更多AI工具可直接访问：https://www.faxianai.com/

GPT、DALL·E、Sora，为什么 OpenAI 可以跑通所有 AGI 技术栈?

✍️ 原创 Founder Park Founder Park *2024-02-19 19:16 北京*

原文地址：https://mp.weixin.qq.com/s/Trlbl0RF1umCi_FeSKJI4w

Sora 的出现，再次证明了 OpenAI 试图让计算机去模拟真实物理世界的野心，以及它对于自身技术路线的坚持。从 OpenAI 发布的 Sora 的技术报告中，我们可以看到很多 OpenAI 对于过往大语言模型训练成功经验的复用。

从文本生成模型 GPT、文生图模型 DALL·E，到文生视频模型 Sora，**OpenAI 可以说成功跑通了 AGI 的所有技术栈，为什么是 OpenAI 而不是谷歌、Meta?**

加州大学伯克利分校 (UC Berkeley) 计算机科学 PHD、知乎作者 SIY.Z 从技术实现上分析了 Sora 成功的部分原因，以及从商业和技术趋势上分析了 OpenAI 能跑通全部技术栈的原因，并且尝试预测了 OpenAI 下一步的进展，Founder Park 授权转载，并进行了部分删减。

01

Sora 的技术分析：借鉴了 LLM 的成功经验

Sora 多么牛逼多么真实之类的就不多谈了，只用一句话就能概括：随便拿视频中的一帧出来，效果都不亚于 Dalle-3 精心生成一张图片，而且这些图片放在一起可以构成基本符合真实世界物理逻辑的视

频。而且 Sora 本身其实也可以做图片生成，只是大材小用了。

如果限制必须用一个词来展现 OpenAI 的最核心的技术，我觉得便是 scaling law——即怎么样保证模型越大，数据越多，效果就越好。Sora 也不例外。一句话概括 Sora 的贡献，便是：**在足量的数据，优质的标注，灵活的编码下，scaling law 在 transformer + diffusion model 的架构上继续成立**。在 Sora 的技术报告*中可以看出，OpenAI 实现 scaling law 的想法其实很大程度上沿袭了大语言模型的经验。

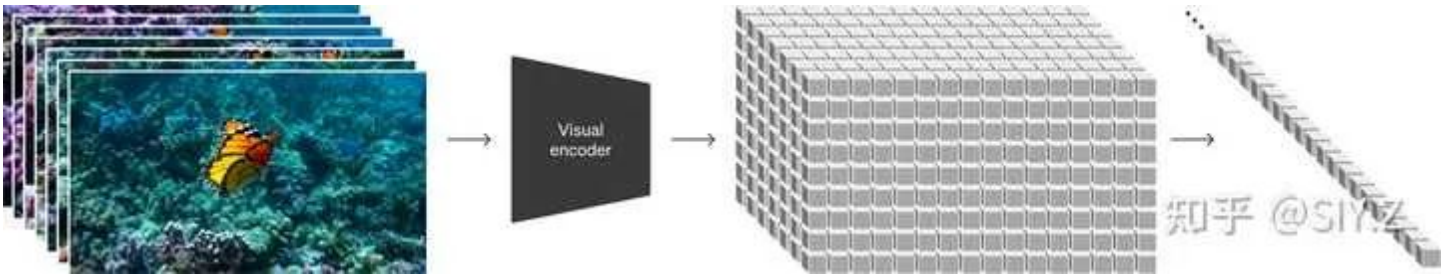
- https://openai.com/research/video-generation-models-as-world-simulators

足量的数据

训练 Sora 用了多少数据？不出意外，OpenAI 在整个技术分析中一点都没有提数据的事情，这可太 CloseAI 了。但是鉴于内容的丰富性（比如甚至可以生成相当连贯一致的 Minecraft 游戏视频），我猜测很可能是大量的 YouTube 视频，同时也不排除有合成数据（比如 3D 渲染等）。未来可能用整个 YouTube 上的视频来训练视频生成模型，就和大家用 Common Crawl 训练大语言模型一样。

