拾象大模型观察思考 - 最新判断猜想

To: Shixiang Flagship Fund LPs
Date: 2023 - 12

几点对模型最新的判断猜想：

I. 闭源模型

1) 复刻 GPT-4 比年初预期要难，Google 举全公司之力也才勉强接近。

2) 首轮模型竞赛决出前三名：OpenAI / Anthropic / Gemini，全球范围仅有这三家公司做出了 GPT-4 能力的模型。很期待接下来半年谁能推出 GPT-4 能力模型，我觉得还有另外 3 家公司具备这个潜力。

3) 然而 2024 年又开始新一轮竞赛，2024 是决定长期格局的关键一年，格局形成后很难再改变。明年 Q1 的 Claude-3 和 GPT-4.5、Q2 的 Google Gemini-2.0 陆续推出，仅上半年又会把模型能力抬升一个台阶，明年 6 月再推出 GPT-4 能力的模型已经不算第一梯队。假如 Claude-3 / GPT-4.5 的训练成本~3 亿美元，再往后 2025/2026 年的下一代模型的训练成本要涨到 10-50 亿美元。

4) 模型竞赛很残酷，类似芯片或 SpaceX，领先的模型能力又强又便宜，后面的玩家很难存活，但因“阵营”抗衡又不会赢家通吃，最后格局很可能只剩 2-3 家。

5) 模型公司的融资和估值几乎全由科技巨头定价和主导，没有大腿很难存活。AWS/Google 支持 Anthropic，Tesla 支持 xAI，接下来 Apple 支持哪家会很重要。

6) 大模型今天还处在实验科学阶段，就像人类对大脑的理解也很有限，更像是“探索发现”而非“发明创造”，提升模型 capability 的路径目前只有一条：Scaling Law，每一代模型至少扩大一倍的参数+Data+数倍 GPU 等等，是否有其他路径也是无法预测的。至于 Scaling Law，今天也没有一个理论支撑，而是大量实验和试错的经验总结，也很难准确判断下一代模型能力涌现如何、什么时候 Scaling Law 就不奏效了。

7) 对于 Scaling law 的信仰，只来源于极少数天才 researcher 科学家，比如 Character AI CEO Noam、Anthropic CEO Dario、OpenAI Ilya 他们三位贡献很大，同时信仰也最强，这是一个由极少数科学家推动的登月时刻。

8) 另外一方面看，大模型算是人类的千亿美元 AI bet，硅谷几家公司未来几年模型训练成本累加肯定超过千亿美元，这个千亿豪赌的投入会给人类带来什么？比如你是否相信这波 AI 能助推未来 10-20 年 double global GDP? 如果美国实现 AGI，会对全球地缘政治格局产生什么影响？

9) 大模型的人才壁垒其实还挺高的，一群天才科学家用“GPU+Data+Power”帮人类做新发现，天才科学家们牛人相吸，全球大概只有 200-300 位天才 researchers 能做出实际大的贡献，其中的 100 多人集中在 OpenAI/Anthropic，20-30 位在 Google，Meta/AWS/Nvidia 里面几乎都没有，缺少天才科学家的聚集效应，因此没那么看好 Big Tech 自身能做好大模型。

10) 模型公司今天还更像 Research lab，除了 ChatGPT 意外爆红，模型的商业模式还不清晰。硅谷 VC 也几乎都错过了大模型投资，大模型公司独立 IPO 也很难，被收购的概率是更高的。

11) 随着继续 Scaling，2024 年 bottleneck 在 Data，2025 年 bottleneck 在于 Power。Data Center 是很大的投资机遇， Power+Cooling+Networking，也是拾象后续 focus 重点。

12) Synthetic Data 可能是明年很多模型的 bottleneck，假如每代模型参数扩大一倍，Data 也要线性扩大一倍，公开数据不够用，如何提升 Synthetic Data 多样性和质量很关键，如何做出无限数据。Google 甚至可能在这里掉队？

13) Post training pipeline 比想象得难，大家都不知道 OpenAI 怎么做的。Google 有着扎实的 pre-training Infra，想要通过每年更多的预训练实验来超车； OpenAI 有着成熟的 post-training pipeline，只要顺利完成了预训练目标就能很快发布。

14) GPT-4 短期壁垒在 Data secret 尤其是 post-training 阶段的数据。全球范围只有 200-300 人知道 GPT-4 data secret，也几乎都在前三家模型公司。未来壁垒更多是综合的，就像登月，全球只有极少数几家公司能参与登月竞赛，未来几年至少准备 100 亿美元的模型训练成本。

