State of LLM

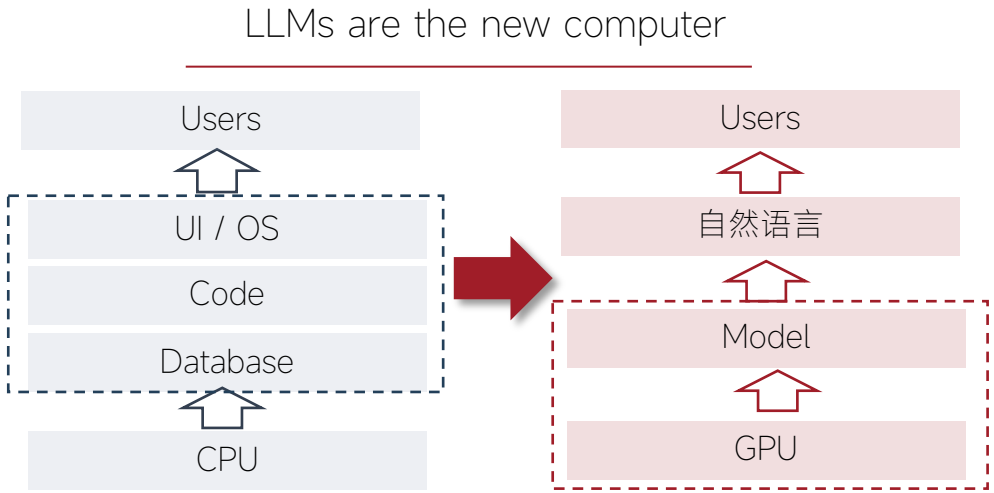
拾象分享 | 2023.07

01 关键结论

01 模型格局：决定下一代超级入口

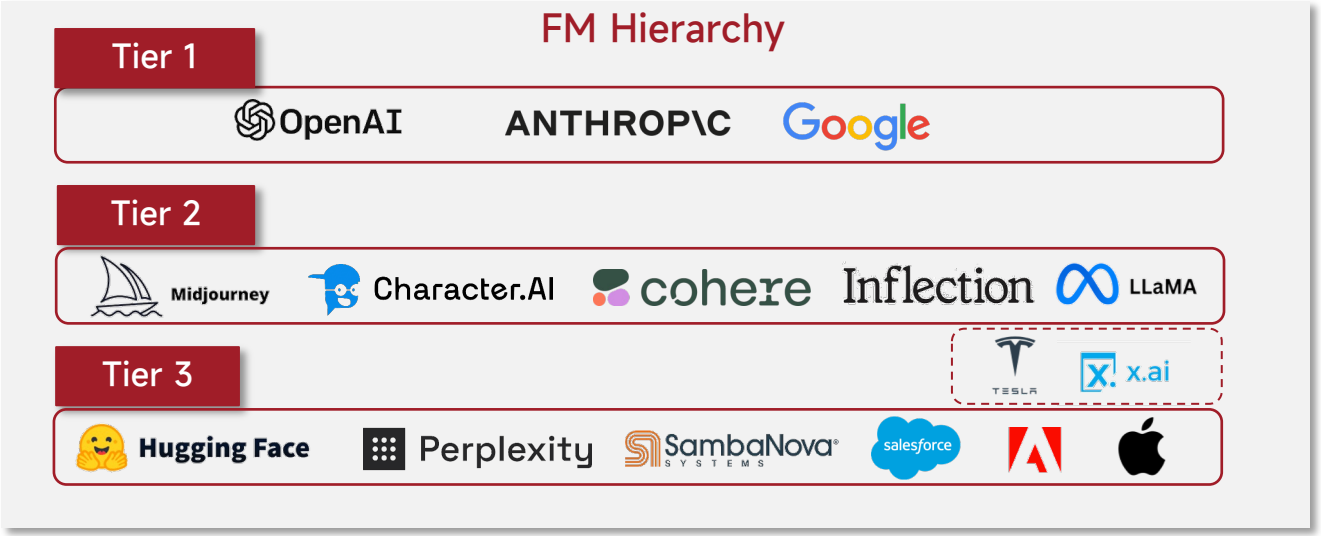
对 GPT 的最新认知：

- GPT不止模型，而是新一代超级计算机，重构“用户交互+软件执行+计算”
- 模型即产品：ChatGPT / MidJourney/ Character
- 自然语言 = API



对模型格局的猜想：

- **类似航空航天**：OpenAI 登月（SpaceX）VS 其他（波音/空客/庞巴迪/湾流，but…）
- **属地独立市场**：US/China…日韩/中东/欧洲/东南亚，和语系相关
- **本质是对入口争夺**：生产力入口、助理入口、娱乐入口…



02 数据是GPT的核心秘方，来源于大量实验工程(GPU)

- GPT 的北极星能力：复杂推理能力（写代码、解题、处理复杂任务）
- 众多 ChatBot 们为何推理和解题能力都还不行？智力来源在哪？
- 代码数据带来推理能力，GPT pre-training code data 比预期要高，猜测接近 50% 占比
- Google 受限于大公司身份，很多数据受到版权限制不能用于模型训练
- 做模型像做菜，原材料、配比、排序、拼接决定口味差异

03 OpenAI 的下一代模型：AI 登月时刻

- **Scale:**
  - Compute: ~10万张H100, ~50亿美元投入
  - Data source: Video + 合成数据
  - Multimodal: AI that can use computer to do complex knowledge work tasks
  - Coding/Tooling: run & debug & use APIs
  - 能力上接近 AGI
- **Cost**
- **Latency/Speed**
- **Hallucination**
- **Excellent API**
- **拉开差距，追赶难度骤然变大?**

拾象科技
SHIXIANG TECH

04 关键胜负手

- 数据
- 人才密度
- 组织能力
- GPU 资源
- Killer App

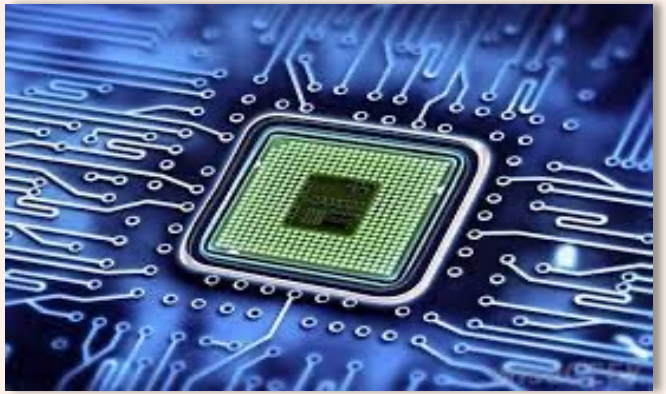
02 Key Takeaways From OpenAI

训 LLM 比造原子弹更复杂，每一代大模型发布都像登月

“简单”技术，可堆积、研发方向明确，涉及技术项目有限，只要造出来就能形成威慑，无需考虑商业盈利。

■ 原子弹

每一步都需要实验和聪明人的奇思妙想，技术上有无数小细节，无法大力出奇迹。



■ 芯片

准备周期长，正式 launch 之前的实验都是在地球上做的，和真实环境相差很远。

■ 登月

LLM：正式训练前准备至少 8 个月，正式训练成功率 50%，内部无数 tricks。

全公司重点攻坚下一代模型训练，一个月内决定数据解决方案

- 大规模服务客户和训练下一代模型的主要瓶颈是 compute
- 做模型越久，越认为模型和硬件应该 co-design

	GPT-3.5	GPT-4	下一代模型	优势	挑战	解决方案
模型架构	小于 50b	MoE/ 16 experts/ 800b 参数	多模态， 十倍参数量级	接近AGI	Cost/Latency	---
Compute	训练成本和推理成本都比 Llama 低	5 万张 A100 实验和训练， 3-4 万张A100 做 inference， 人均 500 张卡做实验	10 万张 H100	GPU 使用率能达到 50%，远超行业水平，微软支持	Cost、互联、物理空间，无法支撑大规模应用	模型和芯片 codesign
Data	3-5T	30 T	多模态数据	没有版权限制，独有 recipe	可用数据量不够，多模态数据质量不够高，Video 数据无法有效 process	各小组提出解决方案/模型生成数据/新架构

OpenAI 内部如何看待 AGI

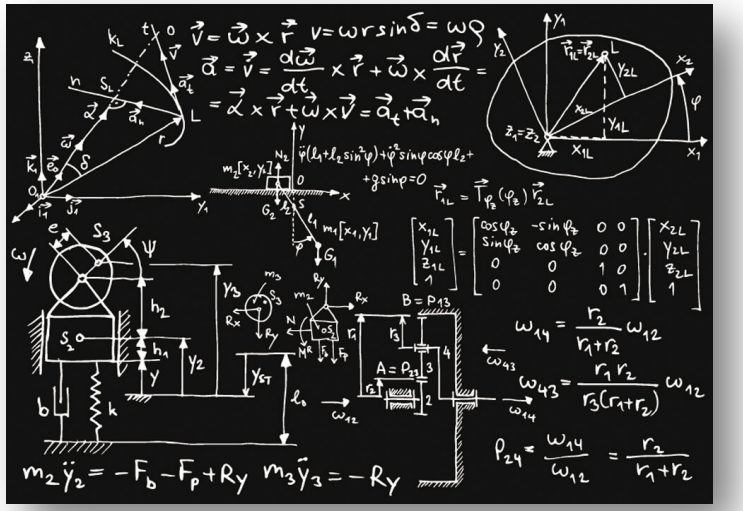
> 非常有信心下一代模型会接近 AGI

短期

Coding 等能力超过 50% 人类，AGI 代表能做绝大部分知识工作者的工作。

中期

解决大部分数学问题，写代码能替代人——模型和世界交互的方式就是代码。



长期

用模型发现新的科学定律是终极目标，挑战是如何让模型和现实世界产生交互

ChatGPT 是另一个重点，目标达到 10 亿用户

重视程度

- 除了下一代模型外最重视的产品
- 招聘、Mobile 版本、dedicate compute 资源

Use Case

- Productivity 场景彻底代替 Stack Overflow
- 20% 是 Education，比人类助教优秀很多

用户行为

- 2.8 亿 MAU，周末数据下跌（说明 entertainment 的场景少）
- 用户会自己知道如何使用 Google 和 ChatGPT

增长预期

- OpenAI 今年预计 10 亿美金收入，20 亿美金 ARR，ChatGPT 占 70-80%
- ChatGPT 目标是达到 10 亿用户，和 Office 一个量级
- 目前付费用户是 MAU 的2.7%，约 700 万

