高校图书馆发展论坛

大模型视域下的大资源共建与共享

黄晨

CADAL项目管理中心
数字图书馆教育部工程研究中心

2023.7 @ 哈尔滨

报告提纲

- “大”时代一瞥
- 大模型的幻觉
- 共建大资源
- 奇点 Vs. 终点

“大”时代—瞥

Philip W. Anderson

972, Volume 177, Number 4047 SCIENCE

More Is Different

Broken symmetry and the nature of the hierarchical structure of science.

P. W. Anderson

The reductionist hypothesis may still be a topic for controversy among philosophers, but among the great majority of active workers in the sciences I think it is accepted without question. The workings of our minds and bodies, and of all the animate or inanimate matter of which we have any detailed knowledge, are assumed to be controlled by the same set of fundamental laws, which except under certain extreme conditions we feel we know pretty well.

It seems inevitable to go on uncritically to what appears at first sight to be an obvious corollary of reductionism: that if everything obeys the same fundamental laws, then the only scientists who are studying anything really fundamental are those who are working on those laws. In practice, that amounts to some astrophysicists, some elementary particle physicists, some logicians and other mathematicians, and few others. This point of view, which it is the main purpose of this article to oppose, is expressed in a rather well-known passage by Weisskopf (*1*):

Looking at the development of science in the Twentieth Century one can distinguish two trends, which I will call “intensive” and “extensive” research, lacking a better terminology. In short: intensive research goes for the fundamental laws, extensive research goes for the explanation of phenomena in terms of known fundamental laws. As always, distinctions of this kind are not unambiguous, but they are clear in most cases. Solid state physics, plasma physics, and perhaps also biology are extensive. High energy physics and a good part of nuclear physics are intensive. There is always much less intensive research going on than extensive. Once new fundamental laws are discovered, a large and ever increasing activity begins in order to apply the discoveries to hitherto unexplained phenomena. Thus, there are two dimensions to basic research. The frontier of science extends all along a long line from the newest and most modern intensive research, over the extensive research recently spawned by the intensive research of yesterday, to the broad and well developed web of extensive research activities based on intensive research of past decades.

The effectiveness of this message may be indicated by the fact that I heard it quoted recently by a leader in the field of materials science, who urged the participants at a meeting dedicated to “fundamental problems in condensed matter physics” to accept that there were few or no such problems and that nothing was left but extensive science, which he seemed to equate with device engineering.

The main fallacy in this kind of thinking is that the reductionist hypothesis does not by any means imply a “constructionist” one: The ability to reduce everything to simple fundamental laws does not imply the ability to start from those laws and reconstruct the universe. In fact, the more the elementary particle physicists tell us about the nature of the fundamental laws, the less relevance they seem to have to the very real problems of the rest of science, much less to those of society.

The constructionist hypothesis breaks down when confronted with the twin difficulties of scale and complexity. The behavior of large and complex aggregates of elementary particles, it turns out, is not to be understood in terms of a simple extrapolation of the properties of a few particles. Instead, at each level of complexity entirely new properties appear, and the understanding of the new behaviors requires research which I think is as fundamental in its nature as any other. That is, it seems to me that one may array the sciences roughly linearly in a hierarchy, according to the idea: The elementary entities of science X obey the laws of science Y.

X	Y
solid state or many-body physics	elementary particle physics
chemistry	many-body physics
molecular biology	chemistry
cell biology	molecular biology
·	·
·	·
·	·
psychology	physiology
social sciences	psychology

But this hierarchy does not imply that science X is “just applied Y.” At each stage entirely new laws, concepts, and generalizations are necessary, requiring inspiration and creativity to just as great a degree as in the previous one. Psychology is not applied biology, nor is biology applied chemistry.

In my own field of many-body physics, we are, perhaps, closer to our fundamental, intensive underpinnings than in any other science in which nontrivial complexities occur, and as a result we have begun to formulate a general theory of just how this shift from quantitative to qualitative differentiation takes place. This formulation, called the theory of “broken symmetry,” may be of help in making more generally clear the breakdown of the constructionist converse of reductionism. I will give an elementary and incomplete explanation of these ideas, and then go on to some more general speculative comments about analogies at

The author is a member of the technical staff of the Bell Telephone Laboratories, Murray Hill, New Jersey 07974, and visiting professor of theoretical physics at Cavendish Laboratory, Cambridge, England. This article is an expanded version of a Regents' Lecture given in 1967 at the University of California, La Jolla.

