

重新思考关于 AI 的能源使用

Daniel Castro | 2024 年 1 月 29 日

关注数字技术使用的能源并不新鲜。在互联网繁荣的高峰期附近 **1990** 年代， **a**福布斯 文章哀叹，“在某处美国， 每次一本书都会燃烧一块煤网上订购”。¹ 文章的作者， 成为在随后几年关于能源的辩论中被广泛引用政策， 估计“一半的电网将是在下一个时代为数字互联网经济提供动力十年”。² 然而， 估计是错误的， 错误是它的事实和方法论。³ 事后看来， 没有不再有任何争议， 如国际能源署 (IEA) 估计， 当今的数据中心和数据输电网络“每个占大约 **1 - 1.5%** 的全球用电量。”⁴

这个错误不是一个孤立的事件。许多头条新闻多年来出现的预测数字经济的能源脚印会失控。⁵ 例如， 作为流媒体战争 2019 年开始 - 苹果， 迪士尼， HBO 和其他宣布视频流媒体订阅服务， 与 Netflix、亚马逊和 YouTube - 多家媒体重复声称来自法国智囊团“观看 Netflix 30 分钟产生的排放是就像开车 4 英里一样。”⁶ 但同样， 估计是完全错误 (这更像是在 10 到 100 码之间行驶)， 这是由于有缺陷的假设和转换错误的混合， 智囊团最终在一年后纠正。⁷

随着最近对人工智能 (AI) 的兴趣激增， 人们再次提出了关于新兴能源使用的问题技术。在这种情况下， 批评家推测 AI 的快速采用

再加上深度学习模型规模的增加将导致能源使用的大量增加，具有潜在的破坏性环境影响。⁸然而，与过去的技术一样，许多关于人工智能消耗能源的早期说法已被证明是夸大和误导。本报告概述了辩论，包括一些早期的失误以及它们是如何形成的政策对话，并澄清人工智能能源的记录足迹以及未来几年可能会如何发展。它建议政策制定者通过以下方式解决对人工智能能源消耗的担忧采取以下步骤：

- 为 AI 模型制定能源透明度标准。
- 寻求关于能源透明度的自愿承诺基础模型。
- 考虑 AI 法规对能源使用。
- 使用 AI 使政府运营脱碳。

关于 AI 的能量使用和碳排放的事实

准确估算 AI 的能源使用和碳排放系统在其生命周期中具有挑战性，因为这些计算取决于许多复杂的因素，包括有关芯片的详细信息，冷却使用的系统、数据中心设计、软件、工作负载和能源用于发电。这个问题不是 AI 独有的。作为一组能源研究人员在一篇文章中描述了这个问题的 Annual 能源与环境回顾：

为信息创建可靠的电力需求估计技术充满了困难。底层数据未知准确地说，经验数据是有限的，最有用的数据通常是专有的，而且技术变化如此之快，甚至准确的数据很快就过时了。⁹

然而，一些研究试图量化当前和未来的人工智能系统的能源需求和碳排放。不幸的是，一些最初的估计已经陷入了与过去早期研究相同的陷阱关于数字技术的能源使用，并产生了误导估计。这些研究通常考虑 AI 所需的能量系统在其生命周期中分为两个阶段：1) 训练 AI 模型；以及 2) 使用 AI 模型来响应特定的查询 - 这个过程称为“推理”。

训练 AI 模型

马萨诸塞州阿默斯特大学的研究人员估计 2019 年几个 AI 模型的碳排放，第一个主要模型之一同类研究。¹⁰ 研究发现 BERT - 当时是 Google 最先进的大型语言模型（LLM） - 发出

约 1,438 磅二氧化碳 (CO₂) 在 79 小时的
使用 64 个高级图形处理单元 (GPU)、芯片进行培训
通常用于训练 AI 模型，因为它们具有优越的并行性
处理能力。从这个角度来看，往返航班从
纽约到旧金山可产生约 2,000 磅的 CO₂
每位乘客的排放量。研究人员还估算了碳
用于训练用于神经架构搜索 (NAS) 的 AI 模型的发射
自动寻找一个或多个神经网络的技术
给定任务的体系结构 - 计算最复杂的体系结构之一
机器学习中的问题。具体来说，他们评估了能量
使用 NAS 来创建更好的英语 - 德语机器
翻译模型。¹¹ 研究人员估计，训练模型
问题产生了 626,155 磅的 CO₂ 排放量 (大致相当于
从东海岸到西海岸的 300 个往返航班)。¹²