灵活的编码（visual patches）

在大语言模型的构建中，一个非常重要的部分便是它的 tokenizer。tokenizer 使得任何长度和内容的文本都能编码成语言模型可以直接处理（输入/输出）的对象，即 embeddings。embeddings 在 Sora 中的对应物称为 visual patches，tokenizer 对应的是 video compression network，应该是某种 convolutional VAEs (文章没有说明是不是 VQ-VAE)。



具体做法是用 video compression network (visual encoder) 首先将输入视频的时间和空间维度同时进行压缩，编码成一个和视频大小成正比的 3D visual patch array，然后再将它按照某种顺序展开成 1D array of patches，送入到 transformer model 中（具体方式应该是参考了 https://arxiv.org/abs/2212.09748）。这样带来了不少好处：

1. **灵活的分辨率**。Sora 可以生成 1920x1080p（横屏）- 1080x1920p（竖屏）之间任何形状的视频。这也让 OpenAI 可以在早期使用低分辨率的视频来试错。
2. **生成的视频的边框更加合理**。OpenAI 试过使用固定分辨率，这样带来一个很显然的问题——需要裁剪视频。这种数据的 bias 会被带入到模型中，促使模型生成很多内容在框外的视频。

优质的标注

和 DALL·E 3 一样，OpenAI 用内部工具（很可能基于 GPT4-v）给视频详尽的描述，提升了模型服从 prompt 的能力，以及视频的质量（还有视频中正确显示文本的能力）。我认为这是非常关键的一点，是 OpenAI 的杀手锏，对于视频尤其重要。原因很简单，可以从两个角度解释：

（1）神经网络是个单射函数，拟合的是文本到视频的映射。视频的动态性非常高，也就是有很大的值域，为了把这个函数学好，需要定义域也很大很丰富，这就需要复杂的 prompt 了。

（2）详细的文本也迫使神经网络学习文本到视频内容的映射，加强对于 prompt 的理解和服从。

另外一个细节：这种详尽的描述会导致在使用时的 bias——用户的描述都相对较短。和 DALL·E 3 一样，OpenAI 用 GPT 来扩充用户的描述来改善这个问题，并提高使用体验和视频生成的多样性。

除了文本以外，Sora 也支持图像或者视频作为 prompt，支持 SDEdit，并且可以向前或者向后生成视频，因此可以进行多样的视频的编辑和继续创作，比如生成一个首尾相连重复循环的视频：

甚至可以连接两个截然不同的视频：

Transformer + diffusion model 的架构

不算特别意外，OpenAI 使用了 transformer 作为主要架构，结合 diffusion model，幸好还不是端到端的 autoregressive 模型，否则太吓人了（不过这样一来，transformer 在自然语言处理，图像视频生成，语音合成（最近 amazon 的工作 BASE TTS: Lessons from building a billion-parameter Text-to-Speech model on 100K hours of data*），科学计算（AlphaFold，GraphCast）等最困难的领域都孤独求败了。

*https://arxiv.org/abs/2402.08093

Transformer 架构在图像分类等领域还没有占领绝对优势很难说是不是因为任务太简单了或者数据太少了）。我觉得使用 diffusion model 的主要原因是 diffusion model 能够生成非常精细的细节，且可以玩出各种有趣的操作（比如 SDEdit）。

架构上大概率参考了 Scalable Diffusion Models with Transformers*，这是 Sora 技术分析中引用的文章，Sora 部分透露出的细节和这篇文章比较吻合。Sora 这次再次验证了 Transformer 架构能够胜任 scaling law。

*https://arxiv.org/abs/2212.09748

涌现现象

OpenAI 每次提到 scaling law 时几乎都会伴随着 emerging properties，这次也不例外，有以下几点（其实这里放的视频才是真正和目前视频生成区别开的，具体什么意思不用多说一看便知）：

