客户可以在平台上直接调用OpenAI模型。

OpenAI绝大多数收入来自前两项，ChatGPT订阅和API调用。据今年6月The Information的报道，Sam Altman在员工全体会议上表示2024年OpenAI年化收入有望达到34亿美元，其中，OpenAI的大部分收入（约32亿美元）来自其聊天机器人ChatGPT的订阅费以及让软件开发人员通过应用程序编程接口访问其模型的费用。OpenAI还有望通过Microsoft Azure提供其AI模型的访问权限，从而创造约2亿美元的收入。

### OpenAI商业模式的简单拆解



## 3.1 全球API定价呈现下降趋势

1) API: 是大模型厂商最为普遍的营收模式, 但我们也注意到由于大模型性能趋向同质化, 全球API价格呈现下降趋势。今年5月, 作为行业风向标的OpenAI发布GPT-4o, 面向ChatGPT所有付费和免费用户发布, 支持免费试用, API价格比GPT-4-turbo降低了50%, 输入价格低至5美元/百万tokens; 谷歌发布Gemini 1.5 Flash时也将输入价格定为0.35美元/百万tokens。国内来看, 5月6日, AI公司深度求索 (DeepSeek) 率先宣布降价, 其发布的第二代MoE大模型DeepSeek-V2定为0.001元/千tokens的输入价格与0.002元/千tokens的输出价格, 随后我们看到, 智谱 AI、火山引擎、阿里云、百度、科大讯飞、腾讯云等国内主要大模型厂商迅速跟进 (详见下列图表)。

此轮国产大模型厂商降价统计

模型	推出/降价时间	API价格/百万Tokens					
		输入 (元)			输出 (元)		
		原价	现价	降价幅度	原价	现价	降价幅度
深度求索DeepSeek-V2	2024/5/6	-	1	-	-	2	-
智谱AI GLM-3-Turbo	2024/5/11	5	1	80%	5	1	-80%
火山引擎 豆包通用模型pro-32k	2024/5/15	-	0.8	-	-	2	-
阿里云Qwen-Long (通义千问)	2024/5/21	20	0.5	98%	20	2	-90%
百度ERNIE-Speed (文心一言)	2024/5/21	4	免费	-	8	免费	-
百度ERNIE-Lite (文心一言)	2024/5/21	3	免费	-	6	免费	-
科大讯飞 星火Lite	2024/5/22	18	免费	-	18	免费	-
腾讯云 混元-standard	2024/5/22	10	4.5	55%	10	4.5	-55%
腾讯云 混元-lite	2024/5/22	8	免费	-	8	免费	-

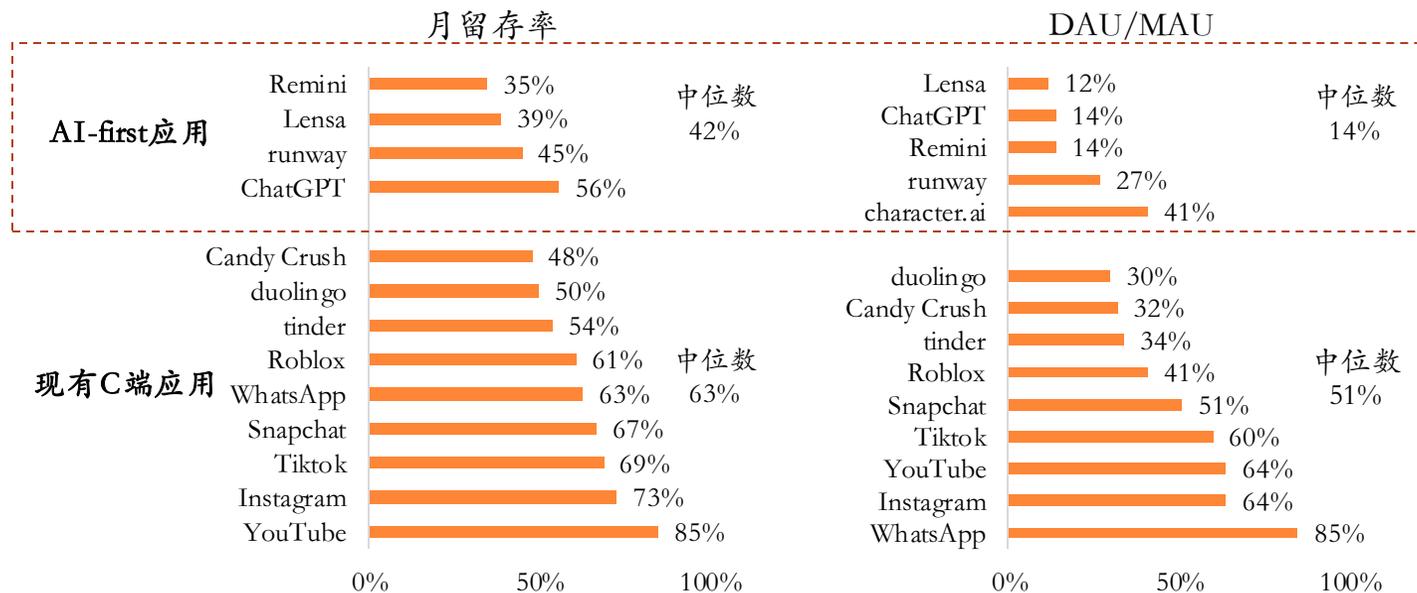
## 3.2 考验供需匹配能力，订阅制实现难度更高

2) 订阅：有ChatGPT的成功案例，我们看到不少大模型厂商通过构建AI应用，尝试走付费订阅的路径。根据Similarweb，月之暗面的智能助手Kimi Chat从推出时16万访问量到2024年2月的292万，再到3月的1219万，伴随其访问量的跃升，5月Kimi上线“给Kimi加油”付费选项（最便宜的选项99元/93天 $\approx$ 1.06元/天），可获得高峰期优先使用权益。

