【知识福利社】你的私人书院，微信：312005650 全网热门课程众筹共享！

第7讲 ChatGPT需要什么资源？

前沿课·吴军讲GPT
2023年5月1日

进入课程 >

时长：10分45秒 大小：14.76MB

00:00 / 10:45

宝木 转述

🍅，你好!

你好，欢迎来到《前沿课·吴军讲GPT》。

ChatGPT走红之后，国内很多媒体和人一下子慌了神。

各种媒体，特别是自媒体上，铺天盖地都是诸如“为什么中国没有诞生ChatGPT”“美国将要爆发一场新的技术革命”之类的文章。很多研究机构和企业都宣称，我们马上就要推出自己的ChatGPT。还有很多人觉得，ChatGPT给自己带来了新机会。

其实，这样全民大炼ChatGPT，既不必要，也不可能。

这是为什么呢？不必要性，我们后面会慢慢展开讲。这一讲，先来谈谈不可能性。

我们前面讲到，ChatGPT的核心是语言模型，而语言模型需要用大量的数据来训练，有了数据后，还需要强大的算力支持，然后还需要有足够高水平的并行计算和机器学习的算法支持。**数据、算力、算法三道坎，只有极少的互联网大厂能够越过去。**

接下来，我们就从这三个方面分别来看一下。

限制一：数据

首先说第一个限制，数据。

早在2012年，我在Google训练自动问答的模型时，就用到了互联网上能找到的几乎所有高质量的数据。今天的ChatGPT所用的数据更多，特别是自媒体上的数据和社交网络上的数据。

这个数据量是极其庞大的。不是每个企业都能得到所有的这些数据。

限制二：算力

而且，就算一家企业今天能得到这些数据，也没有能力存储和处理这些数据，因为处理它们的计算量太大。这就要说到训练语言模型的第二个限制——算力。

前面讲过，早期的算法非常简单，就是对文本做一些统计，然后利用一些简单的数学公式，计算出一个词在不同上下文条件下的概率，事先保存下来。使用的时候，根据这些概率，计算出各种语言现象的概率，比如一句话的概率，或者某个词是名词的概率，或者给定一个问题，某句话是它的答案的概率。

这是截止到上世纪90年代的方法。当时，即使做这样简单的统计，所消耗的算力也是惊人的，只有IBM这类的大公司才能办到。

但是很快，贾里尼克等人发现，如果只是对文本进行简单的统计，再多的数据也无法覆盖所有的语言现象，而且对于很多语言现象，特别是之前没有见过的语言现象的概率，语言模型估计得都很不准确。于是，人们开始考虑深度挖掘语言中所蕴含的信息，包括语法信息和语义信息。这就需要更大的计算量了。

具体计算量需要增加多少呢？100万倍到1亿倍是一个合理的估计。

很多人觉得，摩尔定律可以让计算的成本不断下降，过去做不成的事情今天能做成了。其实，**随着时间的推移，人们对人工智能的要求也在不断提高，需要的算力也在不断增加。**算力，永远是一个瓶颈。

上个世纪90年代末，贾里尼克为我们申请到了一笔美国国家科学基金会（NSF）的基金，让我们有经费建立复杂的语言模型。与此同时，全世界机器学习领域在算法上也有所突破，能够训练更准确的语言模型。此外，贾里尼克还从美国国防部获得了一笔巨额的经费，让我们能够装备上百台超级服务器。

在这些设备的帮助下，我们得以进行其他研究单位没有条件做的语言模型的改进工作。

1998年，我参加IEEE，也就是电气与电子工程师协会的一个重要的学术会议。在我前面做报告的是一个德国著名的科学家，他给出的结论是，更好的机器学习算法似乎对改进语言模型无效；而我接下来的报告则否定了他的结论。我们的研究表明，更好的机器学习的算法很有效，但是需要训练很长的时间，直到模型真正收敛。我们当时能调动的算力大约是他的100倍，因此，他做不了的事情我们能做。

