OpenAI 闭门讨论会 V3【GPT-4】纪要

To：ShiXiang LPs

Date：2023 - 03 - 19

拾象一直关注本轮范式转移的头部公司 OpenAI，以及大模型给行业带来的影响。海外独角兽在社群内举办了一系列优质线上闭门讨论会，**本次为系列讨论会的第三场，主题是在 3 月 15 日发布的多模态预训练大模型 GPT-4。**

围绕 GPT-4，我们集中讨论了以下几个问题：

1. **对模型能力演变和边界的思考**：包括 GPT-4 发布后有哪些新技术导入、解锁了哪些新能力、带来哪些新机会、从应用/算力/infra/研究上的变化，以及未来的演变走向、关键要素、带来哪具体的影响/案例/新机会，还有 LLM 的能力边界；
2. **对 AI Native Apps 的思考**：包括应用 LLM 的什么好案例、什么特点、关键要素是什么、看好哪些垂类应用、壁垒，应该怎么做 AI Native Apps 等；
3. **对模型格局的思考**：OpenAI 一家独大，还是多寡头，模型和应用的关系，垂直应用都要拥有自己的模型，还是基于 OpenAI 开发；
4. **LLM 相关的非共识判断。**

以下为本次会议的详细实录。

注：本纪要为「Generative AI 社群」闭门讨论会的整理总结，文中所有内容为参与讨论的社群成员的观点实录，不代表拾象观点和立场。

Q1：GPT-4 之后，如何看模型能力演变和边界

GPT4 升级、能力提升影响很大，因为通用能力变强，去年夏天爆火的 AIGC 应用如 Jasper.ai, copy.ai 受到很大的挑战。类似当年 iPhone 升级，把应用商店中基础的安全等应用从 code 层面淹没掉了。

讨论 GPT4 有哪些新技术、新能力，从创业做应用、算力、Infra、研究等角度展开；以及从中短期长期，怎么去想大语言模型的演变方向。

A:

1) GPT4 的市场预期

类比 iPhone，Code、系统、基础工具能力层面都是能做，但是做不了 Facebook 网络，Uber 打车/管车，airbnb 等重业务，所以创业要考虑垂直领域。但是它生成能力很强，未来可能更强，GPT4 推理能力变强，并且加上眼睛，可以读基础的图片、做总结。未来多模态升级会有更多的能力。OpenAI 内部的人对于未来的升级预期非常 aggressive，所以应用创业可能不适合长期价值投资，很多应用的生命周期会很短。

2) GPT4 出来后的新想法：

加了图像能力之后，GPT4 拥有视觉信息，一定程度上可以更像人；可以考虑更复杂的事情，比如控制机器人，实现类似 adept 的自动机制。

- **Model 层面有更多的改变，Infra 存在挑战：**

OpenAI 训练了一个更大的模型：175B 的语言模型，加 2B 或者 20B 的视觉模型分支。这意味之前的框架训不动了，国内本身只能用 40G 的 A100，现在要在 40G 的 A100 的前提下负载更多的参数，国内受到更大的挑战。

- **研究方向：**

国内一直在说要做自己的大模型，但 OpenAI 说，language model 是第一阶段，甚至可能是很小的阶段，只是后面的基础而已。国内需要想清楚自己要做什么，是多模态模型，还是之后会出现更复杂的模型，这会带来实际执行和心理上的改变。

GPT4 出来之后，大家会感觉我们大概率短期是追不上的。因为算力、多模态研究都很困难。更务实的做法是模型和应用一起做。

3) 计算 GPT4 有多少参数，可以估计一下他们有多少张卡，算出它训一年的 Tflops，从得到的 Tflops 可以倒推模型有多大，数据有多少。如果这样算的话，它应该要比 175B 大很多，数据和模型都会大很多。

TFLOPS 是 floating point operations per second (每秒所执行的浮点运算次数) 的英文缩写。它是衡量一个电脑计算能力的标准。

4) H100 出来后，OpenAI 用几万张 H100 训模型，模型的能力会有多大的提升？和 Anthropic 和 Cohere 会拉开多大差距？想象边

界在哪？

- **和其他对手的比较：**

从 POE 可以体验，Anthropic 的 claude+ 和 GPT4 并没有差很远，只是 Anthropic 从不宣传。

目前 GPT-4 的变强的能力很多都能预期到。算力拉满后，多模态的涌现能力会加强，GPT-4 的 vision + Language 会有预料之外的涌现能力，之后加上 video 会有更多。因为很多题目是需要眼睛 + 语言才能解答，比如解析几何。

GPT-5 的能力是否会远超 GPT4 要考虑 Alignment text，公开的 GPT-4 是清光了 Align 之后的模型，alignment 本身就会让模型的能力下降，在 GPT4 之前的 GPT4 Early 表现的能力要强一截，所以不用担心能力不上去。现在能观察到的能力只会使你低估它，不会高估它。

5) 能力视角，GPT 会淹没掉哪些公司？围绕 OpenAI 做的应用会不会被淹没 90%，基础能力（理解、推理、生成）最后是不是都基于 OpenAI 本身？做应用的点在哪里？

- **可以以超过人类中的最强者为分界线。**

当模型在某方面的能力超过人类最强，游戏规则会改变。超过人类最强并不是无法达到的目标， AI 超过人类最强是有先例的， 如 Alphago 和 Alphago zero。没有人可以阻止 OpenAI 像训练 Alphago zero 那样训练 GPT。

- **OpenAI 模型本身变强，一定会有很多已有的 APP 受到影响。**

比如 Langchain 把很多模型和外部的东西接在一起，但 Microsoft 365 发布之后，Microsoft 就把这件事情做了，Langchain 就很大程度失去了存在的意义。

第一波 OpenAI 踩中了 local optimal，有了简单的对话能力，第二波会不会 OpenAI 出来几个人把模型 reproduce，说明技术壁垒不是那么高。因为 machine learning 训练的参数不会特别多，不是特别复杂的系统工程，这样给其他竞争者不会有太多的希望。

OpenAI 技术博客里讲，内部有一个非常 Scalable 的训练框架，使它有 predictable scaling 的能力，参数加到多少，**训练能够无缝完成是很重要的。**因为使用数据训练模型需要人为干预，很难自动完成。有很多的细节，比如有可能梯度爆炸，或者 Loss 跑丢了，这时候就需要人为 roll back，把中间的脏训练数据踢掉。所以无缝的自动训练需要很强的认知框架及训练系统。

OpenAI 这次将组织架构公布，告诉大家不是只有一个模型训练组，而是很多组，每个组都有明确分工。有极强的壁垒，短时间很难超过。

Reid Hoffman 公开表示自己去年 8 月的时候就拿到了 GPT-4，原始版本能力更强，这大半年的时间其实是在解决风险。公开 GPT-4 是一个很大的系统工程，能力越强、风险和挑战越大，要避免大家用 GPT-4 作恶，比如教人们制造炸弹，搞破坏等。从这个层面看，OpenAI 领先其他竞争对手很多身位，包括 Google。

6) 从 GPT-3 到 GPT-4 能力的暴涨，从算法、算力、数据三要素来分析：

- **算法：**

底层还是基于 transformer，而 transformer 已经是 2017 年发布的论文。**数据**是大量互联网爬虫， Chrome 占了一大半数据， 维基百科、 reddit 等数据也一直都在。

- 效果的上升应该是**算力**的提升带来的：

V100 已经好几年了，这两年是 A100，未来看起来是 H100。单卡算力提升受摩尔定律限制，每一两年提升两三倍。**这次算力的暴涨来自于大规模分布式训练，**用一台机器和一万台机器，算力暴涨 1 万倍，这是目前是人类最大的分布式计算集群。

下一代能力上升，数据量客观存在、没有革命性的新算法，最多有工程手法的细节调优；算力即使放了 H100，相对 A100 提升 3-5 倍，没有数量级的上升。如果追求更大的算力上升，要做更大规模的分布式。

这次之所以能做更大规模的分布式训练，得益于高速互联的网络，现在的核心网能到 800G。**但是网络传输的上升如果到了上限，分布式规模就上不去，总算力就上不去。**可能快到瓶颈了，网络传输随着规模增大，终究有上限，不会无限暴涨。