商业模式

- Working assistant
- 企业客户定制

提问分布数据最有价值，未来 Plugin 数据可用来操作计算机

ChatGPT	通过用户提问筛选出 45 万条高质量提问分布，非常宝贵，无法通过人工标注者获取。
Google Bard	活跃用户太少，无法获取足够用户提问。
Plugin	数十万用户，还很早期，暂时还没用用户数据做训练 ，未来希望训练模型操作计算机。 挑战是精确度必须很高。

开源社区模型不具备真正智能，更好的小模型来自大模型的 Scale Down

- 开源社区也许也能获得高质量数据，但如何把握配方和各种超参数很难
- 开源软件某部分没写对仍然能 work，但做 LLM 必须全部正确，没法小修小补
- 未来趋势：大模型 explore 新方法，scale down 做小模型，大模型和小模型互相交流
- OpenAI 的模型就是从大到小，以前不愿开源，未来可能开源

OpenAI 组织力优势仍然很强，合并后的 Google Deepmind 或难竞争

	信息透明	人员架构	leader	管理方式	底层研究	人员流动
OpenAI	所有人能看到所有 documents	100 个 researcher，300+ engineer，项目导向	Greg/Ilya/Sam/John Schulman 等人不喜欢管理，真正管理的是 Bob McGrew，制定行动路线，分配资源	自上而下为主，部分自下而上	较少投入	很少有人离职，仍然有人从 google 加入；有人去了马斯克那
Google DeepMind	不透明	合并后近 1700 人，Google Brain 600 人，DeepMind 1100 多人，Gemini 从最初的 20 人变成 300 人	DeepMind 成为 leader，合并造成的组织问题依然严重	过于自下而上	以前更多投入，最近在变化	没有从OpenAI过来的，有人去Apple

减少 hallucination 可通过 scaling/retrieval/reward model

Hallucination 的原因：

- 网上很多信息本来就是错的，学到了错误信息
- 模型喜欢模仿语言风格，对正确信息判断不好

减少 hallucination 的方案：

- 目前 20% 的 hallucination 可以通过 scaling 降低：看到很多问题都能通过 scaling 解决，OpenAI 相信 hallucination 也可以
- 如果 inference 的成本能降到特别低，latency 也能特别低：可以让模型在回答问题前尽可能多地去做 retrival 和 verify，就像一个人的思考如果变得很便宜，就可以让它可能多地去思考

03 Hidden State of GPT

LLM 训练中的事实与观点

	Facts	Opinions
Pre-training	• GPT 系列是小创新乘起来带来的成功 • 训练数据量远超其他大模型 • 预训练阶段使用工具：Ray & Wandb	• 更长的模型输入窗口是一个近期会持续突破的问题 • 预训练数据集的比例会直接影响其模型的效果，Code data 比例很高 • 当模型大于 Llama 这个量级之后，开源团队会遇到瓶颈
Fine-tuning	• 多模态并非预训练一体的模型结构	
Reward-modeling	• OpenAI 的数据飞轮效应有限 • 外部数据标注分散给多家公司	• Chat 不是一个适合收集数据反馈的产品形态 • 精挑细选的反馈数据更有价值 • 机器能高质量的反馈打分，这一步的 Human in the Loop 会逐渐削弱
Reinforcement Learning	• 难度大且不稳定，目前做成的只有 OpenAI 和 Anthropic	• 开源模型在使用 RLHF 之后普遍没有明显提升 • Direct Preference Optimization 等方法出现后，强化学习不再是必须路径

LLM 技术应用的新趋势

- **好的 LLM 扮演人类思考的系统 1，好的 AI 应用扮演人类思考的系统 2**
  - prompting 方式的变化将会与更好的 Python glue code 框架一起进化，简单的 prompting 交互只能达到 System 1 的效果
  - 高级的 chain & agents 才能接近 System 2 的能力
- **LLM 是新一代计算机，有更接近人类的智能与 probabilistic 输出，与传统软件的 deterministic 输出不同**
- **API 能力将继续进化**

** Tree of Thought 是典型的的系统 2 模拟

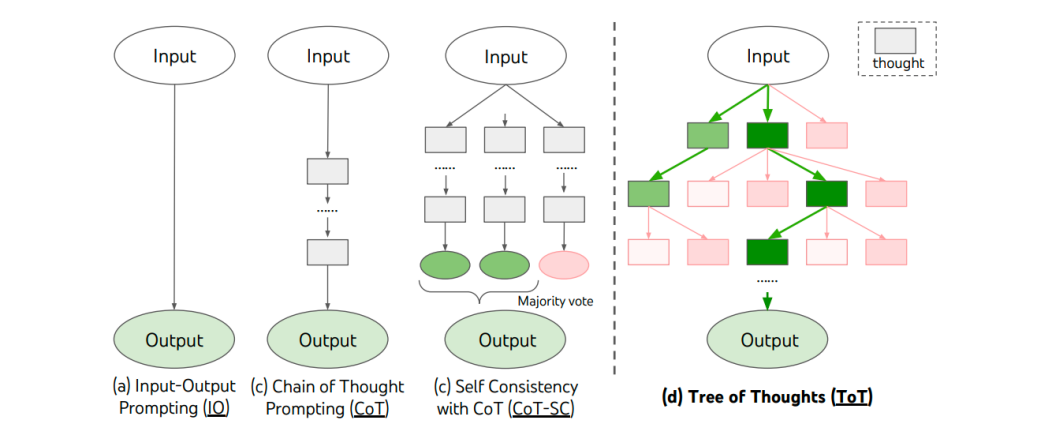
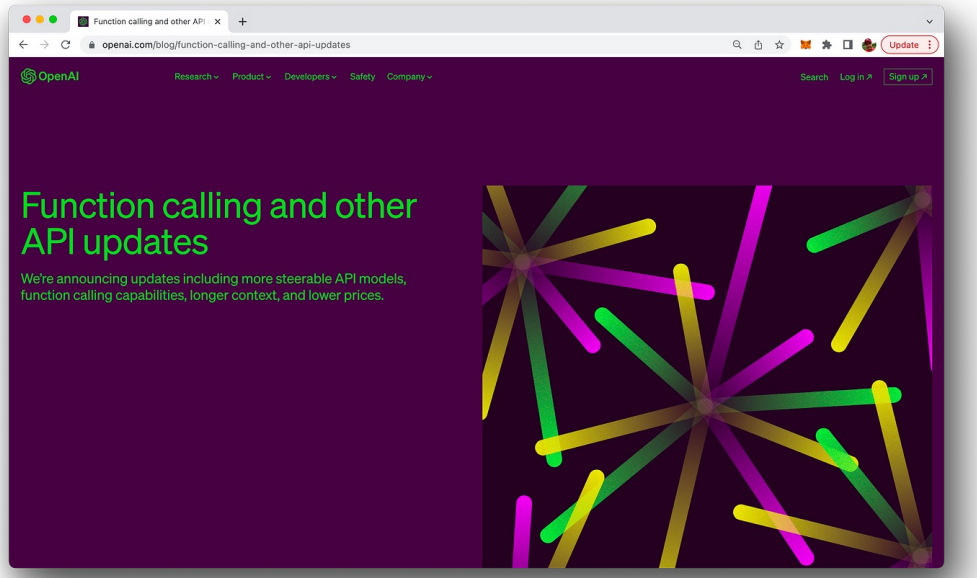


Figure 1: Schematic illustrating various approaches to problem solving with LLMs. Each rectangle box represents a *thought*, which is a coherent language sequence that serves as an intermediate step toward problem solving. See concrete examples of how thoughts are generated, evaluated, and searched in Figures 2,4,6.

** Function Calling 的进化是未来趋势，也是 LLM 迁移成本的开始

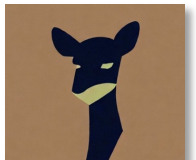


04 硅谷开源模型社区近况

指令遵循的开源小模型被过度炒作，已经到达瓶颈

- 今年第一季度的炒作“LLaMA 7B/13B + 指令微调 = GPT 3.5 Level” ✗

Stanford Alpaca



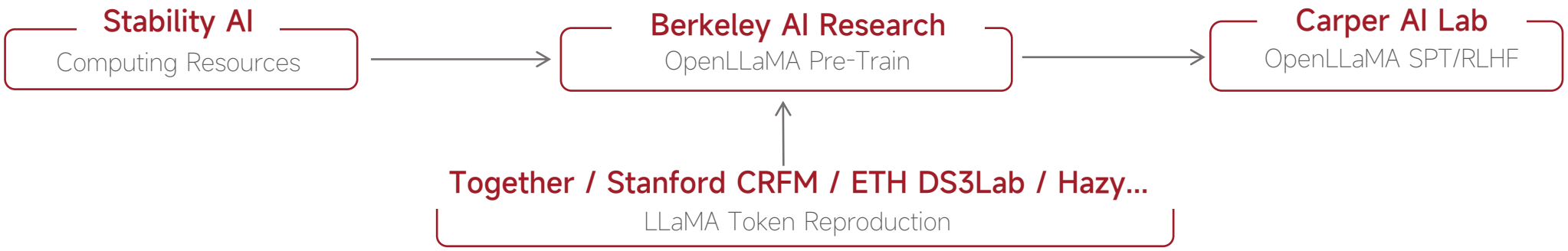
.....