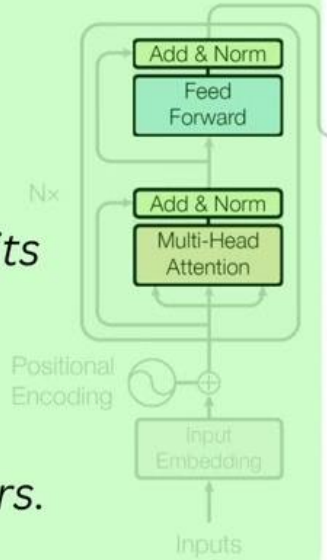
- 诺贝尔物理学奖得主、著名凝聚态物理学家菲利普·安德森于1972年在Science发表了题为“More is Different”的论文，指出还原论假说从来都不意味着建构论（constructionist）假说。

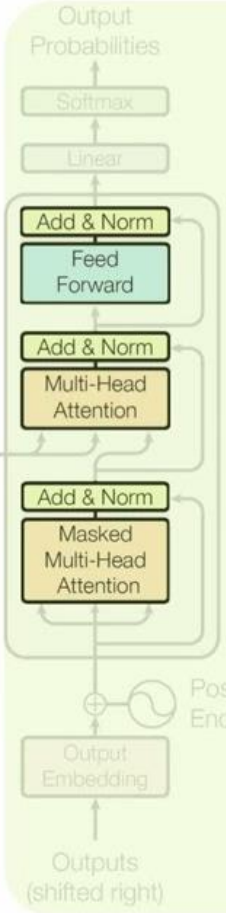
“大”时代—瞥

BERT

Google

use transfer learning to ***continue learning*** *from its existing data when adding user-specific tasks and layers.*





