利奥波德·阿申布伦纳

**局势感知**

——未来十年

献给伊利亚·苏茨凯弗。

虽然我曾在OpenAI工作过，但本文都是基于公开可用的信息、我自己的想法、一般领域知识或旧金山的传闻。

感谢 Collin Burns、Avital Balwit、Carl Shulman、Jan Leike、Ilya Sutskever、Holden Karnofsky、Sholto Douglas、James Bradbury、Dwarkesh Patel 和许多其他人的建设性讨论。感谢许多朋友对早期草稿的反馈。感谢 Joe Ronan 提供图形帮助，以及 Nick Whitaker 提供出版帮助。

situational-awareness.ai

leopold@situational-awareness.ai

更新于 2024 年 6 月 6 日

加利福尼亚州旧金山

你可以在旧金山最先看到未来。

在过去的一年里，这座城市的讨论从100亿美元的计算集群转向1000亿美元的集群，再到万亿美元的集群。每六个月，董事会计划中都会增加一个零。在幕后，有一场激烈的争夺，试图在未来十年内确保每一份仍可用的电力合同，以及每一个可能获得的电力变压器。美国的大企业正在准备投入数万亿美元，进行一场久违的美国工业力量动员。到本十年末，美国的电力生产将增长数十个百分点；从宾夕法尼亚州的页岩气田到内华达州的太阳能农场，数亿个GPU将嗡嗡作响。

通用人工智能（AGI）的竞赛已经开始。我们正在构建能够思考和推理的机器。到2025/26年，这些机器将超过大学毕业生的能力。到本十年末，它们将比你或我更聪明；我们将拥有真正意义上的超级智能。在此过程中，半个世纪未见的国家安全力量将被释放，不久之后，这个项目将启动。如果我们幸运，我们将与中国进行全面竞赛；如果不幸，将是一场全面战争。

现在每个人都在谈论人工智能，但很少有人意识到即将来临的冲击。Nvidia的分析师仍然认为2024年可能接近巅峰。主流评论家们还停留在“这只是预测下一个词”的故意盲目中。他们只看到炒作和业务如常；最多他们考虑的是另一场互联网规模的技术变革。

不久之后，世界将会醒来。但现在，也许只有几百人，其中大部分在旧金山和AI实验室，拥有局势感知。通过命运的奇特力量，我发现自己身处其中。几年前，这些人被嘲笑为疯子——但他们信任趋势线，这使他们能够正确预测过去几年的AI进展。未来几年这些人是否也能正确预测，还有待观察。但这些人非常聪明——是我见过的最聪明的人——他们正在构建这项技术。也许他们将成为历史中的一个奇怪脚注，或者他们将像西拉德、奥本海默和泰勒那样载入史册。如果他们对未来的预见接近正确，我们将迎来一段狂野的旅程。

让我告诉你我们看到了什么。

目录

**引言** …………………………………………………………………………………3

旧金山的历史正在上演。

**I. 从 GPT-4 到 AGI：计算数量级的变化**……………………………………… 7

2027 年实现 AGI 的可能性非常高。从 GPT-2 到 GPT-4，我们在 4 年内从学前儿童能力提升到了聪明的高中生能力。追踪计算能力（~每年 0.5 个数量级或 OOMs）、算法效率（~每年 0.5 个 OOMs）和“解锁”收益（从聊天机器人到智能体）的趋势线，到 2027 年我们应该期待另一个从学前儿童到高中生级别的质变飞跃。

**II. 从 AGI 到超级智能：智能爆炸**…………………………………………… **46**

AI 的进展不会止步于人类水平。数以百万计的 AGI 可以自动化 AI 研究，将十年的算法进步（5+ OOMs）压缩到一年内。我们将迅速从人类水平迈向远远超越人类的 AI 系统。超级智能的力量和危险将是显著的。

**III. 挑战**………………………………………………………………………… **74**

**IIIa. 争夺万亿美元集群**…………………………………………………………**75**

最非凡的技术资本加速已经启动。随着 AI 收入的快速增长，许多万亿美元将在十年末之前用于 GPU、数据中心和电力的建设。包括美国电力生产增长数十个百分点在内的工业动员将是激烈的。