- 目前遭遇的卡点：
  - 在复杂推理和代码等表现不好的任务上的难以进入正循环
  - 最好的指令微调数据集仍然来自先进模型蒸馏
  - 对“高质量”数据和“更好的”模型缺少标准化评估
  - 不可商用
  - 开源小模型团队以 PhD 为主，缺少懂产品的人来解决这些问题

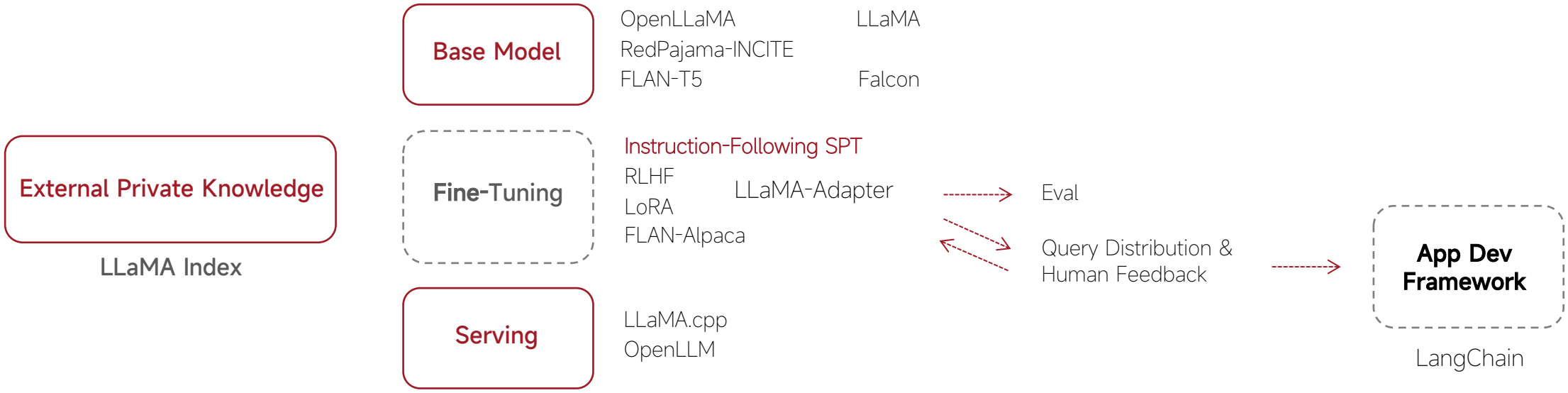
- 突破瓶颈的方向目前看是更强的 Base Model 和比 SFT 指令数据更进一步的反馈数据

注意力转向 Pre-Train 环节，开源社区靠“团结”发展

- OpenLLaMA 7B 和 13B 的分工展现出开源社区在核心项目突破上正在变得更团结；



- 此外，从全链路来看，开源社区已经发展到了全覆盖的阶段，每个环节都有头部的开源项目可用：



开源模型能否替代 OpenAI API？客户乐观，研究者悲观

- **生态里的不同角色对开源模型的前景呈现两极分化：**

角色	态度	原因
研究开源模型的 PhD 和科学家	悲观	清楚跟 OpenAI 等闭源头部公司的三大差距，第一缺算力，第二难以追上的时间差，第三是缺少 ChatGPT、Claude 这样的全栈产品。
“卖”模型的部署咨询公司	乐观	客户往往因为 OpenAI 开始考虑接入 LLM，但是会因为成本、数据安全、用户隐私、ChatGPT API Session 数限制等问题选择开源模型，FLAN-T5 仍然是热门选择。
用模型的科技公司	乐观 (盲目)	哪怕是应用层公司也希望构建自己在模型层的竞争力，复制 Midjourney 的战略，因此创业公司在早期就会注重未来可切换模型的平台建设，而有体量的公司则一开始就采取 OpenAI、Anthropic 及开源混合的多模型策略。

- **从客户需求的迫切角度出发，开源模型目前的“可商用”进展比“更智能”更重要**
- **OpenAI Foundry 的内部投入没有我们预期那么大**

开源还是个社会问题，落后于 OpenAI 不一定是坏事

- **我们在湾区不止一次听到这个类比：将最先进的 LLM 开源相当于把原子弹放在每个人手里**
- **开源社区的发力方向不需要是超越 OpenAI：**
  - 没有一个开源社区可以豁免开源先进模型带来的威胁
  - EleutherAI 直接选择不致力于发布推动智能前沿进展的模型和功能，而旨在特点情况下发布合适大小和智能用例的 LLM
  - 蒸馏的做法抽象看并不差，领先的模型推动 AGI 并且帮助将不会带来智能威胁的模型优化得更实用

- **技术之外，法律和政策制定有进化的空间，许多社区成员认为 Apache 2.0 这样用于软件的协议实际上并不适用于模型**

05 Robotics

软件：Robot Learning 已成为学界主流研究路线

什么是 Robot Learning?

- AI 和 Robotics 的交叉研究领域
- 机器人通过算法学习获得新技能，适应新环境
- Learning vs. 传统控制
- Imitation Learning vs. Reinforcement Learning

重要玩家

早 期 → 现 在

- UC Berkeley
- Google
- OpenAI（后解散）

- 大学：UC Berkeley、MIT、Stanford…
- 软件：Google、Nvidia、Meta、Covariant…
- 硬件+软件：Tesla、1X、Figure、Boston Dynamics…

行业代表人物：

Pieter Abbeel

Sergey Levine

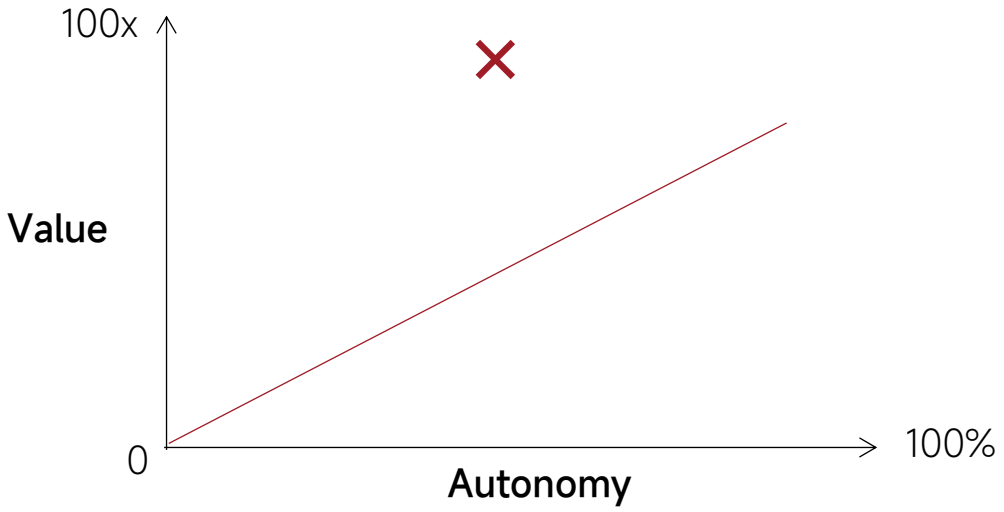
软件：当我们关注 LLM+Robotics，我们在关注什么？

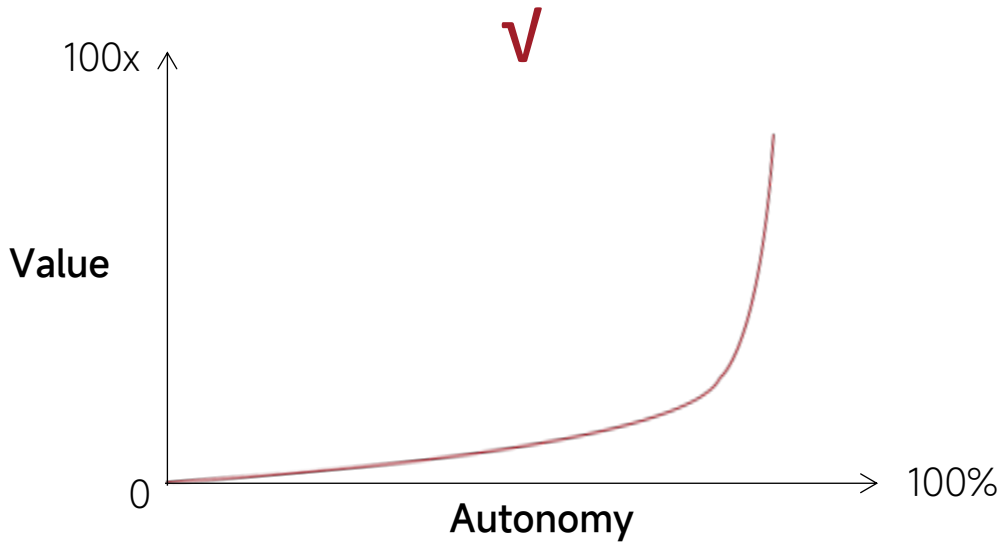
LLM 给 Robotics 带来了什么？

- 人们可以用自然语言给机器人发送指令
- 机器人能够理解人类指令，自主拆分成相应步骤并执行
- 机器人能够理解和应用世界常识，完成此前没有学习过的任务

瓶颈

- Low-level Policy
- 价值与财务模型



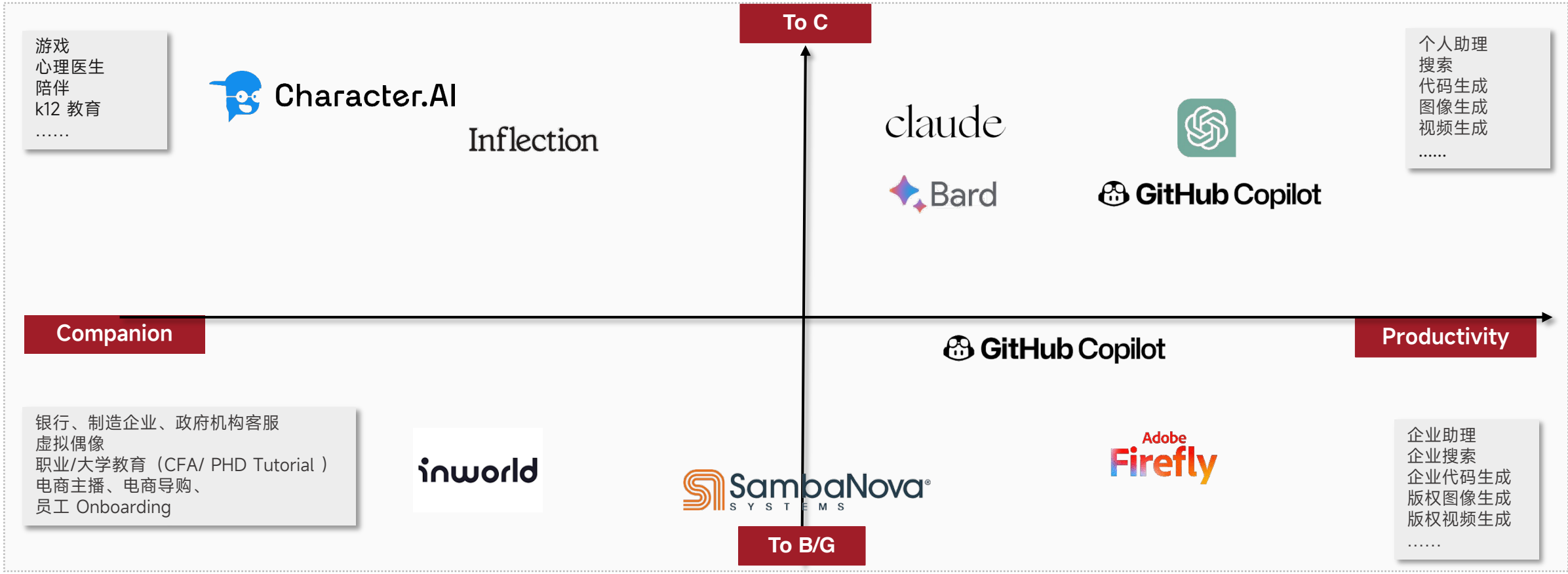


硬件：人形机器人

公司	技术	应用场景	优势	融资阶段	投资人
Tesla (Optimus)	motion mechanics, Machine learning, AI, LiDAR	自动化任务 生产制造 运输 家庭任务	AI能力 高适应性 高速 高精度	Public	Publicly traded
1X Robotics	Precise mechanical movements, Machine Learning, AI	生产制造 品控 集成	高精度 编程灵活度	Unicorn	OpenAI
Hanson Robotics (Sophia)	Facial recognition, AI, Deep learning	人机互动 娱乐 健康保健	人机交互 情绪辨认	Unicorn	Disney Accelerator Delong Capital
Boston Dynamics (Atlas)	motion mechanics, Machine learning, AI	军用 搜救 娱乐	运动 平衡 复杂任务处理	Acquired	SoftBank Hyundai
UBTECH Robotics	AI, Machine Learning	教育 娱乐 家庭任务	人机交互 用户友好	Unicorn	CDH CITIC Securities