GPT

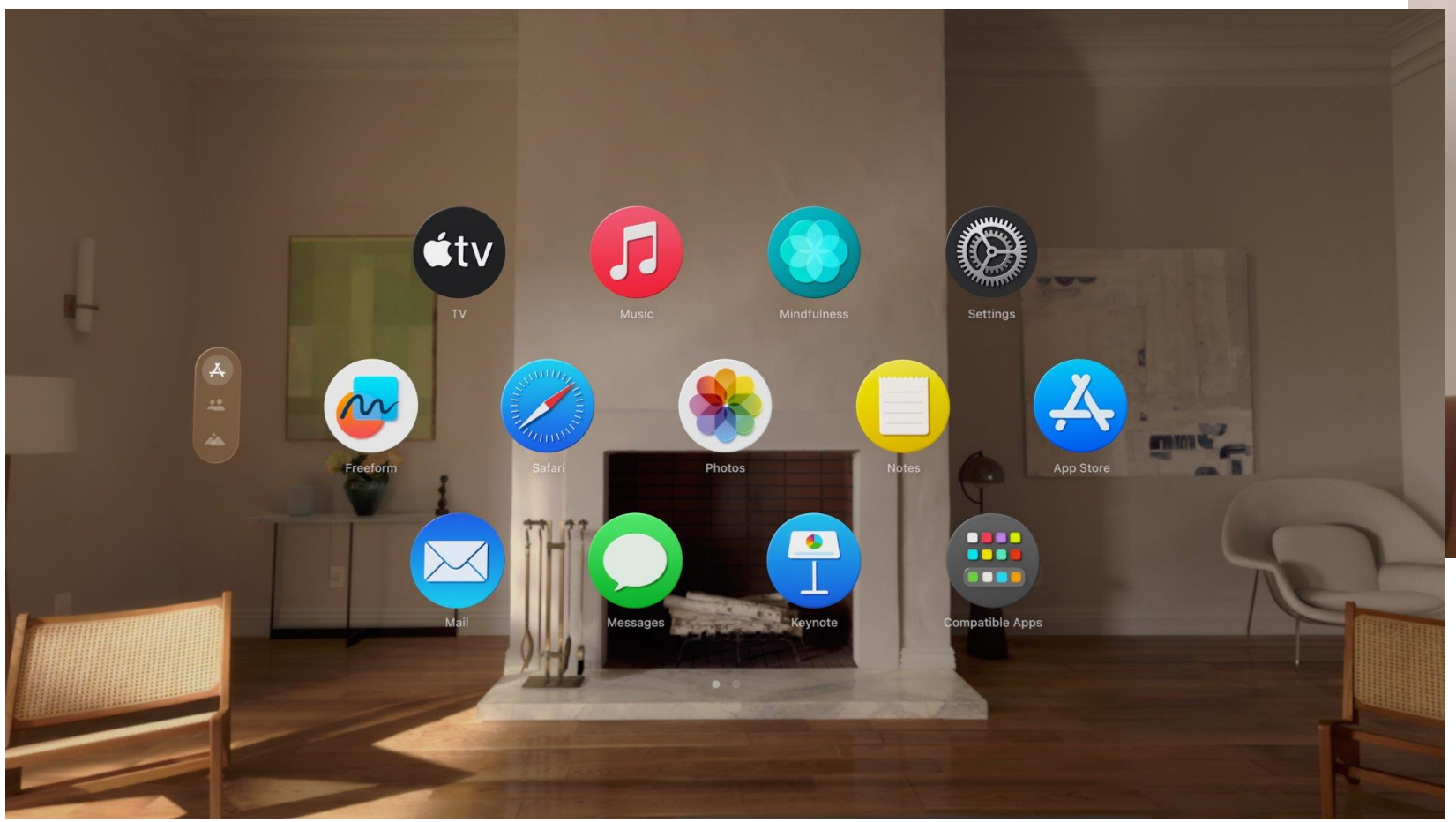
OpenAI

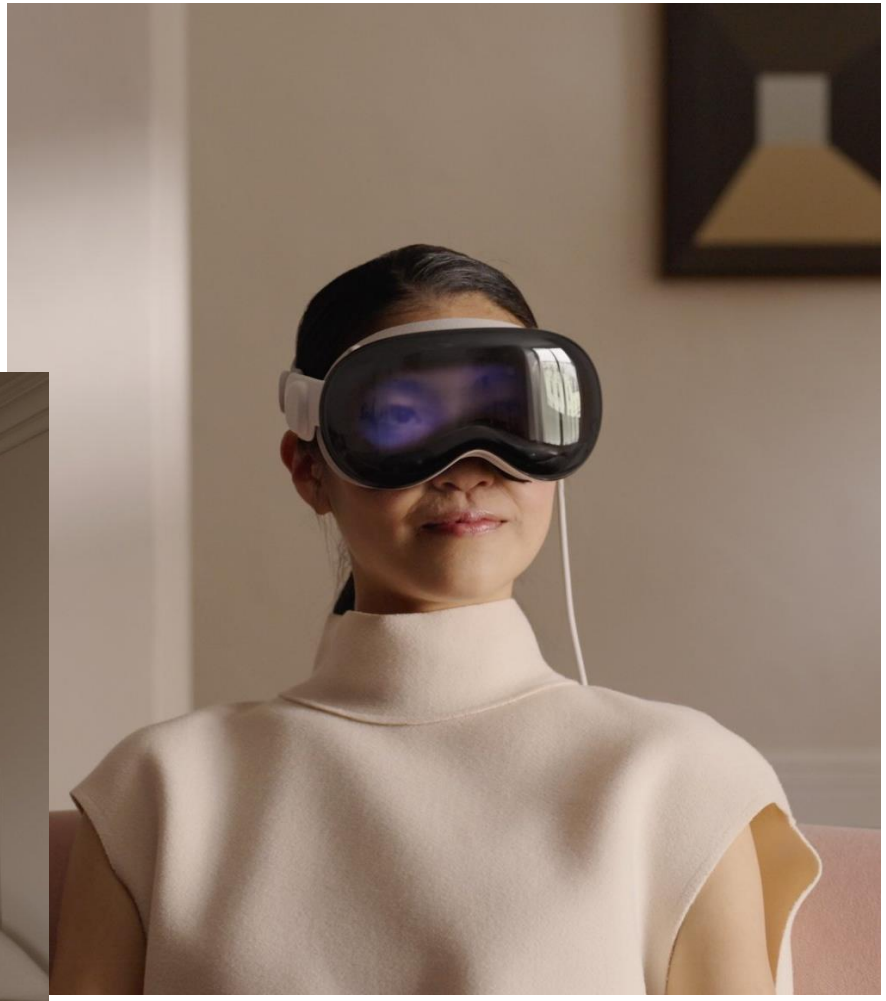
decodes from its massive pre-learned embeddings to present output that matches user prompts. It

