迈向用户拥有的互联网之路

2022 年，我(anna)撰写了一份提案，提出了一个用户拥有的基础模型，该模型使用私人数据而不是从互联网上公开抓取的数据进行训练。我认为，虽然可以使用公共数据（例如 Wikipedia、4Chan）来训练基础模型，但要将它们提升到一个新的水平，您需要高质量的私人数据，这些数据仅存在于需要权限或登录才能访问的孤立平台（例如 Twitter、个人消息、公司信息）中。

这一预测正在开始实现。Reddit 和 Twitter 等公司已经意识到其平台数据的价值，因此他们锁定了开发人员 API（1、2 ），以防止其他公司自由地使用其文本数据训练基础模型。

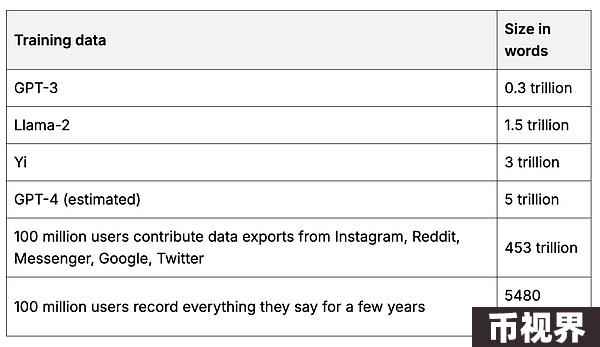
这与两年前相比发生了巨大变化。风险投资人 Sam Lessin 总结了这一变化：“[平台] 只是把这些垃圾扔到后面，没有人看管，然后突然间，你会觉得，哦，该死，那些垃圾是金子，对吧？我们得到了很多。我们必须锁好垃圾箱。”例如，GPT-3 是在 WebText2 上进行训练的，它汇总了所有 Reddit 提交链接中的文本，这些链接至少有 3 个赞成票（3，4）。使用 Reddit 的新 API 后，这不再可能。

互联网变得越来越不开放，孤立的平台筑起更大的墙来保护其宝贵的训练数据。

尽管开发人员无法再大规模访问这些数据，但由于数据隐私法规，个人仍然可以跨平台访问和导出自己的数据（5、6）。平台锁定开发人员 API，而个人用户仍然可以访问自己的数据，这一事实提供了一个机会：1 亿用户是否可以导出其平台数据来创建世界上最大的数据宝库？这个数据宝库将汇总大型科技公司和其他公司收集的所有用户数据，而这些公司通常不愿意分享这些数据。这将是迄今为止最大、最全面的训练数据集，比用于训练当今领先的基础模型的数据集大 100 倍。1

## 表 1. 数据

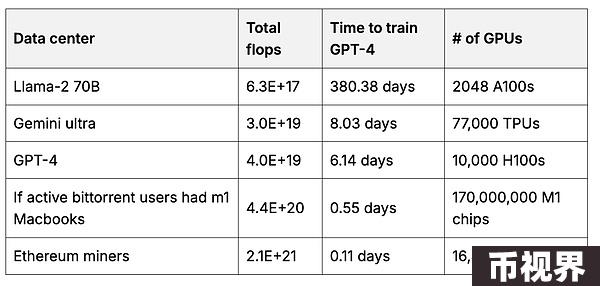
将基础模型训练数据集与示例用户数据集进行比较的粗略估计。来源及计算。



然后，用户可以创建一个用户拥有的基础模型，该模型使用的数据比任何一家公司能够聚合的数据都要多。训练基础模型需要大量的 GPU 计算。但每个用户都可以用自己的硬件帮助训练模型的一小部分，然后将这些部分合并在一起，创建一个更大、更强大的模型（7、8、9 ）。2当激励措施合适时，用户可以汇集大量计算。例如，以太坊矿工的总计算量是用于训练领先基础模型的 50 倍。