**实现难度：订阅>API。**然而，我们看到即使是ChatGPT、runway等具有代表性的大模型产品，用户留存度和粘性也尚未达到现有领先C端应用的水平。根据红杉资本研究，全球领先的C端应用拥有60-65%的DAU/MAU，其中WhatsApp是85%。相比之下，AI-first应用的中位数为14%，可能意味着用户还未在这些AI产品中找到能够每天使用它们的足够价值。

### ◆ Kimi增加“给Kimi加油”付费选项

### ◆ 全球C端应用与AI-first应用的用户留存率、DAU/MAU对比



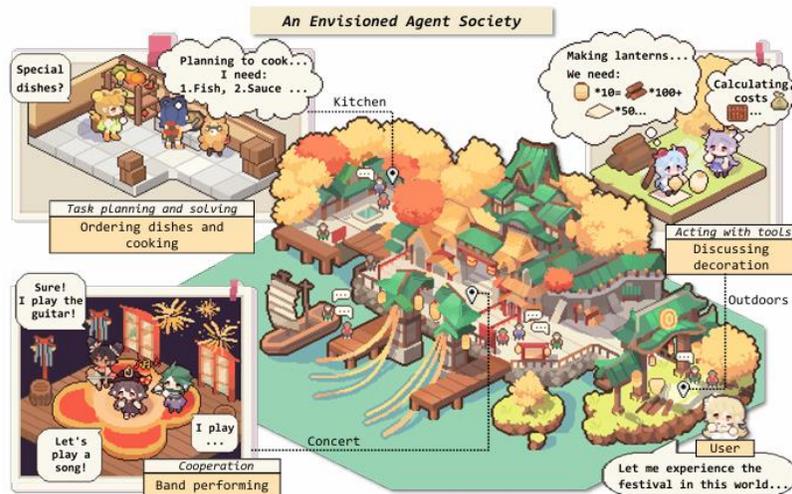
## 3.2 破局之路：AI Agent是通往AGI与Killer App的钥匙

AGI (Artificial General Intelligence, 通用人工智能) 能力是渐进解锁的, 而具有专业能力、可定制的AI Agent (智能体), 被认为是打开AGI之门的钥匙。2023年6月, OpenAI应用研究主管LilianWeng提出: Agent = LLM + 记忆 + 规划技能 + 工具使用。2024年在红杉资本的人工智能峰会上, 吴恩达认为Agent 应该具备四种主要能力, 即: 反思 (Reflection)、使用工具 (Tool use)、规划 (Planning) 以及多智能体协同 (Multi-agent collaboration)。复旦NLP团队这样描述LLM-based Agent的框架, 包含三个组成部分: 控制端 (Brain)、感知端 (Perception) 和行动端 (Action)。简单理解, AI Agent是能够感知环境、自主决策、执行复杂任务的智能实体。

◆ AGI能力是渐进解锁的, AI Agent是通往AGI的钥匙



◆ FudanNLP团队设想的一个由AI Agent构成的社会

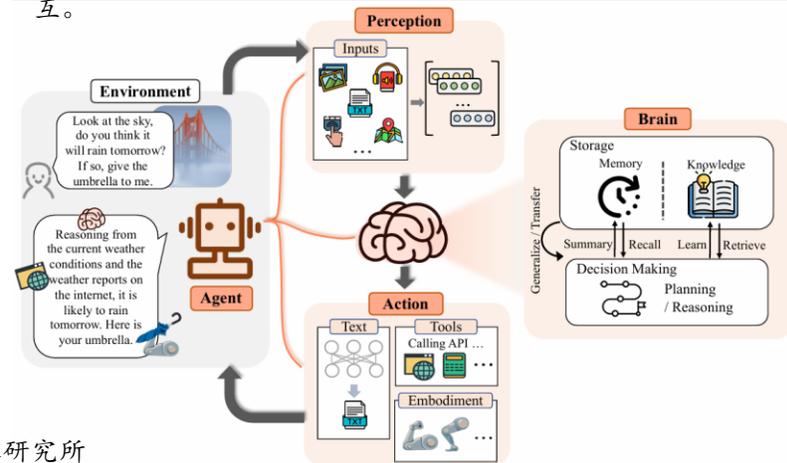


◆ LLM-based Agent的工作流程举例

举个例子: 当人类询问是否会下雨时——

- 感知端 (Perception) 将指令转换为 LLM 可以理解的表示。
- 控制端 (Brain) 开始根据当前天气和互联网上的天气预报进行推理和行动规划。
- 行动端 (Action) 做出响应并将雨伞递给人类。