**IIIb. 锁定实验室：AGI 的安全性**…………………………………………… **89**

该国领先的 AI 实验室将安全性视为事后考虑。目前，他们基本上是在把 AGI 的关键秘密送到中国的手上。保护 AGI 的秘密和权重免受国家行为者的威胁将是一项巨大努力，而我们还没有走上正轨。

**IIIc. 超级对齐**………………………………………………………………… **105**

可靠地控制比我们聪明得多的 AI 系统是一个未解决的技术问题。虽然这是一个可以解决的问题，但在快速的智能爆炸过程中，很容易出现偏差。管理这一过程将非常紧张；失败很容易导致灾难性后果。

**IIId. 自由世界必须胜利**……………………………………………………… **126**

超级智能将带来决定性的经济和军事优势。中国并没有完全退出竞争。在 AGI 的竞赛中，自由世界的生存岌岌可危。我们能否保持对专制势力的优势？我们能否在此过程中避免自我毁灭？

**IV. 项目**………………………………………………………………………… **141**

随着 AGI 竞赛的加剧，国家安全机构将介入。美国政府将从沉睡中苏醒，到 27/28 年，我们将看到某种形式的政府 AGI 项目。没有一家初创公司能够应对超级智能。在某个机密信息设施中，终局将开始。

**V. 临别感言**…………………………………………………………………… **156**

如果我们是对的，会怎样？

**附录**…………………………………………………………………………… **162**

1. **从 GPT-4 到 AGI：计算数量级的变化**

**到 2027 年实现 AGI 的可能性非常高。从 GPT-2 到 GPT-4 的发展，使我们在 4 年内从学前儿童能力提升到了聪明的高中生能力。追踪计算能力（~每年 0.5 个数量级或 OOMs）、算法效率（~每年 0.5 个 OOMs）和“解锁”收益（从聊天机器人到智能体）的趋势线，到 2027 年我们应该期待另一个从学前儿童到高中生级别的质变飞跃。**

看看。这些模型，它们只想学习。你必须明白这一点。这些模型，它们只想学习。

———————————————————————————————————————伊利亚·苏茨凯弗

（大约 2015 年，通过 Dario Amodei）

GPT-4 的能力让许多人震惊：一个能够编写代码和文章，能够推理复杂数学问题并通过大学考试的 AI 系统。几年前，大多数人认为这些是不可逾越的障碍。

但 GPT-4 只是十年来深度学习快速进步的延续。十年前，模型几乎无法识别简单的猫狗图像；四年前，GPT-2 刚刚能够拼凑出看似合理的句子。现在我们正在快速饱和我们能想到的所有基准。然而，这种显著进展仅仅是深度学习规模化趋势的结果。

有些人早已看到了这一点。他们曾被嘲笑，但他们所做的只是信任趋势线。

趋势线非常明确，而且他们是对的。这些模型，它们只想学习；你扩大它们的规模，它们就会学到更多。我提出以下主张：到2027年，模型能够完成AI研究员/工程师的工作是非常有可能的。这不需要相信科幻小说；只需要相信图表上的直线。

在这篇文章中，我将简单地“计算 OOMs”（OOM = 数量级，每增加一个数量级等于 10 倍）：看看以下趋势：1）计算能力，2）算法效率（我们可以将其视为增加“有效计算能力”的算法进步），3）“解锁”收益（修复模型默认受限的明显方式，解锁潜在能力并赋予它们工具，从而带来有用性的飞跃）。我们追踪 GPT-4 之前四年的每一项增长，以及到 2027 年底的接下来四年我们应该期待的变化。