“大”时代—瞥

- 英伟达 5 月 30 日发布Nvidia DGX GH200
- 由256个GH200“超级芯片”组成
- 运算速度达每秒1百亿亿次（ 1 exaflop）

“大”时代——瞥

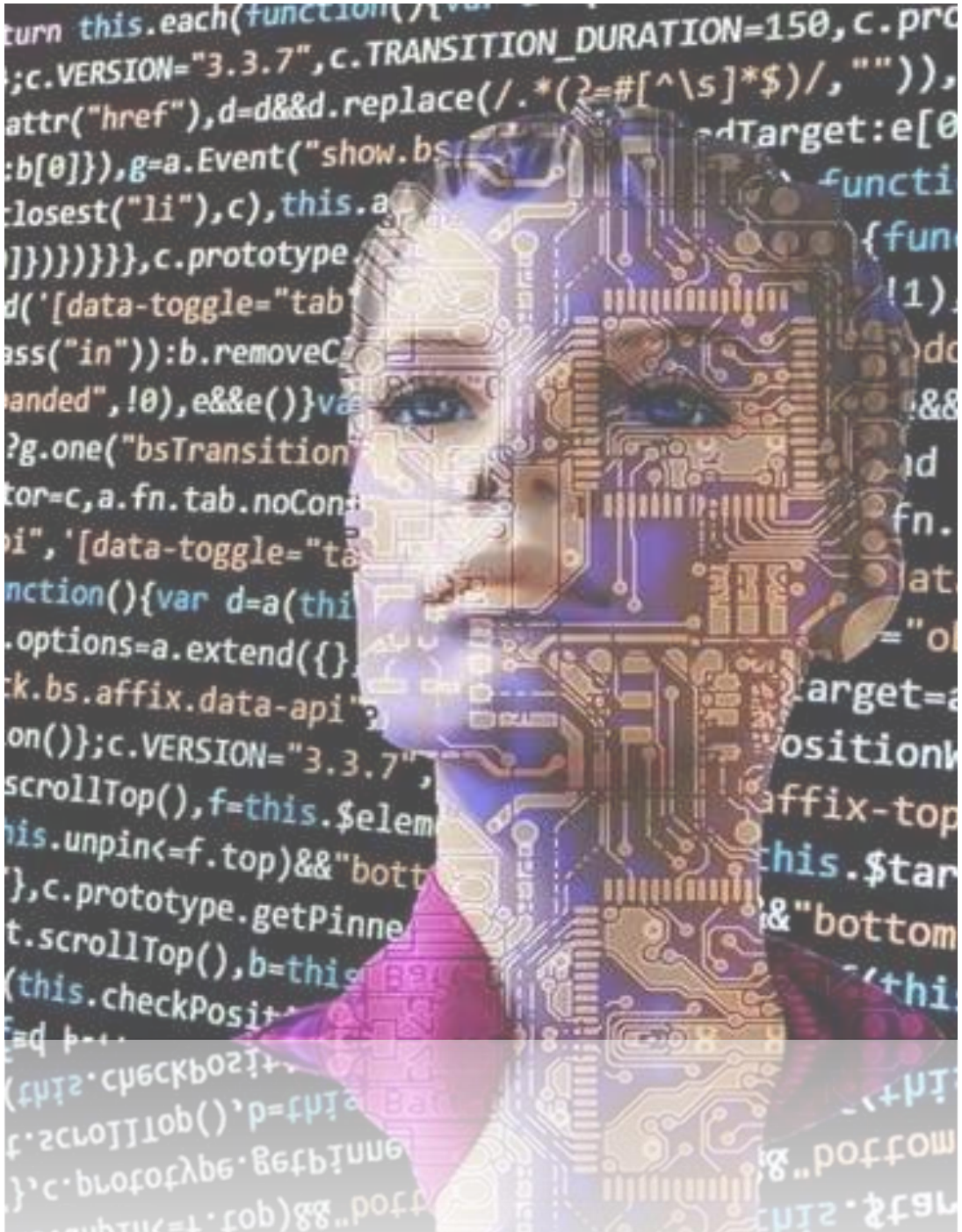




