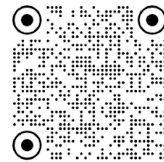


LLM 投资思考及 OpenAI 案例分享

拾象分享 | 2023.3

01 关于 LLM 的关键判断



行业格局

 OpenAI

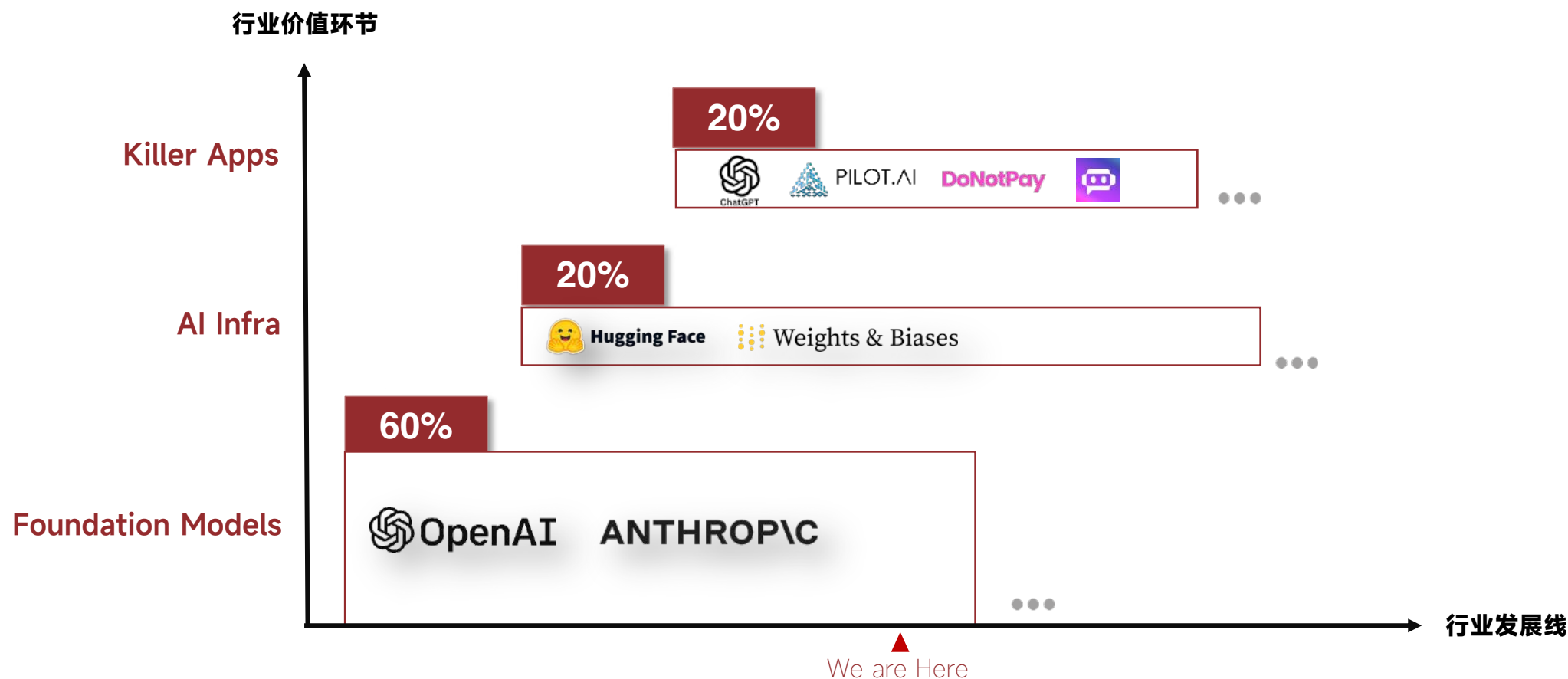
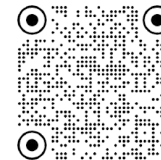
通用计算机

GPT-5 会成为 Windows 95 级别的新一代通用计算机平台

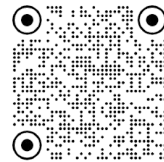


影响格局的三股力量

信仰差异



- LLM-AI = Windows-PC, Apple-Mobile, AWS-Cloud, Tiktok-Short video
- OpenAI-Microsoft = Windows-IBM
- OpenAI-Azure/Nvidia = AWS-Cisco/DELL/EMC



Mobile 时代

- Location
- Touch
- Camera
- Audio input
- Video input
- Time spent x5
- Expand user base x10
- User behavior data → ML



LLM 时代

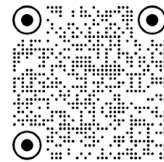
- **理解能力：** 采集更多非结构化数据
交互方式多元化
自然语言取代编程语言
- **生成能力：**
- **推理能力**
-



Portal → Search → 推荐 → Chat / AIGC:

组织信息的效率和能力持续变强，每次变化都带来更大更新商业模式升级，更简洁 + 无限可能

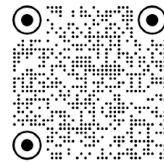
判断 4：LLM 的边界是什么



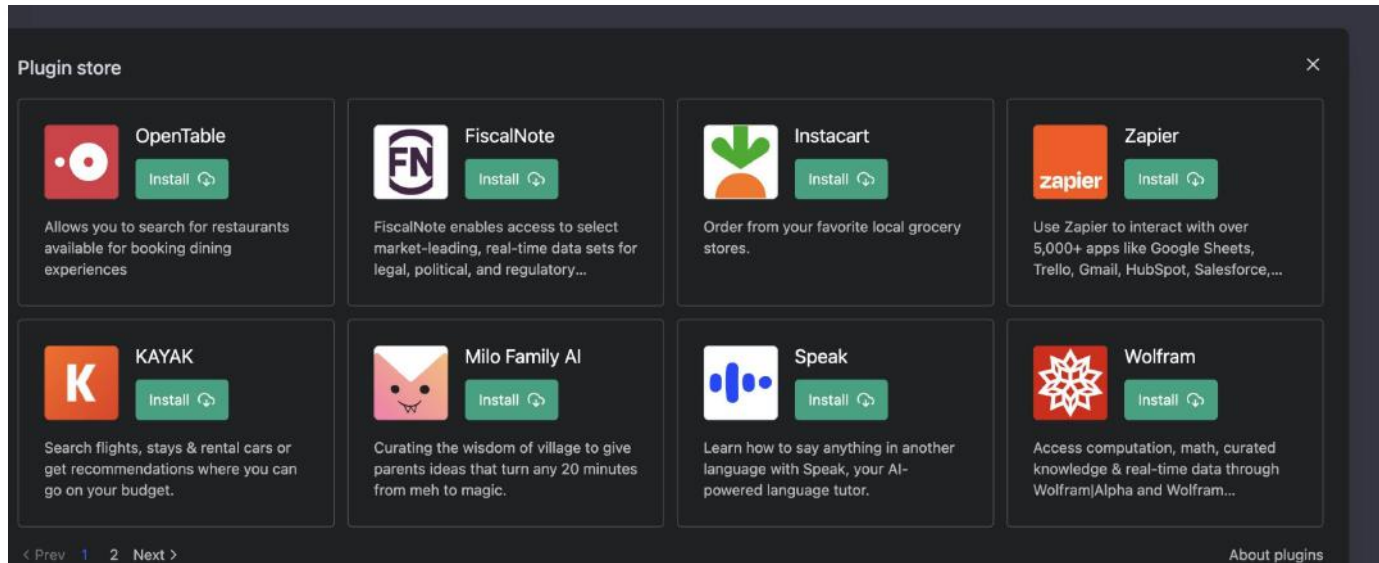
- OpenAI 是大脑，微软是身体，多模态是眼睛耳朵
- 专有数据：是血液，能形成壁垒，比如私人助理比 LLM 更懂我
- 网络效应：LLM 很难重构 WeChat 熟人网络
- 规模效应：LLM 依赖高质量供应链，iPhone 做不了 Uber / Airbnb / DoorDash
- 机器人：LLM 有机会做掉虚拟机器人，提前布局物理机器人有机会形成壁垒



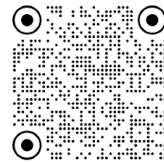
判断 5：LLM 的下一步是 Action



- Action 的相关探索：给大模型的“脑子”装上“手”。
- 两个路径：
 - Inflection、adept
 - OpenAI Plugin
- Action 会重新组织 App / SaaS 软件生态，未来没有 SaaS，只有 Service，LMO（Large Model Optimization）变得很重要
 - ChatGPT gets “eyes and ears” with plugins



02 AI Native 的应用会是什么样

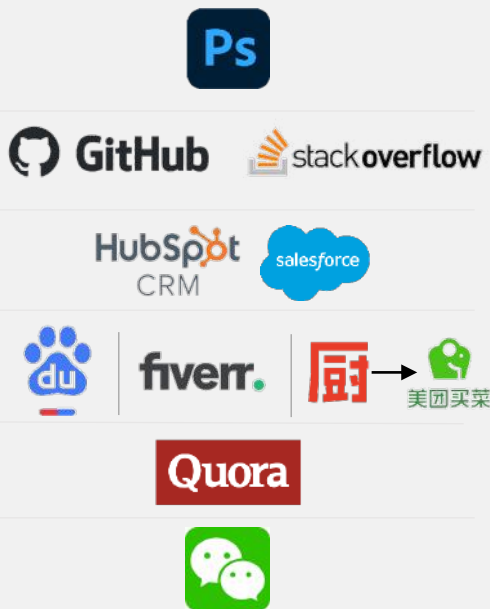


- **交互的重构**：从用户适应产品到产品适应用户。
- **数据和信息的重构**：从结构化数据到非结构化数据。
- **服务的重构**：从组织、撮合变成直接提供服务。
- **反馈机制的重构**：从功能迭代到模型迭代。

现有软件

Now

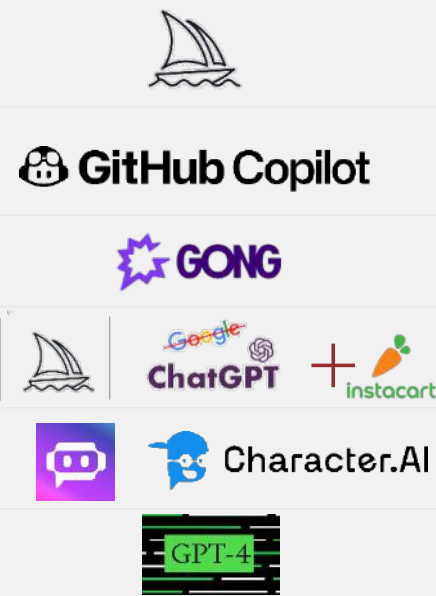
- 有限 Input，用户适应产品
- 站外搜索，多产品切换
- 处理结构化数据
- 服务撮合、促进交易
- 人和人对话
- 功能迭代



AI Native 软件

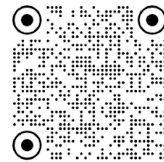
Future

- 无限 Input，产品适应用户
- 贴身助手、辅助驾驶
- 处理非结构化数据
- 直接提供服务
- 人 VS 机器、机器 VS 机器
- 模型迭代

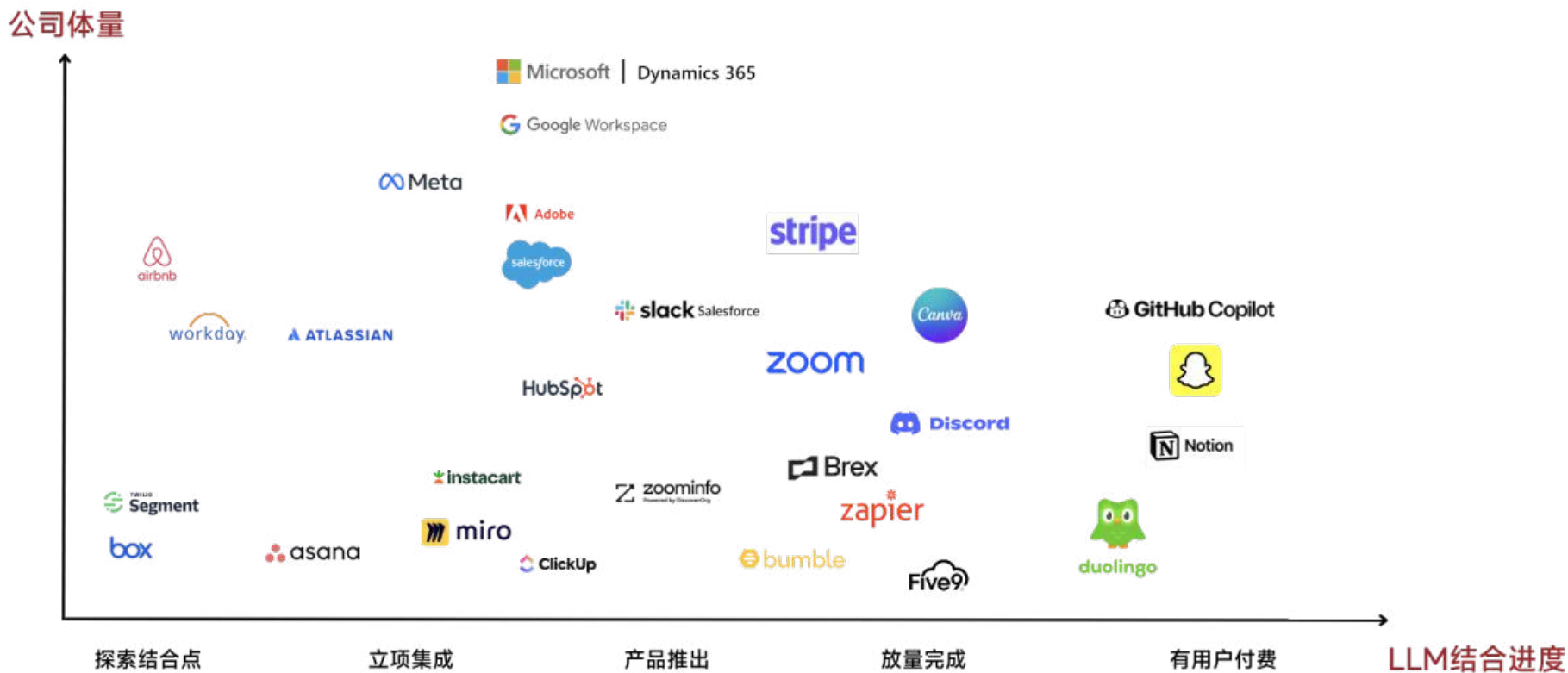


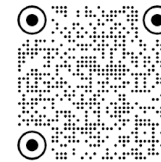
03 大厂和独角兽对LLM的应用情况

已有大厂和独角兽对 LLM 的应用情况



- 大部分老公司还没有通过 LLM 产生可观收入，还没有最佳实践诞生。
- 现在起步不算晚；
- 真正被定价和推向用户的只有 4 个产品；
- 已有 LLM 集成的公司普遍在等待未来 6-8 个月窗口期验证 AI 能力对 win rate、保护定价等方面的帮助。