鉴于深度学习在每个有效计算能力数量级上的一致改进，我们可以用此来预测未来的进展。

自 GPT-4 发布以来，公开信息已经沉寂了一年，因为下一代模型正在酝酿中——导致一些人声称停滞，并认为深度学习正撞上壁垒。但通过计算 OOMs，我们可以窥见我们实际应该期待什么。他们每年都会做出预测，过去十年里他们的预测一直都是错的。结果其实很简单。从 GPT-2 到 GPT-4——从有时能拼凑出几句连贯句子的模型，到能通过高中考试的模型——并不是一次性的提升。我们正在非常迅速地经历数量级的跃升，数据表明我们应该期待在四年内有效计算能力再增加约 100,000 倍——这将带来另一个 GPT-2 到 GPT-4 规模的质变飞跃。而且，这不仅意味着一个更好的聊天机器人；通过“解锁”收益的许多明显的低垂果实，我们应该从聊天机器人转向智能代理，从工具变成更像是远程工作的替代品。

虽然推论很简单，但其含义却非常惊人。再一次这样的飞跃很可能会带我们到 AGI，达到像博士或专家一样聪明的模型，可以作为我们的同事一起工作。也许最重要的是，如果这些 AI 系统可以自动化 AI 研究本身，那将启动强烈的反馈循环——这是本系列下一篇文章的主题。

即使现在，几乎没有人将这些考虑在内。但一旦你退一步看看这些趋势，AI 的局势感知其实并不难。如果你总是对 AI 的能力感到惊讶，只需开始计算 OOMs。

过去四年

我们现在有了可以像人类一样与之交谈的机器。这是人类适应能力的一个非凡证明，这似乎变得正常了，我们已经习惯了进步的速度。但值得退一步看看过去几年的进展。

从 GPT-2 到 GPT-4

让我提醒你在仅仅大约四（！）年内我们取得了多大的进步。

GPT-2（2019）~ 学前儿童：“哇，它可以拼凑出几句看似合理的句子。”它生成的一篇关于安第斯山脉独角兽的半连贯故事的非常精选的例子在当时令人印象深刻。然而，GPT-2 几乎无法数到 5 而不出错；在总结一篇文章时，它只是稍微优于从文章中随机选择 3 个句子。

比较 AI 能力和人类智力是困难且有缺陷的，但我认为这里的类比是有启发性的，即使它极不完美。GPT-2 因其语言掌握能力和偶尔生成一个半连贯段落或偶尔正确回答简单的事实问题而令人震惊。这是对学前儿童来说令人印象深刻的表现。

GPT-3（2020）~ 小学生：“哇，几个少量的示例就能完成一些简单有用的任务。”它在多段文字中保持连贯性更加一致，并且可以纠正语法并进行一些非常基础的算术。第一次，它在一些狭窄的方面具有商业用途：例如，GPT-3 可以生成简单的 SEO 和营销文案。

再次强调，这种比较是不完美的，但 GPT-3 令人印象深刻的地方可能是对小学生来说令人印象深刻的表现：它写了一些基础的诗歌，可以讲述更丰富连贯的故事，可以开始做一些初步的编程，可以相当可靠地从简单的指示和示范中学习，等等。

GPT-4（2023）~ 聪明的高中生：“哇，它可以编写相当复杂的代码并进行迭代调试，它可以智能且复杂地撰写关于复杂主题的文章，它可以推理通过困难的高中竞赛数学题，它在我们能给它的任何测试中击败了绝大多数高中生，等等。”从代码到数学再到费米估算，它可以思考和推理。GPT-4 现在在我的日常任务中非常有用，从帮助编写代码到修改草稿。

在所有从 AP 考试到 SAT 的考试中，GPT-4 的得分都比绝大多数高中生高。当然，即便是 GPT-4 仍然有些不均衡；在某些任务上它远远超过聪明的高中生，而在其他任务上它还不能胜任。尽管如此，我倾向于认为这些限制主要是由于模型仍然受到明显方式的约束，正如我稍后将详细讨论的那样。即使模型仍然受到人为限制，但其原始智能（大部分）已经存在；需要额外的工作来解锁模型，以便能够在各个应用中充分发挥这种原始智能。

深度学习的趋势

在过去十年中，深度学习进展的速度简直是惊人的。仅仅十年前，深度学习系统能够识别简单图像就已经是革命性的成就。如今，我们不断尝试提出新的、更难的测试，但每个新的基准很快就被攻破。过去需要几十年才能攻破广泛使用的基准；现在感觉只需几个月。