但是，即使我们当时有学术界最强的算力，很多想做的事情也还是做不到。当我试图把各种语法信息加入到语言模型中的时候就发现，以我们的算力来做这件事，十年可能也算不完，我也根本

无法毕业。

当时，我们看到一位IBM的科学家做了类似的事情，并且发表了一篇文章，于是就请他来作报告，想了解一下他是怎么做的。其实，他真正了不起的地方不是在方法上有什么独到之处，而是他写了一个程序，将整个IBM沃森实验里闲置的服务器都利用起来，然后计算了整整一年。这个咱们前面也讲到过。

后来，我的导师们希望我超过这位科学家，于是我花了半年的时间从数学上改进算法，将训练的计算量降低了200倍，然后在40台超级服务器上运行三个月，终于训练出一个当时世界上最好的语言模型。如果没有这200倍的改进，我需要计算50年。

举这个例子，不是说我有多厉害，而是让你明白，算力对机器学习的重要性。你看，即便当时大幅改进了算法，减少了计算量，仍然需要40台超级服务器上运行三个月。

今天ChatGPT所采用的语言模型，从参数的数量上讲，是我当时训练的模型的10万倍，比我在Google构建的模型也大了1000倍。开发它所需要的算力，可不只是我在约翰霍普金斯时的10万倍，而可能是上百亿倍。

今天，开发GPT语言模型的公司OpenAI没有讲它使用了多少计算资源，但是它暗示，硬件成本超过一亿美元。如果我们按照1000美元一个GPU计算，它使用了10万个GPU。以32位运算为准，它能提供超100PFLOPS的算力，也就是每秒10亿亿次运算以上。这是整个特斯拉云计算集群计算能力的两倍，是Google最大的数据中心算力的1/3，是阿里云最大的数据中心算力的1/4。

这还只是在它开发GPT-3时的花费。到了升级为GPT-4时，需要的计算资源就更多了，用到了微软云计算大量的计算资源。

即便如此，ChatGPT最近还关闭了它的付费用户的注册，因为为付费用户提供更优质的服务太消耗计算资源了。

因此，如果你也想做一个ChatGPT，先要考虑一下自己的算力够不够。很多人一看到出了个新技术，就觉得机会来了，但是绝大部分时候，那些机会和大家无关，因为大家没有基本的能力实现它。

限制三：算法

那么，如果我们举全国之力建几个超级计算中心，是不是很容易超过ChatGPT呢？

如果只看算力，似乎可以。但其实，算力也分“聪明的算力”和“笨的算力”，对于人工智能问题，只有前一种——聪明的算力——才有意义。

所谓“聪明的算力”，就是说，光有算力还不够，你还得有配套的算法把算力的作用发挥到极致。如果没有配套的优质算法，算力就只能事倍功半。这就是笨的算力。

这就说到了训练语言模型的第三个限制——算法。

2000年，我在约翰霍普金斯构建语言模型时，一次能调动40台服务器同时工作。当时还没有云计算一说，我靠的是加州大学伯克利分校提供的一个并行计算的工具。这让我不需要手工监控每一台服务器的工作，而且能自动分配任务到空闲的服务器上。后来有了云计算，这件事就容易了很多。

但是今天，训练语言模型用到的机器学习算法要复杂很多。

2010年的时候，Google推出了一个基于云计算平台深度学习的工具——Google大脑。采用这个工具，语言模型的效果可以大幅提升，在其它条件不变的情况下，语音识别和机器翻译的相对错误率可以降低10%以上。

今天，深度学习的基本算法和支持它的基础架构，已经是智能化的数据中心，有些地方也称之为“智算平台”的标配。

如果只是用一个由处理器，包括GPU，堆砌起来的数据中心训练语言模型，得到的结果会差很多。

也就是说，**除了算力之外，基础的自然语言处理技术，也就是算法，也是实现ChatGPT这些产品必不可少的条件。**

最直观的例子，你想让计算机回答问题，至少要让它懂得问题。你要想让它从上千亿的文本中知道哪些可能是答案，需要做到问题和答案的匹配。这个工作不是一年半载就能完成的。

今天，很多机器学习的算法已经开源了，有些应用已经有公司和研究机构投入科研力量开发过，比如基本的图形识别和语音识别技术，但是深度的自然语言理解其实还不属于这个范畴。在这些领域所具有的技术积累，其实也是一种资源。