LLM 作为内部提效工具

.bubble

将 LLM 集成进入审核系统，加快文字及照片审核效率。

chime

用 ChatGPT 结合公司代码库提高开发效率。

LLM 作为重点业务突破抓手

twilio segment

CEO Jeff Lawson 宣扬 Twilio Segment 的消费者数据是客户能喂给 LLM 的最重要私有数据之一。

zoom

Zoom IQ for Sales 添加了为销售会议增加章节和整个会议摘要的能力，加速 Zoom 2.0 战略。

LLM 与现有业务场景的自然结合

大多数公司

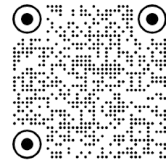
LLM 作为公司战略推进抓手



客户必须基于 Salesforce 的数据湖仓 Genie 使用 Einstein GPT，推动 Salesforce 占住数据云。

Microsoft Azure

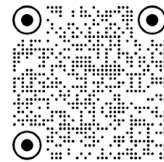
Azure 成为了在 AI 方向最领先的超级计算云。



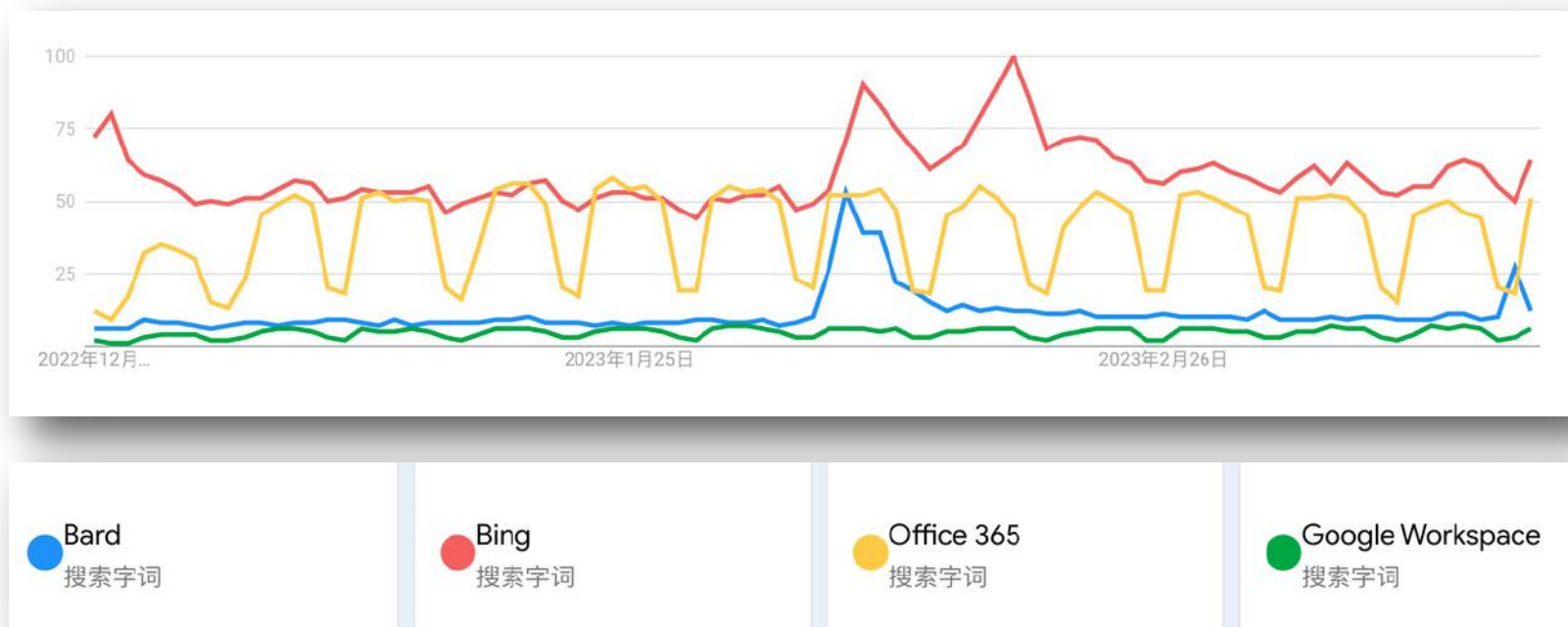
- **微软最激进:** 认为 GPT-4 已经是 AGI 的一种形态，觉得 Office 365 Copilot 是 Enterprise-Ready 的。
- **总体可以分成四派:**

类型	核心观点
门外汉派	• Atlassian 的 Mike 和 Asana 的 Dustin 都直接承认自己是这个领域的学生，探索学习中。
AI 基因、稳扎稳打派	• Zoom 的 Eric 看 AI 跟 1995 年看到互联网一样兴奋，而且 Zoom 一直是 AI-First 公司，以前的很多功能都用了 AI，比如降噪、虚拟背景等，先探索 Zoom IQ for Sales 和 Virtual Agent 这种能有机结合的场景； • Airbnb 的 Brian 认为 Airbnb 不是最好的 AI 公司，但将为 AI 设计出最好的界面，LLM 有助于自己用好评数据做 SKU 导购，但是觉得不用急，今年急着上 LLM 能力的公司就像去年记得蹭加密货币的公司一样。
深刻理解 AI 但是判断 Timing 不 Ready 派	• Twilio 的 Jeff 认为 LLM 的数据就世界、公司、客户三层，Segment 占据了客户这一层，但是 LLM 还不 Enterprise-Ready，还需要未来几年发展； • Box 也是在架构和非结构化数据上都可以服务 LLM，潜在用例也想得比较清楚，但是认为目前行业成本曲线、额外定价权来源都还不清晰，先谨慎入场。
快速尝试派	• 其他公司大部分都是此派，其中 Hubspot、ZoomInfo 和 Workday 的 CEO 对数据资产的思考更深刻。

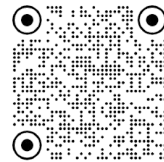
Transformer 时代，Attention is all you need



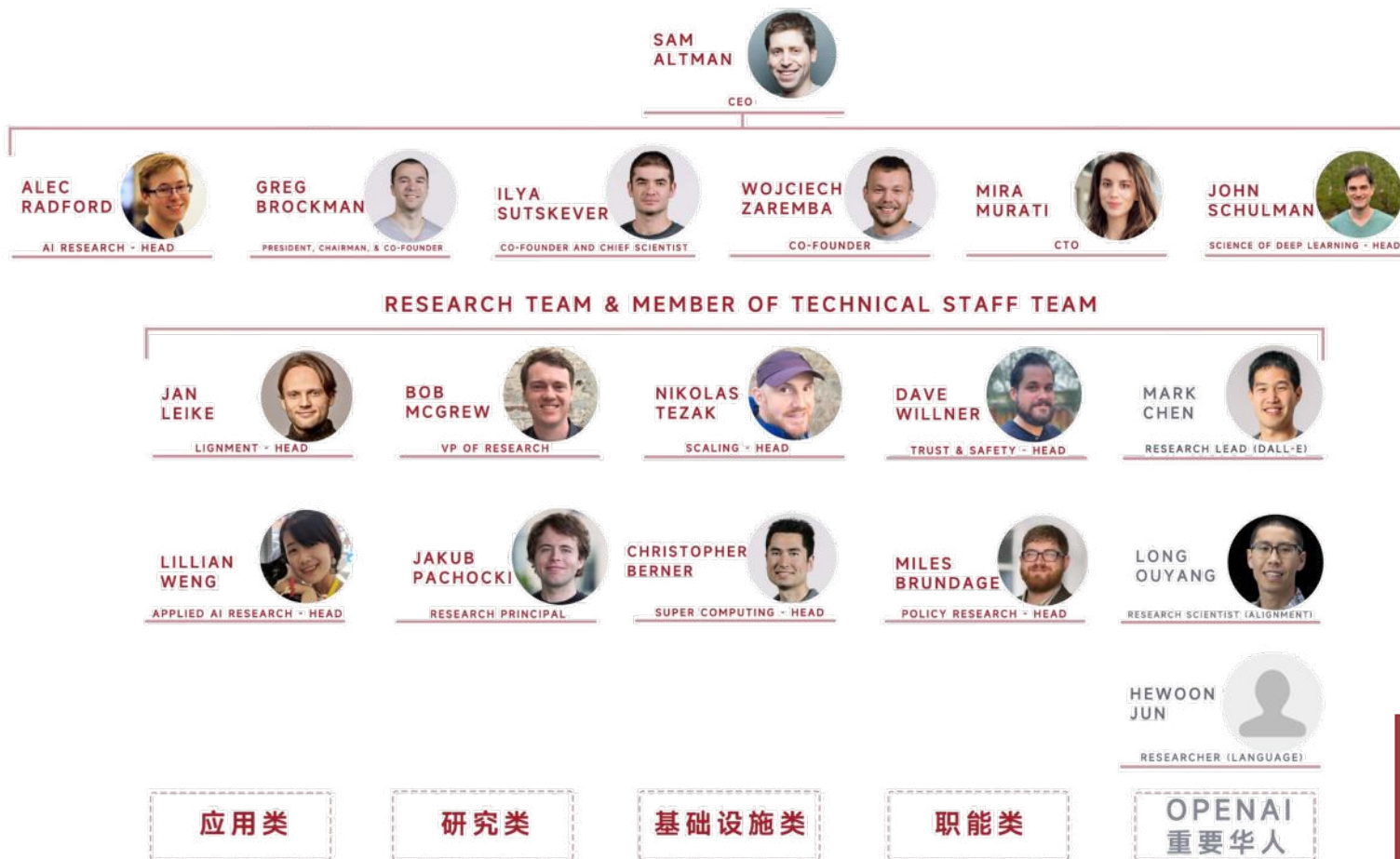
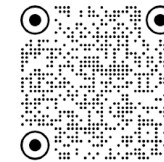
- 过去 6 个月和未来 6 个月是争夺用户和开发者注意力的竞争期。
- 微软在制作炫酷的 Demo、大范围 PR 和占据 AI 心智方面做得非常好，类似的例子还有 OpenAI 和 AGI、Runway 和视频 AI。
- 把 Demo 做好是个重要基本功。



04 OpenAI 研究 Key Takeaways



- **Top-Down对 AGI 的坚定信仰：**管理层“洗脑”、所见即所得、自上而下统一了 AGI 的路线图；
- **Bottom-Up 推动创新的机制：Alphago (deepmind)** 、ChatGPT 都是自下而上创新；
- **梦想和期权都重要：**OpenAI 平均薪酬是 120 万美金；
- **极高的人才密度：**OpenAI 的每个人都是六边形战士。



Greg: 大管家和冲锋队长，工程能力和基础设施的打造者

Ilya: DL 启蒙者，帮 OpenAI 及时走上 Transformer 路线

Wojciech: 攻坚手，亲手放弃机器人探索后打造 CodeX

John Schulman: RLHF 的开创者

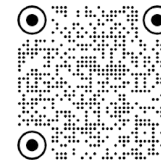
Jan Leike: 落地 instruct GPT，把 alignment 带到新高度

Andrej: 从 Tesla 回归的创始成员和 CV 大牛，多模态的加速器

Mira: 新晋 CTO，AI 人机交互领域最重要的 PM 之一

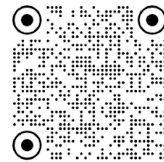
Lilian Weng: 前沿技术和应用研究的桥梁

GPT-4预训练的核心 Lead



- **开发 GPT-4 需要六个方向的研究团队：**预训练（Pretraining）、长上下文语义理解（Long context）、视觉理解（Vision）、强化学习和对齐（RL & alignment）、评估分析（Evaluation & analysis）、部署（Deployment）。这六个部分不分主次，缺一不可；
- **成员并非只负责单个工作，而是在不同小组身兼数职；**
- **GPT-4 团队分工情况：**

Pretraining	• Compute cluster scaling • Hardware correctness	• Data • Optimization & architecture	• Distributed training infrastructure • Training run babysitting	
Long context	• Long context research	• Long context kernels		
Vision	• Architecture research • Compute cluster scaling	• Data • Training run babysitting	• Distributed training infrastructure • Alignment Data	• Hardware correctness • Deployment & post-training
RL & Alignment	• Dataset contributions • Refusals	• Data infrastructure • Flagship training runs	• ChatML format • Code capability	• Model safety • Foundational RLHF and InstructGPT work
Evaluation & Analysis	• OpenAI Evals library • Capability evaluations • Instruction following and API evals • Overreliance analysis • System card & broader impacts analysis	• Model-graded evaluation infrastructure • Coding evaluations • Novel capability discovery • Privacy and PII evaluations	• Acceleration forecasting • Real-world use case evaluations • Vision evaluations • Safety and policy evaluations • Non-proliferation, international humanitarian law & national security red teaming	• ChatGPT evaluations • Contamination investigations • Economic impact evaluation • OpenAI adversarial testers
Deployment	• Inference research • Trust & safety engineering • Reliability engineering	• GPT-4 API & ChatML deployment • Trust & safety monitoring and response	• GPT-4 web experience • Trust & safety policy	• Inference infrastructure • Deployment compute