- **核心网带宽：**

800G 到 1.6T 的 roadmap 非常明确，两三年后就可以。到 3.2T 肯定也可以做到。再往上就是看要性能还是成本，大规模的 AI 集群训练是以前没有过的任务，要求 high performance。云计算考虑成本更多，一些技术不会用。如果只看 performance，10 万颗芯片连起来不会需要很多钱，能用的光互联技术非常多；第二 Nvida 有一个技术叫 Nvida link，它可以在服务器上，或者 pod 里，用比较宽的互联带宽直接把芯片连在一起。所以未来 5 到 10 年内，带宽互联不断上升

应该没有特别大的问题。

这也是为什么美国去年会限制芯片出口到中国，它可能也是看到以后 AI 可能都是比较大的集群，所以把中国以后能用的带宽限死了。这对于大规模的分布式训练是挺严重的问题。

单卡性能越高，效率越高，毕竟分布式会有折扣。换 H100 就 5 倍性能提升，尤其对于 Bert / transformer 的原生算子支持得更好，国外已经全部都在上 H100 了，等今年 3 月 21 号的 GTC（NVIDIA 主办的 GPU 技术交流活动）看会不会发布下一代产品，H100 只用到 4 纳米，但现在 2 纳米基本上已经 ready 了。

- **关于显存的空间换时间：**

用 3D 堆叠就可以解决，AMD 在这点做得比 Nvida 要领先，靠硬件带宽至少再有 10 倍的性能提升问题不是特别大。存储、传输和计算三个叠加在一起，两三年差不多就可以。

> **3D 堆叠**是指把一块芯片从二维展开至三维，AMD 正致力于在 CPU 和 GPU 之上直接堆叠 SRAM 和 DRAM 内存，以提供更高的带宽和性能。3D 堆叠的好处在于缩短了电流传递路径，也就是会降低功耗。

7) OpenAI 能堆到多少张卡，会到什么量级?

- **集群：**

微软全部都是直接买 Nvida 的 DGX，微软的 AI 超算是外包给 Nvida 做的。Nvida 之前的设计包括 NV link scalibility 没有做得特别好。谷歌可以 5 - 10 万张卡直接连在一起训练，英伟达是 2 万张卡，如果这方面多做一些优化，至少集

群可以提高三五倍。

- **单卡性能：**

H100 可以直接提高 5 倍。硬件上能够提高的空间还是比较高的。

再往上就要看有什么新技术。比如芯片间的电互联改成光互联， Nvida 在美国投了一家做光接口的公司。

CXL 是 Intel CPU 的内存可以大家共享。但和 AI 高性能计算的关系不大，在云计算和其他方面会有一些变化。英伟达将来估计也会推一套封闭系统。它的 DGX 其实自己带了几块 ARM 的高性能的 CPU 。

> **CXL:** 是一个开放标准的行业支持的缓存一致性互连，用于处理器、内存扩展和加速器。从本质上讲，CXL 技术在 CPU 内存空间和连接设备上的内存之间保持内存一致性，支持资源共享（或池化）以获得更高的性能，降低软件堆栈的复杂性，并降低整体系统成本。

国内现在主要卡在 NVlink，不仅是 license，还有硬件设备，我们买到了卡，但是连不起来，也是很大的问题。

8) 大家有没有感觉到 GPT 3.5 完全实现不了的新能力？

GPT4 在已有的能力有延续的提升，包括逻辑推理能力的增强、开放性问题的回答、对视觉的支持。但是似乎没有涌现出 0 到 1 的新能力，像 vision，vision transformer 也能做到，现在只是拼到一起了。

视觉处理，现在可以画图片了，并且字数长了 8 倍。

视觉数据进来之后，给了 AI 一些具象的认知，比如有一个热气球，剪开之后热气球会飞上天。有图像输入才能有具象认知。**像人对世界的理解， 80% 靠视觉在具象的过程中理解事物，20% 靠语言输入。所以有了多模态的输入之后，最终使得逻辑增强。逻辑性是一个慢慢演进的过程。**多模态相互加强之后，能够把它的逻辑带到什么样的高度是值得期待的事情。

自动驾驶最早是 CNN，“我见过一个东西，然后我识别出来”，后来到特斯拉，“我没见过的，我可以用占用网络”，现在为多模态的模型，可以对更深层次的语义有一定理解，“我见过这个东西，或者没见过，我觉得它比较危险，就可以提前躲开”。就像这次展示的图片，车后面有一个很奇怪的小丑，它不符合我们通常认知，就可以提前的避开，这种能力是多模态模型赋予的。

GPT-4 的技术文档提到了合成数据的应用，解决封闭领域的 hallucination 的问题，他可以给自己找问题，比如自己写一个回答，再找这个回答中 hallucination 的地方。自我迭代的能力非常重要，这个能力一旦超过人的 benchmark，就可以踢掉数据标注了，就像人作为 boot loader of AI，剩下就自我迭代，会加速整个进程。

- **Boot loader:** 是在操作系统内核运行之前运行的一段小程序。可以初始化硬件设备、建立内存空间的映射图，从而将系统的软硬件环境带到一个合适的状态，以便为最终调用操作系统内核准备好正确的环境。
- **Hallucination:** 幻觉 (hallucination) 是指没有相应的客观刺激时所出现的知觉体验。

9) 如何定义界限？人的界限在哪？大模型未来的发展会不会太快了？超过了人的能力？

GPT-4 在某些领域已经超过 80% 的人的界限了。比如设计图片、图片生成速度。

很多人觉得 GPT-4 没有太多的惊喜，但其实 GPT3.5 已经解决了不错的逻辑问题， GPT4 的 performance 加强了很多。**这恰好反映 OpenAI 不是只追求做出新的东西，而是把已有领域的 performance 进一步加强，在一些细节的地方才能体现出来。这其实是很可怕的，因为做一个新的实验方向，展示一下不是很难，但真的把它做好不容易，所以 GPT-4 最震撼的地方是本来已经挺好了，没想到它还能再好。**

GPT3.5 是 90 分的话，GPT-4 已经是 90 分的水平，从 80 分到 90 分往上再提升一点都很难，这是边际效应的问题。OpenAI 的厉害之处在于把 80 分到 90 分优化的能力界限量化了。

GPT-4 主要解锁了三个能力，第一个多模态，第二个是提高 prompt 数量，第三个 hard task 有更多突破，比方说做那些比较难的题。第二个能力大家没有特别关注，以前是 4K，现在有 8K 和 32K 的版本，这个能力挺强的。**因为 OpenAI 后面的几个接口没有提供 finetune 的能力，只能做 inference 接口**，只有 2021 年之前的知识，如果要反映现在情况，必然会通过 Prompt engineer 。如果说 Prompt 受限，其实受限了不少能力。

另外一个角度，GPT 解锁的是跟人对话的能力，如果能无限的 prompt，模仿人的记忆，理论上可以把对话无限的拉长，真正地去模拟人的一些思考或者情感。受限的话，它会把前面的内容忘记掉。

实现记忆可能不是通过 prompt 实现的。GitHub、copilot 都是通过 script，生成 Embedding 的方式实现，再实时做一个 search，就可以实现人类的短期记忆、中期记忆、长期记忆，人类短期记忆比 GPT4 的 token 少多了，它主要是通过中期记忆实现。Microsoft 和 office copliot 结合的 Microsoft graph 没公布结构，不知道是不是通过 embedding 实现的，但不用通过 prompt 实现。

> **嵌入式（Embedding）**是指一种将输入值（例如文本）映射到向量空间的技术，以便计算机可以识别文本中的含义。它通常用于机器学习和自然语言处理（NLP），并可用于实现搜索引擎、推荐引擎等应用。

OpenAI 提过 embedding 的功能，之前 Text search 的评分只有 50，在句子相似度、code search，可以做到 80、90 分，如果 Text search 在 GPT4 可以从 50 分提高到 80/90 分，说明对知识、深层结构的理解已经超过普通人理性思维的能力。

但图像用 token 化 embedding 恰好说明了其局限性，因为图像用 kernel 生成的像素集合作为 token，其上下左右的语义关系是很不清晰甚至没意义的。