15) 我也在想 AGI 意味着什么，就像共产主义和财务自由，每个人理解都不同，AGI 这个口号可能不是万能的，AGI 能解决的问题很可能也是在特定限定范围内的，人脸识别就是上一波 CV 的的果实。

16) 模型公司今天的核心还是提升 capability，而非产品。capability 只有一个北极星：即 reasoning 推理能力。对产品应用最重要的依次是成本、可靠性、多模态，再其他都是小事。接下来 research focus 重点：Reasoning、Mulltimodal、Coding、Math & Science、Synthetic Data、Reliability。

II. 开源模型

17) 开源模型的使命不是最智能的模型，而是帮助模型能力的 commoditize，在成熟的 use case 上使企业大规模使用。下一个开源模型 commoditize 的重要目标是端侧小模型，可以帮助模型公司分摊部分云端的算力成本。

18) 开源模型在 2024 年内追到 GPT-4 还有不少挑战，LLaMa 团队人才密度不够高。当然不能低估技术开源和人才扩散的力量，有可能明年整个行业都大进步了。

19) 未来大模型覆盖小模型是必然的，大模型是小模型生成器，OpenAI 顺手 train 小尺寸模型只是时间和优先级问题，所以我们还是暂时 pass 了 Mistral。

20) MoE 能提升智能水平吗？MoE 看似可以把参数量做大，但本质上不算 scale up，最后还是看单一 dense 模型大参数量，MoE 能降低成本，但至于能不能提升 capability 还不确定。

21) 模型参数量在 70B 是个分界点，70B 以内能容忍很多错误，在 70B 及向上每扩大一倍遇到的“难度”都是指数级提升的，模型越大，越容易出错，训练大参数量模型失败率很高。

III. 多模态模型

22) 短期内多模态生成和理解还是两条赛道，长期有可能出现同时做好理解和生成的统一模型。

23) YouTube 数据量虽然很大，但很难用到模型训练：视频的信息密度不够 dense，本身没有 language 蕴含的知识密度高。如何高效地处理并提取视频中的关键信息（如忽略视频中的背景，关注关键物体的变化）是关键，Youtube 的用户交互数据可能发挥重要作用。

24) 单一模态模型的长期壁垒还存疑。视频生成未来一段时间 momentum 肯定很好，效果提升显著，但还是侧翼赛道，更没有到商业逻辑清晰阶段。

25) 视频模型生成的技术路线还未收敛：之前 diffusion model 中加入 temporal attention 的技术路线是主流，最近基于 transformer 的 autoregressive model 开始出现。后者很有潜力，因为 transformer 比 diffusion 更适合 scale up，且视频的时间序列结构很适合转化为预测下一帧的任务形态。

IV. 大模型应用格局

26) 为何 LLM-native 应用还没大爆发？

我的感受是：

- 模型即应用，模型能力决定用户体验，
- 模型的能力和成本今天还不足以支持应用大爆发，还在探索期，再等 1-2 代模型，
- 时间问题，需要天才产品经理，后面也很期待认识更多天才 PM。

27) OpenAI 弱点在哪？

我感受有两点：

- ChatGPT 并没有像搜索和推荐一样具备很强的数据飞轮效应，
- OpenAI 并不是一家以用户为导向的公司，而是以模型 AGI 和研究为导向的公司。

如果 Meta/Tiktok/Google 甚至 startup 在产品侧突破，更高效率的数据飞轮，是有机会拿走更大胜利果实的。

28) 假如 Google 输了模型竞赛会如何？

Google 在组织能力、Synthetic Data 和 Networking 的问题都很大。如果 Google 长期落后 OpenAI 一代模型，能否靠超大规模流量补上呢？也许我们不应该把 Google/Meta 当作模型公司，他们更像利用 AI 技术做好自身产品的公司。

29) 大模型公司 storyline 怎么讲？这里有个 big lesson，Character CEO Noam 前几个月用产品 story 融资，被硅谷投资人 challenge 这个产品看不清未来空间多大，现在 Noam 又转回了大模型走向 AGI 的 story，但走 AGI 这条路前面竞争又很激烈，如果你是 Noam 你会怎么

选？

30) 几百家公司都想做 Character AI 方向，但目前进展都很一般，原因是什么？大家严重低估了 Character AI 模型的能力，绝大多数 copycat 的参数量和优化能力比 Character 相差一个数量级

31) 继续相信新摩尔定律：模型能力每 1-2 年提升一代，过程解锁新应用；模型训练成本每 18 个月除以 4，模型推理成本每 18 个月除以 10。