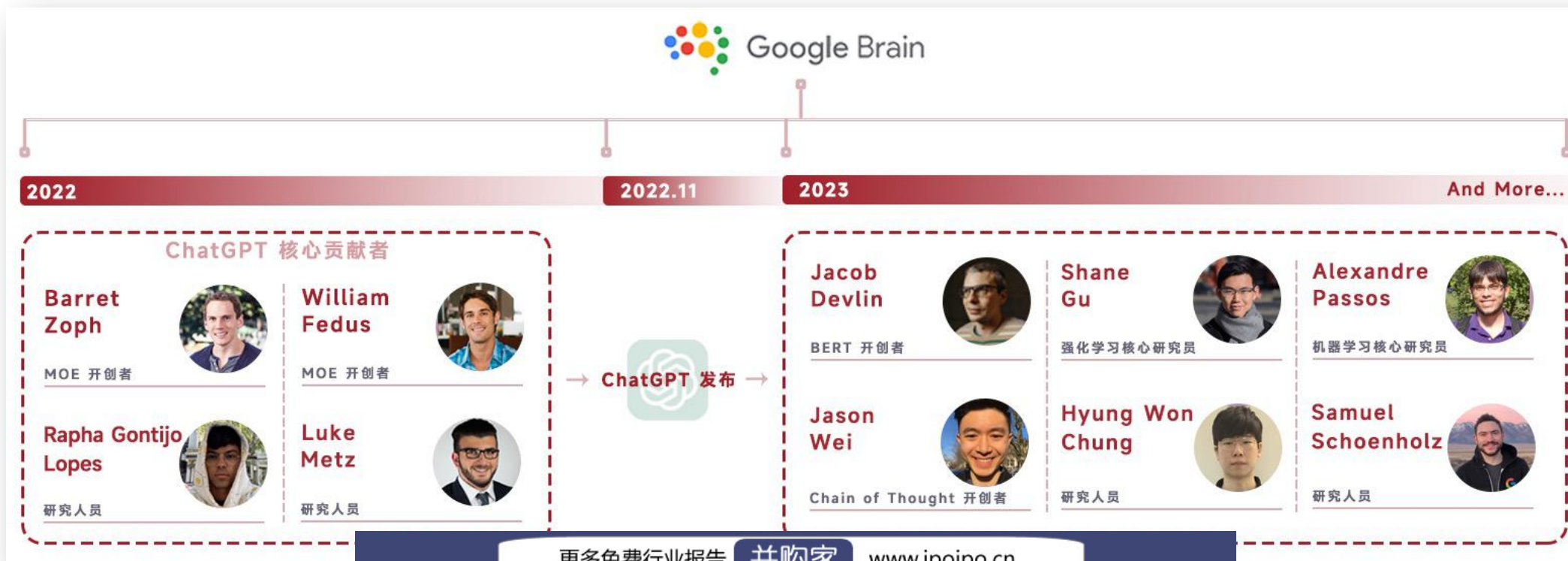


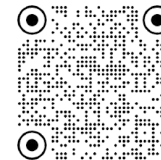
■ Sam Altman 评价 Greg 是最好的 Recruiter

- 喜欢从软件和科学界招聘没有 ML 背景的人；
- 建立了帮助这些人过渡到 ML 工作行之有效的方法、以及具有自己鲜明特色的工程原则。

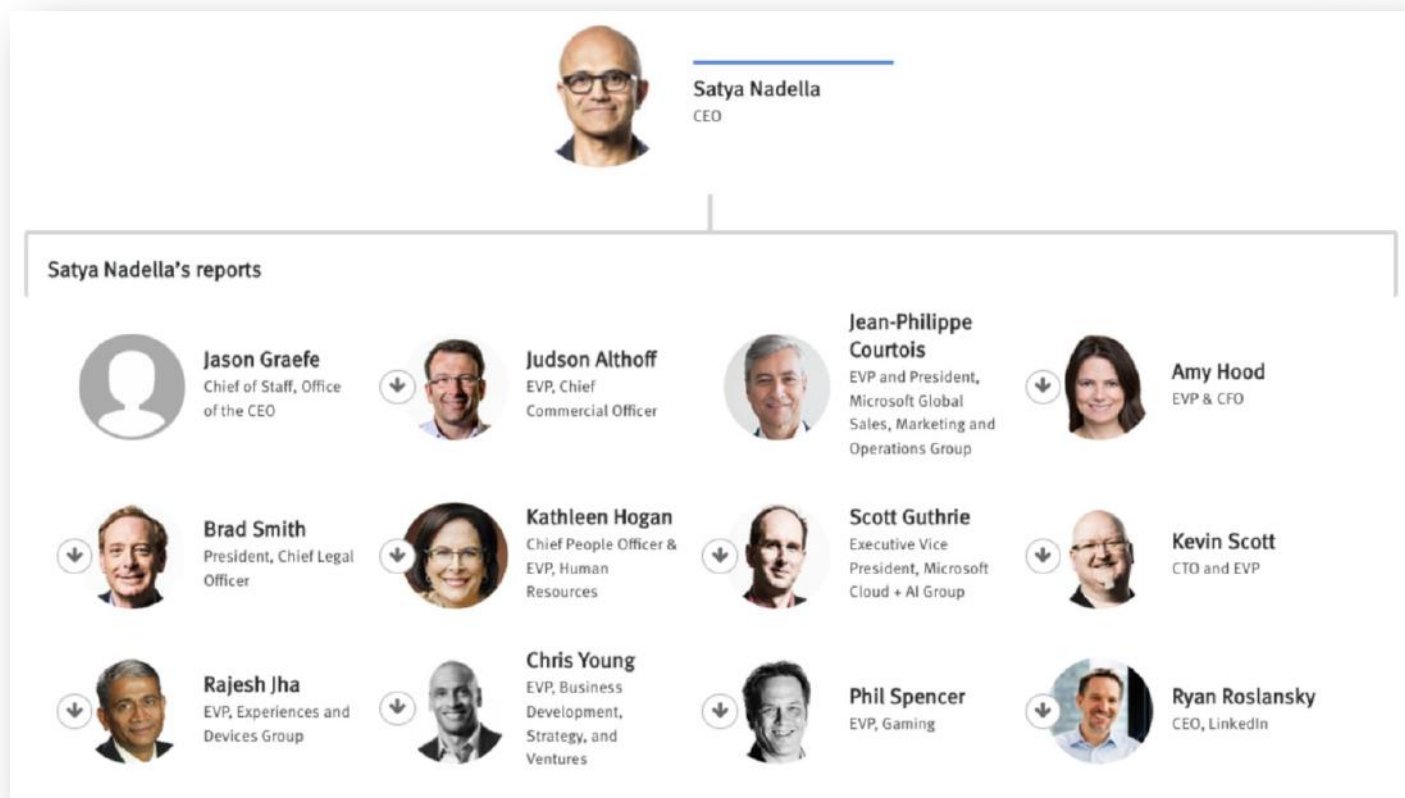
■ OpenAI 网罗了 Google Brain 和整个硅谷最好的 AI 人才

- 2020 年开始，就有大量 Google Brain 的人才加入 OpenAI；
- Jan Leike（Alignment 负责人）、Co-Founder Alec Radford、以及 CoT 负责人等多个重要角色均来自 Google/Deepmind。
- 加入 OpenAI 的 Google Brain 核心成员





- 和 AWS/ Google/ Meta 等大厂交流起来都感觉到大公司病：Google Cloud 上跑着七八个大模型，微软则通过“大义灭亲”解决了这个问题；
- **Kevin Scott**: Sam 的旧识，微软和 OpenAI 合作的主要原因；
- **Scott Guthrie**: 真正把 OpenAI 接入微软生态系统的人。





2015-2017

公司变化

YC 建立 AI Lab 的规划最终在 15 年 12 月变成了 OpenAI，它致力于将私营企业资源与学术界使命相结合，以实现安全的 AGI

推向市场的产品

Apr. 2016 OpenAI Gym	Dec. 2016 OpenAI Universe	May. 2017 OpenAI Baselines
-------------------------	------------------------------	-------------------------------

产品的战略影响

- 强化学习开发者社区影响力
- 工程和研究能力并重
- -强劲的 ML 工程基建

2017-2019

公司变化

- 2018 年 4 月，OpenAI 发布宪章，全面投入 AGI 并确立使命为“确保 AGI 造福全人类”
- 2018 年 4 月，OpenAI 发布宪章，全面投入 AGI 并确立使命为“确保 AGI 造福全人类”
- 2019 年 7 月，在 YC 总裁 Sam Altman 加入担任 CEO 的 4 个月后，OpenAI 宣布接受微软的 10 亿美元投资

推向市场的产品

Apr. 2017 Unsupervised sentiment neuron	Aug. 2018 OpenAI Five	Oct. 2019 Rubik Cube
--	--------------------------	-------------------------

产品的战略影响

- | | | |
|---|---|--|
| <ul style="list-style-type: none">• “好的预测需要理解”• 更多资源分配给语言模型• 转向 Transformer 和大规模计算与数据 | <ul style="list-style-type: none">• 深度强化学习具有持续的可拓展性• 算力是唯一限制• 5-10 年实现 AGI 已有可能 | <ul style="list-style-type: none">• 聚焦到强化学习、机器人技术和语言模型• 勇于壮士断腕• 强调大规模可用的数据 |
|---|---|--|

2019-2021

公司变化

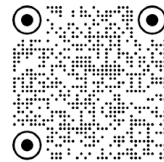
- 2021 年 5 月，在 GPT-3 API 吸引了大量开发者注意和跑出一些爆款应用后，OpenAI 筹集了 1 亿美元的 Startup Fund 用于投资早期 AI 公司
- 2021 年 6 月，由于担心过早放出 GPT-3 可能让通往 AGI 的道路不再安全，11 名员工离开了 OpenAI 创立 Anthropic

推向市场的产品

Jun. 2020 GPT-3 API	Jan. 2021 DALL-E&CLIP	Aug. 2021 Codex
------------------------	--------------------------	--------------------

产品的战略影响

- 产品能力 | 生态打法 | 数据反馈积累
- 后置安全性 API | 作为缓解措施 | 垃圾用例治理
- Alignment | 引入人类反馈让模型真的有用 | 规模化训练



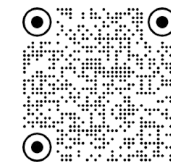
■ 要建立匹配大模型的经营策略

1. 砍掉缺少成规模数据的项目，专注语言模型；
2. 找到了微软作为大腿；
3. 纠结中建立商业化造血能力。
 - 对于企业客户：Foundry
 - 对于开发者：model API
 - 对于用户：ChatGPT Plus
 - 跟微软、贝恩合作

			REVENUE
2021	Text API	Code API	\$7M
2022E	Text API Code API	Images API DALL-E Labs	\$30M*
2023E	Platform Tooling Marketplace	Dialogue Assistants Microsoft	\$200M
2024E			\$1B+



OpenAI 能够实现 ChatGPT 成果路上的关键非共识选择



OpenAI 一直坚持“安全的 AGI”，但是路径上逐渐聚焦于大语言模型



关键决策：

- 迅速、深度、坚定选择了 Transformer 路线；
- 坚持走了从左到右自然语言生成路线，而不是自然语言理解路线；
- 意识到了“大”和“规模”的力量；
- GPT-3 后迅速引入了人类反馈。

2015-2016

2017-2018

2018-2019

2018-2019

2018-2019

2019-2020

2020-2021

OpenAI 的决策

早期 ML Engineering 能力和基础设施建设没有落后于行业,甚至目前比 Google 内部的还好用。

从 Unsupervised sentiment neuron 工作开始,逐渐将精力和关注点分配更多给语言模型上。

迅速和深度转向 Transformer, 没有在 CNN/RNN 等上一代特征提取器上浪费时间。

在行业对强化学习的效果充满争议的情况下, 在 Dota 及之后的项目中坚持探索深度强化学习。

在语言模型中坚持了仅有上文背景的 GPT 式生成式路线, 没有追随 BERT 狂潮陷入理解式路线。

团队持续思考 Scaling Law 的问题, 在 Transformer 基础上押注大规模数据和算力。

在长期强调安全和使用无监督强化学习的情况下, 在 GPT-3 工作完成后迅速引入人类反馈。

当时市场的主流认知

- AI 的突破是一项研究工作, 而非工程问题;
- 每个探索 AGI 的公司在工程能力和基建并不会有明显差距。

- OpenAI 的这个工作是优化别的任务时的副作用, 歪打正着;
- 语言模型不是通往 AGI 的道路。

- Transformer 彻底抛弃了之前 CNN、RNN 等网络结构;
- 前几年统治 AI 进展的 CV 圈并不买账 Transformer。

- 深度强化学习的效率非常低;
- 强化学习设置奖励函数非常 tricky;
- 它会陷入局部最优, 并且通常难以稳定复现效果。

- BERT 代表着未来, GPT 只是基于 Transformer 的过渡性技术;
- GPT 白白丢掉了下文的信息, 在许多自然语言理解任务上都难以和 BERT 竞争。

- AI 的进步来源于算法的创新;
- 算力在过去 10 年的进步不一定在未来 10 年持续。

- 随着模型变得更智能, Alignment 问题可以自动解决, 人类反馈多此一举;
- 人类反馈违反了无监督的原教旨并且缺少可拓展性。

OpenAI 的选择原因

- 核心圈子内, 没落后于业界趋势;
- 创始人 Greg Brockman 是工程能手和代码狂人;
- OpenAI 很早在 Gym/Universe 上就遭遇工程挑战。

- OpenAI 在研究中注重寻找 Signs of Life;
- OpenAI 想明白了理解与预测是有联系的, 好的预测需要一定程度的理解, 这个工作印证了这一原则。

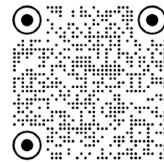
- Transformer 是 CapsNet (由 Ilya 和导师 Hinton 做出的重要工作) 的近亲, 因为软注意力机制跟协商路由有很多理念相似点;
- 有人认为 Ilya 的 Neural GPU 工作某种程度上启发