10) OpenAI 把多模态开放后的影响：

图片只是一个采样，但 video 信息量很大。用图片做机器学起来确实很慢，但人的学习是时间序列，不仅仅有形象，因为时间轴上有动作变化。

NLP 的 token 差不多学完了，数量级已经够了。现在要给神经网喂新的知识。互联网的视频是大量的，AI 真学明白了的话，肯定是又是一个飞跃。

这次加图片看起来比较简单，但是这么大规模的模型，加一个新的模态，从训练上来说是非常复杂的工程问题。类比特斯拉一直只用 vision 做训练，没有用

激光雷达，很大原因是加了激光雷达之后，模型 performance 不一定提升，反而有工程上的问题，OpenAI 现在能把图形加进去，证明它在工程上领先所有人一个段位。

11) OpenAI 继续变强之后，哪些大概率受到冲击，哪些会有一些壁垒？

这类似于滴滴很难被 iPhone 冲击，但应用商店被 iPhone 冲击了。

可能会彻底改变工作生产模式。 OpenAI 去考 GRE、大学课程，都能考到前 5%，也能画 PPT。很可能之后在组织内部，应届大学生只要 20 美金/月、并且供应是无限的，因为 GPT4 已经超越了很多接受大学四年本科通识教育的大学生的水平了。可以想象一个场景，在拥有无限供应的 20 美元/月大学生的情况下，公司和公司的组织架构会变成什么样？

12) GPT-4 的考试能力这么厉害，是什么东西解锁的？

- 把 video、图像的数据给进去，通过多模态的方式迁移过来了。
- 可能用之前的 training data，以及 reinforcement learning 的时候的标注。
- 来源于专业考试的训练数据：

在 text 的角度，**用什么东西去训它，它就有什么样的能力**。GPT-3.5 应该没有在 scientific literature 等论文数据集上面训过。在 2021 年前，大家会倾向把论文的数据集剔掉，因为不好处理，后面意识到要把它加进来。专业知识是用什么训它，它就有什么能力。

13) 随着时间推移会不会跟个体绑定，GPT 对个体意图的理解会更深？GPT 对人类意图的深度理解大概多久会发生？

其实不难，GPT-4 知道用户更多的背景知识就可以。GitHub copilot 无非就是放到 sniggets 的文件夹，把原来写的代码放进去，它就可以复用；

Office copilot 把用户放到 Microsoft graph，现在不知道具体怎么实现。Microsoft graph 就直接可以用邮件和通话记录做 grounding，让两个人变成瞬间熟人，而且远远比两个人交流的速度更快。

14) 技术上，融合图像和文字是怎么实现的？CLIP 是把图片打标，使用对比学习的方式；Pre-training，是前面预测后面。哪种方式更可扩展？图片是只用了像素信息，还是图片可能有一些描述、标签也需要拿到？

VIT 把图片变成 Token，让 transformer 比较容易学习，但图片到底是什么神经网还不知道。还是依靠 diffusion，让图文可以通过另外一个网去关联，diffusion 的学习数量很多，学了 14 亿张图片（甚至更多），所以它知道图是怎么构成的。CLIP 只是映射，出图的质量太差了。

AI 借助了两个网来明白怎么出图，把 diffusion 网搞得更大或许就能实现更好的融合。在 CLIP 上用语言模型还比较弱， Stable Diffusion 考虑了实用性，不把网做太大，否则装不到卡上。Google 已经在把语言理解的模型做大了，但估计实用性有问题。

在文字上引入图片可能不简单，图像和文字在语义上区别很大。把图片 token

化之后，和自然语言的 token 不太一样。自然语言 token 的语义很清楚，但 Pixel 没那么强的语义，所以 GPT4 能把语言和图像的语义层上混起来，就是很大的技术突破。现在出图的语义层上只是映射，真正明白怎么出图要靠另外一个模型做。虽然底层可以让 Transformer 融合，但靠蛮算让 GPT 自己去明白效率很低。

GPT4 的多模态大概率现阶段用的还是 Blip2 和 Flamingo 的方案，也就是拼接 CV model 和 LLM，但后续也许会做 unified model 。

> **具体逆向工程可以参考：** https://thakkarparth007.github.io/copilot-explorer/posts/copilot-internals

15) Transformer 如何融合多模态，机制是怎么样的?

就像在机器翻译中需要有两个语言的对应关系，融合多模态（如图像和文字）时，通过做标注等手段，一定会有两者之间的对应关系。

图像和文字变成 Token 是怎么来的：Token 是通过上下文表达出来的，周围文字的 Embedding 由于注意力机制被发现，因此和文字产生了对应关系。

16) 大模型需要解决什么问题?

1.语言问题 2.知识问题，其中语言问题现在已经得到较好解决。

我们的目的是制造能够理解人类语言的模型，还是制造一个涵盖全世界知识的百科全书?

如果为了后者，则参数越大越好。如果为解决语言问题，只需要 10B 甚至现在 LLaMA 的模型就能够很好解决。所以不能盲目追求参数提升或某项能力提升。

17) OpenAI 组织内部都觉得要做 AGI，知识语言推理，甚至是 motion、抓取，包括很多机器人相关的 task 都可以由 AGI 的方式来做。

AGI 要解决的问题从时间上可分为三部分：

- **过去部分**：世界上所有的知识都纳入大模型会存在很多问题，比如现在让大模型提供几个 URL，但点击进去会发现大部分 404，如果每个 URL 是个“知识”的话，是无法保证所有信息都是正确的，或者需要极大的模型才能保证正确；
- **现在部分**：做项目 DD 及 transaction 的方面还差很多（或者不是发展的方向）；
- **未来部分**：更多涉及到创造，但目前还没有标准的答案，目前在这部分做的很好；

OpenAI 解决这些问题需要一个过程：

- OpenAI 的人才更多（硅谷所有 Super Talent 现在都想去 OpenAI、资源更多有足够的过程来解决上述问题；
- 首先是硬件可以有 10-100 倍的性能提升；
- 模型结构本身可以优化；

- OpenAI 的 GPT 团队现在只有几十人，未来还将涌入更多的 super talent；
- 针对 URL 问题，OpenAI 现在基本达到普通大学生水平，但如果有一天提供的信息全部正确了，这件事还是很恐怖的。

18) 如何定义 GPT 的能力边界？未来几年有哪些高难度任务，是 GPT 最近几年解决不好的？

越无限游戏的越不好解决。无限游戏的核心是不断拓宽边界、制定新的规则，举例自动驾驶就是开放性较强的无限游戏，永远不知道路上会有什么实时路况。

- **有限游戏**：以取胜为目的，拥有明确的开端、终结和界限，在开赛前，参与者需要对游戏规则和获胜条件达成一致，规则在游戏进行当中不可改变；
- **无限游戏**：以延续游戏为目的，因此无限游戏没有明确的开端、终结和界限。为了让游戏延续，规则可以在游戏进行中改变。

但目前可以做到许多之前做不到的事情，比如在游戏中做新的 NPC。NPC 运行的逻辑是一样的，从有限到无限都是从有限的知识中抽象出来结构，应用于新的场景。**需要通过不同思维解决的问题，更难用 GPT 回答。**

在数学上，集合论如果涉及到大基数的问题，针对每一个问题都是全新的，需要不同的数学思维去解决集合论问题。这种方面，利用 AI 解决会比让其解决四则运算要难得多。但 AI 在抽象代数方面已经有六七十分了。

ChatGPT 理性思考能力很强，但涉及到感性、人情世故、办公室政治斗争等方面是不行的。这是因为 objective 中第三点提到的 harmless。因为需要符合一个普世的价值观。**对于自然科学，只要背后有规律，神经网络学起来比较容易。**但社会科学，像勾心斗角，就像股票，背后的规律非常复杂，而且很难学。不

过社会科学可能不是 OpenAI 及大多数公司关心的点。

ChatGPT 在“面”上的工作都可以完成的很好，比如本科和研究生阶段的知识，但 PHD 则是在一个点研究，而且扎的很深，一篇论文有很强的逻辑推理，以及很多的实验，**把某个问题的科研边界向前推进可能是模型短时间内无法做到的。**

19) GPT 能不能“勾心斗角”：

在 GPT3.5 刚刚出来的时候，有人回答了几个职场问题，答得很不错，没有人意识到这不是人工回答的。

我们不应该低估 GPT 做这些的能力。关于 GPT 已经超过 AGI，并不是 GPT 真的那么厉害，而是**人类确实十分 vulnerable，即使是理性的人，碰到具体事情时会变得不客观。**