我们的基准测试几乎用尽了。举个例子，我的朋友 Dan 和 Collin 在 2020 年做了一个叫 MMLU 的基准测试。他们希望最终能够创建一个经得起时间考验的基准，相当于我们给高中生和大学生的所有最难的考试。仅仅三年后，这个基准基本上被攻克了：像 GPT-4 和 Gemini 这样的模型得分约为 90%。

更广泛地说，GPT-4 基本上攻克了所有标准的高中和大学能力测试（图 7）。

或者考虑 MATH 基准测试，这是一组来自高中数学竞赛的难题。该基准测试在 2021 年发布时，GPT-3 只解决了大约 5% 的问题。原始论文指出：“此外，我们发现，如果扩展趋势继续，仅仅增加预算和模型参数数量对于实现强大的数学推理是不现实的。要在数学问题解决方面取得更大进展，我们可能需要来自更广泛研究社区的新算法进展。”我们需要根本性的新的突破来解决 MATH，或者他们是这样认为的。一项对机器学习研究人员的调查预测未来几年会有最小的进展；然而，在仅仅一年内（到 2022 年中），最好的模型从大约 5% 提高到 50% 的准确率；现在，MATH 基本上已经被解决，最新的性能超过 90%。

一遍又一遍，年复一年，怀疑论者声称“深度学习将无法做到 X”，但很快就被证明是错误的。如果我们从过去十年的 AI 发展中学到了一件事，那就是你绝不应该低估深度学习。

现在最难解决的基准测试是类似 GPQA 这样的测试，一组博士水平的生物学、化学和物理学问题。许多问题对我来说像是胡言乱语，甚至其他科学领域的博士花 30 多分钟用 Google 搜索也仅仅得分略高于随机猜测。Claude 3 Opus 目前的得分约为 60%，而领域内的博士得分约为 80%——我预计这一基准测试在未来一两代模型中也会被攻克。

计算数量级

这是怎么发生的？深度学习的魔力在于它确实有效——尽管每一步都有反对者，趋势线却一直令人惊讶地保持一致。

随着每一个有效计算的数量级增加，模型的性能可以预测地、可靠地变得更好。如果我们能计算 OOMs，我们就可以（大致、定性地）推断出能力的提升。这就是少数有先见之明的人看到 GPT-4 即将到来的方式。

我们可以将 GPT-2 到 GPT-4 之间四年的进展分解为三类扩展：

1. 计算能力：我们使用更大的计算机来训练这些模型。

2. 算法效率：算法进步有一个持续的趋势。许多这些进步起到了“计算乘数”的作用，我们可以将它们放在一个增长的有效计算的统一尺度上。

3. “解锁”收益：默认情况下，模型学习了许多惊人的原始能力，但它们在很多愚蠢的方式中受到限制，限制了它们的实际价值。通过简单的算法改进，如从人类反馈中进行强化学习（RLHF）、思维链（CoT）、工具和支架，我们可以解锁显著的潜在能力。

我们可以沿着这些轴线“计算 OOMs”的改进：也就是说，追踪每一个扩展的有效计算单位。3 倍是 0.5 个 OOMs；10 倍是 1 个 OOM；30 倍是 1.5 个 OOMs；100 倍是 2 个 OOMs；依此类推。我们还可以看看从 2023 年到 2027 年在 GPT-4 之上的预期。

我将逐一解释，但结论很明确：我们正在迅速地通过 OOMs 前进。在数据壁垒中可能会遇到阻力，我会讨论这个问题——但总体而言，到 2027 年，我们很可能会期待另一个类似于 GPT-2 到 GPT-4 的跃升，在 GPT-4 的基础上。

计算能力

我将从最近进展中讨论最多的驱动因素开始：向模型投入（大量）计算能力。

许多人认为这只是因为摩尔定律。但即使在摩尔定律的鼎盛时期，其速度也相对缓慢——每十年大约 1 到 1.5 个 OOMs。我们看到计算能力的扩展速度要快得多——接近摩尔定律速度的 5 倍——这是因为巨大的投资。（过去在一个模型上花费一百万美元是一个没有人会考虑的荒谬想法，而现在那只是小菜一碟！）