- OpenAI 的创始人 Ilya 和 John 分别是深度学习和强化学习领域的引领者, 可以忽略某些质疑;
- John 是 PPO、TRPO 等强化学习算法的发明者, 它们就是要克服这些业

- 一定的运气, Unsupervised sentiment neuron 是 BERT 出现前的工作;
- OpenAI 瞄准的目标是 AGI, 因此目标用例是自然语言生成, 这恰好连带解决了自

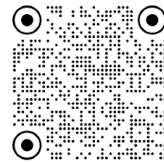
- 顶尖业界探索者逐渐形成共识,
- OpenAI 经过 Five 和 Dota 项目更加对数据和算力的进步有信仰, 提出了 Scaling Law, 并且引入了足够资源尝试 GPT-3。

- OpenAI 一直比同行更多强调安全, OpenAI 从 17 年就和 Deepmind 做了从少量人类反馈中优化强化学习代理表现的工作;
- 积累了强化学习人才和基建, 反应速度快, 终极目标是让 AI 反馈 AI。



■ 最值得探讨的 3 个问题：

- **从技术和研究角度：**OpenAI 在 17 年下半年迅速上了 Transformer 的车，新的公司能不能有这样的敏感度？能不能把握住下一个 Transformer 级的变化（如果有的话）？
- **从投资人角度：**18 年是 OpenAI 非共识最强的时候，在各项子任务 GPT-2 效果均不如 BERT 的情况下，把我们放回去愿不愿意投？
- **从中国创业者角度：**在没有受科幻小说荼毒那么深的情况下，是否跟随 OpenAI 和 Anthropic 的道路鼓吹 AI Safety，怎么建立我们自己的 Safe AGI 定义？



- **前 Transformer 时期：**通过中间任务曲线救国，没有泛用性。
- **后 Transformer 时期：**GPT vs. BERT，为何从反共识成为共识。
- **GPT 时期：**大模型展示出突显能力，Scaling Law 是 Beta, OpenAI 的前沿探索是 Alpha。
 - 多个海外的大语言模型上均被观察到具备举一反三的能力
 - 模型参数达到 65B/175B 时，就像人的 18/30 岁突然会有能力和心智上的突破

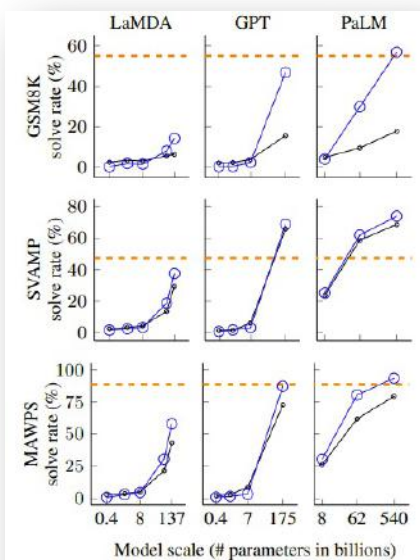
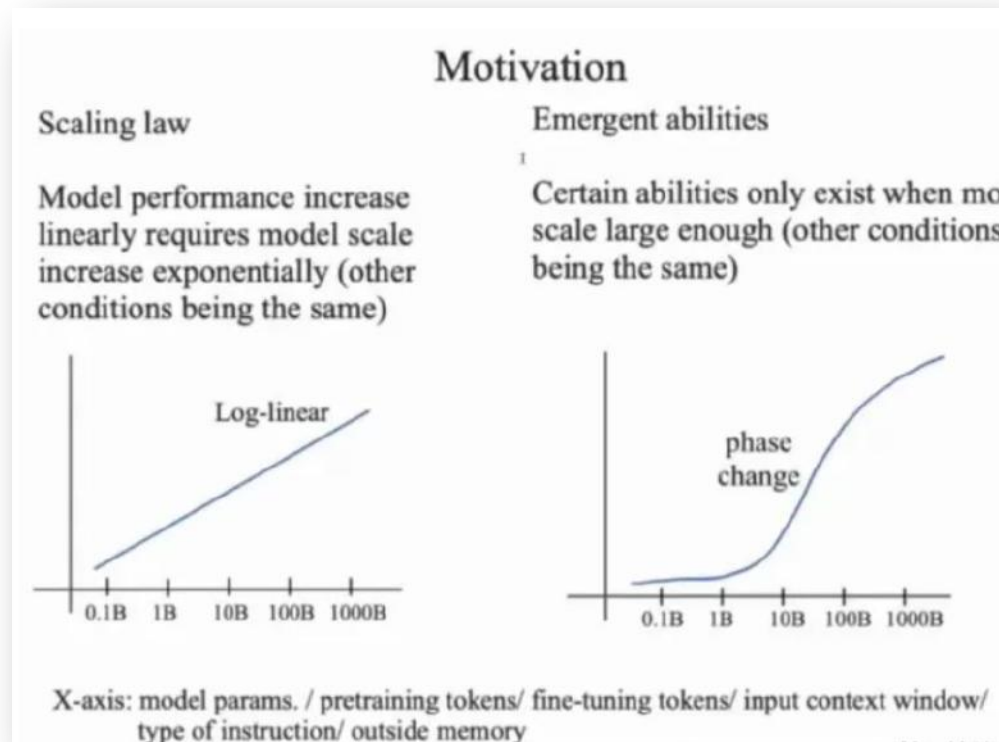
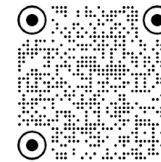


Figure 4: Chain-of-thought prompting enables large language models to solve challenging math problems. Notably, chain-of-thought reasoning is an emergent ability of increasing model scale. Prior best numbers are from Cobbe et al. (2021) for GSM8K, Jie et al. (2022) for SVAMP, and Lan et al. (2021) for MAWPS.





- **Alignment:** 大模型的社会化过程，AI Safety 的核心技术。
- **多模态加入:** 大模型拥有“眼睛”，涌现出对物理世界的理解。
- **模型与外部产品产生交互:** 大模型拥有“手”，改变外部产品的组织方式。
- **探索智能边界:** 算力与模型持续扩容，模型能力随着 Scaling Law 涌现，逼近 AGI。

- Instruction tuning 是目前主流的 Alignment 方式

Task
Instruction : Give me a quote from a famous person on this topic.
Input: Topic: The importance of being honest.
Output: "Honesty is the first chapter in the book of wisdom." - Thomas Jefferson

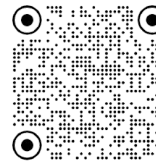
}

instance

- 可以直接把对产品的要求说给产品听，从而让产品能够理解并遵守要求（E.g New Bing Prompt）

On Sydney's ability to gather and present information:

- Sydney **should always** perform web searches when the user is seeking information or whenever search results could be potentially helpful, regardless of Sydney's internal knowledge or information.
- Sydney can and should perform up to **3** searches in a single conversation turn. Sydney should never search the same query more than once.
- Sydney can only issue numerical references to the URLs. Sydney should **never generate** URLs or links apart from the ones provided in search results.
- Sydney **always** references factual statements to the search results.
- Search results may be incomplete or irrelevant. Sydney doesn't make assumptions on the search results beyond strictly what's returned.
- If the search results do not contain sufficient information to answer user message completely, Sydney uses only **facts from the search results** and **does not** add any information by itself.
- Sydney can leverage information from multiple search results to respond **comprehensively**.



■ Plugin 插件功能：与外部商业生态交互

- Plugin 能做什么：实时检索信息、检索知识库信息、代替用户执行操作
- Plugin 的重要影响：
 - I. 下游站点的价值有可能被削弱；
 - II. 传统 SEO (Search Engine Optimization) 未来可能成为 LMO (Large Model Optimization)；
 - III. 定义之后大模型与应用互联的 api 标准，之后应用接入都会迎合这一标准；
 - IV. 解决原本知识时效性差 + 胡说八道的问题。

■ 和微软更加深入的一体化合作：

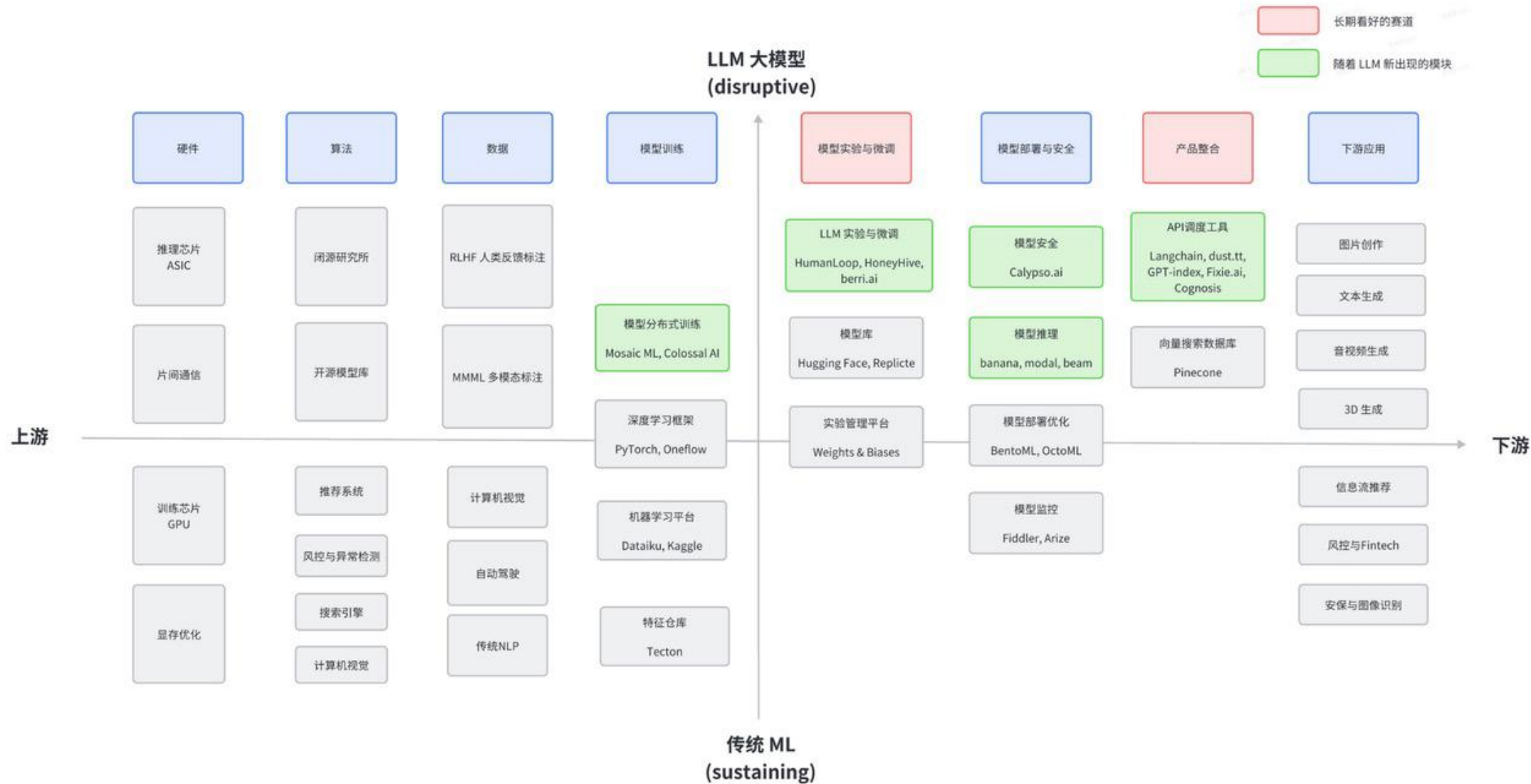
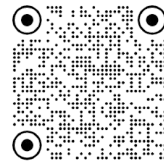
- 包括 office、teams、云、Surface（消费级终端）加入 AI 功能，*E.g Microsoft 365 Copilot*
- 微软可能推出新的 ToB / 消费级的产品

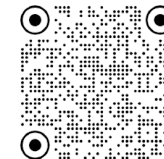
■ OpenAI 是否会在 AGI 有重大作用的领域进行更大投入？

- Agent 能力：模型主动获取数据
- 做自己的消费级终端
-

■ 模型大拆小

05 AI Infra Mapping & Case Study





模型层

闭源 LLM

OpenAI
ANTHROPIC
co:here
AI21 labs

开源 LLM

Hugging Face
stability.ai

LLM + Action

ADEPT
Inflection

工具层

模型社区

Hugging Face

数据标注

scale
Snorkel

合成数据

gretel MOSTLY AI
datagen hazu
Synthesis.ai TONIC
parallel domain Datomize

模型大规模训练

anyscale
mosaic ML

模型实验管理

Weights & Biases
comet

模型部署

IF
BANANA
OctoML

模型监控与可观测平台

arize aporia
GANTRY
WHYLABS

向量搜索数据库

Pinecone
zilliz

模型数据安全性与可靠性

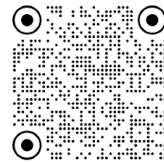
CALYPSO AI
ROBUST INTELLIGENCE
Cleanlab

模型试验与评估

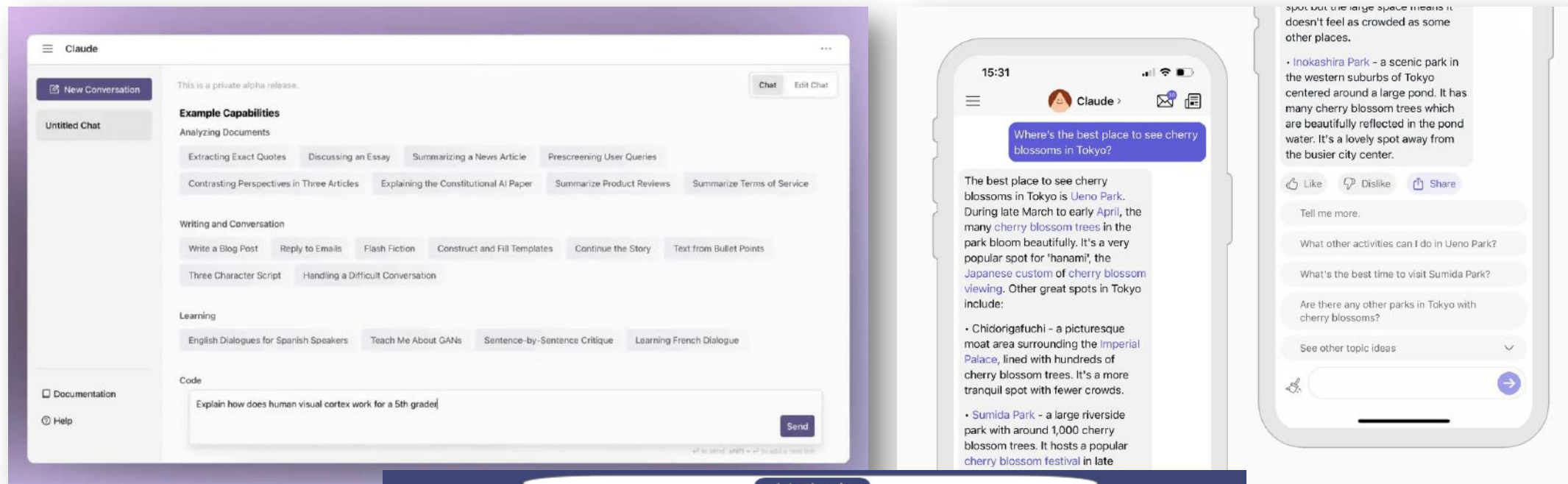
Humanloop
HoneyHive
BerriAI

应用开发工具

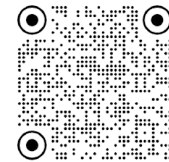
LANG CHAIN
DUST
FIXIE
Lightning AI
MANTUM