大家都会说 GPT 有时候会 hallucination，说胡话，但实际上 15%-20% 的语言是 human nature 的 hallucination。

勾心斗角其实是 alignment 的问题。在类似实验中，**让模型来做 Negotiation，看能否让他的话术更强、更有说服力，模型很轻易就可以做到。**所以研究社区并不是讨论模型是否具备勾心斗角能力（因为肯定具有很强的能力），而是如何让其更加 honest。**现在的模型看起来很直男，实际上也是被调出来的，不让它有太多谈判或战略讨论的能力。**

GPT 其实还可以避免了人的冲动，是非常理性的，社交水平还可以。

20) 在 Ilya 的一期访谈中，提到了很多观点，其中一个是 hallucination 问题：

一开始模型需要很多数据，随着后面提升，需要的数据变少了。因此未来发展很关键的一点是，如何用更少数据去训练模型。

第二点是在 GPT 给出不符合要求的回答后，告诉他不对，**不断加强反馈，很大程度能在不改变模型的前提下解决 hallucination 问题**。

Ilya 说很多人质疑用统计学概率模型，是不是不能做到 AGI，但实际上从 Microsoft Sydney (bing 中 chat 机器人）的反馈来看，通过统计学习，能够产生真正的理解能力，最终可以实现 AGI。

本期播客是对于 OpenAI 联合创始人兼首席科学家 Ilya Sutskever 的采访，他在创建 GPT-3 和 GPT-4 方面发挥了关键作用。Sutskever 介绍了他在机器学习领域的背景以及对计算机如何学习的兴趣。他讨论了大型语言模型的局限性，包括它们无法理解语言所涉及的一些基本现实，但同时也指出 OpenAI 正在进行研究来解决这些问题。Sutskever 还强调了了解生成模型中的统计规律的重要性。此外，讨论还涉及到机器学习模型在未来可能变得更少依赖于大量数据的潜力。谈话随后转向了 AI 在民主中的应用以及创建高带宽民主的可能性，在其中公民提供数据给 AI 系统。

上述内容是基于 summarize.tech 总结、ChatGPT 进行翻译及润色生成

Q2: AI Native Apps： 如何思考 AI native 和 LLM 应用的前景？AI-native Apps，最终会呈现什么形态？未来看好哪些垂直应用？

A：

1) 如果我们都采用了 OpenAI 的模型，我们该如何建立壁垒？

LLM 最终需要细分场景的数据及规则，如果浅层地在前端组合应用，那么短时间就会被超越，但是快速拿到细分场景的业务流或者 knowhow，做出细分场景下很好用的应用会是一个方式。

在**交互模式上会发生巨大改变**，现在类似 Adobe 等等图形化的交互模型，或是需要人去按按钮的模型，都会被改变。**未来可能就是一个输入框似的模型**，多种交互方式共存、交互方式更宽广、交互的效率大幅提升。

OpenAI 与以前的移动互联网，包括 PC、Mobile，在范式上既有相同也有不同。移动互联网是基础设施，因为需要 location、数据传输，解决基础设施上网在线的过程。GPT4 代表着移动应用的方向，更多地是解决人在情感上的诉求。比如教学方面跟人的沟通，甚至可以看延伸到律师等职业场景的变化。**下一个改变应该是在情绪上。可以变为人类的工作助理或陪伴者。**

2) 针对未来 AI native 产品形态的看法

- **产品方面**：ChatGPT 展现了很好的对话能力及推理能力，但本质上还是集合了世界海量的知识；
- **技术方面**：大家设计技术的思路会发生变化。AI native 可能是**基于 Large Language Model First 的设计思路**。比如推荐引擎系统，可能会有 Pre-define 的架构，有了 LLM 之后能帮助我们做更多联想，大量取代一些模块、产生更多的机会，甚至产生一些特征空间帮你显著提高商业化能力及用户

数量、以及对不同场景的认知。

AI 带来的改变不是 PC 到移动互联网这一跨度，而是类比到 PC 发明出来的那种改变。**我将这次的 AI 科技理解为一个 50 年甚至 100 年跨度的结构性的科技变化。PC 对人类的改变：比如之前用笔记笔记，后来有了 Excel、Spreadsheet，是这样的飞跃。**

这场科技革新更应该类比个人电脑的 IBM 时代，OpenAI 会被比作超级通用计算机。

APP 和网站是把统一的超文本呈现在不同的浏览器框架上，壁垒不在渲染呈现而是框架上附带的粘性和广告。Chatgpt 实际上更**类似于超文本而不是浏览器框架。**

3) AI native 能在企业中完成哪些工作?

AI native 影响现在创业公司的人员架构，应用 LLM 可以提升企业执行效率。

AI native 能改变软件开发的工作流程，**现在的链路：业务-产品-技术中，每个人在接受信息也在传递信息，应用 LLM 的链路上不需要有人存在，全部以机器理解的形式存在。**

- 程序编写可能未来以**信息流、数据流**来驱动整个产品的设计。过去表单式的交互高频刚需，现在的对话式更加低频、长尾。因此在设计程序的时候，在前端需要加入长尾式的对话交互方式。可能 prompt 工程师会取代已有工程师。

如果应届大学生类似劳动力以很低的人员成本提供，自己该如何优化组织架构。中国 SAAS 企业人员效率低，如果用 AI 替换是否就可以得到解决。

无代码及低代码是期待销售人员能直接根据用户需求调整软件。GPT4 给架构师带来了很好的体验，因为基础的工作，对于需求文档、API 的理解，都可以通过 LLM 生成，而且足够灵活。**OpenAI 的产品能否带来一位“能听懂需求的工程师”。**

通过 LLM 可以快速理解需求，完善产品功能及用户流程，最终输出一份 PRD，还有技术实现的文档，最终基于文档优化代码结构。不过要求 LLM 将所有完整代码按照架构梳理出来，会有很大误差。

PRD: Product Requirement Document 产品需求文档，是产品项目由“概念化”阶段进入到“图纸化”阶段的最主要的一个文档。

GPT 能取代技术、蓝领、机械性的逻辑工作。但自我认知、共情力是未来在招人的时候是非常重要的，机器无法取代。

GPT 会改变现在的金字塔形状的组织架构，组织的边界在快速扩大，变为网络化。传统的雇佣模式也在发生变化，AI 更像是公司中层。

最早 Copilot 出现的时候没有太大颠覆性，因为相比逻辑思维，编程语言并不困难。如今我们可以教 ChatGPT 逻辑思维。

需要生产一个什么样的 APP 可以拆分成，比如 UI 界面的业务逻辑，后端的数据库 API 连接，后端的运维环境。当然这里面有有一部分问题在上一波的

SAAS 的自动化中得到解决。

如果重新定义整个生态流程，然后把方法论交给 GPT，它就可以去实现一整套的生态逻辑。

Native AI 可以将许多专业用户的工具软件革新，对于白领来说，获取知识的方式原来是被动接收式的，那未来可以用订阅制的方式变成了一个专门面向某一类人群的整理好的，且具有偏好性的。

今天的生产方式在发生变化，基于新的生产方式下，可以做出来的应用的基数也比原来扩大了很多倍。

Native AI 弥合了专业团队与非专业团队之间的巨大 gap（好莱坞大片和抖音视频）接下来普通的创作者只要有 vision 和 idea，也能做出这种高级作品，全世界的内容生产质量会迅速的拟合到一个平均水平。

自然语言到每个人可用的编程工具是巨大的生产力革命，给每一个人赋予了一个跟数字世界的一个接口，natural language to everything，整个生产力的解放和革命，也是逐渐下沉的一个过程。比如从 Adobe 到 Figma，工具的易用化带来可用人群的扩大，都会创造一些百亿美金的公司。而 Native AI 将这一过程可以延伸到许多领域。

两个设想：如果所有的知识数据都能够记录到大模型中的话，那么就不需要数据库来做 SaaS 模式的记录。如果一切问题都在大模型得到解决，那么也就不需要新软件的开发。但目前的问题是大模型还需要参数，而且不能实时更新和反馈。未来如果这个问题得到了解决，实际上只需要一个公用大模型和一个私人大模型。

不需要再后端数据处理，后端团队会变成数据集梳理团队，更高效和准确的喂给 LLM 得到最高效的结果。甚至可能最终技术活都是 AI 做。

不一定只抽象出高频刚需的需求，复杂和长尾需求 LLM 也可以帮助解决。

很多工程师就像一个 Pre-trained Model，理解需求其实很难。语言是广义的，定义正确的问题和需求是第一位的，产品经理会更重要。之后的中小企业或许只需要 CTO 和少量的 PM，未必需要工程师。