我们可以使用来自 Epoch AI（一个因其对 AI 趋势的卓越分析而广受尊敬的来源）的公共估计来追踪 2019 年到 2023 年的计算能力扩展。GPT-2 到 GPT-3 是一次快速的扩展；计算能力从一个较小的实验扩展到使用整个数据中心来训练大型语言模型。随着 GPT-3 到 GPT-4 的扩展，我们过渡到现代模式：必须为下一个模型构建一个全新的（更大）集群。然而，戏剧性的增长仍在继续。总体而言，Epoch AI 的估计表明，GPT-4 的训练使用了比 GPT-2 多大约 3,000 倍到 10,000 倍的原始计算能力。

从大体上看，这只是一个长期趋势的延续。在过去的十五年里，主要由于广泛的投资扩展（以及以 GPU 和 TPU 形式为 AI 工作负载专门设计的芯片），前沿 AI 系统的训练计算能力以大约每年 0.5 个 OOMs 的速度增长。

从 GPT-2 到 GPT-3 的一年内计算能力的扩展是一个不寻常的现象，但所有迹象表明，长期趋势将继续。旧金山的谣言充满了关于巨大 GPU 订单的戏剧性故事。所涉及的投资将是非凡的——但它们正在进行中。我将在本系列的 IIIa. 章节“争夺万亿美元集群”中进一步讨论这一点；基于该分析，到 2027 年底，额外增加 2 个 OOMs 的计算能力（价值数百亿美元的集群）似乎非常有可能发生；甚至更接近 +3 OOMs 的计算能力集群（价值超过 1000 亿美元）也似乎是可能的（据传正在微软/OpenAI 的计划中）。

算法效率

虽然计算能力的巨额投资吸引了所有的注意力，但算法进展可能是同样重要的进步驱动因素（并且一直被大大低估）。

为了了解算法进步的重大影响，请考虑以下插图（图 12），展示了在两年内达到 MATH 基准测试（高中数学竞赛数学）大约 50% 准确率的价格下降情况。（相比之下，一个不特别喜欢数学的计算机科学博士生得分为 40%，所以这已经相当不错了。）推理效率在不到两年的时间里提高了近 3 个 OOMs，即 1,000 倍。

尽管这些数字只是针对推理效率的（可能与训练效率的改进相对应，也可能不相对应，因为从公开数据中推断这些数字更难），它们清楚地表明，算法进步有着巨大的潜力，并且正在发生。

在这篇文章中，我将算法进步分为两种类型。首先，我将讨论“范式内”的算法改进——这些改进仅仅是带来了更好的基础模型，直接起到计算效率或计算乘数的作用。例如，一个更好的算法可能让我们在训练计算量减少 10 倍的情况下实现相同的性能。反过来，这相当于有效计算量增加 10 倍（1 个 OOM）。（稍后，我将讨论“解锁”，你可以将其视为“扩展范式/扩展应用”的算法进步，解锁基础模型的能力。）

如果我们退一步看看长期趋势，我们似乎以相当一致的速度发现新的算法改进。个别发现看起来是随机的，在每个转折点上似乎都有不可逾越的障碍——但长期趋势是可预测的，在图表上是一条直线。相信趋势线。

我们有关于 ImageNet 的最佳数据（算法研究大多是公开的，我们有数据可追溯到十年前），在 2012 年到 2021 年的 9 年期间，我们一致地提高了大约每年 0.5 个 OOM 的计算效率。

这非常重要：这意味着四年后，我们可以用大约 100 倍更少的计算量实现相同水平的性能（并且对于相同的计算量，性能会显著提高！）。不幸的是，由于实验室不公开这方面的内部数据，过去四年中前沿大型语言模型（LLM）的算法进步难以测量。EpochAI 有新的研究，复制了他们在 ImageNet 上的结果用于语言建模，并估计从 2012 年到 2023 年，LLM 的算法效率趋势大约为每年 0.5 个 OOMs。（不过，这些估计的误差范围更大，而且未能捕捉到一些最近的进展，因为领先的实验室已经停止发布他们的算法效率数据。）