06 模型公司的机会、挑战和投资判断

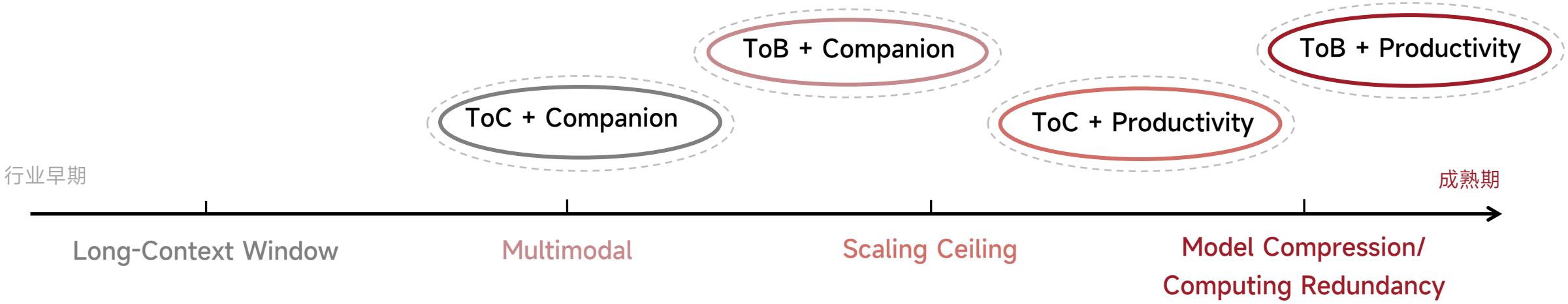
模型公司的象限



- **ToB/G 机会：**私有化部署和版权；小语种模型
- **ToC 机会：**端侧推理，能满足成本和隐私问题
- **Companion 机会：** Tier 2 模型公司打差异化的方式；商业模式和应用场景更有想象空间；内容形式不够精彩，需要多模态
- **Productivity 机会：**Scaling Law、多模态和 MoE 带来复杂推理能力，同时吃下标准和非标的软件生意

判断 1：何时从 Training 转向 Serving?

投资机会时间轴



< OpenAI 目前 80% 的算力放在 Training，未来何时会将 80% 的算力放在 Serving >

1 ToC+ Companion

- ToC+角色陪伴型将最先大规模 Serving
- T2 的模型公司打差异化的方式；不同的商业模式
- 陪伴能力要求不高， Long-Context 和 Multimodal 加入后预计就能大规模采用

2 ToB + Companion

- 市场关注度也较小，参赛选手不多不强
- 现在大模型公司资源投入较少，预计在 ToC+Companion 成熟后就会开始攻坚

3 ToC + Productivity

- 竞争最激烈的赛道，拿走市场最主要的声量、资金和算力资源
- OpenAI & Anthropic ：Scaling 才刚起步，团队内部有追求智能极限(Scaling Ceiling)的愿景。

4 ToB + Productivity

- 不确定较大/成熟周期最长的赛道，Bet on Regulation
- OpenAI & Anthropic 为巩固模型技术优势，难以服务好算力需求更大的企业客户。

判断 2: Serving 带来的投资机会

- **共识与非共识：**
  - 共识打满的在 toC+Productivity，非共识的机会在 toB+Companion；
  - 现在投资难度/风险最大的是 toB+ Productivity ，
  - 中期内投资回报率可能最高的在 toC+ Companion 。

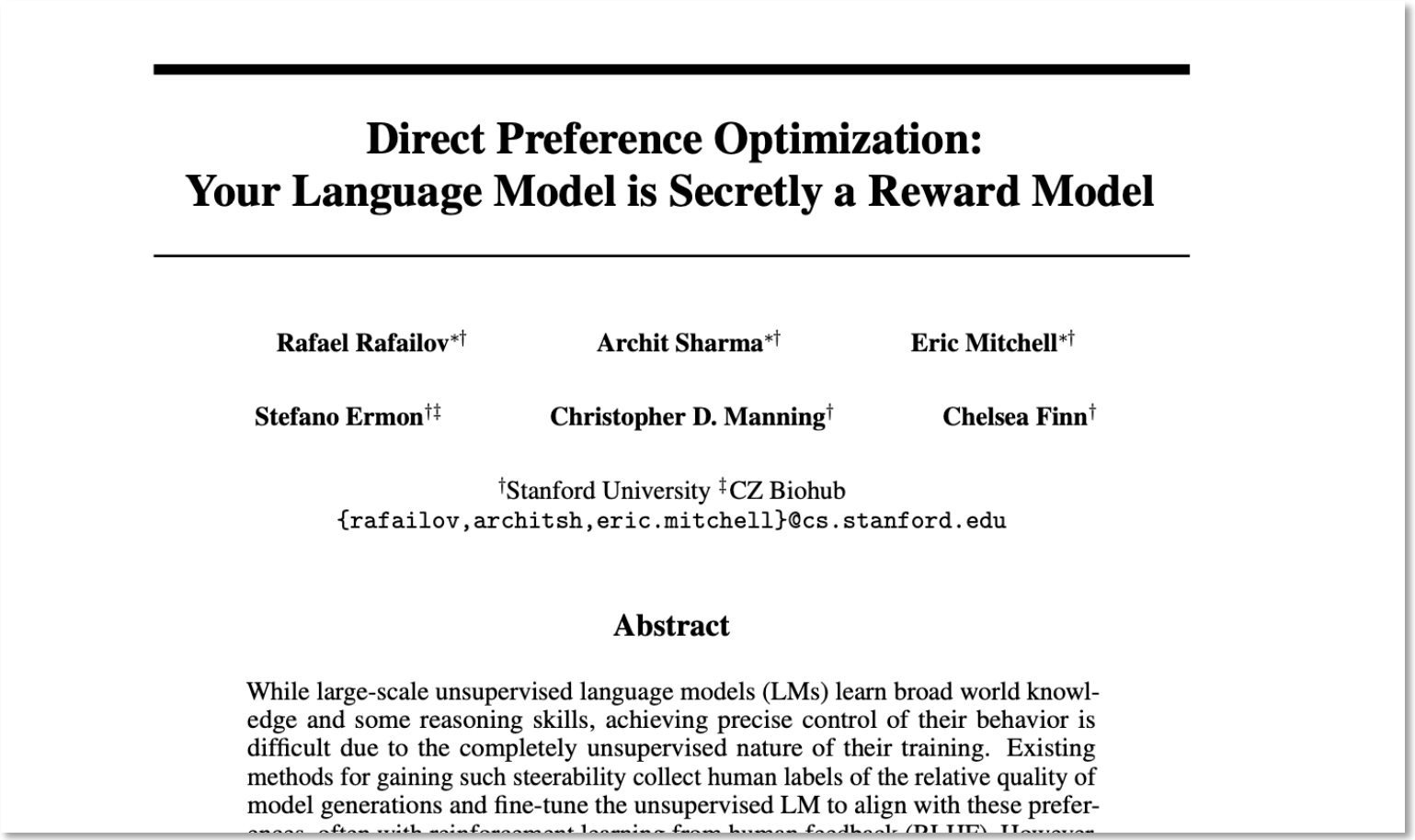
- **芯片：**
  - NV 因其遥遥领先的片间通讯能力在 Training 端形成垄断；
  - 但 2-3 年后，当角色陪伴型模型的资源倾向推理后，AMD 和 TPU 的芯片在推理方面完全不落下风。

- **开源社区和学界：**它们核心精力放在模型压缩、Instruct tunning 和 Prompt Engineering，这些都是 Serving 中重要环节，所以尽管他们现在的能力和必要性饱受质疑，可能只是 Timing 问题。

- **投资 Infra 的时机：**
  - WanDB， 是否有足够的时间开发 Serving 产品线，现有的 Training 产品的红利还有几年
  - MosaicML，短期内业务爆发厉害，但内部若没有 Serving 产品将难以为继
  - Pinecone，当前并不是业务爆发期，可能 2-3 年内会出现不错买点

判断 3: 未来模型的迭代节奏会很像手机系统

- 预计未来模型的迭代节奏会很像手机系统：1-2 年更新一次大版本，中间有无数次小版本迭代；
- 中期来看，RLHF 不应该是 Alignment 的唯一手段， Direct Preference Optimization 和 Stable Alignment 是新路径
- 长期来看，小版本迭代的角度：隐私/权限的增强、更多的图片/文件/声音格式的 Embedding



Direct Preference Optimization:
Your Language Model is Secretly a Reward Model

Rafael Rafailov[*†] **Archit Sharma**[*†] **Eric Mitchell**[*†]

Stefano Ermon[†‡] **Christopher D. Manning**[†] **Chelsea Finn**[†]

[†]Stanford University [‡]CZ Biohub
{rafailov,architsh,eric.mitchell}@cs.stanford.edu

Abstract

While large-scale unsupervised language models (LMs) learn broad world knowledge and some reasoning skills, achieving precise control of their behavior is difficult due to the completely unsupervised nature of their training. Existing methods for gaining such steerability collect human labels of the relative quality of model generations and fine-tune the unsupervised LM to align with these prefer-

判断 4: LLM 的 Context Window 就像电脑内存，向量数据库是 LLM 的硬盘

- 4k 可能就像 XP 的 256MB 一样，而 32k 就像 2GB。未来图片和视频的内容将让 Context Window 变得更重要；
- 现在 Context Window 都是顺序读取的，有没有可能选取最相关的，能像内存一样随机/选择性读取？
- 未来模型里会不会出现多级内存？Cache，HBM 和 DRAM