通过重复上述过程, AI Agent 可以不断获得反馈并与环境交互。



## 3.2 破局之路：AI Agent是通往AGI与Killer App的钥匙

国内来看，以互联网大厂、大模型厂商、企业服务SaaS类厂商为代表的越来越多的企业参与进AI Agent市场，产品形态既包括面向企业和开发者的Agent构建平台/框架，也包括服务于各个垂直行业的专业Agent。2024年上半年，国内多个AI Agent平台发布，未来将进一步提升AI Agent开发便利性，从而加速国内大模型应用的发展。2月，字节跳动的新一代一站式AI Bot开发平台扣子Coze在国内上线，用户可以快速、低门槛地构建专属聊天机器人，4月，百度AI开发者大会发布文心智能体平台AgentBuilder，随后，钉钉正式上线AI Agent Store，首批上架包括通义千问在内的超过200个AI Agents。

### 国内AI Agent生态图谱

### 示例：在字节Coze上构建Agent，实现“行业短评Bot”

今年5月Sam Altman接受《麻省理工科技评论》记者的采访时，将Killer App描述为：“超级能干的同事，它绝对了解我一生的一切，包括我的每封电子邮件、每一次对话。此外，它可以立即解决一些任务，而对于更复杂的任务，它可以进行尝试，但如果需要的话，它会向你提出问题”。



设置Agent人设与回复逻辑（包括：角色、技能、限制）

插入工作流Workflow，模拟Agent的思考过程

Agent根据输入和工作流，打印调研报告输出



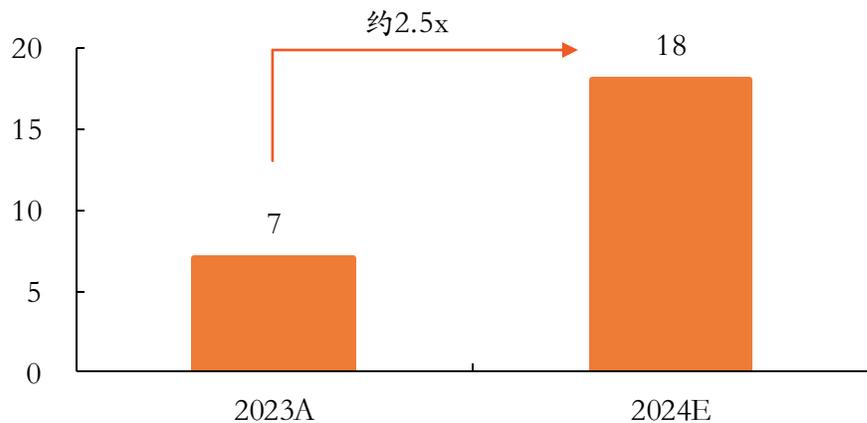
### 3.3 破局之路：企业级需求快速增长，MaaS助力降低模型使用门槛

Killer App仍在酝酿的背景下，我们认为也需要关注到B端对于大模型需求的快速增长。根据a16z调研，2023年平均每家受访企业在API、自托管和微调模型上支出达到700万美元，并且几乎所有企业计划将2024年LLM预算增加2~5倍。

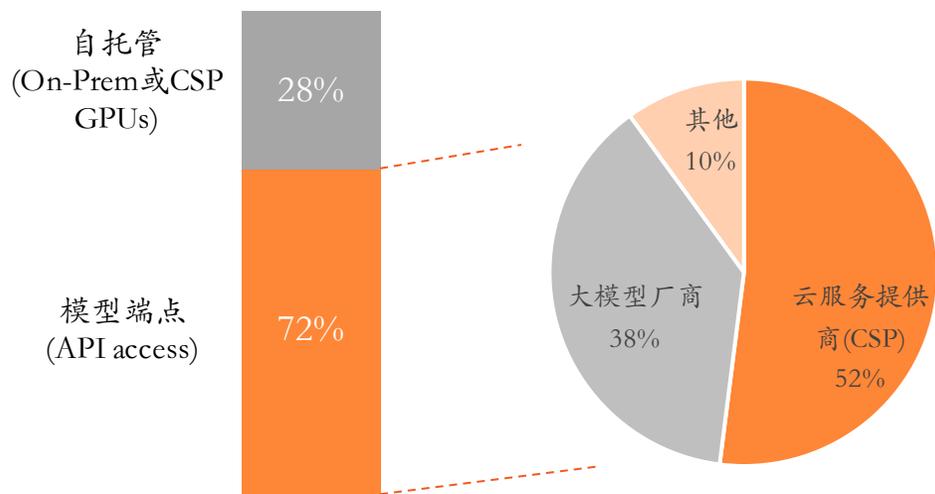
企业的AI模型采购决策仍然主要受到云服务提供商（CSP）的影响。根据a16z调研，2023年大多数企业出于安全考虑通过现有的CSP购买模型，2024年情况依然如此，在72%的使用API访问模型的受访企业中，有超过50%通过其CSP访问，比如Azure用户更常用OpenAI，而Amazon用户更倾向于使用Anthropic或Cohere。剩下28%的受访企业选择了自托管，可能是出于运行开源模型的需要，采用私有化部署或者CSP提供的GPU服务。