更直接地来看过去四年，GPT-2 到 GPT-3 基本上是一个简单的扩展（根据论文），但自 GPT-3 以来，已有许多公开已知和可推断的进展：

• 我们可以从 API 成本中推断出收益：

• GPT-4 在发布时的成本与 GPT-3 发布时大致相同，尽管性能有着绝对巨大的提升。（如果我们根据扩展定律进行一个简单和过于简化的粗略估算，这表明从 GPT-3 到 GPT-4 的有效计算能力增加中大约有一半来自算法改进。）

• 自从一年前 GPT-4 发布以来，OpenAI 的 GPT-4 级别模型的价格在发布 GPT-4.0 时又下降了 6 倍/4 倍（输入/输出）。

• 最近发布的 Gemini 1.5 Flash 提供了介于“GPT-3.75 级别”和 GPT-4 级别的性能，同时成本比原始 GPT-4 低 85 倍/57 倍（输入/输出）（非凡的收益！）。

• Chinchilla 扩展定律带来了 3 倍以上（0.5 OOMs 以上）的效率提升。

• Gemini 1.5 Pro 声称取得了重大的计算效率提升（在使用“显著更少”的计算量的情况下，性能超过了 Gemini 1.0 Ultra），其中专家混合（MoE）是一个突出的架构变化。其他论文也声称 MoE 带来了计算量的显著增加。

• 在架构、数据、训练堆栈等方面一直都有许多调整和改进。

综合来看，公开信息表明，从 GPT-2 到 GPT-4 的跳跃包含了 1-2 个 OOMs 的算法效率提升。

在 GPT-4 之后的四年里，我们应该预期这一趋势将继续：计算效率平均每年约 0.5 个 OOMs，即到 2027 年相比于 GPT-4 增加约 2 个 OOMs 的收益。随着我们采摘低垂的果实，计算效率将变得越来越难以发现，但 AI 实验室在资金和人才上的投资以找到新的算法改进正在迅速增长。（至少从公开可推断的推理成本效率来看，似乎并没有任何放缓的迹象。）在高端方面，我们甚至可能会看到更多基础性的、类似 Transformer 的突破，带来更大的收益。

综合来看，这表明到 2027 年，我们应该预期算法效率将增加 1-3 个 OOMs（相较于 GPT-4），或许最佳估计是增加约 2 个 OOMs。

数据瓶颈

这其中有一个潜在的重要变量：我们正在耗尽互联网数据。这可能意味着，很快，使用更多抓取的数据来预训练更大的语言模型的简单方法可能会开始遇到严重的瓶颈。

前沿模型已经训练了大量的互联网数据。例如，Llama 3 就是用超过 15 万亿个标记进行训练的。Common Crawl 是一个包含大量互联网内容的数据集，原始数据量超过 100 万亿个标记，尽管其中大部分是垃圾邮件和重复内容（例如，相对简单的去重处理后，剩下 30 万亿个标记，这意味着 Llama 3 已经基本上使用了所有的数据）。此外，对于代码等更特定的领域，标记数量更少，例如公开的 GitHub 仓库估计只有低万亿级的标记。

通过重复数据可以稍微进一步扩展，但学术研究表明，重复只能带来有限的收益，发现经过 16 轮（16 倍重复）后，收益会急剧下降至无效。某种程度上，即使有更多（有效的）计算能力，由于数据限制，提升模型性能也会变得更加困难。

这不可小觑：我们一直在沿着扩展曲线前进，乘着语言建模-预训练范式的浪潮，如果没有新的突破，这一范式（至少在表面上）将会耗尽。尽管有大量的投资，我们也会达到瓶颈。

所有的实验室据传都在押注新的算法改进或方法，以绕过这一问题。研究人员据说正在尝试许多策略，从合成数据到自对弈和强化学习方法。业内人士似乎非常乐观：Anthropic 的 CEO Dario Amodei 最近在一档播客节目中说：“如果你非常天真地看待这个问题，我们离耗尽数据并不远……我的猜测是这不会成为阻碍……有很多不同的方法可以解决这个问题。”当然，任何关于这一问题的研究结果都是专有的，最近也没有公开发表。