## 表 2. 计算

与以太坊矿工 GPU 相比，对用于训练基础模型的数据中心的总浮点运算次数（每秒浮点运算次数 = 所有 GPU 的“思考”速度总和）进行估计。3带有计算的来源。



为该模型做出贡献的用户将集体拥有并管理该模型。他们可以在使用模型时获得报酬，甚至可以根据他们的数据对模型的改进程度按比例获得报酬。集体可以制定使用规则，包括谁可以访问该模型以及应该实施哪种控制。也许每个国家的用户都会创建自己的模型，代表他们的意识形态和文化。或者也许一个国家并不是正确的分界线，我们将看到一个世界，每个网络国家都有自己的基于其成员数据的基础模型。

我鼓励您花时间思考一下您希望拥有哪些基础模型的一部分，以及您可以从使用的平台贡献哪些训练数据。您可能拥有的数据比您意识到的还要多——您的研究论文、未发布的艺术品、您的 Google 文档、您的约会资料、您的医疗记录、您的 Slack 消息。将这些数据整合在一起的一种方法是通过个人服务器，这使您可以轻松地将您的私人数据与本地 LLM 一起使用。将来，您的个人服务器还可以训练您拥有的用户基础模型的一部分。

基础模型倾向于垄断，因为它们需要在数据和计算方面进行大量的前期投资。我们很容易选择简单的选项：尽我们所能地使用落后几代的开源模型，即大型人工智能公司的残余。但我们不应该满足于落后几代，只吃剩饭剩菜！作为用户，我们应该创建我们自己的最佳模型——我们拥有实现这一目标的数据和计算能力。

随着人工智能越来越有能力完成有价值的经济工作，一场巨大的经济转变正在发生。大型科技公司已经根据您的公开工作、写作、艺术作品、照片和其他数据以及其他人的数据训练了人工智能模型，并开始每年赚取数十亿美元（1）。他们现在正在追逐您在公共互联网上无法获取的数据，从 Reddit 等公司购买您的私人数据，这样他们就可以将人工智能的收入增加到每年数万亿美元（2、3 ）。

## 您难道不应该拥有由您的数据帮助创建的AI 模型的一部分吗？

这就是数据 DAO 的作用所在。数据 DAO 是一个去中心化的实体，允许用户汇集和管理他们的数据，并用代表特定数据集所有权的数据集特定代币奖励贡献者。它有点像数据的工会。这些数据集可以复制甚至超越大型科技公司以数亿美元出售的数据集 (4)。DAO 对数据集拥有完全控制权，可以选择将其出租或出售匿名副本。例如，Reddit 数据甚至可以用来播种新的、用户拥有的平台，包括好友、你过去的帖子和其他数据，这些数据可以在新平台上随时使用。

如果您对技术细节感兴趣：数据 DAO 有两个主要组成部分：1）链上治理，通过数据贡献获得代币；2）安全服务器，使用公钥-私钥对进行加密，社区拥有的数据集驻留在该服务器中。要做出贡献，您首先要验证数据以证明所有权并估计其价值。然后，使用服务器的公钥在浏览器中加密数据，并将加密数据存储在云中。只有当 DAO 批准授予访问权限的提议时，数据才会解密。例如，它可以允许 AI 公司租用数据来训练模型。您可以在此处阅读有关 Vana 网络架构的更多信息，该网络旨在实现数据集和模型的集体所有权。

数据 DAO 不仅使用户受益，还推动了 AI 的发展，使像开源软件一样构建 AI 成为可能，让所有做出贡献的人受益。开源 AI 正在努力寻找可行的商业模式：支付 GPU、数据和研究人员的费用非常昂贵。而且，一旦模型训练完成，如果它是开源的，就无法收回这些成本。数据 DAO 的技术架构可以应用于模型 DAO，用户和开发人员可以贡献数据、计算和研究以换取模型的所有权。

当今社会的默认选项是允许大型科技公司获取我们的数据，并用它来训练为我们工作的人工智能模型。他们从这些人工智能模型中获利，因为我们被用我们的数据训练的模型所取代。这对社会来说是一笔非常糟糕的交易，但对大型科技公司来说却是一件好事。防止这种情况发生的唯一方法是采取集体行动。数据就是货币，集体数据就是力量。我鼓励你参与：世界上第一个专注于 Reddit 数据的数据 DAO今天在 Vana 网络上上线。通过打破少数特权阶层控制的数据护城河，数据 DAO 开辟了一条通往真正用户拥有的互联网的道路。