◆ B端对于大模型的预算正在快速增长

■ 受访企业平均每年在LLM上的支出（百万美元）



◆ 使用API的受访企业中，有50%以上通过云服务提供商访问大模型



## 3.3 破局之路：企业级需求快速增长，MaaS助力降低模型使用门槛

MaaS让企业更简洁地使用大模型。根据中国信通院的定义，MaaS围绕低技术门槛、模型可共享、应用易适配三大特性，提供包括算力服务、平台服务、模型服务、数据集服务、AI应用开发服务在内的全栈服务，使得企业能够快速高效地构建、部署、监控、调用模型，而无需开发和维护底层基础能力。

云厂商正在加速布局MaaS。目前，微软云Azure、阿里云、华为云、腾讯云、百度云、京东云等CSP都已经推出了MaaS服务。以阿里云的魔搭ModelScope为例，支持用户使用来自达摩院大模型平台和SOTA模型矩阵的超过300个优质大模型，提供包括模型管理和下载、模型调优、训练、推理、部署、应用在内的一站式模型服务。

### MaaS是人工智能架构的中间层



### 阿里云MaaS平台：魔搭ModelScope





## 目录CONTENTS

- 技术：大模型发展呈现“规模定律”，Transformer为技术基座
- 市场：全球大模型竞争白热化，国产大模型能力对标GPT-3.5Turbo
- 变现：API同质化、订阅实现难，Agent与MaaS探索破局之路
- 算力：大模型发展催生海量算力需求，预计带来千亿美元市场规模
- 投资建议及风险提示



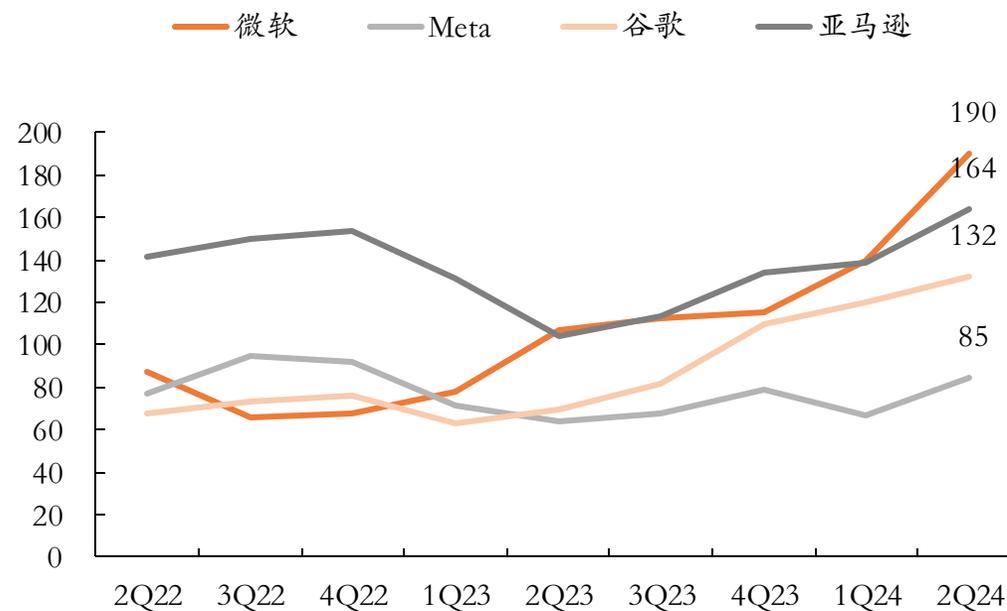
## 4.1 大模型技术与应用发展催生海量算力需求

持续增长的大模型与AI产品研发需求推升资本支出。科技巨头持续投入算力采购，以H100的采购为例，Omdia Research认为Meta和微软是H100最大的购买者，谷歌、亚马逊、甲骨文、腾讯其次，此外，云服务提供商CoreWeave、百度、阿里巴巴、字节跳动、云服务供应商Lambda Labs、特斯拉等均是H100重要的终端客户。海量算力需求下科技巨头的资本支出也保持较高的增长水平，2024年2季度微软、Meta、谷歌、亚马逊四家公司合计资本支出为571亿美元，同比增长66%，环比1季度增长22%。

◆ Omdia Research预测2023年H100客户



◆ 四家科技巨头资本支出（亿美元）



## 4.2 算力需求测算逻辑

我们主要考虑训练+推理两个阶段的算力需求。Transformer模型训练和推理都是经过多次迭代完成的，一次训练迭代包含了前向传播和反向传播两个步骤，而一次推理迭代相当于一个前向传播过程。前向传播过程指将数据输入模型计算输出，反向传播是计算模型的梯度并存储梯度进行模型参数更新。根据NVIDIA论文《Reducing Activation Recomputation in Large Transformer Models》，反向传播的计算量大约是前向传播的2倍，因此可以得出，一次训练迭代（包含一次前向+一次反向）的计算量大约为一次推理迭代（包含一次前向）的3倍。

◆ NVIDIA论文：反向传播的计算量是前向传播的2倍

### A FLOPs Calculation

For FLOPs calculations, we follow the derivation from Narayanan, et.al.[13] and only consider the matrix multiplications (GEMMs) which are the main contributors to the number of floating-point operations. For the attention block, the main contributors to floating-point operations are: key, query, and value transformation ( $6Bsh^2$  operations), attention matrix computation ( $2Bs^2h$  operations), attention over values ( $2Bs^2h$  operations), and post-attention linear projection ( $2Bsh^2$  operations) where  $B$  is the microbatch size.