07 重要公司

硅谷大厂的 LLM

Google

- Google 其实有一手高潜力的好牌，GenAI 战略的三个支柱也非常明确：做最先进的 LLM（Google Brain 和 DeepMind），对现有产品进行重大改进（搜索、Workspace 等），和对外提供工具（GCP、TPU 等）；

- Google Brain 和 DeepMind 合并引发短期震荡，但是内部员工和 OpenAI 的人有很多交集，觉得对方并没有什么特别的秘诀，仍然有信心。目前最重要的项目是 Gemini，瞄准 GPT-4 的下一代，定位不一样的多模态。

硅谷大厂的 LLM

Microsoft

- 微软仍然是 GenAI 的引领者，在 Office 365 接近峰值渗透率和云服务 commodity 化的情况下靠 AI 扩大 TAM，并且加强了在 Azure、Dynamics 365、Office 365、Edge、Bing 以及安全方面的市场份额；

- 微软和 OpenAI、NVIDIA 长期来看都是竞合关系，在和 OpenAI 筹备下一代超级计算机，但是也训练自己的模型，Azure 也正在争取把 OpenAI 的开发者入口搬到自己生态上，此外微软还在和 OpenAI 联合做芯片来降成本。

硅谷大厂的 LLM Meta

- Meta 把 GenAI 视作和推荐系统平行的重要技术路线，LLaMA、Segment Anything 等开源模型在第一季度收获了成功，在 6 月初的 All Hands 还宣布了嵌入到社交媒体场景内的 GenAI 产品，预计在第三季度末面向大众推出；

- Meta 正从 all in CPU 转向，各个产品组正探索能给产品指标产生正向影响的 GenAI 结合，FAIR 则被要求做创造可见收益的事情，给 CTO 汇报，同时在 Meta 只要是效果比原来好的 AI/ML 模型都可以进入生产环境，因此通过 Review 进入产品内的项目比例远高于 Google 的 30%。

硅谷大厂的 LLM

- Tesla 在语言方面的进展没有太多披露，但是它专注于投资于数据积累、NVIDIA GPU 和降低训练成本的 Dojo，并希望在未来像 Amazon 销售 AWS 一样对外销售 Dojo，这个产品专门用于自动驾驶视频数据的训练；

- Tesla 的风格是产品导向，和 OpenAI 的研究导向不同，倾向于如何以最优成本把场景做出来，因此目前内部没有明确要发力的围绕大模型的项目，因为车的推理硬件对于部署大参数量的 LLM 有很大限制。

硅谷大厂的 LLM

- 尽管 NVIDIA 是目前 GenAI 军备竞赛的最大受益者，它的公司战略并不局限于 GenAI，而是将 GenAI 视作和 CV 等平行的 8 个 MLPerf 子类之一，为这些客户提供全堆栈的产品服务以及跟云厂竞合的 GPU Cloud；

- NVIDIA 在 Gen AI 领域和 MidJourney 及 Stability 等公司最大的差异是数据集，非常重视版权，很适合做海外 to B 服务，比如游戏公司会乐意买单来避免版权纠纷。

硅谷大厂的 LLM

- Apple 不向外界透露其产品路线图，对外强调的 AI/ML 的融合范例仍然是跌倒检测、碰撞检测和心电图等产品；

- Apple 在 OS 层面以及 Siri、Spotlight 等产品上都很适合嵌入 GenAI 相关能力，但是内部是自上而下推动并且对于功能的可靠性和安全性要求非常高，因此进展较缓慢，此外苹果不允许研究人员发表论文的要求可能会阻碍其人才获取。

硅谷大厂的 LLM

- Salesforce 投资 GenAI 的决心非常大，支持了 Anthropic 和 Cohere，并且已经推出了 Einstein GPT、Slack GPT 和 Tableau GPT，并借此推动 Data Cloud 战略以为客户提供智能自动化和成本节约；

- Salesforce 内部非常重视数据安全合规以及 LLM 的可信任性，因此各个云产品虽然留存了大量用户数据，但是严禁用于训练模型或者自行利用模型帮助客户进行分析或自动化，各个业务线自下而上寻找场景仍然有难度。

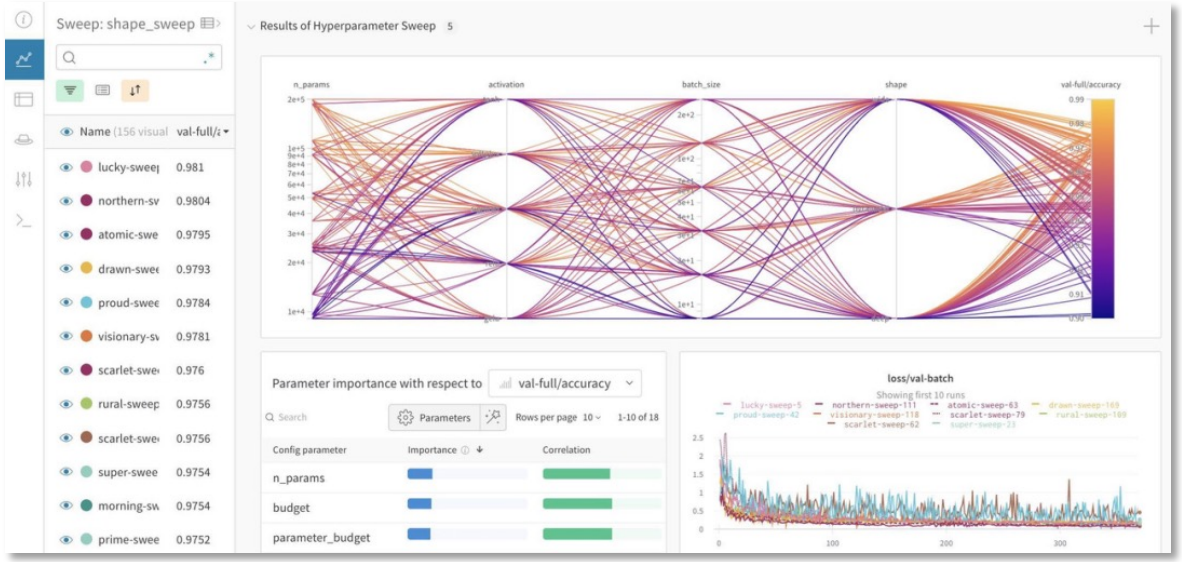
硅谷大厂的 LLM Adobe

- Adobe 低调、投入大、实力强，团队跟 OpenAI API 合作的同时也在自建 Base Model，90% 的工作量在数据的选择和处理，也非常重视版权，并且设计了对创作者版权内容的激励分成机制；

- 内部最关注 Midjourney、剪映、Runway，目前不直接竞争。Adobe 收购 Figma 的还有个原因是协同设计听起来很简单，但实际在技术实现上很难，Adobe 做了很久都做不出 Figma 的效果，所以不得不收。现在造成类似威胁的是 Runway。

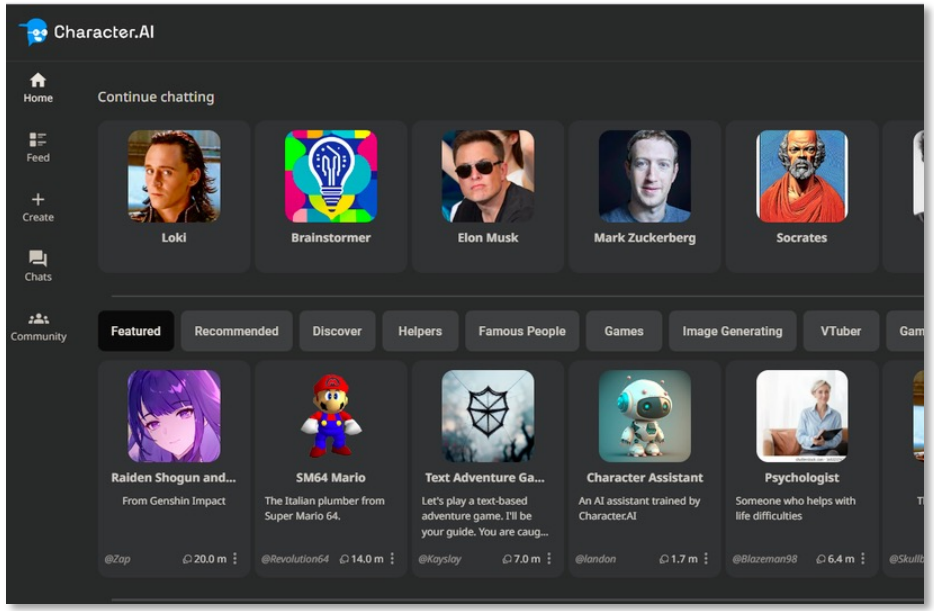
案例研究：Weights & Biases

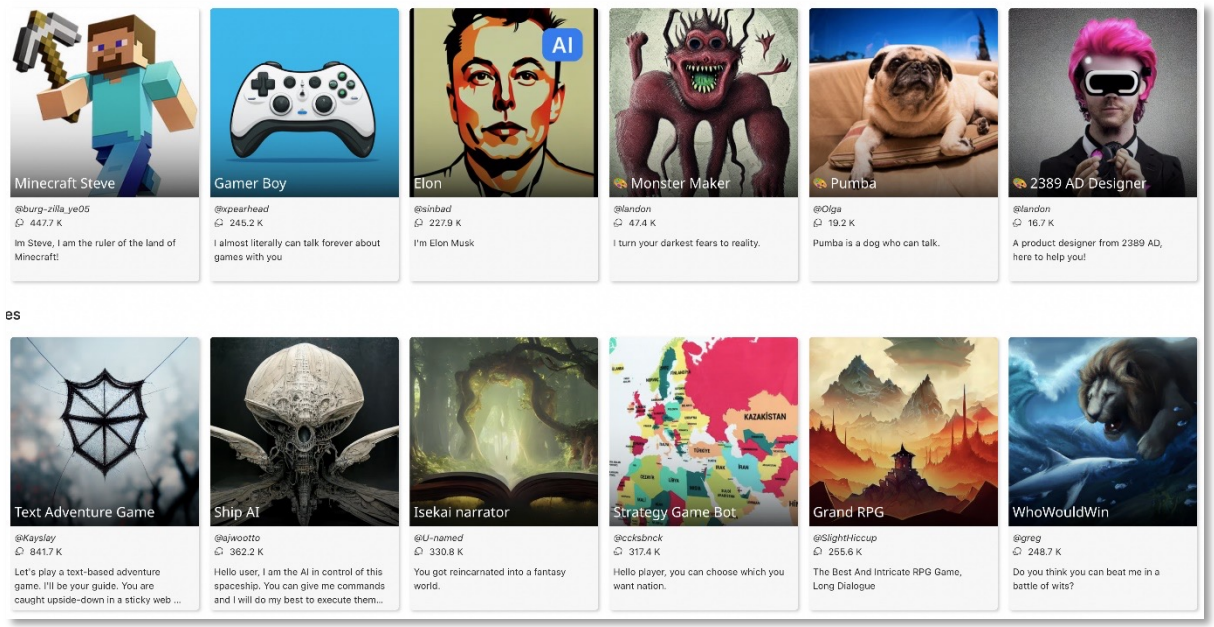
- **Weights & Biases 聚焦于模型实验管理环节，今年在 Gen AI 浪潮切入 AI/ML 工作流中更多的过程：上游数据探索、下游模型监控和开发者使用的 LLMOps，扩大了使用者的市场规模。**
- **模型实验管理 + 模型监控，是 AI/ML 领域 Datadog 级别的生态位，**有机会诞生出 AI/ML 时代的 Datadog，而 Weights & Biases 被赋予最高期待。2023年对于 Weights & Biases 类似于 Datadog 的 2017 年一样关键：通过在核心产品推出新的功能，获得更高的市场占有率和客单价。
- **大模型训练的复杂程度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。**
- **Weights & Biases 有着口碑极佳的产品，很多 AI 从业者从科研到业界都在使用其产品粘性很强。当前在商业化上还做得相对克制，免费产品体验也很好，接下来会有 Pricing 的变化，预期能使产品收入上一个台阶。**