因此，一些长期在这个领域有投入的公司，比如Google、微软，中国的百度、字节跳动等，其实是有一定的技术积累的，它们可能通过对现有的技术的集成，在较短的时间里，做出类似ChatGPT的产品。

但是，对于那些完全不具备应有的技术资源的人来说，跟风ChatGPT，要么是无知者无畏，要么是炒作。

我们前面讲了，像GPT这样的语言模型技术，除了处理自然语言，还可以在很多领域找到应用场

景，比如在生物医疗领域、在视频编辑方面，这就需要具有相应的算法，这里我们就不一一列举了。因此，没有足够的算法资源，也不建议大家都去打造ChatGPT。

下节预告

好，这一讲，我们说了**训练语言模型的三个限制，分别是数据、算力和算法。这三道坎，只有极少的互联网大厂才能越过去。**

其实，即便跨过了三道坎，做出了比ChatGPT更好的产品，它也不可能像很多人想象的那样无所不能。因为ChatGPT，乃至人工智能，本身就是有能力边界的，并不能解决所有问题。这个我们下一讲再说。

我是吴军，我们下一讲再见。

划重点　　添加到笔记

训练语言模型的三个限制，分别是数据、算力和算法：
1、数据：训练语言模型所需的数据量是极其庞大的。不是每个企业都能得到所有的这些数据。
2、算力：随着时间的推移，人们对人工智能的要求也在不断提高，需要的算力也在不断增加。
3、算法：除了算力之外，基础的自然语言处理技术，也就是算法，也是实现ChatGPT这些产品必不可少的条件。

《前沿课·吴军讲GPT》
成为搞懂GPT的少数人
版权归得到App所有，未经许可不得转载
作家，计算机科学家
吴军

最近更新: 2023年5月4日
首次发布: 2023年5月1日

上一篇　第6讲 ChatGPT有哪些固有缺陷？
下一篇　第8讲 今天人工智能的边界在哪里？

可免费赠送给20位朋友

微信　朋友圈

用户留言

最新　最热　　只看作者回复

海清
5月1日　　＋ 关注

吴军老师好！许久没有来留言了，但你的课程我都是实时跟进学习的，包括最新的《世界文明史》和反复学习的《硅谷来信N》。没能留言的原因很多，并不是没能坚持早起早睡，而是突然离开了生活工作的舒适区，让我着实知道了什么是舒适区😁。好在个人个性能够适应快并调整好心态，非常认可米兰昆德拉的那句话：除

了生病以外，你所感受到的痛苦，都是你的价值观带来给你的，而非真实存在。

最近真是全民炼GPT，刚巧今天看到《金融帝国贝莱德》，说贝莱德“黑岩”早就拥有的“阿拉丁”系统进行投资交易决策。正像老师前文说到的GPT其实很早就应用在一些关键领域了，譬如这个管理着22万亿美元的资产管理公司。这套基于蒙特卡罗模型的“阿拉丁”有着2500万行代码其代码行数与脸书可较高低了。

这些年我们的世界已经出现了部分的赛伯格（CYBORG），一种智能，部分是人类，部分是技术。量化交易和买卖双方的信息都是通过阿拉丁的算法算力和非垃圾数据来实现的，我们看到的信息也是被筛选加工过的，而这些信息差是否会被GPT和各种技术先发人群占有，并使我们的社会加速到达2%神人和98%人的未来简史中描述的新世界呢？

对此担心，仿佛杞人忧天，因为天幕下已尽是电磁波👣

敬颂
时绥

收起

1 1 13 分享

Levy
5月1日
+ 关注

第7讲ChatGPT需要的资源：数据、算力和算法，这个能力只有顶级大公司才能参与。对于大多数普通个人而言，想办法利用ChatGPT提升自己工作效率和学习效率才是应该重点考虑的。