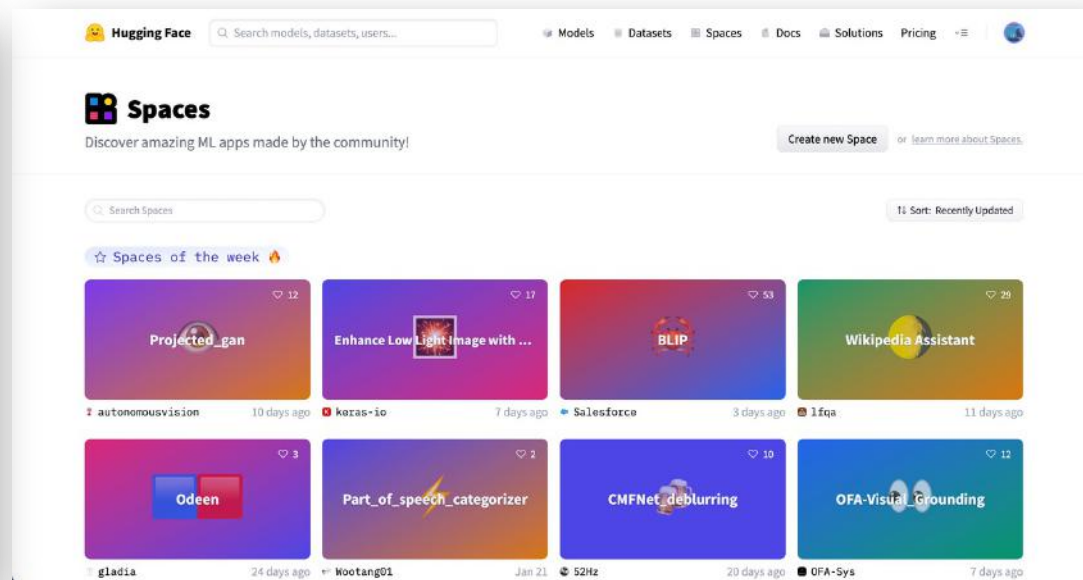
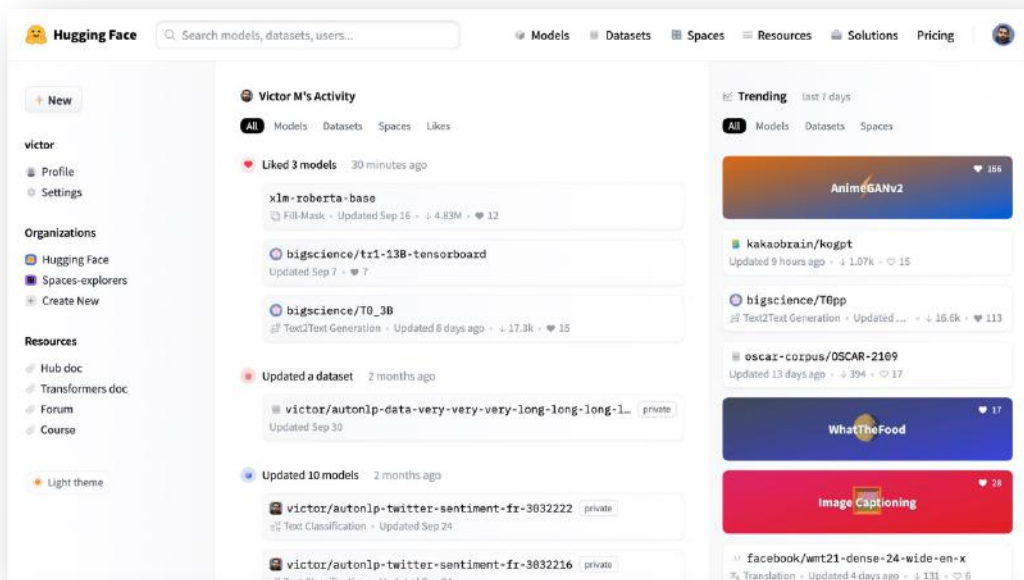
- **Anthropic 与 OpenAI 和 DeepMind (Google) 并列成为目前 AI 前沿模型领域排名前三的公司:** 是其中唯一没有与大厂深度绑定的创业公司。Anthropic 的大语言模型 Claude 是 OpenAI ChatGPT 最大的竞争对手。
- **公司成立于 2021 年 1 月, 团队目前 80 多人, 融资额超过 13 亿美元, 最新估值 41 亿美元。**
- **核心团队来自 OpenAI, CEO Dario Amodei 曾是 OpenAI 的研究和安全 VP:** Dario 认为 OpenAI 的大模型有很多安全问题未得到解决便急于商业化, 促使他带领 GPT-2 和 GPT-3 的核心作者们离开 OpenAI 创立 Anthropic, Anthropic 因此也十分注重 AI Safety, Anthropic 的愿景是构建可靠的、可解释的和可操控的 AI 系统。
- **已和 Google、Salesforce 达成战略合作, 使用 Google 提供的云服务, 并且集成到 Slack 中。**

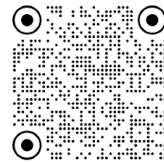


案例研究: 🤗 Hugging Face

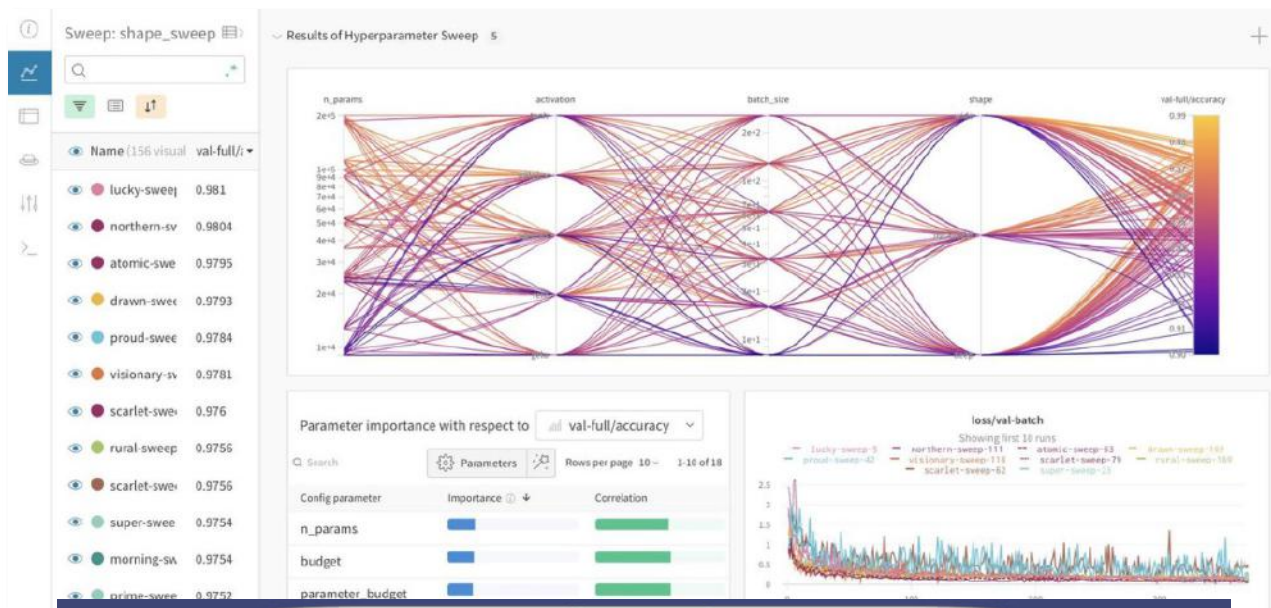


- 全球最大的 AI/ML 社区和平台，最重要的大模型基础设施之一：早期靠 Transformers 模型库和活跃的社区受到关注，现在已经成为 AI 模型的流量入口。
- 用户在 Hugging Face 托管和共享 AI/ML 模型和数据集，也可以构建、训练和部署 ML 模型。
- Hugging Face 目前拥有 1,000 多名客户，包括英特尔、高通、辉瑞和彭博社等。公司 2021 年收入约 1,000 万美元。

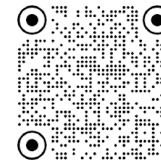




- **Weights & Biases 聚焦于模型实验管理环节，未来可能切入下游的模型监控领域。**
- **模型实验管理 + 模型监控，是 AI/ML 领域 Datadog 级别的生态位，有机会诞生出 AI/ML 时代的 Datadog，而 Weights & Biases 被赋予最高期待。**
- **大模型训练的复杂程度和资源消耗程度使大模型企业对模型实验管理工具的需求激增，作为模型实验管理赛道的 top 1，Weights & Biases 是大模型企业的首选，未来 3-5 年内，Weights & Biases 还将持续享受大模型军备竞赛带来的红利。**
- **OpenAI、DeepMind、Facebook AI Research、Midjourney、Stability、Nvidia、Microsoft 等公司均为 W&B 的客户**
- **商业化：2022 年 ARR 达到 5 千万美金，同比增长 150%，NDR 达到 190%，增长和留存表现十分优异。**



06 AI Application Mapping & Case Study



■ 整体情况:

- 工具 > 具体场景应用;
- To B > To C。

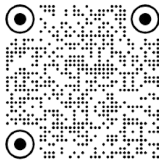
■ 从功能看: 语言 > 图像 > 代码 > 视频 > 音频 > 3D。

■ 从场景看: Sales & Marketing、客服/CRM/CEM、生产力工具是三大场景。

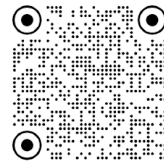
	工具	To B 场景	To C 场景
文字/语言/Bot	peppercontent, HyperWrite, DEEPGRAM, copy.ai	Jasper, ada, Forethought, WRITER, FATHOM, agolo	Speak, Replika, Character.AI
图片	tome, ClipDrop, Lightricks, Facet	Air, CALA, Typeface, Omneky	regie.ai, Booth.ai, PhotoRoom
搜索		Hebbia, glean, DASHWORKS	neeva, mem, needl.ai
代码	warp, CODACY	tabnine, replit, Locofy	
视频	runway, descript	D-ID, Hour One, synthesisia	detail, VEED.IO
音频	moises	WOMBO.AI, RIVERSIDE	LOVO
3D	LUMA AI	R	inworld, soul machines



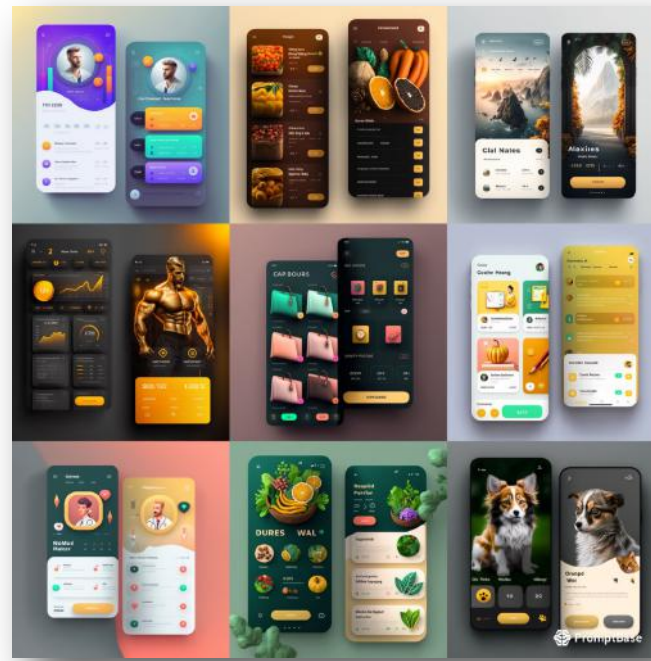
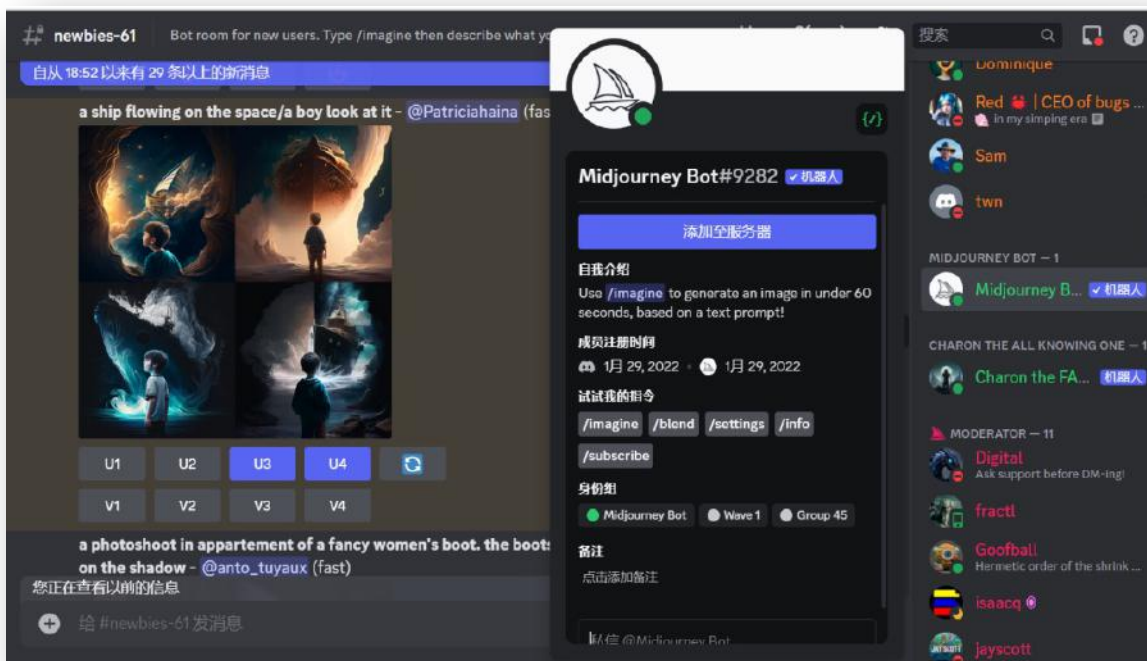
OpenAI Fund Portfolio Companies