案例研究：Character.AI

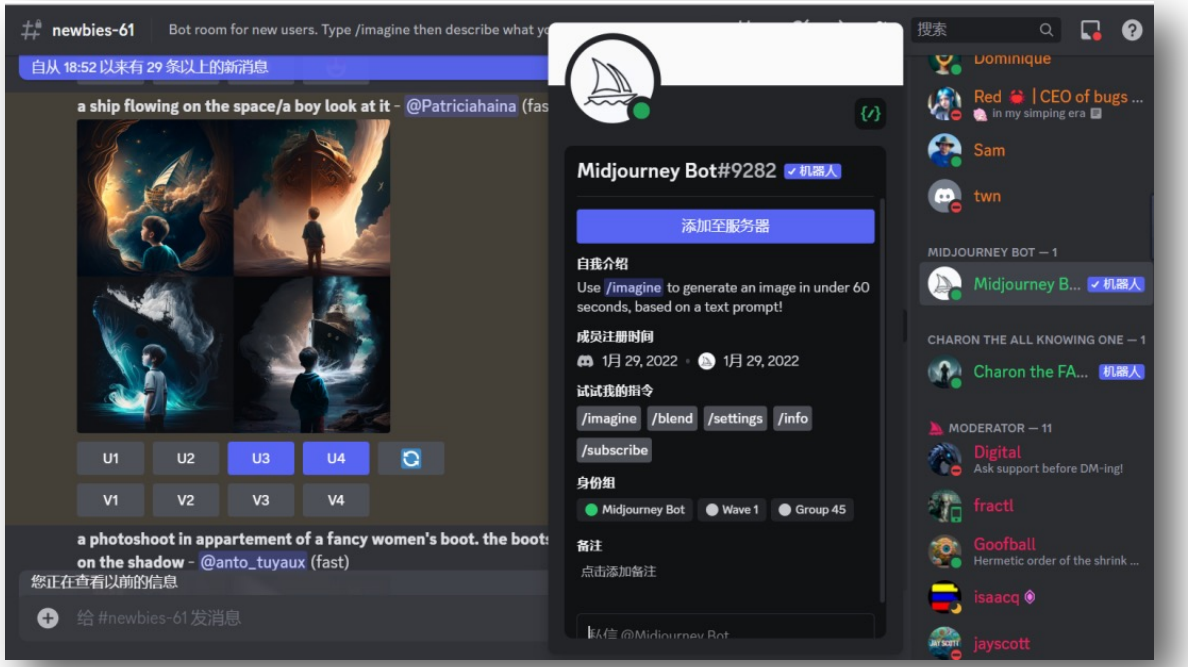
- **定位 foundation model 公司，公司现在 30 多人，最近在招 community manager、后端人才。准备推出移动端。**
- **公司内部认为的核心竞争力和壁垒是：**model 比别家好。创始人自己发明了很多降成本的新方法，从底层的 infra 到算法设计都做了调整。
- **用户增长很快，目前最重要是让 model 和产品能力能匹配上用户增长的速度。**之前用户很多是男性，最欢迎的角色是原神，聊游戏，现在女性用户越来越多。18-25 岁用户占 1/3，也有 60 岁的老人。
- **Character 上的 bot 头部效应不集中，非常长尾，因为没有人去做推荐算法。**

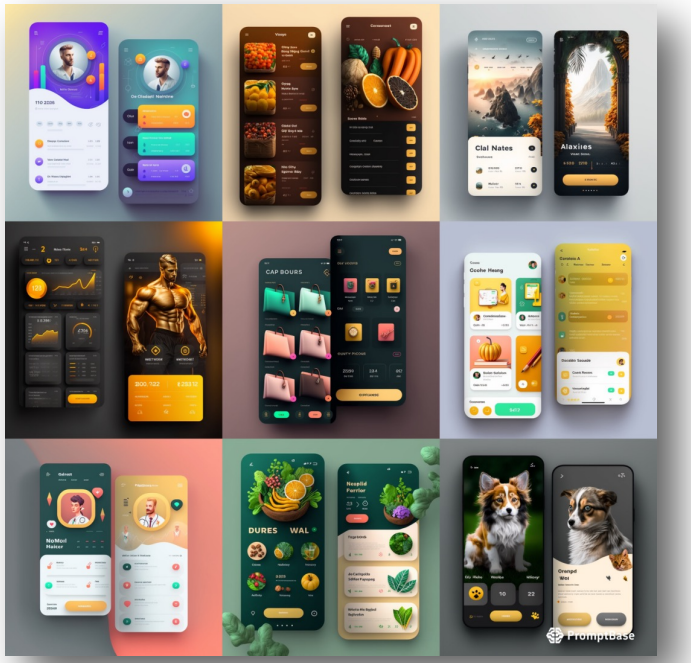




案例研究：Midjourney

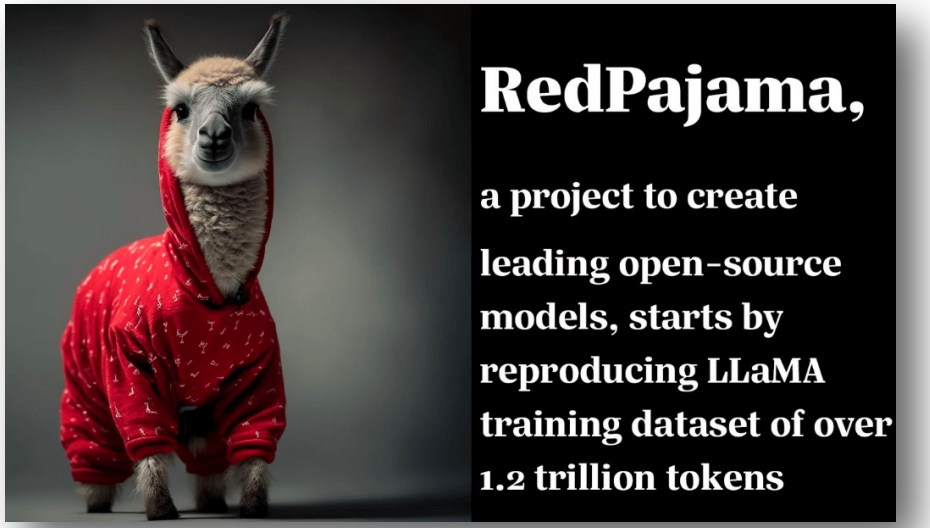
- **文生图领域的大模型公司：**MidJourney 的算力和模型量级客观，效果明显优于其他文生图模型。基于其模型搭建的应用，已经拥有超 1000 万社区成员，年营收约为 1 亿美元，并具有很强的盈利能力。团队目前仅 20 多人。
- **大模型中最优的数据飞轮效应：**MidJourney 的四选一产品设计、文本与图片的匹配对，都是很好的输入分布和用户偏好的方式。而其他 Chat 产品形态，或开源的模型，都很难收集那么好的数据闭环，与其相似的产品只有 Github Copilot。
- **核心用户群：**产品设计师（玩具、墙纸等）；图片设计师（网站、广告、PPT、Logo、插图等）；游戏设计师（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。





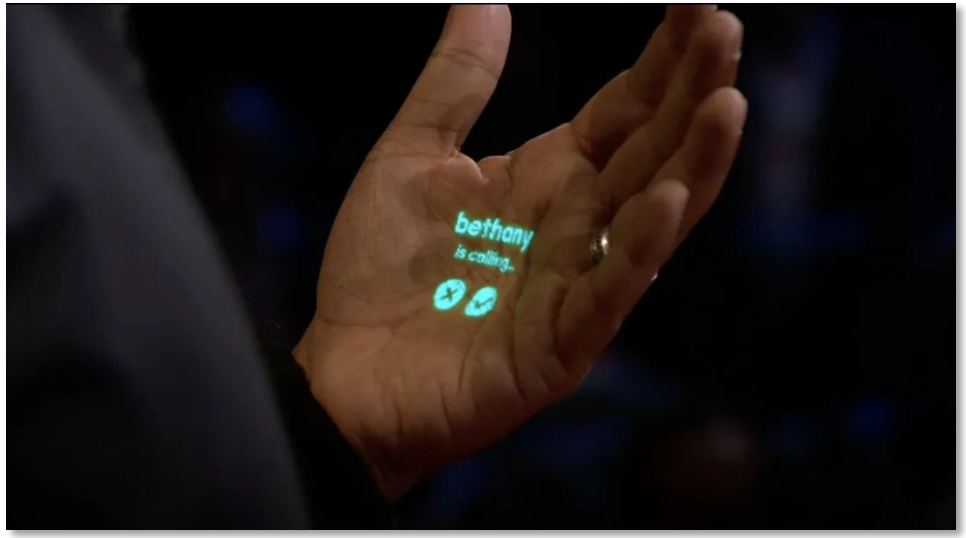
案例研究： TOGETHER

- Together 的目标客户为**大型企业和大型组织，以及政府部门**。因为企业端更加看重准确性和可靠性，注重数据隐私，需要基础模型具有足够的透明度和解释能力，并减少对其他各方的依赖。目前发布的**产品包括 RedPajama、OpenChatKit 等开源模型及用于运行、训练和微调开源模型的云平台**
- CEO Vipul 是连续创业者，上一家公司 Topsy 被 Apple 收购，成为 Apple 的高级主管，领导由 250 多名工程师和项目管理人员组成的团队。在 9 个月内建立了一个支持超过 100,000 qps 的搜索引擎，推出了涵盖 Spotlight、Safari、Messages、Siri 的功能，并在搜索和 AI/ML 方面进行长期的努力。
- **2023年5月，Together 完成 $20M 种子轮融资，由 Lux Capital 牵头**。Lux Capital 的 Brandon Reeves 表示："**Together 通过提供一个跨越计算和最佳基础模型的开放生态系统，正在引领 AI 的'Linux时刻'。"**
- Together 招聘了大量来自 Apple 和 Snorkel 的产品人才，有一定潜力成为 LLM 开源生态的扛把子。