转发 评论 8 分享

清晨落幕
5月1日
+ 关注

Gpt岗位替代这个问题一直争议不断，主要分为两派。一派是替代的岗位会造成失业大于就业；另一派是使用新技术会让人效率提高，创造出新的职位覆盖原有的旧岗位。

技术发展有一个特点是，用最少的能量去提高效率。相当于用技术替代人力是必然的，人力成本降下来，收益也归机器背后的使用者所拥有。这是人性的行为，不会因为自己钱太多而把钱分出去。

收起

转发 评论 5 分享

徐浩
5月1日
+ 关注

📚感谢吴军老师的分享，学完今天的课程，想到3月份得到头条，南师也讲过类似的话题。
📚核心是过去一提到技术，我们都觉得，决定它发展的首要因素，是科学水平。但事实上，硬件资源才是更底层的刚性基础。
📚AI发展，到底都需要哪些硬件资源呢？
1⃣芯片。GPT对芯片的需求是极大的。大到什么程度？比如上一代的GPT3.5，训练它需要用到一种英伟达的高端GPU芯片，叫A100。一颗A100的成本，大概是8万人民币。而完成整个训练过程，大概需要三万颗A100芯片。而且GPT在长大的过程中，它的胃口可不是线性增长的，而是指数级增长的。
2⃣能源。我记得根据头条课程里，说AI对电的消耗，

非常大。ChatGPT，仅仅是训练一次，就要消耗90多万度电。相当于1200个中国人，一年的生活用电量。要知道，训练它可不是一次就够，需要很多次。而现在，它仅仅是维持日常运转，每天的电费大概就要5万美元。

3⃣ 数据标注师。AI训练依然需要大量的数据标注师。比如ChatGPT，它的训练分三个阶段，只有第三阶段不需要标注，前两个阶段都需要大量的人工标注。而且因为ChatGPT的数据是从网上采集的，里面什么数据都有，也可能包含一些暴力、犯罪或者反人类的内容。怎么把这些内容过滤掉？也得靠人来手工标注。

收起

# GPT学习小组

转发　评论　21　分享

---

webmin

5月1日

＋ 关注

4月具备数据、算力和算法的互联网头部大厂向外推出了计算引擎，把现在宝贵的算力拿出来当商品卖，用以帮助缺少这部分能力的厂商练训大模型，从这就可以窥觑大厂对这场竞赛也是存在分歧的，参考加州淘金的历史，提供挖金矿工具的人可能才是最后真正挣到钱的人。

注：“字节跳动旗下的火山引擎推出支持单任务GPU万卡集群大模型训练引擎” - https://www.thepaper.cn/newsDetail_forward_22752555

收起

# GPT学习小组

转发　评论　7　分享

赵棹

5月1日

＋ 关注

2023年5月1日 吴军老师讲GPT

GPT实现了人机对话，是技术技术的进步，同时。也带来了新的问题，如，行业狂热和被替代的担忧。

当然，任何一项技术的发明发展，都有积极和消极两面性，只是随着互联网和智能化的发展，GPT与人类的幸福关系更为紧密、直接，涉及的范围更加广泛，此时，我们更需要理性看待:

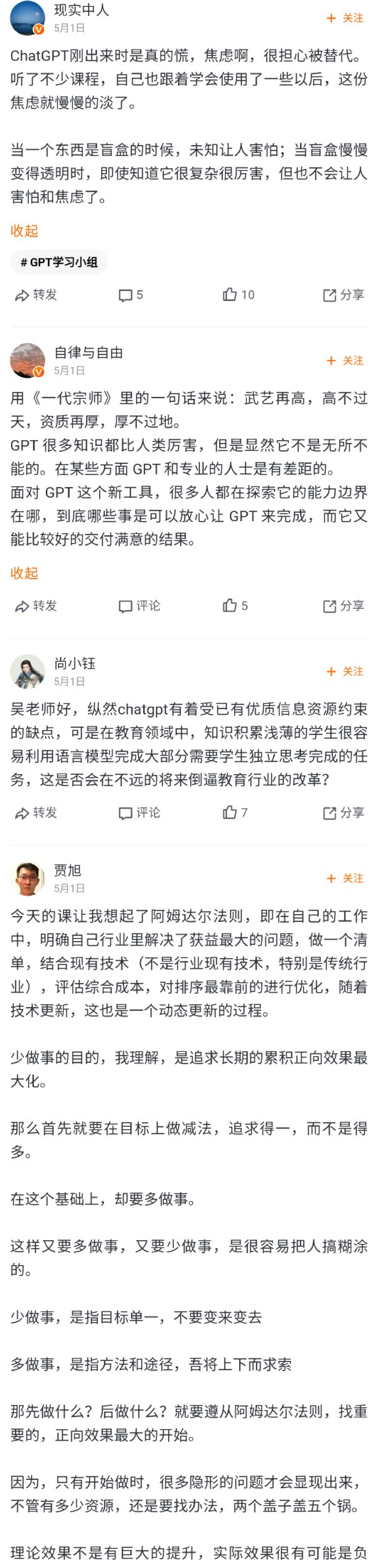
一是行业发展没有想象的那么简单，数据、算力和算法是GPT的三大瓶颈。经济学上有有效市场假说，是说市场价格反应了所有已知的信息，但我们往往会问，那为什么还有人做多有人做空？那是因为有信息不对称的因素，信息就在那，不是所有的信息都可得，都能建立无缝、互动链接。相对而言数据还是最简单的问题，算力和算法则更是无限递进的瓶颈和竞争压力，因此，投资人要正确看待自身的资源和能力边界，择一域，而非择全域。

二是GPT的终极替代很可怕，但它的发展不是一蹴而就的事，人工替代也不会铺天盖地的发生。算法是它的发展瓶颈，恰恰是机器难以替代能力，语法、语义和思维，是人的特有优势，学习力，以及不断在主动思维的行业嵌入我们的作用，替代在我们身上就很难发生。

三是GPT有不足，又有对人类替代的消极因素，我们要不要抵制？答案是不要。我们该利用它在无情绪干扰或重复计算等的优势，为我们所用，提升效能，提升人的职业幸福感。

收起

转发　评论　8　分享



现实中人
5月1日

ChatGPT刚出来时是真的慌，焦虑啊，很担心被替代。听了不少课程，自己也跟着学会使用了一些以后，这份焦虑就慢慢的淡了。

当一个东西是盲盒的时候，未知让人害怕；当盲盒慢慢变得透明时，即使知道它很复杂很厉害，但也不会让人害怕和焦虑了。

自律与自由
5月1日

用《一代宗师》里的一句话来说：武艺再高，高不过天，资质再厚，厚不过地。
GPT 很多知识都比人类厉害，但是显然它不是无所不能的。在某些方面 GPT 和专业的人士是有差距的。
面对 GPT 这个新工具，很多人都在探索它的能力边界在哪，到底哪些事是可以放心让 GPT 来完成，而它又能比较好的交付满意的结果。

尚小钰
5月1日

吴老师好，纵然chatgpt有着受已有优质信息资源约束的缺点，可是在教育领域中，知识积累浅薄的学生很容易利用语言模型完成大部分需要学生独立思考完成的任务，这是否会在不远的将来倒逼教育行业的改革？

贾旭
5月1日

今天的课让我想起了阿姆达尔法则，即在自己的工作中，明确自己行业里解决了获益最大的问题，做一个清单，结合现有技术（不是行业现有技术，特别是传统行业），评估综合成本，对排序最靠前的进行优化，随着技术更新，这也是一个动态更新的过程。