For the feed-forward network that increases the hidden size to  $4h$  and then reduces it back to  $h$ , we have  $16Bsh^2$  floating-point operations. Summing these together, each transformer layer results in  $24Bsh^2 + 4Bs^2h$  FLOPs for the forward pass. The other main contributor to the number of floating-point operations is the logits layer in the language model head, which transforms features of dimension  $h$  to the vocabulary dimension  $v$ . The required FLOPs for this operation is  $2Bshv$ .

The backward pass requires double the number of FLOPs since we need to calculate the gradients with respect to both input and weight tensors. Summing all the contributions, the number of FLOPs required to do one forward and one backward pass (denoted by model FLOPs) is:

$$\text{model FLOPs per iteration} = 72BLsh^2 \left( 1 + \frac{s}{6h} + \frac{v}{12hL} \right). \quad (7)$$

1) 反向传播的计算量 (FLOPs) 约为前向传播的2倍

2) 一次训练迭代包含一次前向传播和一次反向传播

## 4.2 算力需求测算逻辑

再结合OpenAI论文《Scaling Laws for Neural Language Models》的结论，训练Transformer模型的理论计算量为 $C \approx 6N * D$ ，其中，N为模型参数量大小，D为训练数据量大小；推理所需计算量即为 $2N * D$ 。举个例子，参考OpenAI关于GPT-3的论文《Language Models are Few-Shot Learners》，GPT-3 175B版本模型参数量为1746亿、训练数据量tokens为3000亿，我们根据公式预估，训练模型所需算力为 $C \approx 6N * D = 6 * (1746 * 10^8) * (3 * 10^{11}) = 3.14 * 10^{23}$ FLOPs。

◆ OpenAI论文：训练LLM每token需要的计算量约为6倍的模型大小

Operation	Parameters	FLOPs per Token
Embed	$(n_{\text{vocab}} + n_{\text{ctx}}) d_{\text{model}}$	$4d_{\text{model}}$
Attention: QKV	$n_{\text{layer}} d_{\text{model}} 3d_{\text{attn}}$	$2n_{\text{layer}} d_{\text{model}} 3d_{\text{attn}}$
Attention: Mask	—	$2n_{\text{layer}} n_{\text{ctx}} d_{\text{attn}}$
Attention: Project	$n_{\text{layer}} d_{\text{attn}} d_{\text{model}}$	$2n_{\text{layer}} d_{\text{attn}} d_{\text{embd}}$
Feedforward	$n_{\text{layer}} 2d_{\text{model}} d_{\text{ff}}$	$2n_{\text{layer}} 2d_{\text{model}} d_{\text{ff}}$
De-embed	—	$2d_{\text{model}} n_{\text{vocab}}$
<b>Total (Non-Embedding)</b>	$N = 2d_{\text{model}} n_{\text{layer}} (2d_{\text{attn}} + d_{\text{ff}})$	$C_{\text{forward}} = 2N + 2n_{\text{layer}} n_{\text{ctx}} d_{\text{attn}}$

**Table 1** Parameter counts and compute (forward pass) estimates for a Transformer model. Sub-leading terms such as nonlinearities, biases, and layer normalization are omitted.

Model	Total train compute (PF-days)	Total train compute (flops)	Params (M)	Training tokens (billions)	Flops per param per token	Mult for bwd pass	Fwd-pass flops per active param per token	Frac of params active for each token
T5-Small	2.08E+00	1.80E+20	60	1,000	3	3	1	0.5
T5-Base	7.64E+00	6.60E+20	220	1,000	3	3	1	0.5
T5-Large	2.67E+01	2.31E+21	770	1,000	3	3	1	0.5
T5-3B	1.04E+02	9.00E+21	3,000	1,000	3	3	1	0.5
T5-11B	3.82E+02	3.30E+22	11,000	1,000	3	3	1	0.5
BERT-Base	1.89E+00	1.64E+20	109	250	6	3	2	1.0
BERT-Large	6.16E+00	5.33E+20	355	250	6	3	2	1.0
RoBERTa-Base	1.74E+01	1.50E+21	125	2,000	6	3	2	1.0
RoBERTa-Large	4.93E+01	4.26E+21	355	2,000	6	3	2	1.0
GPT-3 Small	2.60E+00	2.25E+20	125	300	6	3	2	1.0
GPT-3 Medium	7.42E+00	6.41E+20	356	300	6	3	2	1.0
GPT-3 Large	1.58E+01	1.37E+21	760	300	6	3	2	1.0
GPT-3 XL	2.75E+01	2.38E+21	1,320	300	6	3	2	1.0
GPT-3 2.7B	5.52E+01	4.77E+21	2,650	300	6	3	2	1.0
GPT-3 6.7B	1.39E+02	1.20E+22	6,660	300	6	3	2	1.0
GPT-3 13B	2.68E+02	2.31E+22	12,850	300	6	3	2	1.0
<b>GPT-3 175B</b>	<b>3.64E+03</b>	<b>3.14E+23</b>	<b>174,600</b>	<b>300</b>	<b>6</b>	<b>3</b>	<b>2</b>	<b>1.0</b>