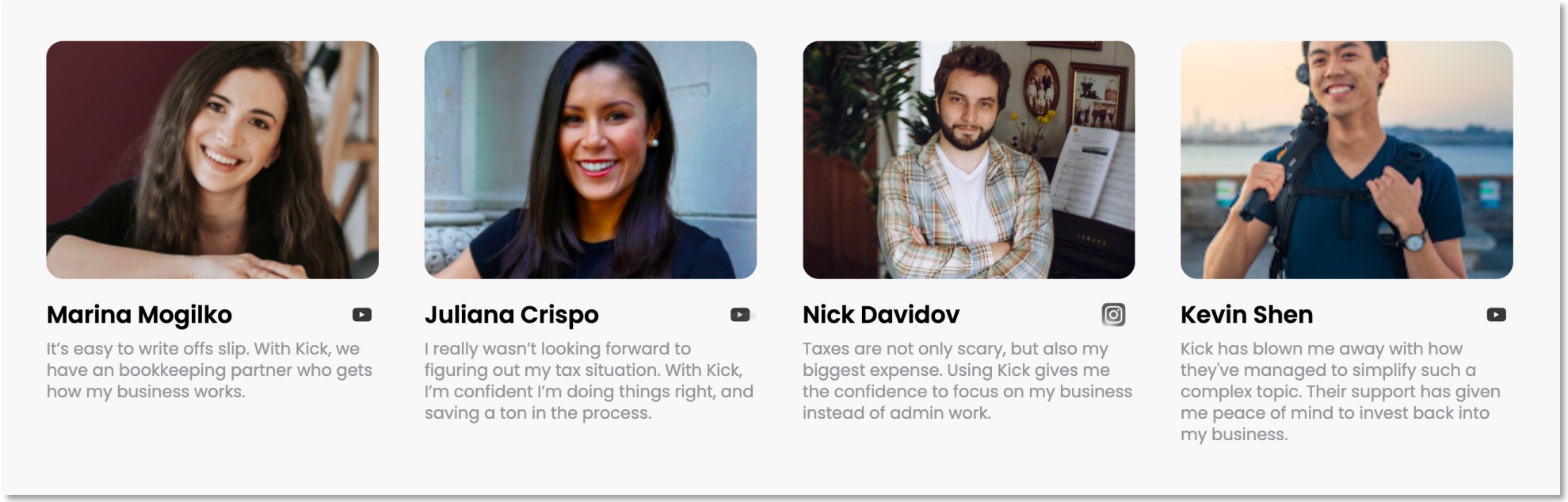
案例研究：hu.ma.ne

- **Humane 由前 Apple 员工 Bethany Bongiorno 和 Imran Chaudhri 创立**，Chaudhri 设计了初代 iPhone 的交互和界面，1995 年至 2017 年，他在 Apple 参与了 Mac、iPod、iPhone、iPad、Apple TV、Apple Watch、AirPods 和 HomePod 等产品的设计。Bongiorno 作为当时的 Apple 软件工程总监协助监督了初代 iPad 的发布。Humane 团队目前 200 多人，40 多人苹果背景。
- **与微软和 OpenAI 在技术上有合作，微软和 Sam Altman 也是 Humane 的投资人。**
- **美国投资人试用产品评价：**很喜欢 Humane 的 demo 版产品，它能看到你看到的东西，比如指着一个海报说，帮我买这个电影票（不用说具体是什么电影）它就能理解并且帮你买。
- **内部员工评价：**很多员工都体验了 demo 版产品，反馈不错。但公司内部也知道 target 下一代产品风险很大。



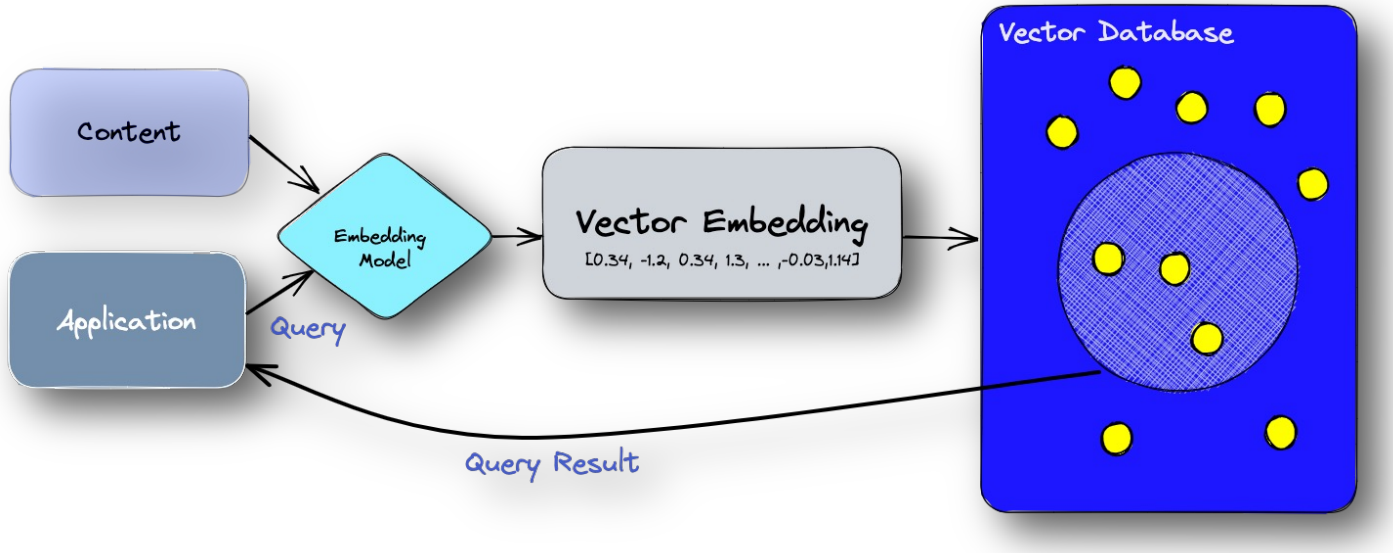
案例研究：Kick

- Kick 帮助企业主进行日常 Bookkeeping 的自动化，通过 Plaid 连接到客户的银行和信用卡账户，然后自动对收入和支出进行分类，并帮助客户提高报税效率
- Bookkeeping 是 LLM 应用最火的方向之一。**最早使用软件自动化 Bookkeeping 流程的产品是 1983 年诞生的 Quickbooks，跑出来了现在 1300 亿美元市值的 Intuit，是第二年轻的千亿美元金融服务公司。**过去 10 年里不断有公司希望颠覆 Intuit，比如 Bench 和 Pilot 等 Bookkeeping Marketplace，但都没有成功
- **OpenAI 生态基金最早投的公司，还在非常早期阶段，受到了很多关注，产品预计 9 月发布**
- 希望先做细分 + SMB/创作者经济客户打磨好产品，把金融财会领域的复杂知识跟 AI 结合好



案例研究：Pinecone

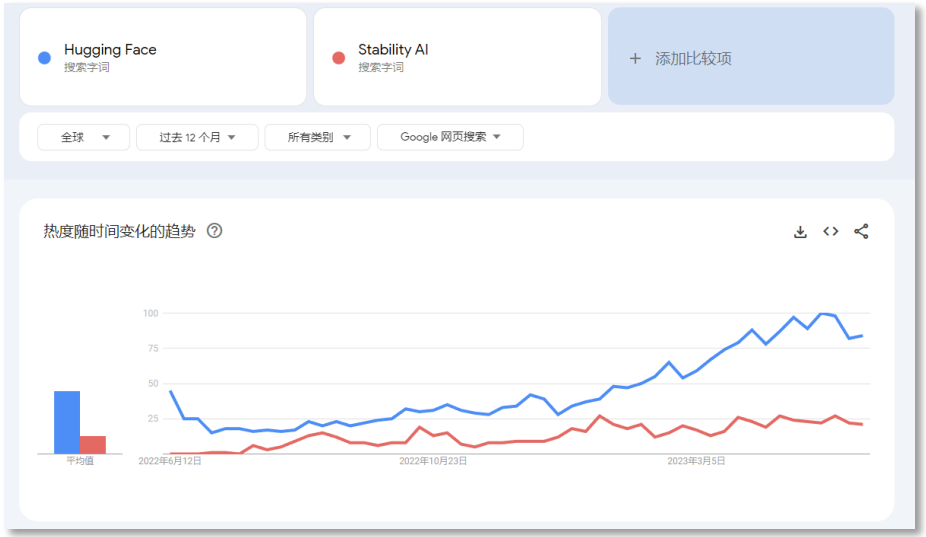
- **如今开发 AI 应用的团队，80% 以上都使用向量数据库，来提供用户个性化记忆和企业专有数据的召回能力。Pinecone 由于其开箱即用的产品体验，占有了众多应用开发者的心智。**
- **LLM 本身是缺失记忆更新能力的大脑，向量数据库是大脑中的海马体提供记忆。LLM 的 Context Window 就像电脑的内存，向量数据库则是 LLM 的硬盘。**
- **Pinecone 的 ARR 增长比较快，**从去年底的 2 mil 到 5月的 7mil，年底预期能够达到 15 mil。最新一轮向量数据库估值普遍由于最近需求增长偏高，Pinecone 在今年3月融了B轮估值达到 750 mil。
- **向量数据库技术栈同质化比较明显，竞争激烈。**Pinecone 是以闭源状态称为领跑者，其他竞争者都以类似的产品特点开源竞争。大公司如 Databricks, ElasticSearch, Redis 和云厂商都在这一领域进行布局。
- **语义搜索和向量召回本身并没有技术供给上的突破，而是 LLM 的火热带来了需求层面的飞速增长。**