除了业内人士的乐观态度外，我认为还有一个很强的直觉理由说明找到方法用更好的样本效率来训练模型是可能的（算法改进可以让它们从有限的数据中学到更多）。想想你或我如何从一本非常密集的数学教科书中学习：

• 现代 LLM 在训练期间所做的，基本上是非常快速地浏览教科书，文字飞快地闪过，并没有花费太多脑力。

• 相反，当你或我阅读那本数学教科书时，我们会慢慢地读几页；然后在脑海中进行内部独白，并与几个学习伙伴讨论；再读一两页；然后尝试一些练习题，失败，再以不同的方式尝试，获取这些问题的反馈，直到我们做对问题；如此反复，直到最终材料“点击”。

• 如果我们只能像 LLM 那样快速浏览，我们从一本密集的数学教科书中几乎学不到什么。

• 或许，有一些方法可以结合人类消化密集数学教科书的某些方面，使模型能够从有限的数据中学到更多东西。

简化来说，这种事情——对材料进行内部独白，与学习伙伴讨论，尝试并失败直到成功——是许多合成数据/自对弈/强化学习方法试图做的。

以前训练模型的状态是简单而天真的，但它有效，所以没人真的努力破解这些样本效率方法。现在它可能会成为更多的限制，我们应该预期所有实验室会投入数十亿美元和他们最聪明的头脑来破解它。深度学习的一个常见模式是，需要大量努力（和许多失败的项目）才能把细节做对，但最终某种明显和简单的方法就会奏效。鉴于深度学习在过去十年中已经成功地突破了每一个所谓的壁垒，我的基本假设是这里也会类似。

此外，破解其中一个算法赌注（如合成数据）实际上可能会显著改进模型。这里有一个直觉上的解释。当前的前沿模型如 Llama 3 是在互联网上训练的——而互联网大部分都是垃圾，如电子商务或 SEO 等。许多大型语言模型（LLMs）在训练计算中大多数时间都花在这些垃圾数据上，而不是在真正高质量的数据上（例如，处理困难科学问题的推理链）。想象一下，如果你能在完全极高质量的数据上使用 GPT-4 级别的计算能力——这可能会是一个更强大的模型。

回顾一下 AlphaGo——这是第一个击败围棋世界冠军的 AI 系统，比人们预想的要早几十年——在这里也很有用。

• 在第1步，AlphaGo 通过模仿学习对专家的围棋对局进行训练。这为它奠定了基础。

• 在第2步，AlphaGo 自我对弈了数百万局。这让它在围棋上达到了超人水平：记住在对阵李世石的比赛中著名的第37手，这是一个人类永远不会下的极其不寻常但又极其聪明的一手。

为大型语言模型（LLMs）开发相当于第2步的方法是克服数据瓶颈的关键研究问题（而且最终将是超越人类水平智能的关键）。

所有这些表明，数据限制似乎在预测未来几年的 AI 进展时引入了较大的误差范围。有很现实的可能性是事情会停滞不前（LLMs 仍可能像互联网一样重要，但我们不会真正达到疯狂的 AGI）。但我认为，实验室破解它的可能性是合理的，并且这样做不仅会保持扩展曲线的持续，还可能使模型能力实现巨大提升。

顺便说一下，这也意味着未来几年不同实验室之间的差异可能会比今天更大。直到最近，最先进的技术都是公开的，所以每个人基本上都在做同样的事情。（而新的初创公司或开源项目可以轻松地与前沿竞争，因为配方是公开的。）现在，关键的算法思想变得越来越专有。我预计实验室的方法将更加多样化，有些会比其他的进展更快——即使是现在看似在前沿的实验室也可能卡在数据瓶颈上，而其他实验室则取得突破，使他们能够领先。而开源项目将更难竞争。这无疑会使事情变得有趣。（如果某个实验室成功破解，这一突破将成为 AGI 的关键，成为超级智能的关键——这是美国最珍贵的秘密之一。）