一个需求被解决，除了编码，还有构建、部署和运维，这三方面 AI 能解决吗？尤其是部署和运维，会对云计算和云原生的使用方式会有变革吗？

所有不带 Critical thinking 的岗位和人才都有可能会消失，LLM 负责执行，人负责 critical thinking 和 judgement。

4) 五百强企业的数字化转型问题：

财富前五百大多数公司数字化能力与 Google Facebook 这样的科技巨头相比要差一个数量级，而且过去十年基本上 80%-90%优秀的工程师，尤其是尖端人才，都是在 Google Facebook 或者那几个大厂里面。大模型广泛应用以后，可以拉平双方的差距。

许多世界 500 强，尤其外企，做数字化转型是为了提高其人效。除了解决招聘早期的 hr 筛简历，电话，面试等等以外，上线数字化的 AI 工具，还可以提升

用户的体验。提高内部组织数字化转型，让整个组织变得更加敏捷，降低沟通成本，实现业务最大化。

国内数字化转型速度是远远落于欧美的，但是能够关注到 AGI 特别火了之后，对于很多企业来讲，大家都在疯狂的去使公司业务上线，将公司的系统做相关的数字化转型。

GPT 没开源，但私有的模型这件事已经发生了，今后每个人都可能有很多模型。

5) 应用大模型的好案例：

notion AI 可用到可卖实现得不错，Jasper 也是商业化较好的案例，Copilot 正在与 20 多家客户共创。除了这几个之外，其实没有什么特别好的案例。或者说现在 MVP 产品只有 1%，甚至没几家产品实现了商业化。

AI 发展出的共情力也可以有一些衍生的可能性：

- 局限于在于输入太少。Mobile 普及之后，多模态的普及之后，共情力能够得到发展。（Mobile 设备上传感器更多）；
- Character.AI 介于娱乐、社交、游戏之间。其聊天机器人可以表现出一些共情能力；
- AI 劝说了一个马上要退学的学生，是一个令人惊讶的点，目前来看 AI 的共情能力还没有天花板。

6) 如果短期来看，AI 产品能做好，其特点可以总结为以下几点

不改变原来的使用习惯，这代表了它的替换成本比较低。

确实能够解决一个比较痛点的问题。或者说不太痛点的问题，但是有多个类似小场景的叠加来提供比较大的增量价值。

定位一定是做 0 到及格线成绩的事情。是一个助手辅助的定位，而不是全面替代。

举例：

- Notion：不改变原来的习惯，也提供了一些比较小的价值，但是它的场景比较多，加之交互比较轻便，易用性就好；
- Github 的 Copilot：就是实现一些小模块的代码快速完成，零到 60 分。因此我觉得现阶段就能够出来跑出来的 AI 产品是符合这三个特点的。

7) 长期来看，AI native 的 app 如何判断其未来的趋势：

当年 Excel 时代的时候，就是因为大家都在使用，所以出现了很多围绕 Excel 的模板 VBA，许多生产制造管理系统，但是后面 Excel 时代慢慢过渡到了软件 SAAS 时代。这个过程其实是逐渐迁移的，差不多花了七八年时间。

关于 AI native，其发展进度还不足以令我们做出一个假设。但是微软的 Copilot，谷歌的 workspace 都是指向一个方向的。

国外 AIGC 的公司有大约 610 家， AI native 的 app 一定会从这 610 家中跑出来。所以一个思路是去持续跟踪、试用这些 AIGC 公司的产品，然后从中不断获得启发。

8) 垂类应用的机会：

垂类应用中，像客服、写作、儿童陪伴、心理咨询、员工服务这些垂类应用是值得看好的，是因为这件事情本身就是重复机械的知识推理，用 GPT 可以实现直接覆盖。

像 SEO 内容写作，儿童陪伴，则是需要一个长久持续性的过程，也需要有不断的新内容去产生，因此 GPT 也可以适用于这方面。提到心理咨询，现在数据标注的公司会接收到很多心理咨询类文本的标注需求，可能代表这个市场是会优先被孵化出来的。

输入法作为掌握握所有内容输入的入口，可能会有一些比较大的变化。现在所有的内容的输入都是大脑构思，然后通过键盘打出来的。国外一家输入法公司，将其产品接入 AIGC 的能力，根据用户以往打出来的字进行自动高级联想。

触摸屏让硬件具备了快速适应软件的能力，软件因为 LLM 可能会具备快速适应功能/人的能力。生产流程很可能会发生变革，创业者的机会在于重新构建生产流程工具。

或许应用的数量和质量都会井喷，人人都可以做 App，但 middle layer 的价值会变大

模型把 Engineer 和算法的能力民主化了，抹平了很多行业和公司的积累，大家起跑线相差很小。

9) 垂直应用的一些问题：

可用到可卖存在巨大的鸿沟。现在许多产品都是 MVP，能够非常快速的完成 0

到 60 分的事情。但是如何从 60 分变到 90 分，没有很好的方法，即便是 Midjourney，对于一张图的阐述，实现到 70 分很快，但是 70 分到 90 分的调优方法，用户是完全不知道的，一些客服工具，也是类似。

有一些产品使用感比较割裂的，它是一个点状的小工具，难以融入到整个 Workflow 里面。

运用这些应用的厂商如何去建立壁垒？

- 首先是从企业内部去建立壁垒，内部效率提升，比如说服务环节，服务成本应用了 AIGC 后五个员工能完成十个员工的工作量，带来成本优势。比如美团，每单外卖能够做到便宜五毛钱，在这个规模效应下，成本优势就会变得非常强；
- 第二个是面向用户的私有化的一个模型，这个模型里边的差异化就来自于自己的客户，或者说是调出来的参数。如果有这个数据，调教出来的模型就是更强大，用户体验就是更好。**壁垒可以通过数据飞轮的形式建立起来。**

10) 未来 AI-native 的 APP 架构两种可能：

- 很多人希望会把原有的 SaaS、工作流重新做一遍，比如 Typeface，Adobe 的 CTO 出来说要用 AI-Native 的方法把整个工作流重新做一遍，目前硅谷很多这样的机会，VC 也像打了鸡血一样疯狂投钱；
- 工具软件，比如微软，会有一个知识图模型作为一个过滤器，这边获得业务的数据、商业的流程，会先通过这个模型输出一个 Prompt，再输入大模型中。

11) 长远来看 AI-Native，比较难想象：

按照 Sam 的说法，模型的性能每 18 个月提升两倍，这件事延续十年问题不大，问题应该是不大的，如果 OpenAI 最后的模型能做到比现有的模型能力提升十倍百倍，最后的优化到底是什么样的，比较关键。

比如现在 Midjounery 画得很多图可能只能到 60 分，但一个 OpenAI 的朋友说今年应该会出来一个 DALL-E 升级版，模型效果远超 Stable Diffusion 和 Midjounery，大家可以拭目以待。

Midjourney 的模型语言理解能力不太行，OpenAI 的技术进步方向应该是把图文相结合，但自然语言的理解还有些难，怎么和图像映射起来是一个挑战。

如果想看长远，也不应该只盯着现有的用户，可以考虑目前还有哪些用户没有接入互联网：

- 比如说很多老年人，还没有很好地使用现在的互联网服务，比如打车，随着自然语言处理技术的进步，只需要告诉 AI 目的地，然后就可以打车；
- 下一个时代的张小龙、一鸣、宿华应该长什么样子？首先这几个人都很熟悉互联网 mobile 的特点，其次抓住了关键要素，比如头条把商业化、公司组织、机器学习技术的引入抓得非常好，下一代产品经理，可能是对大语言模型某个被大家忽视的关键点上能做得特别好。

Q3： 大模型的未来格局：可能是多寡头格局，并且中美之间应该分开讨论 OpenAI 一家独大，还是多

寡头战场?