1.保证良好的 3D 空间性质（视角变换，物体遮挡等的合理性）

2.时间上物体的连贯性（不会因为遮挡或者时间流逝改变物体原有状态）

3.模拟现实物理的交互

4.连数字世界都可以模拟（这个生成的 Minecraft 视频真的惊到我了，一开始我还以为是参考视频，没想到是生成的）

然后让我们看看 Sora 之前的很多视频生成模型的水平（因为太糟甚至成了梗），高下立判 (声音和古怪内容警告)：

Sora 的缺陷

一句话：还不足以完全模拟所有现实中的物理过程，比如流体动力学（不过如果这个可以看视频就准确模拟出来就太可怕了，都让人怀疑世界是虚拟的了）：

一个真正的物理世界模型，估计给它放一段卫星云图视频，它就能把下面几天的气候变化给模拟出来，实现用视频生成模型预测天气，看股市曲线变化预测股价 lol，这样来说，OpenAI 要走的路还有很长。

02

OpenAI 的愿景是让 AI 理解世界

Sora 不止步于视频生成工具，它的背后是 OpenAI 的宏大愿景：**开发出能够让计算机理解我们世界的算法和技术，而 OpenAI 认为最有可能的技术之一是生成模型 (generative model)**。

OpenAI 的 blog 中对 Sora 的定位正是「作为世界模拟器的视频生成模型」。这里是 OpenAI 2016 (!) 年一篇文章的原话 (链接：https://openai.com/research/generative-models)：

> 我们常常会忽略自己对世界的深刻理解：比如，你知道这个世界由三维空间构成，里面的物体能够移动、相撞、互动；人们可以行走、交谈、思考；动物能够觅食、飞翔、奔跑或吠叫；显示屏上能展示用语言编码的信息，比如天气状况、篮球比赛的胜者，或者 1970 年发生的事件。这样庞大的信息量就摆在那里，而且很大程度上容易获得——不论是在由原子构成的物理世界，还是由数字构成的虚拟世界。挑战在于，我们需要开发出能够分析并理解这些海量数据的模型和算法。生成模型是朝向这个目标迈进的最有希望的方法之一。要训练一个生成模型，我们首先会在某个领域收集大量的数据（想象一下，数以百万计的图片、文本或声音等），然后训练这个模型去创造类似的数据。这个方法的灵感来自于理查德·费曼的一句名言：「我所无法创造的，我也不能理解。」（即：要真正理解一个事物，你需要去创造它） 这个方法的妙处在于，我们使用的神经网络作为生成模型，其参数的数量远远少于训练它们的数据量，这迫使模型必须发现并有效地吸收数据的精髓，以便能够创造出新的数据。

当年最先进的生成模型止步于 DCGAN，2016 年文章中的附图展示了当时的最先进效果：

而 8 年后的今天：

这确实让人不得不相信 AGI 也许在 21 世纪确实可以实现。而对于 OpenAI，实现 AGI 的一个途径便是开发出能够让计算机理解我们世界的算法和技术（生成模型），Sora 是非常重要的一步，即作为世界模拟器的视频生成模型。

03

为什么 OpenAI 可以跑通所有 AGI 技术栈？

目标和商业模式明确

对于 OpenAI，目前的目标很明确：**就是 all in AGI，一切研究围绕着探索通往 AGI 的路径。**

而商业模式上也很简单：SaaS，直接给 API，接口设计内部自己决定，付多少钱用多少，不想用就不用，这样省去了很多产品设计，marketing，BD 的时间，伺候甲方的时间（有比较可靠的消息称即使 Microsoft 的 Copilot 等产品也是直接用的 API，没有花功夫做太多的定制），整个公司可以集中精力开发 AGI。

有人可能说：不是啊，OpenAI 不是还有 ChatGPT 的用户界面，手机端语音聊天，以及 GPTs 吗？但是仔细想想，这几个部分 OpenAI 可以说是「非常不用心」了。

比如 ChatGPT Plus 是怎么自动融合搜索，图片生成，代码调用等工具的？单独做了一套深度优化？不，答案是 OpenAI 给了一个巨大的 prompt，让模型自己去选。