少做事的目的，我理解，是追求长期的累积正向效果最大化。

那么首先就要在目标上做减法，追求得一，而不是得多。

在这个基础上，却要多做事。

这样又要多做事，又要少做事，是很容易把人搞糊涂的。

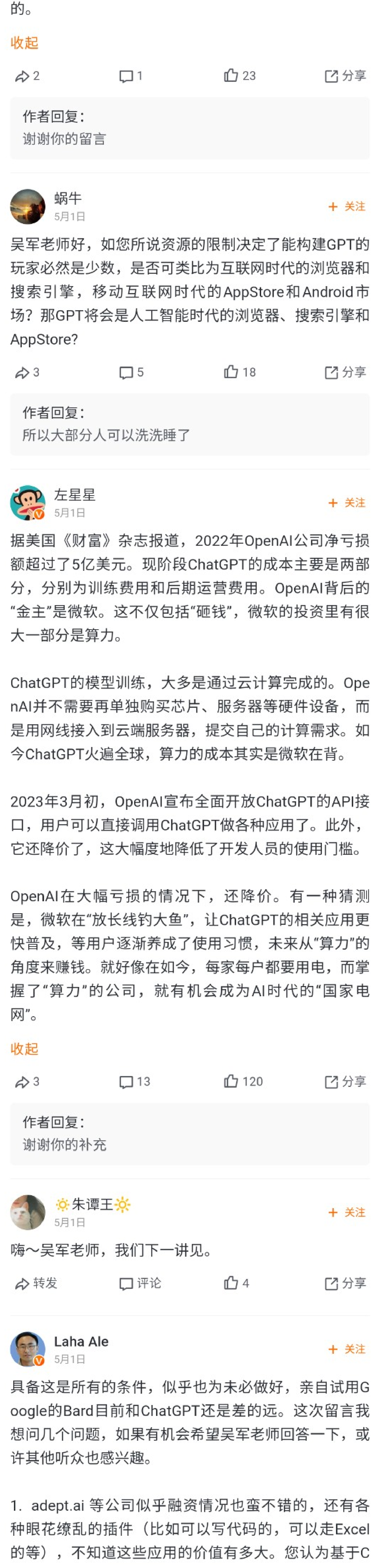
少做事，是指目标单一，不要变来变去

多做事，是指方法和途径，吾将上下而求索

那先做什么？后做什么？就要遵从阿姆达尔法则，找重要的，正向效果最大的开始。

因为，只有开始做时，很多隐形的问题才会显现出来，不管有多少资源，还是要找办法，两个盖子盖五个锅。

理论效果不是有巨大的提升，实际效果很有可能是负



的。

收起

作者回复：
谢谢你的留言

蜗牛
5月1日

吴军老师好，如您所说资源的限制决定了能构建GPT的玩家必然是少数，是否可类比为互联网时代的浏览器和搜索引擎，移动互联网时代的AppStore和Android市场？那GPT将会是人工智能时代的浏览器、搜索引擎和AppStore?

作者回复：
所以大部分人可以洗洗睡了

左星星
5月1日

据美国《财富》杂志报道，2022年OpenAI公司净亏损额超过了5亿美元。现阶段ChatGPT的成本主要是两部分，分别为训练费用和后期运营费用。OpenAI背后的“金主”是微软。这不仅包括“砸钱”，微软的投资里有很大一部分是算力。

ChatGPT的模型训练，大多是通过云计算完成的。OpenAI并不需要再单独购买芯片、服务器等硬件设备，而是用网线接入到云端服务器，提交自己的计算需求。如今ChatGPT火遍全球，算力的成本其实是微软在背。

2023年3月初，OpenAI宣布全面开放ChatGPT的API接口，用户可以直接调用ChatGPT做各种应用了。此外，它还降价了，这大幅度地降低了开发人员的使用门槛。

OpenAI在大幅亏损的情况下，还降价。有一种猜测是，微软在“放长线钓大鱼”，让ChatGPT的相关应用更快普及，等用户逐渐养成了使用习惯，未来从“算力”的角度来赚钱。就好像在如今，每家每户都要用电，而掌握了“算力”的公司，就有机会成为AI时代的“国家电网”。

收起

作者回复：
谢谢你的补充

☀朱谭王☀
5月1日

嗨～吴军老师，我们下一讲见。

Laha Ale
5月1日

具备这是所有的条件，似乎也为未必做好，亲自试用Google的Bard目前和ChatGPT还是差的远。这次留言我想问几个问题，如果有机会希望吴军老师回答一下，或许其他听众也感兴趣。