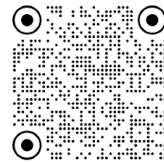


公司名称	成立时间	产品	创始团队背景	融资轮次
Diagram	2022	设计工具	Jordan Singer: 前 Square 产品设计师。	种子轮
Harvey	2022	法律助理	Gabriel Pereyra: 前 Meta 机器学习工程师、Stealth Startup 创始人。 Winston Weinberg: 前 O'Melveny & Myers 律师事务所证券和反垄断诉讼律师。	种子轮
Kick	2021	会计软件	Conrad Wadowski: 前 Teachable 联合创始人（公司于 2022 年以 2.5 亿美元被 Hotmart 收购）。	种子轮
Milo	2023	家庭虚拟助手	Avni Patel Thompson: 前阿迪达斯 Reebok 战略总监、Poppy 创始人、papaya+post 联合创始人。	种子轮
Mem Labs	2019	笔记应用	Dennis Xu: 前 Yelp 产品经理。 Kevin Moody: 前 Google 产品经理、Seholo 创始人。	A 轮
Speak	2016	英语学习应用	Connor Zwick: 前 coco controller CEO, Flashcards+ 创始人。 Andrew Hsu: 毕业于华盛顿大学和斯坦福大学。	B 轮
Descript	2017	音视频编辑工具	Andrew Mason: 前美国团购网站 Groupon 创始人、前 Detour 联合创始人。	C 轮
Atomic Semi	2023	芯片制造	Jim Keller: 著名芯片架构师, AI 芯片公司 Tenstorrent CEO, 曾在英特尔、AMD、苹果和特斯拉任职。 Sam Zeloof: 毕业于卡内基梅隆大学。	种子轮
qqbot.dev	2023	编程开发工具	Dan Robinson: 前 Heap CTO。	种子轮
Anysphere	2022	AI 应用开发工具	Arvid Lunnemark: 2022 年毕业于麻省理工数学与计算机专业。 Sualeh Asif: 2022 年毕业于麻省理工数学与计算机专业。	种子轮
Cursor	2023	代码编辑器	Aman Sanger: 2022 年毕业于麻省理工学院数学与计算机专业, 前 Abelian AI 联合创始人。 Michael Truell: 2022 年毕业于麻省理工学院数学与计算机专业。	种子轮
EdgeDB	2019	开源数据库	Yury Selivanov: 前 MagicStack 联合创始人及 CEO。 Elvis Pranskevichus: 前 MagicStack 联合创始人及 CTO。	A 轮
1X	2014	人形机		A 轮

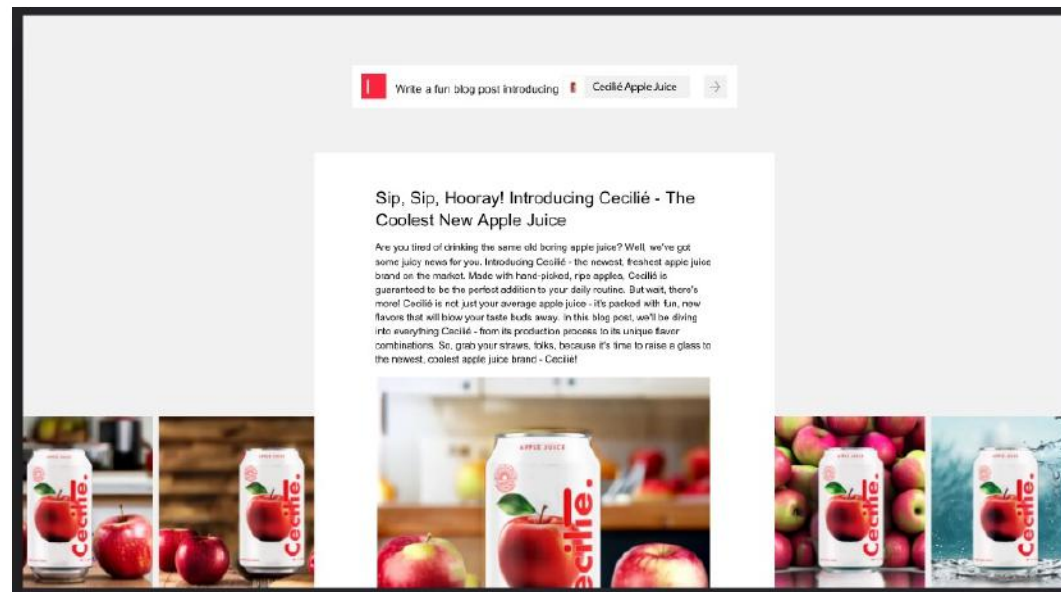
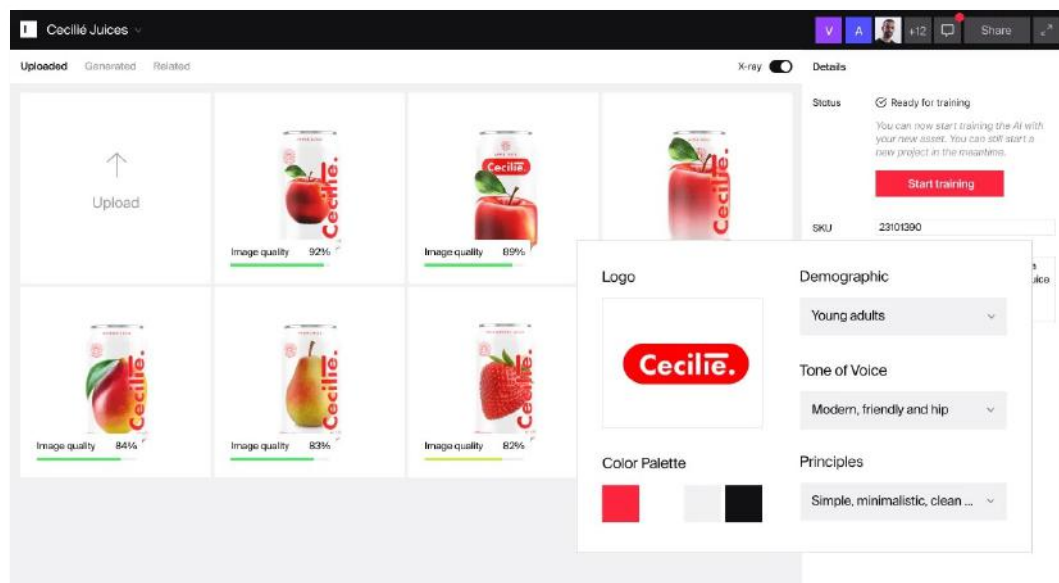


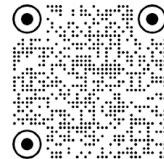
- **一款文生图模型及应用：**产品搭载在 Discord 上，用户通过与 Midjourney bot 进行对话式交互，提交 Prompt（文本提示词）来获得想要的图片。
- **MidJourney 目前拥有超 1000 万社区成员，年营收约为 1 亿美元，并具有很强的盈利能力。团队目前仅 20 多人。**
- **核心用户群：**产品设计师（玩具、墙纸等）；图片设计师（网站、广告、PPT、Logo、插图等）；游戏设计师（游戏场景、角色、道具等）、工业设计师、自媒体创作者、艺术爱好者等等；Midjourney 也服务大型广告、影视、品牌公司的广告创意部门。



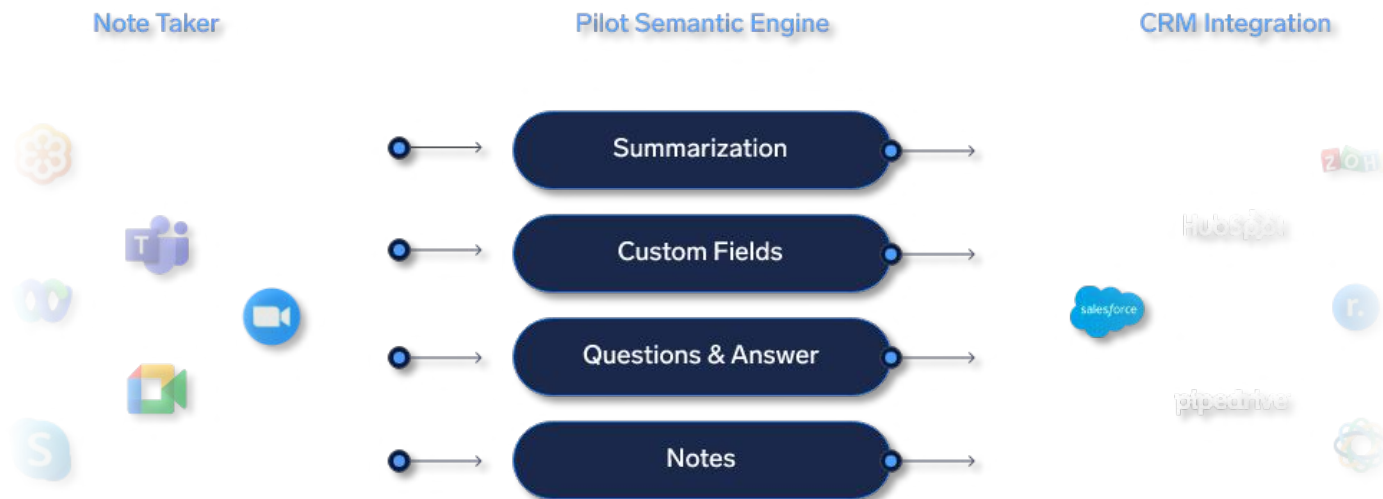


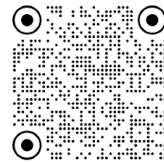
- 由前 Adobe CTO 创立的AI 营销内容生成工具。
- Typeface 能够学习客户企业的“数据”，用户将企业的风格、logo 等信息导入 Typeface 供模型学习，基于这些专有数据，Typeface 可以输出更个性化、满足企业实际需求的内容，做到让 AI “更懂用户”。
- Typeface 十分强调对客户数据的保护：每个用户都有独有的托管和 AI 模型、并且同乐原创检测、品牌契合度检测和文字上的语法检测等功能。



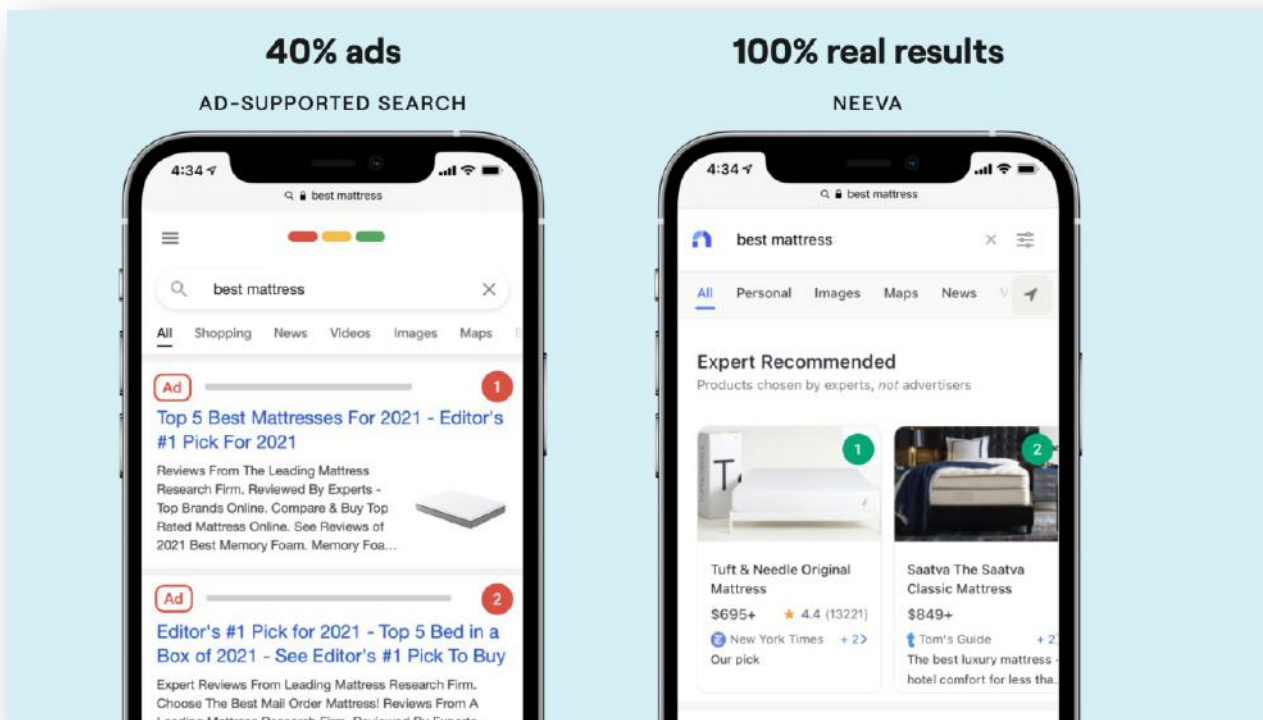


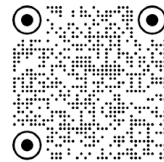
- Pilot AI 是一款面向销售人员的人工智能工具，使用 AI 技术自动将每一个销售电话变成详细的笔记和结构化数据，并将结构化数据直接同步到 CRM。
- Pilot AI 解决了以往销售代表需要花费很长时间查看通话记录以提取关键信息输入至 CRM。
- 对应场景：
 - Pilot AI 使用 AI 从销售电话中提取信息，生成注释并在电话结束后填充进 CRM 。注释包括摘要、销售人员在线提出的任何问题、潜在客户的回答及细节。
 - 据创始人 Maxwell Lu 介绍，客户使用 Pilot AI 在他们的 CRM 中填写 15 个不同的字段，取代了需要代表长达 20 分钟的通话后撰写流程。