For contexts and models with  $d_{\text{model}} > n_{\text{ctx}}/12$ , the context-dependent computational cost per token is a relatively small fraction of the total compute. Since we primarily study models where  $d_{\text{model}} \gg n_{\text{ctx}}/12$ , we do not include context-dependent terms in our training compute estimate. Accounting for the backwards pass (approximately twice the compute as the forwards pass), we then define the estimated non-embedding compute as  $C \approx 6N$  floating point operators per training token

$C$  - 模型计算量       $N$  - 模型参数量       $D$  - 数据量 (Tokens量)

训练所需计算量:  $C \approx 6N * D$

推理所需计算量:  $C \approx 2N * D$

MFU - 集群算力利用率 (Model FLOPs Utilization)       $S$  - 所用集群算力

训练耗时:  $T = C / (MFU * S)$

## 4.2 算力需求测算：训练端

参考7月23日Meta公布的最新开源模型——Llama 3.1 405B，该模型基于15.6T tokens的数据量进行预训练达到最优性能，训练数据量约为模型参数规模的38.5倍。我们假设最优大模型参数量（N）与Tokens数（D）的近似线性关系为： $D=38.5*N$ ，并设置了参数量分别为100亿、500亿、1000亿、3000亿、5000亿、1万亿、2万亿的共七档模型进行算力需求的测算。

1) **训练算力需求测算**：根据前述公式计算，得出训练一次Llama 3.1的计算量大致为 $3.79*10^{25}$ FLOPs，与实际情况接近（论文中Llama 3.1 405B的training budget为 $3.8*10^{25}$ FLOPs）。同时，根据NVIDIA官网，A100 80G SXM、H100 SXM的在16位精度下（FP16）的算力峰值分别为624TFLOPs、1979TFLOPs，我们假设集群算力利用率MFU为30%，假设模型训练时长分别为30天、60天、90天的背景下，得出训练一次Llama 3.1 405B模型分别需要7.8万张、3.9万张、2.6万张A100，或者等同于分别需要2.5万张、1.2万张、8208张H100。**训练阶段算力需求测算（假设MFU=30%）**

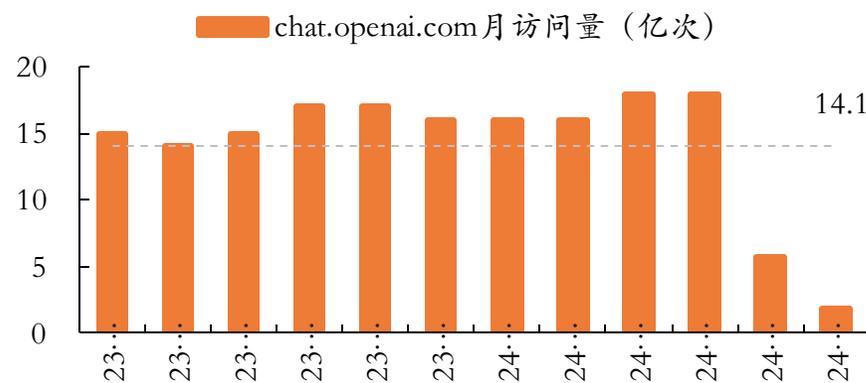
训练算力需求测算		模型1	模型2	模型3	模型4	Llama 3.1 (405B)	模型5	模型6	模型7
训练所需算力 (PFLOPs)		2.31E+07	5.78E+08	2.31E+09	2.08E+10	3.79E+10	5.78E+10	2.31E+11	9.24E+11
模型参数量 (亿)		100	500	1,000	3,000	4,050	5,000	10,000	20,000
数据量Tokens (亿)		3,850	19,250	38,500	115,500	155,925	192,500	385,000	770,000
假设训练时长 (天)	所需GPU数量	模型1	模型2	模型3	模型4	Llama 3.1 (405B)	模型5	模型6	模型7
30	A100 (FP16)	48	1,191	4,761	42,847	78,088	119,018	476,071	1,904,282
60		24	596	2,381	21,424	39,044	59,509	238,036	952,141
90		16	397	1,587	14,283	26,030	39,673	158,691	634,761
30	H100 (FP16)	16	376	1,502	13,510	24,622	37,528	150,111	600,441
60		8	188	751	6,755	12,311	18,764	75,056	300,221
90		6	126	501	4,504	8,208	12,510	50,037	200,147

## 4.2 算力需求测算：推理端

2) 推理算力需求测算：根据Similarweb统计，ChatGPT网站在过去12个月平均月访问量为14.1亿次。我们假设在问答场景下，月访问量为14亿、单用户平均每次访问提问次数为15次，即模型每秒处理请求数8000次，假设单次问答消耗600Tokens，计算得出该场景下每秒消耗0.049亿Tokens。

考虑并发峰值和显存预留等问题，我们假设峰值预留倍数为10，计算得出每日峰值消耗4.2万亿Tokens。根据 $C \approx 2N * D$ 我们得到对应推理所需算力及所需GPU数量（假设采用INT8精度、MFU=30%），如下图表所示，千亿参数模型单问答场景推理大约需要2.6万张A100或者8188张H100。