1. adept.ai 等公司似乎融资情况也蛮不错的，还有各种眼花缭乱的插件（比如可以写代码的，可以走Excel的等），不知道这些应用的价值有多大。您认为基于C

hatGPT 或者 LLM的应用在多大程度上会改变我们使用计算机？哪些方面应用改变会最大？您这么看理由是什么？

2. 把范围再扩大一点，目前Generative AI相关应用似乎很多，不久前我看到吴恩达他们整理的一个报告说，云服务都快支撑不起这些产品了，目前国内AIGC也同抄的样火热。您觉得Generative AI会是下一个区块链或者元宇宙吗？理由是？

3. 所谓顺势而为，很多人也相信“站在风口上，猪都能飞起来”，但是我好像从来不在风口。我本人经历了，移动互联网，云计算，和人工智能等几波浪潮，虽然自己从一个普通一个程序员，来美国读博，然后也到了哈佛做博后（这事还感谢吴军老师当时回答我问题的时候鼓励我去哈佛而不是去那小公司，来了我没有后悔，在做学术方面学到了不少），除了在学术上我蹭了一点AI，但商业上这些好像都和我无关，有时候看到一些消息我就心急。比如，最近看到李开复也蹭这波热度再亲自组建 Project AI 2.0, 我看了一下他的网站，基本上也是准做 AI 大模型、NLP、Multi-modality相关的产品，看到这些我就急。反观我导师R语言之父 Robert Gentleman，似乎就从来不蹭热度，踏踏实实做自己，甘于去做大家不愿意做的一些事情，也能成就非凡的事业，看到他这个样子我也在想是不是自己太心急，但是担心自己成为只是低头拉车，不会看路的人。 您觉得这是一个Generative AI是个趋势或者风口吗？您是如何去看的？我们该如何培养对这些机会洞察力和敏感度呢？又该如何去把握，利用好这一些。 这是我们见识和平时踏踏实实积累专业知识吗？