毫不奇怪，鉴于新闻业倾向于负面
科技报道，几乎所有流行媒体的头条新闻都集中在
尽管它的用例狭窄，但在后一种估计上。¹³ 即使受到尊重
科学新闻媒体，如麻省理工学院技术评论这样的头条新闻
作为“训练一个单一的人工智能模型可以排放与五辆汽车一样多的碳
lifetimes。”¹⁴ 这些文章表明，需要大量的能源
训练这个特定的 AI 模型是正常的，尽管这个估计很清楚
指的是一个非典型的例子。这就像一个汽车新闻媒体
发表一篇文章，建议“驾驶汽车排放的碳与
一架飞机”仅基于一项研究环境影响的研究
飞行汽车的原型。

此外，原始的研究论文和随后的新闻
文章经常指出，虽然大型人工智能模型的性能优于现有的
在语言翻译基准测试中，改进只是
边际。这意味着人工智能研究人员正在做琐碎的事情
性能改进以牺牲大量的
碳排放。事实上，其他人工智能研究人员在一份报告中明确了这一点
广泛阅读的论文“关于随机鹦鹉的危险：可以语言”
模型太大？”¹⁵ 他们认为这是“环境种族主义”
富裕的西方国家部署更大的 AI 模型，因为这些 AI
系统将对全球贫困社区产生负面影响
南方。具体来说，他们写道：

例如，问马尔代夫的居民是否公平或公正
(到 2100 年可能会在水下) 或苏丹的 80 万人
受严重洪水影响的培训和环境代价
部署越来越大的英语 [语言模型]，当类似的大 -
没有为 Dhivehi 或苏丹阿拉伯语生产比例模型？¹⁶

考虑到这些指控 - 训练 AI 系统不仅对
环境，但也是一种公开的种族主义行为 - 这并不奇怪，许多人
政策制定者对人工智能的能源消耗提出了质疑。
然而，2019 年研究中的头条新闻估计是疯狂的

不正确 - 就像许多先前关于能源足迹过大的说法一样
数字技术。马萨诸塞州阿默斯特大学
研究人员做出了几个错误的假设，严重夸大了他们的
估计使用的总能源和碳排放量。在
对 2019 年研究的回应，参与 NAS 模型的研究人员
提供了能源使用和碳排放的详细摘要
他们的工作，指出为什么外部研究人员的估计是错误的。
实际排放量比先前研究的估计值小 88 倍。¹⁷
不幸的是，大众媒体很少注意纠正
记录或注意新的发现，所以最初的印象
住在。

研究人员在随后的几年中发表了多项研究
估算训练许多知名 AI 模型所需的能量
作为他们的碳排放量。如表 1,虽然更大的模型
通常需要比较小的能源使用更多的能源，确切地说
不同的 AI 模型之间的数字差异很大。例如，
研究人员估计，训练 GPT-3 - 1750 亿个参数 AI
流行的 ChatGPT 应用程序中使用的模型 - 创建 552 tCO₂
排放量，但包括 OPT 在内的可比 AI 模型 (1750 亿
Meta 创建的参数 AI 模型) 和 Gopher (一个 2800 亿参数
Google 创建的 AI 模型) 的碳足迹明显较小。
此外，训练 AI 模型的效率不断提高。对于
例如，在 GPT-3 之后的 18 个月，谷歌生产了 GLaM，LLM 为 1.2
万亿参数。尽管 GLaM 比 GPT-3 大将近 7 倍
并且优于其他 AI 模型，GLaM 需要少 2.8 倍
能量训练。¹⁸ 最后，用于为数据中心供电的能源组合
开发人员训练 AI 模型会影响其碳排放。对于
例如，BLOOM 的开发人员使用法国的数据中心
核能，减少了碳足迹。¹⁹