A：

1) 美国：未来格局会有点像公链。

虽然以太坊是最大的公链，但实际上，每个大模型会有自己的生态，也会用自己的基金啊，或者各种各样的方式去驱动它，所以有可能是多寡头格局，大家会提出一些自己的优势领域，不论是模型更大，运算效率更高，用更少的数据，还是能在某个垂类市场，比如阿里就能在商业的场景上做得非常好。

除非有一种情况，模型能力差别非常大，OpenAI 跟大家拉开了一个代际差。

2) 未来格局会可能像芯片：

最开始微处理器有很多家，英特尔领先，但没有领先特别多，摩托罗拉、飞利浦、AMD 都有自己的微处理器，收敛到最后只剩英特尔和 AMD，再到最后只剩英特尔一家。

不管怎么样，有量才是基础，对于其他玩家，本身要有场景，比如当年的 IBM，有自己的场景，无论技术如何，拿来用就可以了，小公司不管怎么样，都要把量先坐上去，后面再慢慢切，就像苹果最开始是比较独特的处理器，后来用英特尔，最近又开始自己做。

后面未来不管中国团队还是美国团队，可能都会放弃做大模型，因为发现这个

根本就追不上，所以大家都会好好去搞应用，就是怎么用好模型能力，可能就是一个非常重要的能力。

自己做模型成本太高了，微软现在是直接买 DGX 的，就是服务器所有的软件硬件，端口盒子打包在一起，大概有 60--70 的毛利，除了 Google 目前有能力自己搭建，其他都是直接买英伟达，非常非常贵，然后现在芯片还在类似于摩尔定律的迭代速度，新的硬件出来，整个系统要跟着升级。

OpenAI 目前能招到全世界最优秀的人，一年就几百人、几千人，长期看小公司自己做个大模型，ROI 应该是算不过来了。

2022 年 5 月，NVIDIA 宣布推出第四代 NVIDIA® DGX™ 系统，这是全球首个基于全新 NVIDIA H100 Tensor Core GPU 的 AI 平台，它也是全球最先进的企业级 AI 基础设施。

DGX H100 系统能够满足大型语言模型、推荐系统、医疗健康研究和气候科学的大规模计算需求。每个 DGX H100 系统配备八块 NVIDIA H100 GPU，并由 NVIDIA NVLink® 连接，能够在新的 FP8 精度下达到 32 Petaflop 的 AI 性能，比上一代系统性能高 6 倍。

DGX H100 系统是新一代 NVIDIA DGX POD™ 和 NVIDIA DGX SuperPOD™ AI 基础设施平台的构建模块。新的 DGX SuperPOD 架构采用了一个全新的 NVIDIA NVLink Switch 系统，通过这一系统最多可连接 32 个节点，总计 256 块 H100 GPU。

3) 中国：多个大模型的格局，首先要分清两个概念：

第一个概念是做大模型的目是说为了探索 AI 能力的天花板，还是为了自己应用落地？

- 如果是前者，确实都没必要做了，或者说很难，需要找到更加 Scalable 的方法，因为目前完全是堆钱堆人嘛，未来的多模态、Action 算力成本都是现在的百倍前辈，肯定跟不了；
- 如果是后者，那还是值得做的，训练大模型就像培养大学生，很多中型公司，比如百亿美金几十亿美金量级，都会培养自己的大学生，培养出来的成本也不像大家想象得那么高，每个公司对大学生能力的要求也都不一样，就像我不可能全部都招清华，北大，哈佛耶鲁；
- 未来比如中国可能有 50 家，他有自己的大模型推到自己的应用落地场景，以 OpenAI 为最通用，但不代表自己不需要拥有；
- 这个事情发生过，2012 年语音识别系统，现在中国至少有二，三十家语音识别系统，比如小红书是有自己的语音识别系统的，在 2012 年的时候，当时做云语音识别系统，可能中国就是 3、4 家，所以我觉得大模型也会找到这一步。

第二个是我们不应该把大模型看成芯片，芯片是计算机结构体系最底层的，大模型比较偏上层的。

- 它不一定是操作系统，虽然它很通用，但也是在应用层面的通用，不会说最后就是全世界就两个芯片；
- 至少中国的 BAT、字节肯定有自己的大模型，然后 50 亿美金左右或者 100 亿美金左右的公司也会有自己的大模型，比如说小红书也会自己搞，知乎也会，其实也可以基于开源的大模型，训练一个自己能掌握的模型；
- 再下面包括所有 AI 公司，不可能放弃做 AI 模型，只用 OpenAI 的接口，包括大的金融机构也不可能用 OpenAI 的模型，所以就中国而言，两年之后应该有 50 家左右拥有自己大模型能力，包括 pre-training、Fine-tune、RLHF，美国、欧洲、日本应该也会有。

4) 可能是国外寡头、国内很多模型：

国外语境下，因为互联网、云计算普及，AI 的传播的边际成本为零，所以国外的任何一个角落，甚至比方说越南，巴基斯坦，只要美国愿意让你用的地方，你都可以用到。

国内其实是更封闭更自主可控的，包括我们能够用到的硬件，跟国外有代差，在有代差的情况下，可能很难做出一个像 OpenAI 那么优秀的通用模型。在这种情况下，针对一些垂直领域的 Fine-tune，是有必要的，只需要做到 GPT 70、80 水平的能力就好。

5) 其实国外也会有很多模型：

哪怕在美国，未来也不会只剩下两三个大模型，甚至内部可能两三大模型，一个做商品搜索，一个做云服务，主要是不是非要去比拼那个 AI 能力的天花板，成本难度没有大家想象的那么高。

前面提到的芯片格局，即使是 AMD、Intel 共分天下的 20 年，其实也有成百上千家芯片，规模不大，做垂直、小的应用场景。大模型可以和芯片类比，但不是 AMD、Intel 两家，芯片公司挺多的。

特别细的垂直场景肯定是有价值的，比方说 Biotech，在很多垂直领域，Data 并不容易爬，也不像文字、图片这种 Task 去 hold，这些场景还有非常多的机会。

其实 OpenAI 目前根本没有把自己的 Platform、Data 打开，只开放了 API，可以做一些 General 的事情，短期内这个公司没有能力和 DNA 去开放，Microsoft 可

能帮助到它很多，但这家 AI 公司，是没有可能性开放能力，让你用的很舒服，去做很多垂直场景。

Q4：垂直场景是否有训练自己模型的必要性?

A：垂直场景是有必要自己训练模型的：

- **成本角度**：很多人担心国内很多公司，投了 3 亿美元下去，最后发现 Meta、Google 开源了，被直接碾压，其实并不会：
  - 如果做落地，不需要投那么多钱，但如果你要学 OpenAI，这个钱都不够；
  - 就像 5000 万是门票，现在训练一个 300 亿、500 亿参数的模型，足够像 Notion 在里面写文案了，花个一两千万美金也能搞出来；
  - 做一个低级版的大模型，门槛还是蛮低的，做一些具体的事情，还是有用的。我儿子才读初中，今天早上已经把 Stanford 那个 aparker 在他的 Mac 上 Run 起来了；
- **应用开发者角度**：他其实最大的诉求就是好用，另外也包括成本
  - 一些垂直行业，在任务的这个难度级别上也会有模型的分层，现在大模型很大的问题就是吃算力，吃数据，成本的花费是比较高的；
  - 那就可以简单的问答，调用用户本地的 CPU 的算力，就可以放一些开源的模型出来，包括像 LLaMA 模型，像 Stabford 的 LLaMA 模型，然后未来还有很多开源模型可以跑在本地或者更低一级别的算力上；

- 然后遇到一些复杂问题，放在 GPT，或者是更贵的模型上跑，就这个分层关系，感觉已经是目前的一个现状了

- **人才角度**：世界上能做这种大模型或者高级人工智能的人才应该不会都在 OpenAI，估计外面可能有几十倍于 OpenAI 的人才，他们属于高校科研机构，或者其他一些公司，还有开源社区。

  - 这些人可能靠他们的聪明才智，想出一种方法，可以在一个中等模型上达到大模型的效果；
  - 毕竟 Transform 不见得是终极答案，而且 OpenAI 的大模型里面其实相当大的参数量是用来做知识记忆的，就是他要把很多知识背下来，但如果可以现翻书，比如找外包的模型，专门负责逻辑推理和数值计算，模型可能就没那么大了，这种知识外挂的模式非常适合开源社区去做；
  - 全球的开源社区，共同建设一个庞大的知识库，然后用一个中等模型实现阅读理解和知识推理，内核开发组非常少的人，几个人几十个人就够，这种模式，以开源社区的力量，去对抗微软，这在过去曾经有人做到了，在大模型阶段可能也有人去做，比如 HuggingFace；
  - 开源社区也分很多种类，有一些是乌合之众，另外一种是依托于某个大公司，比如说依托 Google 或者 Facebook，这更靠谱，因为它毕竟还是一个系统工程，跟以前的很多开源不一样，它的投入比较大，对长期规划要求比较高，还有一定的 Service 成分，不是纯粹的一个离线的代码，可以拿来即用；

    - 刚才所说的中国会有四五十家，甚至更多团队，基于这个开源大模型去改，能不能做成真的是还是取决于这些大厂的开源模型；
    - 这几个大厂也都会想制衡微软，就会有更多的开源社区冒头出现，尤其像 Meta，Google 被这件事搞得这么痛苦，估计也会大笔投入

开源社区建设制衡微软。

Q5：ToB 场景的垂直大模型会向什么方向发展?