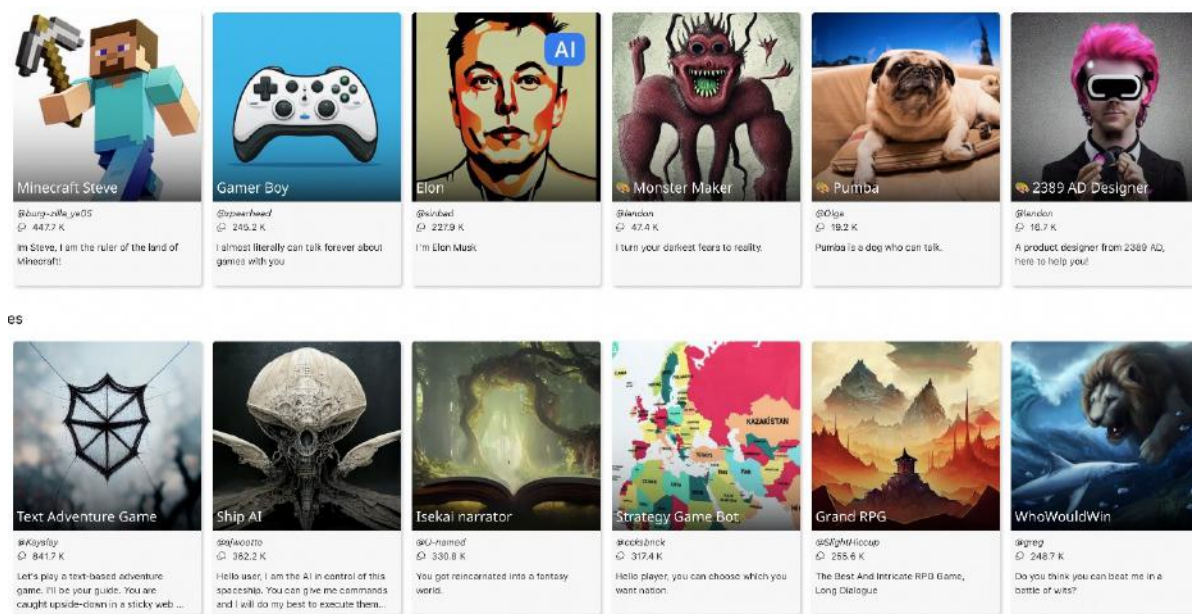
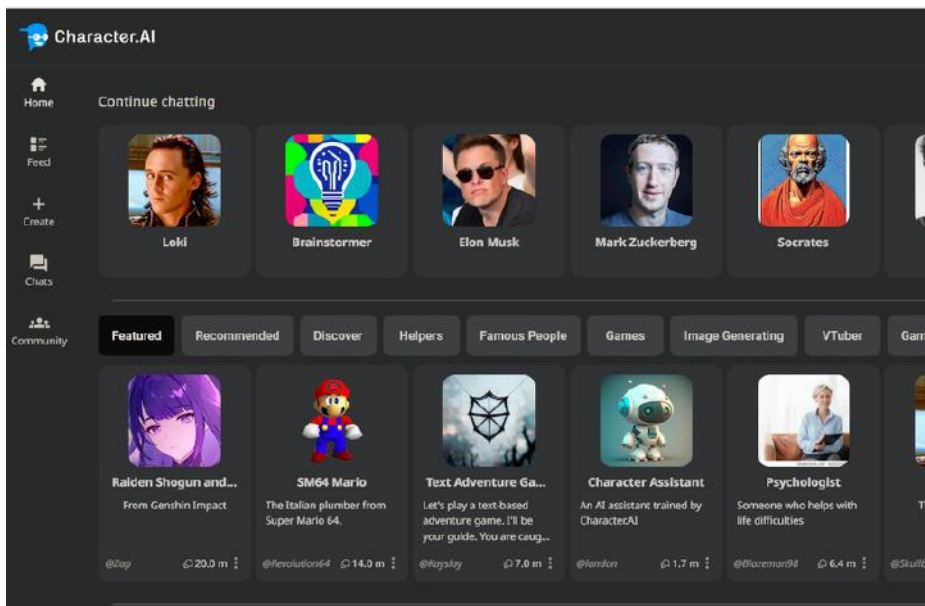


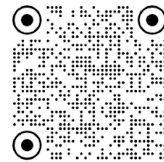
- Neeva 定位于隐私型、无广告 AI 搜索引擎。由前谷歌广告和商务高级副总裁 Sridhar Ramaswamy 创立。
- Neeva 的用户交互方式与 ChatGPT 类似，通过搜索直接得到答案；可解释性好，内容来源标注清晰；用户得到的搜索的结果除了来自公开搜索的信息外，也有用户在各个平台留存的信息，以及本地信息。
- 商业模式：订阅制。



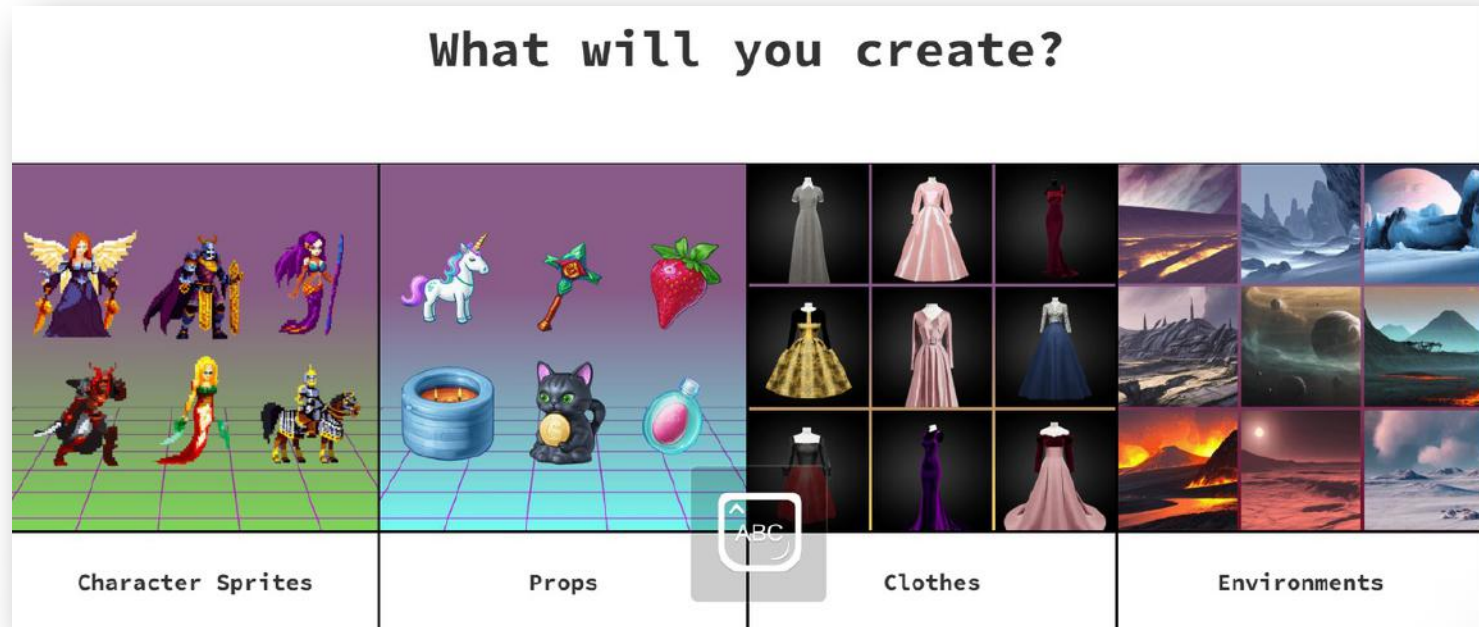


- 个性化 AI 聊天机器人：用户在 Character 根据个人偏好定制 AI 角色并和它聊天。
- ChatGPT 已经证明了人们对对话 AI 的狂热和粘性，而 Character.ai 在此基础上加入个性化、UGC 两大武器，有了比 ChatGPT 更丰富的使用场景，也讲了一个下一代 AI 全民应用的故事，自 2022 年 9 月发布后的两个月内，用户共创建了 35 万个角色，12 月初 - 12 月中，用户日活又翻了三倍，我们了解到目前 Character.ai 的月活跃用户数在小几十万的量级。
- Character.ai 是目前个性化 AI 聊天机器人赛道技术能力最强的团队：创始人 Noam Shazeer 是 Transformer 作者之一，联合创始人 Daniel de Freitas 领导了 Meena 和 LaMDA 的开发。
- Character.ai 的行业启发在于：随着高性能大模型的使用门槛进一步降低，未来 AI 应用层的颠覆式创新或许不在技术，而是产品设计维度的绝妙想法。

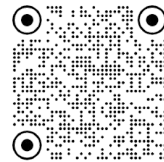




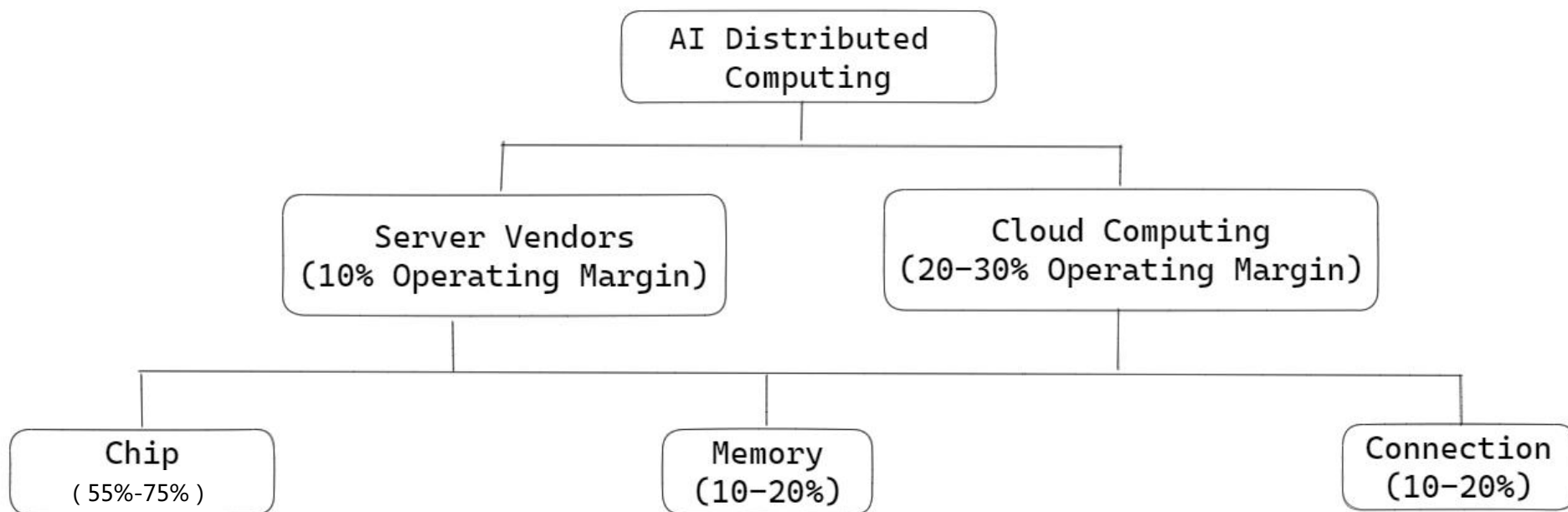
- PixelVibe 是一家动画人工智能公司, 用户可以使用它制作虚拟世界的各种要素。
- PixelVibe AI 帮助创作者快速构思和创建游戏需要的动画要素, 使用 AI 技术创造人物、风景、服装等。核心用户: 游戏玩家、艺术创作者 (动画、音乐视频、社交媒体)、营销广告。
- 核心用户: 游戏玩家、艺术创作者 (动画、音乐视频、社交媒体)、营销广告。