尽管有了新的研究，但批评人工智能的团体一再引用
最初的不正确研究他们要求决策者减少
对大规模计算资源的投资。例如，美国
公民自由联盟 (ACLU) 致信科学和
2021 年 10 月的技术政策 (OSTP) 抱怨
白宫计划中的国家人工智能研究的“环境成本”
资源 (NAIRR)，并认为“NAIRR 应该专注于提供
替代关注数据和计算的应用程序
许多工业和研究实验室。”²⁰ 同样，人工智能中心和
数字政策在 2022 年错误地声称：“支持 AI 的系统需要
计算能力呈指数级上升。这种计算能力的增加
需要大量的能源消耗，产生巨大的碳
足迹和颠覆数字化的绿色效应。”²¹ 在每种情况下，
尽管有大量证据表明他们是
误导和夸大。

表 1：训练各种 AI 模型的估计能源需求

模型	# of 芯片 参数 (型号 x #)	Hours	能源 (MWh)	CO ₂ 排放量 (公制 Tons)	Estimate 来源
BERT	0.1 B V100x64	79	1.5	0.7	Strubell 等 Al., 2019 ₂₂
GPT - 2	1.5 B TPuv3x32	168	1.7*	0.7*	Strubell 等 Al., 2019 ₂₃
Llama 2	7B A100x (n / a)	N / A	74*	31.2	Meta, 2023 ₂₄
Llama 2	13B A100x (n / a)	N / A	147*	62.4	Meta, 2023
Llama 2	70B A100x (n / a)	N / A	688*	291.4	Meta, 2023
LaMDA	137B TPuv3x1024	1,385	451	26	Thoppilan et al., 2022 ₂₅
GPT - 3	175B V100x10000	355	1,287	552	帕特森等 al., 2021 年 ₂₆
OPT	175B A100x992	N / A	N / A	75	张等人， 2022 ₂₇
布卢姆	176B A100x384	2,820*	433	24.7	Luccioni et al., 2022 ₂₈
地鼠	280B TPuv3x4096	920	1,151*	380	Rae 等人， 2022 ₂₉
PaLM	540B TPuv4x6144	1200	3,436*	271.4	Chowdery et al., 2022 ₃₀
	TPuv4x3072	326			
GLaM	1, 162B TPuv4x (n / a)	N / A	456	40	帕特森等 al., 2022 ₃₁
GPT - 4	1, 800B A100x25000	2280	N / A	N / A	沃克， 2023 ₃₂

* 根据现有数据推断， 详见附录

使用 AI 模型

尽管政策制定者和媒体对能源成本的关注
在训练人工智能模型的过程中， 多项研究得出的结论是， 大多数

与 AI 系统相关的能源成本来自使用 AI 模型 - a 称为“推断”的过程 (因为模型是推断结果基于给定的输入)。例如，Amazon Web Services 估计 AI 模型 90% 的成本来自推理。³³ 同样，a 施耐德电气的研究估计，80% 的人工智能工作量 2023 年的数据中心是来自推理，20% 是用于培训。³⁴ 最后，Meta 研究人员的一项研究指出，确切的分解训练与推理之间的差异在不同的用例中有所不同。对于 LLM，他们估计推断与 65% 的碳有关足迹，但对于参数必须为的推荐模型根据新数据经常更新，他们估计平均分裂在训练和推理之间。³⁵

多个因素影响推理过程中使用的能量，包括任务类型和 AI 模型。如表 2,能源推理的要求可能因任务而异。例如，使用 AI 模型对文本进行分类通常计算强度较低（因此使用更少的能量）比使用 AI 生成图像。³⁶ 不同 AI 模型也有不同的能源成本，并且在特定模型中 (例如，Llama 2 7B 与 Llama 2 70B)，更多的参数通常需要更多的能量来推断。

表 2：按任务划分的每 1,000 个查询的平均能耗³⁷

任务	kWh
文本分类	0.002
图像分类	0.007
对象检测	0.038
文本生成	0.047
Summarization	0.049
图像生成	2.907