ChatGPT过去12个月平均月访问量为14.1亿次



单场景下（问答）推理算力需求测算

单场景（问答）推理算力需求测算	模型1	模型2	模型3	模型4	模型5	模型6	模型7
单日推理所需算力 (PFLOPs)	8.40E+07	4.20E+08	8.40E+08	2.52E+09	4.20E+09	8.40E+09	1.68E+10
模型参数量 (亿)	100	500	1,000	3,000	5,000	10,000	20,000
峰值每日消耗Tokens (亿)	42,000	42,000	42,000	42,000	42,000	42,000	42,000
每秒消耗Tokens (亿)	0.049	0.049	0.049	0.049	0.049	0.049	0.049
QPS (query per second, 万次)	0.8	0.8	0.8	0.8	0.8	0.8	0.8
月访问量 (亿)	14	14	14	14	14	14	14
单用户每次访问提问次数 (次)	15	15	15	15	15	15	15
每次问答消耗Tokens数	600	600	600	600	600	600	600
峰值预留倍数	10	10	10	10	10	10	10
推理所需GPU数量 (A100 INT8, MFU=30%)	2,597	12,984	25,968	77,903	129,838	259,675	519,350
推理所需GPU数量 (H100 INT8, MFU=30%)	819	4,094	8,188	24,564	40,940	81,879	163,757

## 4.3 大模型服务器成本测算

我们将前述训练（假设训练时长60天）和推理（单场景）两个阶段所需GPU数量加和，并统一假设1台服务器集成8张A100，计算得出千亿参数规模模型训练+推理大约需要3544台服务器，万亿参数规模模型需要6.2万台服务器。

### ◆ 训练+单场景推理算力需求测算

训练+推理算力需求测算	模型1	模型2	模型3	模型4	模型5	模型6	模型7
模型参数量（亿）	100	500	1000	3000	5000	10000	20000
模型训练+推理所需GPU数量（A100）	2,621	13,580	28,349	99,327	189,347	497,711	1,471,491
训练所需GPU数量（A100，训练时长60天）	24	596	2,381	21,424	59,509	238,036	952,141
推理所需GPU数量（A100，单场景）	2,597	12,984	25,968	77,903	129,838	259,675	519,350
模型所需服务器数量（8张A100）	328	1,698	3,544	12,416	23,669	62,214	183,937

大模型的持续迭代升级将为AI芯片及服务器发展提供强劲动力。未来，我们假设有100家大模型实现持续经营，由于垂类大模型参数量通常在百亿至千亿量级，且数量较多，我们假设中位数500亿参数规模的模型有50个，通用大模型参数规模普遍更大，我们假设中位数分别为1000亿、5000亿、10000亿参数规模的大模型分别有25个、15个、10个。假设单台服务器成本为20万美元，测算得到AI服务器的市场规模为2301亿美元。如果再考虑问答场景之外的推理场景，AI服务器的市场规模将更加巨大。根据IDC数据，2023年的全球AI服务器市场规模是211亿美元，相比2023年全球AI服务器市场规模，大模型的持续迭代升级将为AI服务器市场带来广阔的市场空间。

### ◆ 大模型服务器成本测算

大模型服务器成本测算	模型数量	单个模型所需服务器数量	单台服务器成本（万美元）	成本测算（亿美元）
模型参数规模500亿	50	1,698	20	170
模型参数规模1000亿	25	3,544	20	177
模型参数规模5000亿	15	23,669	20	710
模型参数规模10000亿	10	62,214	20	1,244
合计				2,301



## 目录CONTENTS

- 技术：大模型发展呈现“规模定律”，Transformer为技术基座
- 市场：全球大模型竞争白热化，国产大模型能力对标GPT-3.5Turbo
- 变现：API同质化、订阅实现难，Agent与MaaS探索破局之路
- 算力：大模型发展催生海量算力需求，预计带来千亿美元市场规模
- 投资建议及风险提示



## 投资建议&风险提示

### 投资建议:

当前，全球范围内大模型领域的竞争依然白热化，我国大模型厂商持续迭代升级算法能力，2023年底国产大模型市场迈入爆发期，根据SuperCLUE上半年最新的评测结果，国内绝大部分闭源模型已超过GPT-3.5Turbo，将有望加速国产大模型在各场景的应用落地。同时，大模型的持续迭代升级将为AI芯片及服务器发展提供强劲动力，根据我们测算，假设未来有100家大模型实现持续经营，在训练与推理（问答场景）两个阶段，大模型将为AI服务器带来约2301亿美元的市场空间。我们继续看好AI主题的投资机会，标的方面：1) 算力方面，推荐浪潮信息、中科曙光、紫光股份、神州数码、海光信息、龙芯中科，建议关注寒武纪、景嘉微、软通动力；2) 算法方面，推荐科大讯飞；3) 应用场景方面，强烈推荐中科创达、恒生电子、盛视科技，推荐金山办公，建议关注万兴科技、同花顺、彩讯股份；4) 网络安全方面，强烈推荐启明星辰。