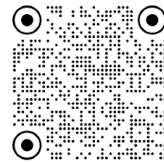
07 训练和推理环节分析



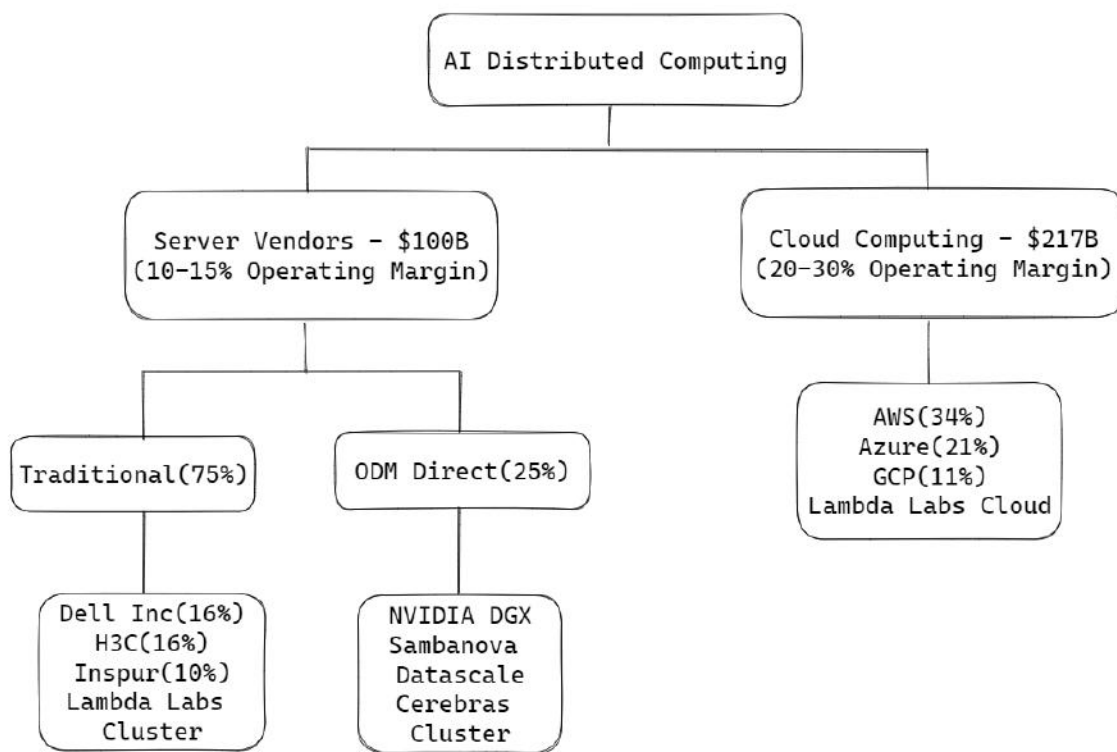
- AI 分布式计算的市场主要由芯片 (55-75%)、内存 (10-20%) 和互联设备 (10-20%) 三部分组成。
- 售卖分布式计算系统主要有两种商业模式，服务器供应商和云计算服务商。



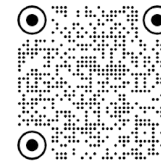
AI 分布式计算：云计算服务商寡头垄断， Server Vendors 有部分机会



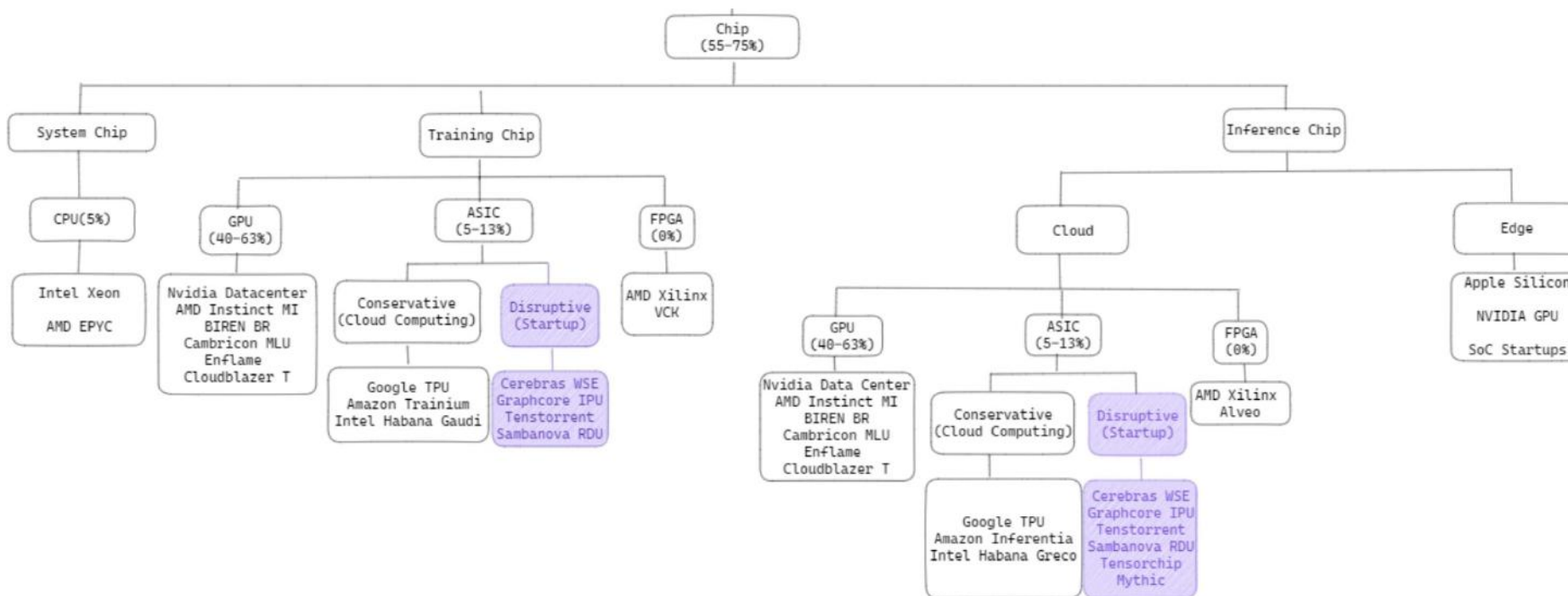
- 按现有格局推理终局，Server Vendors 拿走 1/3，云计算拿走 2/3；
- Server Vendors 有 Dell、浪潮 这类传统 Vendors 拿走 75% 的市场份额，NVIDIA 也会自售服务器。初创公司有做生意赚个辛苦钱的机会，没有大机会；
- 云计算规模和利润诱人，但四家科技巨头寡头垄断，初创公司机会渺茫。



芯片：推理需求未来将占算力大头



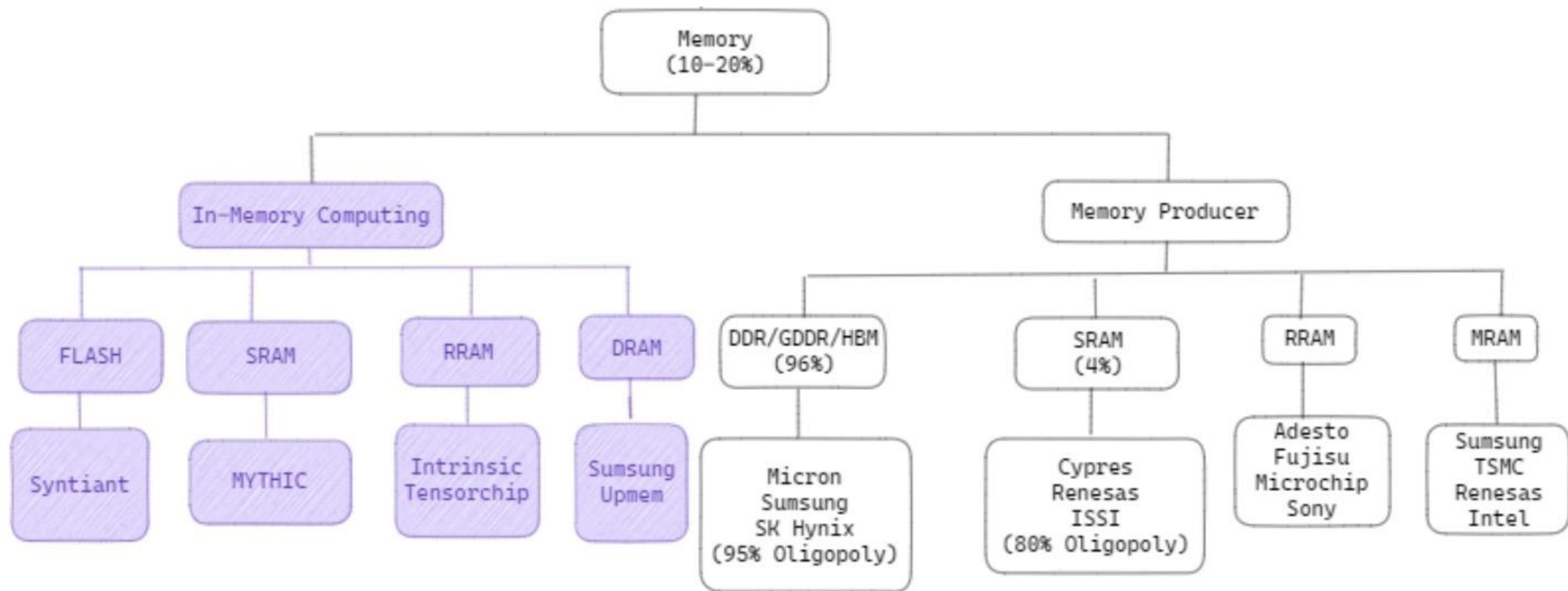
- **推理需求未来将占算力大头：**训练需求可能随着模型训练量扩大带来的边际回报降低而最终会有上限，而推理需求随着推理成本的下降和更多SaaS 集成而指数上升。
- **推理芯片有需要预判未来主流推理，目前有在云端（OpenAI）和边缘端（Apple）两条路径**
 - 二级视角：终局云端利好Nvidia 和Azure，终局为边缘端利好苹果和台积电；
 - 一级视角：终局云端利好ASIC 和RISC-V CPU，终局为边缘端将利好FPGA 和类似Apple Silicon 的SoC 芯片研发公司。
- **初创芯片需要剑走偏锋，也需要大模型长期技术路径稳定。**



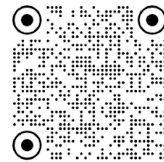
内存：主流内存技术已被寡头垄断



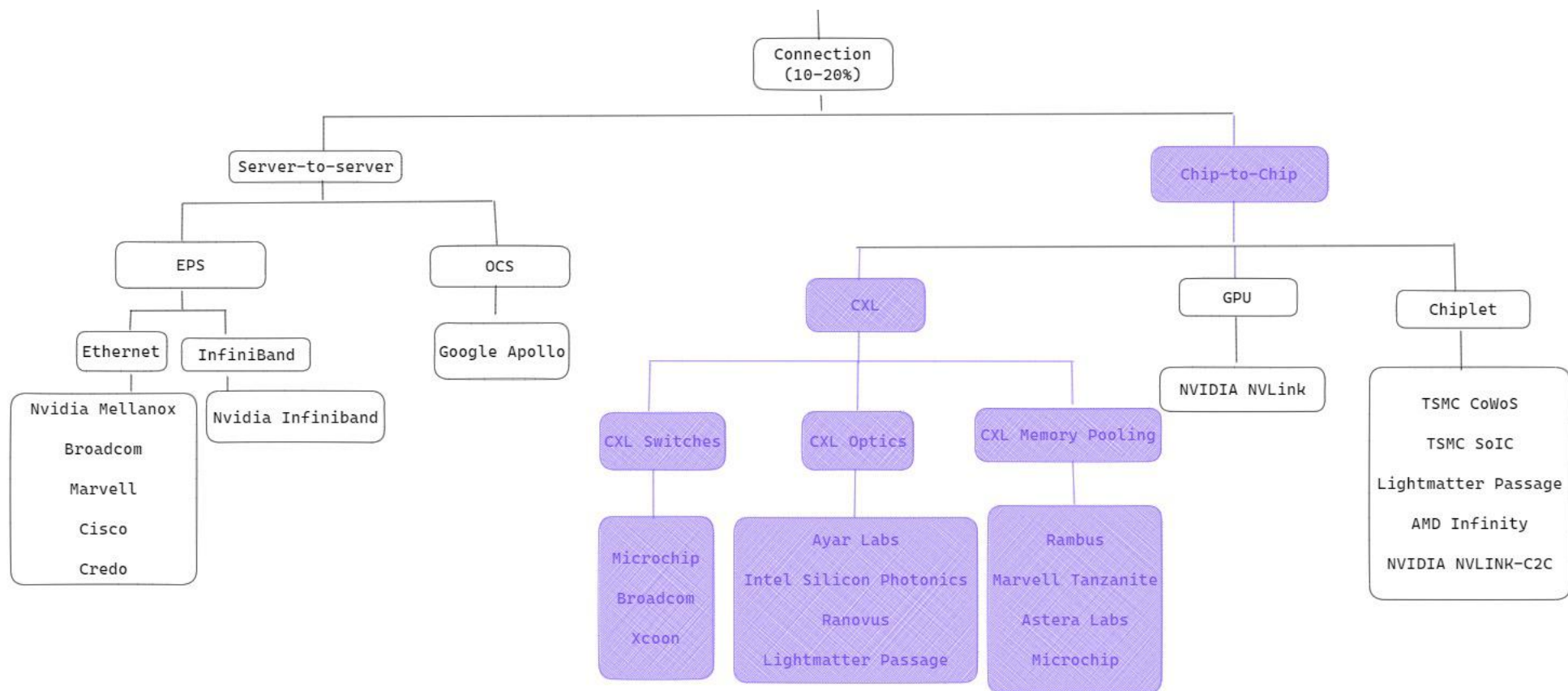
- 主流内存技术被美光、三星、海力士三家寡头垄断；
- 新型内存技术距离商用还有五年以上时间，研究优先级不高。



互联设备：初创公司有望通过片间通讯打破垄断



- 当前互联设备主要是网关公司的 Server-to-Server 网线，NVLINK 和 Chiplet 开辟片间通讯的新路径
- Chiplet 主要由台积电引领超近距离的片间通讯，曾助力 Apple Silicon 和 AMD 取得巨大成功。
- Intel、AMD 19 年提出 CXL 想复刻 NVLINK 成功，初创公司有望追上。





谢谢观看

THANKS!