A:

1) 大家讨论对 AI-Native 的接受程度，跨越度很大

这个礼拜有美国财富 500 的很多 Chief Data Officer 开会大家的思维还是很不一样的，我们这边大多数人还是在讲科技怎么去改变生活、改变工作，有很多人，其实他的想法并不是这样子的，他们只会找出来一堆的问题。

最保守的一家表达的意思：我们不可能使用 OpenAI 的产品，不只是数据隐私，还有版权之类的问题，他们要求所有的软件公司提供的代码里面不准使用 Copliot，但是这件事他们又不好去溯源

不只是金融公司，硅谷有些科技公司也不许人在公司里面使用 OpenAI 也是怕内部的 data 或者 codebase 泄露。

OpenAI 内部正式在做一个 Foundry，针对私有化部署，之前很多人一直针对他，说只给微软，只给被投，目前不确定是否已经发布。

2) 针对开源模型做 Fine-Tune 的成本没有大家想象得那么高：

数据量可能是 Pre-Training 的万分之一，算力需求和数据量成正比，预训练花

一亿美金，Fine-Tune 也就需要十万美金，需要的时间也很短。

不需要上万张卡，一台机器，8 张 A100，640 GB 得内存，给一些大模型做 Fine-tune，就足够了，也就是 1.5 万美金就足够 host 一个节点，而且成本会迅速下降，平均每年下降到原来的 1/3 到 1/5。

用基础能力比较好的开源模型去 Fine-tune，数据也不会外流，同时可以兼顾推理中速度、安全性各种问题，这是充分的。

平安保险的技术能力比较强，18 年 BERT 出来的时候那部就已经铺开，其实也是根据开源模型自己微调，甚至做了一个模型去调参数，去刷 squad 榜，还能刷到全球第一。

但如果这么做，第三方服务机构，似乎没有什么壁垒。

- 中国给各个金融机构做金融外包服务公司有上百家，如果大家都基于开源做 Fine-tune，方法基本都差不多的，最后也收敛得差不多大公司最后都会用起来大模型；
- 对大公司来说，可控性排第一，它不用把数据给别人，他可以可控地去优化推理。，跟自己的业务、流程深度整合；
- 另外肯定还会有一些二三线的公司是需要第三方服务的，那这些 ToB 的公司可以去帮助那些能力没这么强的公司。所以或许它没有巨大的壁垒，但做第三方服务的创业公司是有生存空间的。

Q6：开源模型是个伪命题吗？大模型其实具备网络效应和先发优势与规模优化，拥有极强网络效应，我

们未来是否真的需要开源模型?

1) 大模型时代开源可能是个伪命题:

就算真正的开源时代，谷歌不会开源他的 Page Rank 算法，微软不会开源 它的 windows 系统。

所以 Open AI，如果大模型成为了它的核心，它也不会开源，剩下的可能就是 Tier two，这些公司会去开源。

在 AI 的时代，很多问题可能解决不了:

- Tier two 的能力无法提高到跟 Tier one 同一水平，在过去开源的时代，大量的极客愿意去玩开源产品，可是在大模型时代，没有几个极客能在上面再去做这个 Fine-Tune;
- 虽然算力不是很多，但也需要很多钱，估计没有几个 Individual 的人能够去给开源的大模型做出贡献，我也怀疑大家拿着 Tier two 的模型去调是否真的要比购买 OpenAI 的服务更好;
- 云时代时期 Google 作为 Tier two，为了和 AWS 对抗，就把 K8S 开源了，这也是非常核心的容器调度系统，但当时并非是 google 的核心业务，所以 Meta 很有可能开源，因为本来这不是它的核心。

Kubernetes 也称为 K8s，是用于自动部署、扩缩和管理容器化应用程序的开源系统。它将组成应用程序的容器组合成逻辑单元，以便于管理和服务发现。Kubernetes 源自 Google 15 年生产环境的运维经验，同时凝聚了社区的最佳创意和实践。

云原生技术的三架马车：

- 微服务 易伸缩，高可用，快速迭代；
- 容器化 资源隔离，跨平台环境一致性，借助于 K8s 的强大编排部署能力；
- DevOps 应对微服务的发布、运维和监控。

2) K8S 一开源，国内一下就冒出了很多做私有云部署或者云项目公司，也热闹了很多年，最后发现还是大的供应厂商拿走了很大的利润。

核心技术都是有时间点的，Google 搞出分布式三驾马车的时候，绝对是它的核心技术，但是过个一年雅虎的一个团队就搞了个 Hadoop，大家也可以基于论文去实现，然后来做各种内容；

不会所有人都去用 Google 的开源工程，反而是 Google 把自己的技术当做一个宝贝，不愿意让大家去更改底层，最后服务搞得一塌糊涂。

但是 Google 在安卓时代第一天就开源，然后安卓自己有个核心版本，也有个对外版本，自己的版本和 Gmail 全家桶绑定，所以这个时候它的核心早期是操作系统，但后面其实是整个 GMS，全家统的商业变现模式。

就是你的事情刚开始是核心，但到后面出现了竞争，它不再是核心，这也是这竞争最美妙的地方，所以总结下来，开源的模型有可能是靠谱的。

Google App Engine， 这是一种平台即服务 (PaaS) 产品，可让软件开发人员访问 Google 的可扩展托管。开发人员还可以使用软件开发工具包 (SDK) 来开发在 App

Engine 上运行的软件产品。

2003 年和 2004 年，Google 公司先后发表了两篇著名的论文 GFS 和 MapReduce，这两篇论文和 2006 年发表的 BigTable 成为了现在著名的"Google 三大论文"。Doug Cutting 在受到了这些理论的影响后开始在雅虎开发 Hadoop。

HDFS 是一种分布式文件系统，用于处理在商业硬件上运行的大型数据集。 它用于将单个 Apache Hadoop 集群扩展到数百 （甚至数千）个节点。

Q7：公司会将核心能力开源吗?

A：

1) 刚刚说的三个东西可能都不是核心，虽然外面的人都认为是核心。

虽然分布式的技术是核心，如果当年 他们不写那三篇论文，我觉得他至少领先业界 3 到 5 年，论文发布以后，外面搞了一个开源，可能只领先一两年了，但分布式应该不是核心赚钱的业务，它背后的技术有很多，那未来大模型赚钱的业务就是模型本身，模型本身就是产品，他的 subscript 就是他的模型本身。

安卓开源最初大家也不看好，iPhone 成功之后才成功，未必是核心：

- GMS 是 Google 拿来做 Service 来收钱，Tensorflow，Deep Mind 都是作为 Show Case，其实业界也有分析，就是 Google 无法推出类似

ChatGPT 的产品，因为影响核心业务广告收入；

- 所以最后 OpenAI 会在预训练模型变成商业体系的边缘地带时候，选择将其开源，反过来说，现在把最核心的大模型开源，结果开源模型的效果不如现有模型，别人就会怀疑你还有一些核心的东西没有开源，最后可能两边不讨好；
- 开源的时候，工程问题、经验问题，Domain 的问题，包括硬件，训练架构，可能真的是一个非常难做的决定。

2) 我们需要重新梳理一下问题：

- 第一个问题是，预训练模型做得好，选不选择开源，我的结论就是说我觉得一定会有人选择开源；
- 第二个问题是基于预训练大模型，自己去做 Fine-tune，是否能满足用户的需求，肯定不是所有人都能做好，但是一定有几十家是可以做的；
- 第三点是我们基于开源模型做 Fine-Tune，“好”的标准是什么，它的能力比不过 Chat GPT，它要在很多场景，无论是 toc 还是 tob，是否能服务好客户，这大概率是可以做好的。