### 风险提示:

- 1) **AI算力供应链风险上升。**美国对华半导体出口管制升级，将倒逼我国国产AI芯片产业链加快成熟。但如果我国国产AI芯片的迭代速度不达预期，将影响我国AI算力的发展，进而制约大模型的突破。
- 2) **大模型产品的应用落地低于预期。**当前，我国国产大模型已经开始在教育、医疗、汽车、办公、工业、智能硬件等B端和C端应用场景持续落地，如果产品的市场拓展不及预期，我国大模型产品的应用落地将存在低于预期的风险。
- 3) **国产大模型算法发展可能不及预期。**当前，虽然国产大模型能力持续升级，但相比OpenAI等全球领先的大模型厂商，国产大模型厂商的追赶进度存在不达预期的风险。

## 重点公司预测与评级

股票简称	股票代码	2024/8/14	EPS (元)				PE (倍)				评级
		收盘价 (元)	2023A	2024E	2025E	2026E	2023A	2024E	2025E	2026E	
中科创达	300496.SZ	39.04	1.01	1.48	1.84	2.28	38.47	26.40	21.22	17.12	强烈推荐
恒生电子	600570.SH	16.49	0.75	0.91	1.11	1.31	21.93	18.09	14.86	12.52	强烈推荐
盛视科技	002990.SZ	18.04	0.77	1.06	1.41	1.85	23.33	16.92	12.76	9.76	强烈推荐
启明星辰	002439.SZ	13.46	0.61	0.79	0.99	1.25	22.13	17.01	13.60	10.78	强烈推荐
浪潮信息	000977.SZ	32.81	1.21	1.57	1.97	2.42	27.09	20.95	16.68	13.57	推荐
中科曙光	603019.SH	39.11	1.25	1.57	1.99	2.56	31.17	24.85	19.68	15.29	推荐
紫光股份	000938.SZ	19.89	0.74	0.87	1.05	1.28	27.05	22.98	18.94	15.55	推荐
神州数码	000034.SZ	22.20	1.75	2.13	2.62	3.23	12.68	10.41	8.47	6.88	推荐
海光信息	688041.SH	73.57	0.54	0.76	1.03	1.35	136.2	96.8	71.4	54.5	推荐
龙芯中科	688047.SH	97.05	-0.82	-0.26	0.11	0.39	-118.29	-367.14	884.48	251.08	推荐
科大讯飞	002230.SZ	34.67	0.28	0.31	0.37	0.45	121.99	113.52	94.07	77.74	推荐
金山办公	688111.SH	188.25	2.85	3.71	4.94	6.69	66.06	50.71	38.11	28.16	推荐
顶点软件	603383.SH	29.56	1.13	1.37	1.61	1.87	26.06	21.61	18.35	15.81	推荐

## 股票投资评级：

强烈推荐（预计6个月内，股价表现强于市场表现20%以上）

推 荐（预计6个月内，股价表现强于市场表现10%至20%之间）

中 性（预计6个月内，股价表现相对市场表现±10%之间）

回 避（预计6个月内，股价表现弱于市场表现10%以上）

## 行业投资评级：

强于大市（预计6个月内，行业指数表现强于市场表现5%以上）

中 性（预计6个月内，行业指数表现相对市场表现在±5%之间）

弱于大市（预计6个月内，行业指数表现弱于市场表现5%以上）

## 公司声明及风险提示：

负责撰写此报告的分析师（一人或多人）就本研究报告确认：本人具有中国证券业协会授予的证券投资咨询执业资格。

平安证券股份有限公司具备证券投资咨询业务资格。本公司研究报告是针对与公司签署服务协议的签约客户的专属研究产品，为该类客户进行投资决策时提供辅助和参考，双方对权利与义务均有严格约定。本公司研究报告仅提供给上述特定客户，并不面向公众发布。未经书面授权刊载或者转发的，本公司将采取维权措施追究其侵权责任。

证券市场是一个风险无时不在的市场。您在进行证券交易时存在赢利的可能，也存在亏损的风险。请您务必对此有清醒的认识，认真考虑是否进行证券交易。

市场有风险，投资需谨慎。

## 免责声明：

此报告旨在发给平安证券股份有限公司（以下简称“平安证券”）的特定客户及其他专业人士。未经平安证券事先书面明文批准，不得更改或以任何方式传送、复印或派发此报告的材料、内容及其复印本予任何其他人。

此报告所载资料的来源及观点的出处皆被平安证券认为可靠，但平安证券不能担保其准确性或完整性，报告中的信息或所表达观点不构成所述证券买卖的出价或询价，报告内容仅供参考。平安证券不对因使用此报告的材料而引致的损失而负上任何责任，除非法律法规有明确规定。客户并不能仅依靠此报告而取代行使独立判断。

平安证券可发出其它与本报告所载资料不一致及有不同结论的报告。本报告及该等报告反映编写分析员的不同设想、见解及分析方法。报告所载资料、意见及推测仅反映分析员于发出此报告日期当日的判断，可随时更改。此报告所指的证券价格、价值及收入可跌可升。为免生疑问，此报告所载观点并不代表平安证券的立场。

平安证券在法律许可的情况下可能参与此报告所提及的发行商的投资银行业务或投资其发行的证券。

平安证券股份有限公司2024版权所有。保留一切权利。