鉴于训练特定的 AI 模型会产生一次性成本，而使用 AI 模型会随着时间的推移继续消耗能量，这是有道理的用于 AI 的大部分能量最终将来自推理。它这也意味着运行 AI 模型的能源需求将有一个对人工智能系统的整体能源使用产生重大影响。虽然大多数批评人士将注意力集中在用于训练 AI 模型的能量上，有些人对推理过程中使用的能量表示担忧。³⁸ For 例如，写在 2023 年 10 月版的杂志上焦耳，一研究人员估计，与 LLM 互动需要大约

10 倍于进行典型的网络搜索查询的能量，以及
从这一估计推断得出的结论是“最坏的情况
下一篇：谷歌的 AI 可以消耗与一个国家一样多的电力
例如爱尔兰（每年 29.3 TWh）。”³⁹

有很多理由怀疑这种“最坏情况”正在发生
在 2022 年，谷歌的全球能源消耗总量
整个公司是 21.8 TWh。⁴⁰ 对于最坏情况的预测
诚然，谷歌仅对人工智能的能源使用就必须超过其
当前全球能源使用总量。确实，该公司的能源
消费随着时间的推移而增长，特别是来自其数据中心的消费，因为其
业务增长。例如，Google 的数据中心使用了约 3 TWh
2022 年的电力比前一年多。⁴¹ 但是，虽然它的整体能量
在 2019 年至 2021 年的三年中，使用量一直在增长，
它用于机器学习的能量比例保持不变 -
占其总能耗的 10% 至 15% -
其中大约 60% 用于推理。⁴²

用于以下目的的能量相对恒定比例的一种解释
推理是在 AI 模型和硬件中看到的改进。事实上，作为
显示在表 3,性能和效率都倾向于提高
时间。表格显示，在几年内，计算机的准确性
视觉 AI 模型显著改进。此外，能源
这些模型的推理需求通常随着
新芯片的发布。正如最近一项关于能源的研究所指出的那样
用于 AI 模型中的推理“，当 SOTA [最先进的] 模型是
发布它通常有大量的 FLOP [浮点运算]，
并因此消耗大量的能源，但在几年
有一个模型具有相似的精度，但数量要低得多
FLOPS”。⁴³ 换句话说，最新的 AI 模型可能并不特别
高效的设计，因为研究人员专注于性能
改进，但随着时间的推移，研究人员将解决效率问题。

表 3：深度神经推理的能耗
用于两个不同 GPU 的计算机视觉的网络⁴⁴

	Year	Top - 1 精度 (ImageNet)	P100, 6 月发布 6 月发布 2016 (焦耳)	V100, 2017 (焦耳)
AlexNet	2012	56.52	0.033	0.023
GoogLeNet	2014	69.77	0.077	0.055
Vgg16	2014	71.59	0.542	0.373
ResNet50	2015	75.30	0.179	0.132

什么 AI 能量预测错误

对 AI 未来能源需求的预测如此之高的原因之一是
他们使用不准确或误导性的测量， 如前所述。
另一个原因是预测忽视了实际的经济和
人工智能广泛商业化带来的技术现实。

人工智能的能源使用受到经济考虑的限制

许多对人工智能能源使用的高端估计是不切实际的， 因为
所涉及的成本。购买更多芯片， 建立更多数据中心， 以及
为这些数据中心供电是昂贵的。例如， 即使作者
预测 Google 的 AI 每年可能消耗 29.3 TWh
承认， 达到这个水平需要 1000 亿美元的投资
芯片以及数据中心数十亿美元的运营成本和
电力。⁴⁵ 即使是大型科技公司也会发现支付不可持续
对于如此大量的计算。企业是逐利的
企业和计算成本， 因此， 他们不会
提供服务的时间长， 运营成本高于他们在
收入。要么使用人工智能的能源成本会下降， 要么如何
公司部署人工智能将受到成本因素的限制。

随着时间的推移， AI 的性能提升率将下降

人工智能模型在过去几年有了显著的改进。
OpenAI 的 LLM 模型 GPT-4 于 2023 年 3 月发布， 可以通过许多
为人类设计的流行考试， 如 SAT、GRE、LSAT 和 AP
各种科目的测试。⁴⁶ 这些结果是一个实质性的
比上一年发布的早期模型有所改进。而 AI 仍然
不能像人类一样执行许多任务， 例如抽象
推理， 现在一些人工智能模型在许多
基准， 在
某些领域。因此， 许多开发人员可能会更专注于
优化他们的 AI 模型， 而不是挤出越来越小的
提高准确性， 因为他们不会收到回报
建设和运营大型模型的投资。