收起

GPT学习小组

转发 3 25 分享

作者回复：
你有一个好导师！

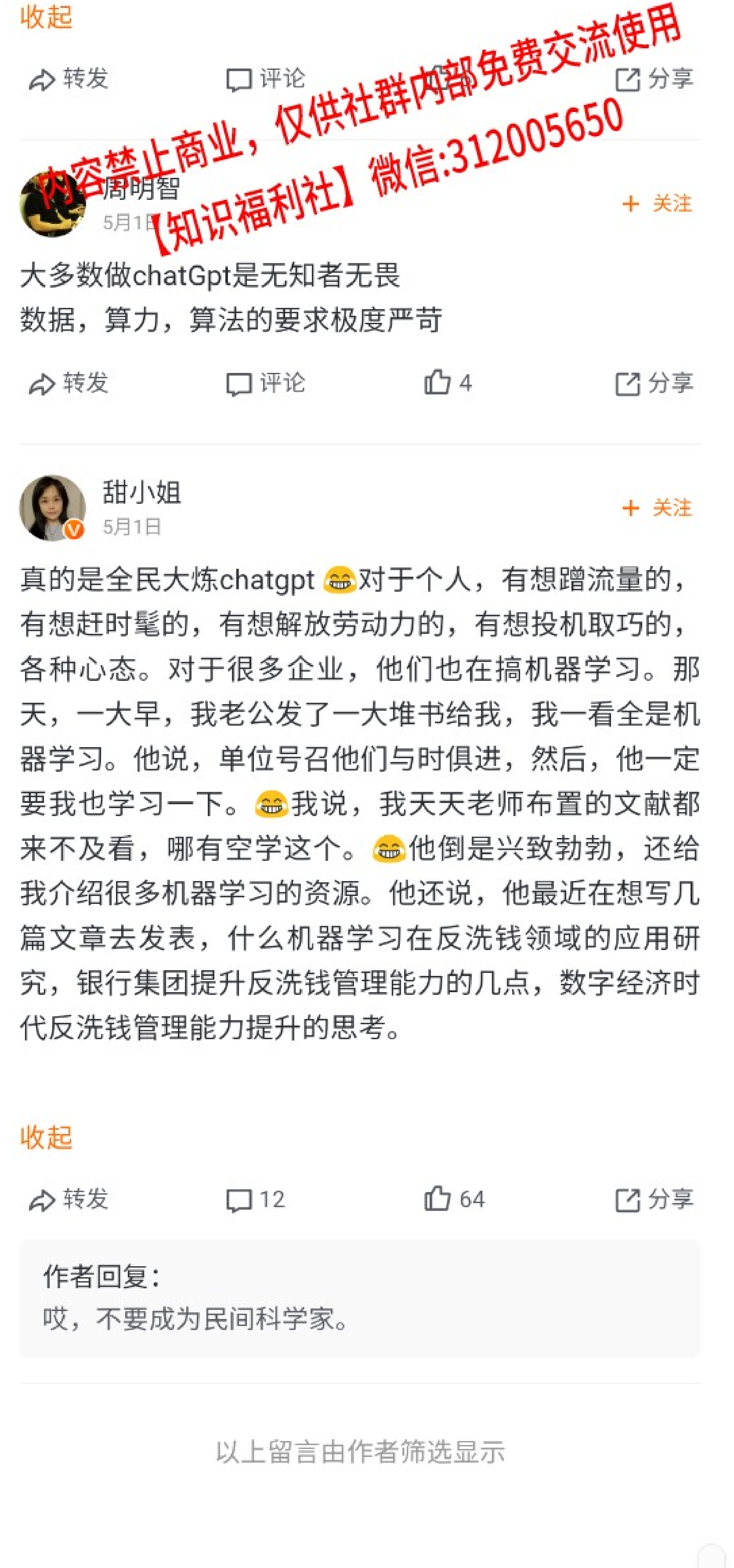
莫听
5月1日
+ 关注

数据、算力和算法构成了搭建chatgpt的三道门坎，而能够真正做好这三方面工作的，只有极少数公司。个人想要搭建一个高质量的chatgpt简直难如登天。

转发 评论 5 分享

得大自在佛
5月1日
+ 关注

ChatGPT现在还是在网络学习阶段，也就是人工智能的初级阶段，在问与答，通过大数据分析，也就是信息处理量，信息处理准确度上，和信息处理的延展性上下功夫，可能第二阶段是在第一阶段的基础上在添加人脸情绪情感识别判断与分析，与景象，图像识别做出判断与分析，做出反应，到了这一步就很危险了，人工智能的情感问题与编程人情感表达有关系，对于事物在人内心的反应是每个人是不一样的，而这个事物与情感的搭建标明人工智能可以用程序来表达，但编程人是低俗的或者是抑郁的，那人工智能的情感反应也是低俗的或者抑郁的。如果编程人员是佛系的那人工智能也是佛系的。我可能考虑的过了，当然人工智能的情绪情感反应应该是由语言算法来选择，呵呵，路漫漫其修远兮，……。



内容禁止商业，仅供社群内部免费交流使用
【知识福利社】微信:312005650

周明智
5月1日

大多数做chatGpt是无知者无畏
数据，算力，算法的要求极度严苛

甜小姐
5月1日

真的是全民大炼chatgpt😁对于个人，有想蹭流量的，有想赶时髦的，有想解放劳动力的，有想投机取巧的，各种心态。对于很多企业，他们也在搞机器学习。那天，一大早，我老公发了一大堆书给我，我一看全是机器学习。他说，单位号召他们与时俱进，然后，他一定要我也学习一下。😁我说，我天天老师布置的文献都来不及看，哪有空学这个。😁他倒是兴致勃勃，还给我介绍很多机器学习的资源。他还说，他最近在想写几篇文章去发表，什么机器学习在反洗钱领域的应用研究，银行集团提升反洗钱管理能力的几点，数字经济时代反洗钱管理能力提升的思考。

作者回复：
哎，不要成为民间科学家。

以上留言由作者筛选显示

【知识福利社】你的私人书院，微信：312005650 全网热门课程众筹共享！