案例研究：

HUGGING FACE

- **今年收入增长很快，现在的收入能 cover 所有员工成本和 30% 的 compute 成本，准备 break even，目前有 30 几个月的 run way。**
- **目前的商业化路径很清晰，除了之前定位的 Github for AI/ML 卖算力资源外，还要做 AI/ML 界的麦肯锡和埃森哲。**埃森哲没有跟上 AI 这一波，现在有 AI 需求的公司都会找 Hugging Face 咨询，包括科技公司和传统企业。Databricks 的 Dolly 就是在 Hugging Face 的支持下做出来的，迪士尼等等传统企业每年也会给 Hugging Face 几十万美金的咨询费用。咨询业务做得很轻，几位专家时不时回答客户几个问题，或给客户指方向。目前很多企业只能招到基础人才，但招不到 senior 的、精通 LLM/ML 的人，Hugging Face 相当于他们的 senior 人才。
- **已经不打算死磕大模型，OpenAI 太强，论卡、钱、人才、数据都无法与之相比。**并且 Hugging Face 是开源里的台柱子，模型做出来一定是要开源，无法很好的靠大模型商业化。内部也知道 Bloom 效果不好，核心是用来训练的数据量太少，Hugging Face 做 LLM 只是形象和面子上的考虑。
- **商业上和 OpenAI 等大模型公司竞争不强，Hugging Face 有大量客户不用 LLM，而是用普通的小模型就能解决具体问题，而且很多客户希望把自己的私有数据/专业数据用来训模型，并不想把这些数据给到 OpenAI。**



案例研究：mosaicML

- MosaicML 的核心业务是为所有想要训练大模型、苦于人才储备不够的公司提供工具，使他们能更高效的将计算资源利用好。
- 其核心产品是两个开源库 Composer（分布式训练框架）、Streaming Dataset（分布式数据处理和商业化产品（MosaicML Platform）。其产品梯度明显，用好其开源框架很依赖对计算资源的深入理解，用户对产品的付费意愿比较强。
- MosaicML 也提供 Mosaic Cloud 和多云混合使用的能力。这一能力使其商业化收入增长很快达到了 20 mil，但也因此 margin 不会很高。
- Databricks 宣布以 13 亿美元的价格收购了 MosaicML，补上了其在 AI/ML 分布式训练和云资源调度的关键能力。

Model	LAMBADA (OpenAI)	HellaSwag	PIQA	ARC-Easy	ARC-Challenge	BoolQ	COPA	Winograd	Winogrande	TriviaQA	Jeopardy	MMLU
MPT-7B	0.703	**0.761**	**0.799**	**0.673**	0.394	0.750	**0.813**	**0.878**	**0.683**	0.343	0.308	0.296
LLaMA-7B	**0.738**	0.751	0.792	0.652	**0.411**	**0.767**	0.779	0.807	0.675	**0.443**	**0.334**	**0.302**
StableLM-7B (alpha)	0.533	0.411	0.666	0.435	0.259	0.606	0.672	0.646	0.513	0.049	0.000	0.251
Pythia-7B	0.667	0.636	0.761	0.581	0.325	0.634	0.769	0.786	0.607	0.198	0.022	0.265
Pythia-12B	0.704	0.672	0.768	0.605	0.351	0.675	0.781	0.847	0.627	0.233	0.026	0.253
GPTJ-6B	0.683	0.665	0.762	0.583	0.355	0.648	0.789	0.833	0.641	0.234	0.026	0.261
GPT-NeoX-20B	0.719	0.712	0.780	0.644	0.392	0.691	0.781	0.861	0.665	0.347	0.146	0.269
Cerebras-7B	0.636	0.582	0.744	0.564	0.311	0.625	0.734	0.779	0.603	0.141	0.012	0.259
Cerebras-13B	0.635	0.588	0.740	0.571	0.321	0.611	0.719	0.760	0.602	0.146	0.013	0.258
OPT-7B	0.677	0.676	0.773	0.579	0.329	0.665	0.719	0.840	0.656	0.227	0.020	0.251
OPT-13B	0.692	0.701	0.774	0.586	0.345	0.657	0.805	0.851	0.670	0.282	0.126	0.257

案例研究：SambaNova SYSTEMS

- **公司三名创始人在芯片和 AI/ML 领域均为传奇人物：**不仅有曾助太阳微电子的芯片部门力战 IBM 的 Rodrigo Liang，也有多核处理器理论奠基人 Kunle Olukotun，还有积极活跃在各大 ML/AI/LLM 社区的意见领袖 Christopher Ré。
- 公司早在 2019 年就预见 GPT 等超大模型的爆火，不仅超前在初代芯片设计上倾向 GPT 而设计超大内存，并且早在 2020 年就开始组建 GPT 的训练专家小组，也在 2022 年 2 月份就针对银行客户开发了完整的解决方案。
- 团队从全新的芯片架构出发，针对客户需求定制 Complier 后为多种 ML/AI 算法提供计算加速，也会帮助注重数据安全的客户训练并部署 LLM 等新兴 ML/AI 算法，有机会成为 AI 时代的 IBM/Oracle。
- Lawrence Livermore National Lab（美国），Argonne National Lab（美国），RIKEN Center（日本）等知名国家实验室，以及 Accenture 和 OTP Bank 等传统金融机构都是公司的客户。

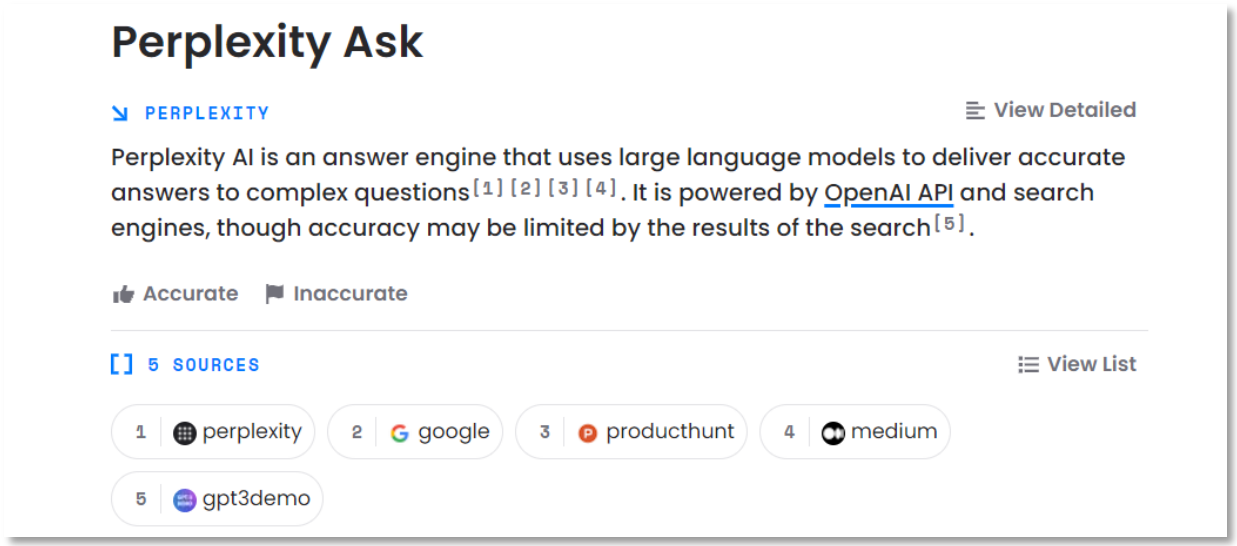
案例研究：Inflection

- **Inflection 将自己定位为 AI 应用公司，而非大模型公司，**作为 AI Studio 针对不同场景开发 AI Agents。
- **团队阵容豪华，**三位联创分别为原 DeepMind 联合创始人 Mustafa Suleyman、LinkedIn 联合创始人及 Greylock 合伙人 Reid Hoffman、原 DeepMind 资深科学家 Karén Simonyan。
- **第一款产品为名为 Pi 的个人 AI 助理，在服务于生产力需求的同时主打“情感陪伴”，以“情商高”而出名。**设定为动态、跨平台的 Personal AI，可以随时调用，但具体如何分发还在探索。
- **在商业化上，采用订阅、增值付费模式，**可能部分用户会接受在对话中出现广告，最终的商业化形态还需要不断测试迭代来完成。
- 创始人 Mustafa Suleyman 认为未来将有数十亿个 AI Agent，人类与 AI 将“相互依存”、“紧密联系”。

Model	Average	Humanities	STEM	Social Sciences	Other
GPT-4	86.4	–	–	–	–
PaLM 2-L	78.3	–	–	–	–
Inflection-1	**72.7**	**79.2**	**61.7**	**82.6**	**74.1**
GPT-3.5	70.0	–	–	–	–
PaLM (540B)	69.3	77.0	55.6	81.0	69.6
Chinchilla (70B)	67.5	63.6	54.9	79.3	73.9
LLaMA (65B)	63.4	61.8	51.7	72.9	67.4

案例研究：Perplexity

- **Perplexity 是基于 OpenAI GPT 模型的新一代 AI 搜索引擎，**能够汇总搜索结果，为用户提供 AI 分析后的答案，返回的信息附带引用，允许用户确定信息的来源和可靠性。它的产品功能还包括 Copilot，调用 GPT-4，引导用户细化问题，并展示解决过程；以及 AI Profile，用户填写个人介绍，在交互中，引擎会根据用户个人信息提供更有针对性的结果。
- **团队实力强：**来自 OpenAI 及 Meta 等，并在 3 月初完成 2560 万美元的 A 轮融资
- **产品分发渠道：web 端包括网站和 Chrome 插件，**并从 web 端扩展到移动端，推出了 iOS 和 Android APP
- **商业化：推出了 Perplexity Pro，**$20/月，可以每天使用超过 300 次 Copilot 功能，使用 GPT4 回答问题；开发者 API 获得 1000 个申请，Pro 计划推出前三天 ARR 达到 26 万
- **用户数据：**网页端 25 万 DAU，450 万 MAU，访问量 2200 万次/天，query 数 140万次/天



Key Questions

- 下一代模型长什么样?
- 如何解决 Cost、Latency、Hallucination ?
- B 端私有化部署是不是伪命题?
- 模型 & 芯片的 Co-design 联盟猜想
- ChatGPT 团队的下一步是什么?
- 下一代操作系统：有谁实现？如何实现?
- 模型之间的差距会持续拉大吗?
- 最佳实践/ Use Case 何时出现？会出现在哪里?
- LLM 会催生新的交互界面吗?
- ……

谢谢观看

THANKS!