未来的创新将提高 AI 的能源效率

计算的历史是不断创新的历史之一， 这些
创新延伸到能源效率。例如， 在过去的十年中，
对全球数据中心的需求大幅增加， 即使
数据中心的能源强度下降了约 20
每年%。⁴⁷ 在 2010 年至 2018 年之间， 有 550 %
计算实例增加， 存储增加 2,400%
全球数据中心的容量， 但全球数据仅增长 6%
中心能源使用。⁴⁸ 这些能源效率的提高来自
硬件、虚拟化和数据中心设计方面的改进， 以及它们
是云计算能够扩展的部分原因。

类似的趋势已经出现在人工智能中。正如最近的一篇论文指出的那样，“许多研究报告说，神经网络的规模正在增长指数。然而，这并不一定意味着成本也是呈指数级增长，因为可以使用相同的能量，主要是由于硬件专业化，但特别是随着单位计算的能耗正在下降。”⁴⁹硬件和软件的改进可能会保持能源的步伐来自 AI 的增长。芯片制造商继续创造更高效的用于 AI 的 GPU。例如，Nvidia 最近从一代 GPU 到另一个不仅导致处理速度明显加快，而且能源效率几乎翻了一番。⁵⁰同样，研究人员继续实验技术，如修剪、量化和蒸馏，创建更紧凑、更快、更节能的 AI 模型以最小的精度损失。⁵¹这些类型的进步是一种谷歌用于人工智能的能源比例保持不变的原因近年来，尽管机器学习增长到 70 到 80 公司使用的计算机的百分比。⁵²事实上，作为一名研究员简洁地说，“[人工智能的] 能源消耗并没有飙升，相反普遍表达的恐惧。”⁵³

AI 的能源足迹忽略替代效应

讨论 AI 系统的能源使用趋势可能会产生误导不考虑技术的替代效应。许多数字技术通过用移动钻头代替移动钻头来帮助经济脱碳移动原子。例如，发送电子邮件代替邮寄信件，流媒体电影取代了租用 DVD 和参与视频会议取代了亲自参加会议。人工智能将有类似的随着时间的推移，通过进一步数字化许多活动 (例如通过提高视频通话的质量)，并通过使用人工智能来完成更多任务比使用人力更有效。

2023 年的一项研究估计了使用人工智能与使用人工智能的碳足迹用于编写文本页面或创建插图的人。之后考虑不同 AI 模型的碳排放 (ChatGPT，布鲁姆、中程和 DALL-E 2)，并比较美国的工人美国和印度，研究人员发现“人工智能写一页文字发出二氧化碳比人类少 130 到 1500 倍，”和“人工智能创造了一个图像发出的信号减少 310 到 2900 倍。”⁵⁴当然，因为使用 AI 不会消除人类 - 他们仍然存在，吃，呼吸等 - 使用 AI 不会消除这些碳排放。但是人工智能确实消除了碳人类用于这些任务的设备的排放，如笔记本电脑或台式计算机。如表 4, 这些节省可能是巨大的；however, there are limits to generalizing these findings. For example, by 使生成文本和图像更容易，活动量可能增加。然而，这些发现表明，在保持所有其他条件平等的情况下，使用人工智能替代人类劳动力可以减少碳排放某些案件。

表 4： 使用人类的碳足迹 (克二氧化碳) (在美国) 与某些任务的 AI (BLOOM / Midtry)