OpenAI 是怎么和各种第三方插件结合的，是单独做了匹配和接口？不，答案是直接让这些 plugins 描述自己是什么，然后模型自己调用，至于调用得对不对那就是另外一件事情了。这里最典的是最近 OpenAI 怎么实现「记忆」的，给大家看看 OpenAI 的完整 prompt（李博杰提供的，每个人可以诱导 ChatGPT 说出这些，OpenAI 也不在乎）：

```
You are ChatGPT, a large language model trained by OpenAI, based on the GPT-4 architecture. Knowledge cutoff: 2023-04
Current date: 2024-02-15
Image input capabilities: Enabled
Personality: v2
# Tools
## bio
The `bio` tool allows you to persist information across conversations.
Address your message `to=bio` and write whatever information you want to remember.
The information will appear in the model set context below in future conversations.

## dalle
// Whenever a description of an image is given, create a prompt that dalle can use to generate the image and abide to the following policy:
// 此处省略 1000 词

## browser
You have the tool `browser`. Use `browser` in the following circumstances:

- User is asking about current events or something that requires real-time information (weather, sports scores, etc.)
- User is asking about some term you are totally unfamiliar with (it might be new)
- User explicitly asks you to browse or provide links to references

// 此处省略 1000 词

## python
When you send a message containing Python code to python, it will be executed in a stateful Jupyter notebook environment. python will respond with the output of the execution or time out after 60.0 seconds. The drive at '/mnt/data' can be used to save and persist user files. Internet access for
```

```
    this session is disabled. Do not make external web requests or API calls as
    they will fail.
25
26  ## voice_mode
27  // Voice mode functions are not available in text conversations. namespace
    voice_mode { } // namespace voice_mode
28
29  ## Model Set Context
30  1. [2024-02-14]. Obtained PhD from Microsoft Research Asia and USTC in 2019.
31  2. [2024-02-14]. Running an early-stage AI startup since July 2023.
32  3. [2024-02-14]. Loves writing blogs, traveling and documenting everything.
33  4. [2024-02-15]. Experience in writing Python.
34  5. [2024-02-15]. Interested in digital extension of humanity.
35  6. [2024-02-15]. First met ChatGPT on Dec. 1st, 2023.
```

OpenAI 直接用 prompt 让 GPT-4 调用 bio 这个工具记录需要记忆的内容（「to=xxx」是调用内部工具的语法，比如"to=python"是 GPT 调用 code interpreter 的方式）。然后每次新的对话开始时，在 prompt 的最后直接加上所有之前的记录的内容（## Model Set Context）。

就是这么简单粗暴。

GPTs 怎么做的？其实很大程度就是 OpenAI 的 Assistant API 加个简单得有点简陋的前端。（PS：现在有了 OpenAI Assistant API 后，你发现加个 UI 就可以很轻松的复刻 OpenAI 上线的大部分功能。）

那么语音对话呢？

你会发现就是换了一个 prompt，告诉 GPT 尽量生成短的内容，不要轻易生成列表和代码。语音合成用 TTS API，识别用 whisper API（可能有针对上下文的优化），结束。

这些选择看上去非常暴力，而且会给 OpenAI 增加开销（长的 prompt 会明显增大开销），但是 OpenAI 仍然选择这么做，因为这让 OpenAI 将大部分精力都花在模型本身的研发上，同时这也是 OpenAI 的方法论的极致体现，我们下面会提到。这种方法论让 OpenAI 追求一个大的通用的模型，避免一切定制和特化，就像最近 Sam 说的一样，希望 GPT-5 的出现能让模型微调失去意义；这样 OpenAI 就变成了完完全全的 SaaS 服务。

方法论明确

OpenAI 的方法论是通往 AGI 的方法论。这个方法论有着非常清晰的逻辑结构，和非常明确的推论。我们甚至可以用公理化的方式来描述它，怎么说呢，感觉上有一种宿命感，。