Q8：关于 GPT-4、大模型有哪些非共识？

1) 前面说到中国会有 50 个大模型，但或许中国应该算自己的账，比如如果大模型的颠覆成都真的和原子弹一样，那还是要举国体制去做一个对标 OpenAI 的，而不是先去做场景。

大模型很有可能像原子弹一样重要，以后打仗可能都是用 AGI 相互打，飞机大炮都不用人去操作。

- 人的指挥是跟不上了，你看三体描述的末日之战的场景，到后面复杂的战场环境，人类已经没有办法做判断了，只能靠 AI;
- 中美目前处于一个对抗的时期，所以我们都说中国要有自己的大模型，会不会就是中美两国的大模型就形成了对抗，只有在有一个和 OpenAI 同样强大的模型相对抗的情况下，才有可能让大家都坐下来一起谈，确保 AI 向善;
- 当然如果两个超级大国处在相互对抗的状态，那怎么防止 AI 作恶，也会成为一个问题，这个问题不适合展开，但可以参考曼哈顿计划的历史，看一下当年超级计算机 Cri 的研发历史，包括中美建交之后中国建设自有超算的历史，都是有启发的。

国家之间肯定要竞争，不管是商业还是政治上，Open AI 首先是商业化公司，还不是政府行为，我们现在可能也就落后两三年，此时此刻我们不去做这件事情，将来被卡脖子的概率可能性很大：

- 因为我们现在从基础软件到基础芯片全面落后，基础芯片作为硬件实体，我们可能通过各种渠道拿到；
- 但软件是拿不到的，虽然说有开源，我们可以在很多垂类小环境做得比较好，但在大模型领域会有持续长期的落后；
- 其实类比下来，现在要做的国产替代的芯片，是上一次工业革命的产物，我们目前在欧美主导的国际秩序下想办法自主替代，结果人家已经进入到下一次工业革命了，如果我们现在不奋起直追的话，那可能等到第五次工业革命的时候，我们还在替代第四次工业革命的东西；
- 这个到后面会提到国家高级别的战略，比如报告已经提到了很多算力补贴的方案了，政府会像补贴新能源一样，甚至力度会超过新能源，因为这件

事情是可预期的。

过去一段时间，大家搞自动化、机械化用来对抗廉价劳动力，防止后发国家弯道超车，OpenAI 这样的产品，也有可能是对抗后发国家的工程师红利，像我们上亿的大学生劳动力进入市场，也会面临一场战争。

去年 10 月份去美国的时候，会觉得中国去美国做 SaaS 很有机会，因为中国像对美国，研发和运营的 ROI 特别高，除了客户关系、BD 不行以外，我们的成本，甚至服务都可以比美国强很多，但是 ChatGPT 出来之后，中国这一板块的优势瞬间消失，甚至变得更差，因为观察中国公司员工的协作精神、线上的开源精神，距离美国的差距比较大，将其用到极致的能力也弱于美国。

2) 把大模型的作用限制在 "生成式 AI" 这个词汇当中是不是一种误导？

微软最近发布的 Copilot 里面最吸引人的部分，是基于过往用户生成的数据进行汇总，并完成个性化的分析和推荐，但内部“生成成分”很少，包括 ChatGPT 的语言理解、对话部分也不属于严格意义上的“生成”，这或许也是概念意义上的反共识理解。

需要进一步明确的是，AIGC 是一个很典型的中国叙事，算作传统 PGC、UGC 的延续，国外很少用到，也就是在红杉美国官网发布的文章 *Generative AI: A Creative New World* 的爆火，将生成式 AI 引入了公众视野，才引导了媒体、公众很大程度上的错误理解。相对而言，“AGI”的概括要全面许多，只不过它还处在比较初级的阶段。

3) 国内是不是将大模型应用窄化为了聊天工具？

大模型的应用场景应该远远大于聊天、对话，微软 Copilot、Notion.ai 等软件都是很好的场景范例，而国内很多做 AI 大模型应用的企业显然视野不够开阔。生成能力显然只是其中的一种作用，国内许多 VC、媒体都将其窄化了。实际上大模型背后的推理、理解能力也至关重要。

Q9: 算力底层和芯片的行业机遇?

如果沿着工业革命讲，针对大模型训练，是否会将整个芯片工具链整体颠覆？最近看到一些美国的公司它们内部的战略也比较摇摆。

除了当前能看到的 GPU、CPU 这类芯片以外，一些低功耗的 FPGA 和 ASIC 芯片也是有很大潜力完成进一步突破的。一方面，现在的 GPU, 包括谷歌的 TPU 等在内做完大训练模型之后，放进本地化部署，整个过程是一定需要低功耗的 FPGA 和 ASIC 芯片来辅助的。

在未来可能会形成这样一种格局：大的算力中心训练好大模型之后，将其分布到各个边缘节点，然后通过 FPGA 和 ASIC 芯片来完成，而不光只是当前看到的 GPU 芯片，也不只是像 NVIDIA 和 AMD 这样的公司能够参与其中。将来谷歌的 GPU，甚至可能包括国内比特大陆、嘉楠耘智等其他一些做过 ASIC 的企业都可以进入赛道。

也应该考虑到这样一个问题，尽管当前底层硬件的创新很多，但五年之内去跟 NVIDIA 比拼软件生态仍是十分困难的，因为最终还是要将大模型应用到这些硬件当中，并完成模型优化、模型剪枝等后续处理工作。NVIDIA 近期收购了很多优质公司，来巩固它在硬件生态领域的市场地位。但可能 5 到 10 年后，比方说摩尔定律对模型算力的预测真的没办法持续的时候，才可能会有一些新的

机会。

Q10: 未来是否会有大模型边缘端推理的芯片需求?

以国内的电力调度来举例说明，现在的潮流调度，包括 N+1 的潮流计算在内，实际上还是基于大部分的云节点和少量的边缘计算节点来进行支撑的，这对于未来智能电网的调度而言是远不够的。但假设在大模型预设之后，将简单计算放在边缘节点，做成一个类似智能电表之类的网关，完成简单的推理计算，将会大幅度减少云端和边缘-中心端的计算压力，模型的数据优化和调度方面也会有质的进展。

太大的模型放在边缘端确实是比较头疼的问题。大模型的内部链接很多，因此想要把它拆分为基础和应用两部分，还是存在很大难度的。尽管有部分模型在中间是有比较窄的，可以把计算分成两头，但是对于一些中间链接较多的模型，数据带宽过大，想要硬性切开就非常复杂。

其实模型内部还是会切割的，每个企业需要利用模型来解决具体问题，因此没必要用大模型来运算，他们就会切一些小模型完成相应任务。

Q11: 大模型在未来会出现技术周期拉长，乃至终止的问题吗?

这个问题是始终伴随着人类价值观和风险应对的，因此 Alignment 的处理至关重要，也会是影响技术周期的潜在因素。要想解决 AI Alignment 的问题，就必须对涌现的基本原理有所了解，但目前还未能对其有更好的解释，因此 AI 模型的安全性始终存在着不可控的风险。

大模型技术的对齐和可解释性和自动驾驶在很大程度上是相似的，大模型的最终价值在于和各行各业深度链接，但人类想法实在是太复杂了，所以这个部分始终没有很好的方法论来解决。“解释”始终是一个分布式的表示，想要对所有参数进行完全解释几乎是不可能的。

目前的可解释性主要表现在语言逻辑层面，比方说在给定答案的基础上告诉用户为什么，但是神经元层面的解释依旧是迟滞的。做工程和模型结构的人员肯定希望能够寻找到这种可解释性，否则就会影响到模型的调整方向，语言逻辑的解释对于目前仅能满足外围用户的需求必然是不够的。

未来在可控性上应当投入更多的时间，因为它通过实验是可以解决的，在某种程度上也可以暂时规避可解释性的问题，帮助各个行业完成更多的事情。例如 Stable Diffusion 将它的模型开源之后，基本上 70%-80% 的工作都是在可控性层面做文章。由于 SD 的语言理解能力确实太弱，大家转而在 prompt 工程上加码，但这显然不是一个长期的事情。但现在 Stable Diffusion 开源后，大家都开始关注模型内部每一个部分能干什么，能不能服从控制，将控制信号调整得更强，并辅以神经网络更好的控制，但对于模型核心的改动并不大。