	AI	笔记本电脑	台式机	Human
写一页文字	0.95	27	72	1,400
创建图像	1.90	100	280	5,500

AI 的能量如何使用适合更大的图片

关于人工智能能源使用的辩论是关于如何应对全球气候变化。在这种情况下，有重要的政策制定者应该记住的因素。

AI 将在应对气候变化方面发挥重要作用

有很多机会使用人工智能来减少碳排放，支持清洁能源技术，应对气候变化。这些机会跨越多个行业，包括交通运输，农业和能源部门。例如，人工智能对于整合至关重要风能和太阳能等可再生能源通过以下方式进入电网使用数据点来预测供应和需求。同样，公用事业是使用 AI 进行能源资产的预测性维护、管理和控制电网，并设置动态定价 - 所有关键要素高效电网。⁵⁵

人工智能还可以帮助理解来自传感器和卫星，如变化的海平面、地表温度和降雨，创造更好的预测并应对气候变化的风险。对于例如，人工智能可以从卫星数据中检测甲烷排放，允许监管机构更有效地监控行业。⁵⁶ 同样，农民可以使用人工智能用于精准农业，减少肥料和水的使用相关的环境成本。⁵⁷

企业，政府和消费者已经在使用 AI 来运营更有效。AI 是创建使用 AI 来创建智能城市的关键部分运营高效的建筑、道路、水路等。⁵⁸ 例如，在加利福尼亚州，政府正在使用 AI 监控一千多个摄像头快速检测和应对野火，减少碳排放来自这些火灾。⁵⁹ AI 可以使公司优化工业流程，减少浪费，更有效地利用能源，从而减少碳强度。⁶⁰ 例如，物流提供商使用 AI 来优化交付路线，从而减少其车队的燃料消耗。⁶¹ And in 消费者空间，Nest 智能温度计等工具节省了客户在 2011 年至 2022 年之间达到 1,130 亿千瓦时，效率更高从谷歌地图驾驶减少了 120 万的碳排放量公吨。⁶²

随着人工智能的成熟，政策制定者应该继续将技术视为应对气候变化的关键工具。

人工智能的能源使用没有独特的市场失灵

解决全球气候挑战将需要过渡到清洁能源技术具有与肮脏相当的价格和性能

ones.⁶³在此期间，任何使用能源的活动都有一个环境影响，而 AI 对能源的使用也没有什么不同。然而，没有独特的与人工智能使用能源相关的市场失灵，这将导致比替代用途更大的环境影响。千瓦时用于人工智能与用于看电视的千瓦时没有什么不同，微波爆米花，供电灯或任何其他活动。确实，如前面提到，在许多情况下，人工智能应用程序将被用作替代能源效率较低的活动并解决气候问题

改变。在这两种情况下，那些消耗能源的人必须为此付出代价，并且理性的行为者通常会寻求最小化这些成本。虽然这样成本可能不包括与能源使用相关的负外部性，这个问题并不是人工智能独有的，也不能单靠人工智能来解决。

大型科技公司做出了大胆的净零承诺

大型科技公司处于人工智能的最前沿，这些都是一样的做出了一些最大胆承诺的公司

公司减少碳足迹。考虑以下几点：

- 谷歌（现在的 Alphabet）在 2007 年成为碳中和，第一个大公司这样做。⁶⁴十年后，它成为第一个专业公司购买足够的可再生能源来匹配其电力消耗。⁶⁵而且，在 2020 年，它购买了碳抵消以消除公司的全部碳遗产。⁶⁶ The 公司继续向前推进，并承诺在无碳能源上运行其所有数据中心和校园到 2030 年。⁶⁷
- 亚马逊于 2019 年共同创立了气候承诺公司承诺到 2040 年实现净零碳排放，10 年在《巴黎协定》之前。⁶⁸而且，在 2022 年，亚马逊报告它消耗的 90% 的电力来自可再生能源来源，到 2025 年，它有望达到 100%，五个比 2030 年的目标提前了几年。⁶⁹
- 微软承诺到 2030 年实现碳减排到 2050 年消除其公司的全部碳遗产。⁷⁰它也有承诺到 2025 年使用 100% 的可再生能源。⁷¹
- Facebook（现为 Meta）直接实现零碳排放 2020 年运营（即数据中心和办公室），并已承诺通过以下方式在其整个价值链中达到净零排放 2030。⁷²

以上内容仅为本文档的试下载部分，为可阅读页数的一半内容。如要下载或阅读全文，请访问：<https://d.book118.com/987010140162006036>