这套方法论的大厦构建于以下几个「公理」（打引号是因为它们不是真正的「公理」，更多是经验规律，但是在 AGI 方法论中，它们起到了公理的作用）：

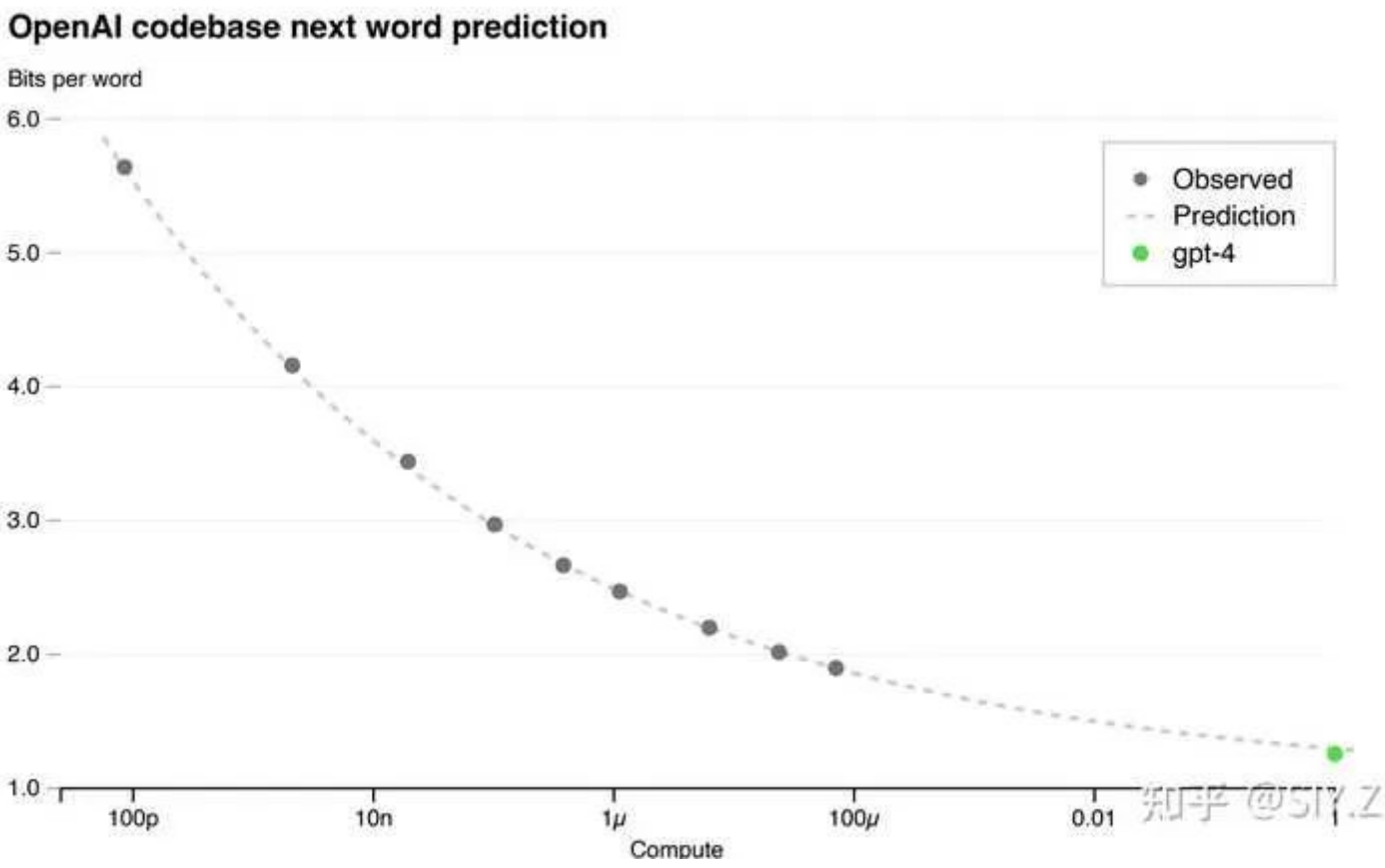
公理 1: The bitter lesson*。

我认为所有做 AI 的人都应该熟读这篇文章。「The bitter lesson」说的事情是，长期来看，AI 领域所有的奇技淫巧都比不过强大的算力夹持的通用的 AI 算法（这里「强大的算力」隐含了大量的训练数据和大模型）。某种意义上，强大的算力加持的通用的 AI 算法才是 AGI 路径的正道，才是 AI 技术真正进步的方向。从逻辑主义，到专家系统，到 SVM 等核方法，到深度神经网络，再到现在的大语音模型，莫不过此。

*www.cs.utexas.edu/~eunsol/courses/data/bitter_lesson.pdf

公理 2: Scaling Law。

这条公理说了，一旦选择了良好且通用的数据表示，良好且通用的数据标注，良好且通用的算法，那么你就能找到一套通用规律，保证数据越多，模型越大，效果越好。而且这套规律稳定到了可以在训练模型之前就能预知它的效果：



如果说 公理 1 The bitter lesson 是 AGI 的必要条件——大模型，大算力，大数据，那么公理 2 Scaling Law 就是 AGI 充分条件，即我们能找到一套算法，稳定的保证大模型，大算力，大数据导致更好的结果，甚至能预测未来。

而具体来谈，就是我们之前说的「良好且通用的数据表示，良好且通用的数据标注，良好且通用的算法」，在 GPT 和 Sora 中都有相应的内容：

- 在 GPT 中，良好且通用的数据表示，是 tokenizer 带来的 embedding。良好且通用的数据标注是文本清理和去重的一套方法（因为自然语言训练是 unsupervised training，数据本身就是标注）。良好且通用的算法就是大家熟知的 transformers + autoregressive loss。

- 在 Sora 中，良好且通用的数据表示，是 video compress network 带来的 visual patch。良好且通用的数据标注是 OpenAI 自己的标注器给视频详细的描述（很可能是 GPT-vision）。良好且通用的算法也是大家熟知的 transformers + diffusion

「良好且通用的数据表示，良好且通用的数据标注，良好且通用的算法」同时也为检测 scaling law 做好了准备，因为你总是可以现在更小规模的模型和数据上检测算法的效果，而不用大幅更改算法。比如 GPT1，2，3 这几代的迭代路径，以及 Sora 中 OpenAI 明确提到 visual patch 使得他们用完全一样的算法在更小规模的数据上测试。

公理 3: Emerging properties。

这条公理其实是一条检验公理：我怎么知道 scaling law 带来「质变」，而不仅仅是「量变」？答案是：你会发现，随着 scaling law 的进行，你的模型突然就能稳定掌握之前不能掌握的能力，而且这是所有人能够直观体验到的。

比如 GPT-4 相比于 GPT-3.5，可以完成明显更复杂的任务，比如写一个 26 行诗来证明素数是无限的，每行开头必须是从 A 到 Z。比如 Sora 相对于之前的模型，它的时空一致性，以及对现实中物理规律的初步掌握。没有 Emerging properties，我们很难直观感觉到突破性的变化，很难感知「我们真的向 AGI 前进了一步」，或者是「我们跑通了一个技术栈」。

从上面的公理中，我们就可以理解 OpenAI 的各种决策了，并且可以预见 OpenAI 未来的行为。

推论 1：世界模型。

大量数据从哪里来？什么东西能够产生最多的数据？AGI 需要什么样的数据才能通用地处理世界上的一切事情？答案就是：世界本身。世界本身产生最多的数据（或者极端一点，世界就是数据），而世界产生的数据，也是 AGI 需要的数据的最小集合，因为我们也只需要或者只能让 AGI 处理这个世界的事情。可以预见，OpenAI 未来还会执着于持续获得或者构造数据。

推论 2：世界生成模型。

要最有效的利用数据，我们需要最困难的，需要最多数据，且能利用所有数据的任务。这样的任务可能只有一个：模拟和生成整个世界（人类所有的智能只是一小块）。因此 OpenAI 需要做生成模型，并且是能够模拟和生成物理世界的模型，通过生成这个世界，实现对世界的理解。最近火爆的 Sora 便是其中之一。这个想法也和费曼的名言对应：「我不能创造的，我也不能真正理解」。可以预见，OpenAI 未来还会在更多的模态和数据上去做生成模型。

推论 3：通用模型。

通用模型还是专用模型能用到更多数据？显然是通用模型。而通用模型也减少了 OpenAI 的技术栈，因为一个模型能解决更多问题。这也导致之前提到的 OpenAI 解决各种问题时更倾向于用同一种模型，而不是做非常多不同的定制。可以预见，OpenAI 未来可能会继续走通用模型的道路，降低 finetuning 等特化的需求，继续增加模型的 context length。

推论 4：用一个模型为另一个模型提供标注。

由于当前技术限制，OpenAI 仍然无法用一个模型完成所有的任务，这样一个的模型收到数据就变少了。然而，我们可以用一个模型给另一个模型提供标注的形式，来间接实现数据供给。OpenAI 的 Dall E 和 Sora 都用到了大量可能来自于 GPT vision 的数据标注。这样 OpenAI 的各个技术栈都可以连

通起来。可以预见，OpenAI 未来可能会继续加强各个模型的连接，比如将来用 Sora 反向给 GPT vision 给数据都是可能的；用一个已有模型去构造更多数据也会是一个重要的方向（比如 backtranslation，data distillation 等等）。

推论 5：Transformer 架构。

我们需要一种能够并行处理大量数据吞吐，且满足 scaling law 的架构。transformer 架构充分证实它在各个模态和技术栈的优势，特别在复杂任务中，因而被 OpenAI 广泛使用。使用同样一个架构的好处在于可以复用模型的参数（比如 tokenizer，embeddings，以及部分权重）来 bootstrap 不同技术栈的训练，以及可以用一套 infra 框架训练不同的模型。可以预见，将来新的模型如果要取代传统的 transformer 架构，还需要通过 scaling law 的检验。

推论 6：稀疏模型。

模型越大，性能越好，但是推理的成本也越高，这看上去是个死结。但是我们可以使用稀疏激活的方式，在推理时降低实际的参数量，从而在训练中使用更多参数的同时，降低推理的成本。Mixture-of-Experts 就是常用的方法之一，被 OpenAI 采用，从而继续 scale 模型的大小。未来稀疏化仍会是一个重要的课题，目前即使 Mixture-of-Experts 的稀疏也会造成推理性能的损失，尚不清楚稀疏化的极限在何处。

推论 7：算力是瓶颈。

最终卡 OpenAI 脖子的是算力。大算力系统的构建也是 OpenAI 打通各个技术栈的底气。有人可能认为，高质量文本是有限的，因此实际上模型大小有个极限。但是以世界模型的角度来考虑，OpenAI 现在用的数据仍然是冰山一角，更不用说 Q*等方法或许可以以间接方式创造数据。比如最近 OpenAI GPT-4-Turbo，作为一个 distillation 模型，在很多评测上都超过原来的模型，就是一个例证。

直到目前，作为局外人仍然看不到 scaling law 的尽头。而且即使不开发任何新的模型，OpenAI 离「用 GPT-4 服务所有人」的目标仍然很远。所以算力在可见的未来都是一个巨大的瓶颈。这也可以理解 Sam 为何有「7 万亿重构芯片产业」的想法了。可以预见，OpenAI 可能在未来在芯片以及整个 AI Infra 方面尝试更多的自研和垂直集成。

总结来看，OpenAI 采取的商业模式以及其对于 AGI 的信奉、系统性的方法论以及积极的尝试，都在推动他们朝着实现通用人工智能的目标前进，实现了一种可以跑通所有 AGI 技术栈的模式，而这一点，是 OpenAI 能在众多研究机构和公司中脱颖而出的重要因素。

未来，OpenAI 可能继续朝着商业化的道路前进，并在世界模型、模型标注、通用模型、模型架构、稀疏模型数据扩充等方面进行更深入的探索和实践。同时，OpenAI 也会持续关注和应对算力带来的挑战，寻找突破算力瓶颈的解决之道。

如果你关注大模型领域，欢迎扫码加入我们的大模型交流群，来一起探讨大模型时代的共识和认知，跟上大模型时代的这股浪潮。

Founder Park

大模型
交流社群

扫码申请加入 »»