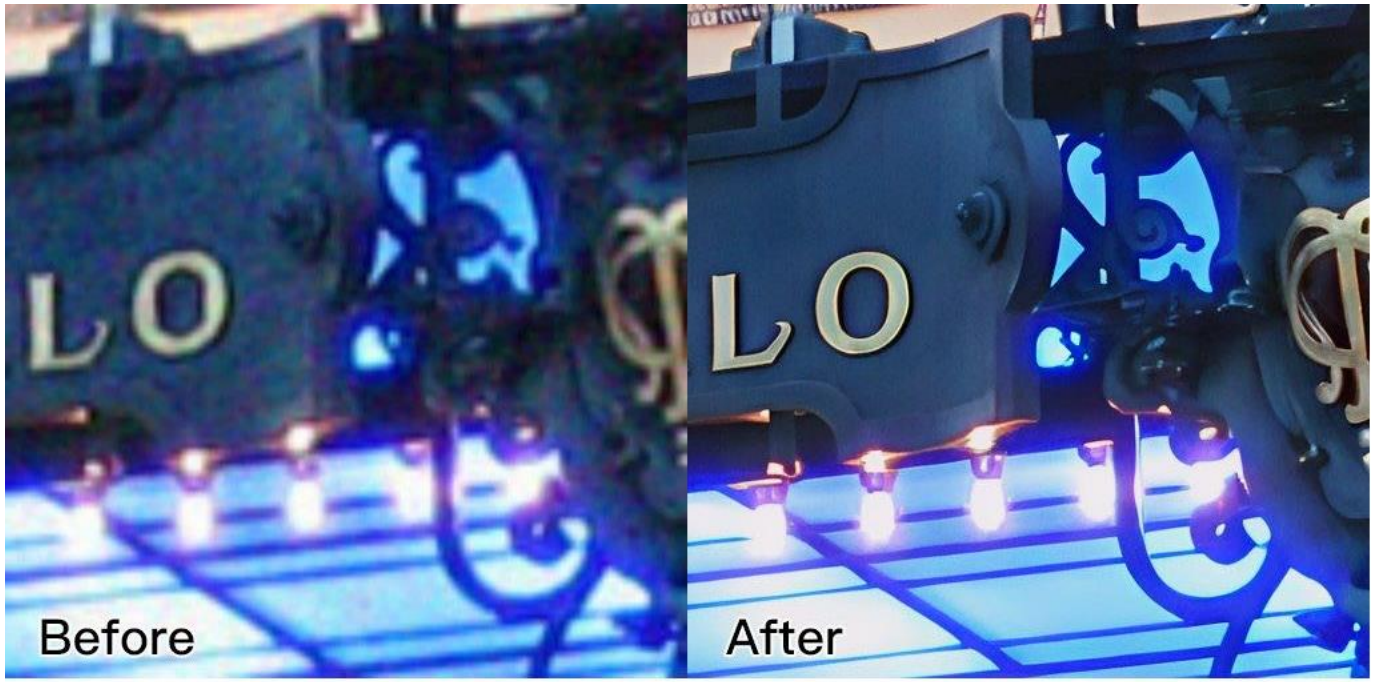
6 月 6 日，Apple 在 WWDC23 上发布了空间计算设备： Vision Pro

“大”时代——瞥

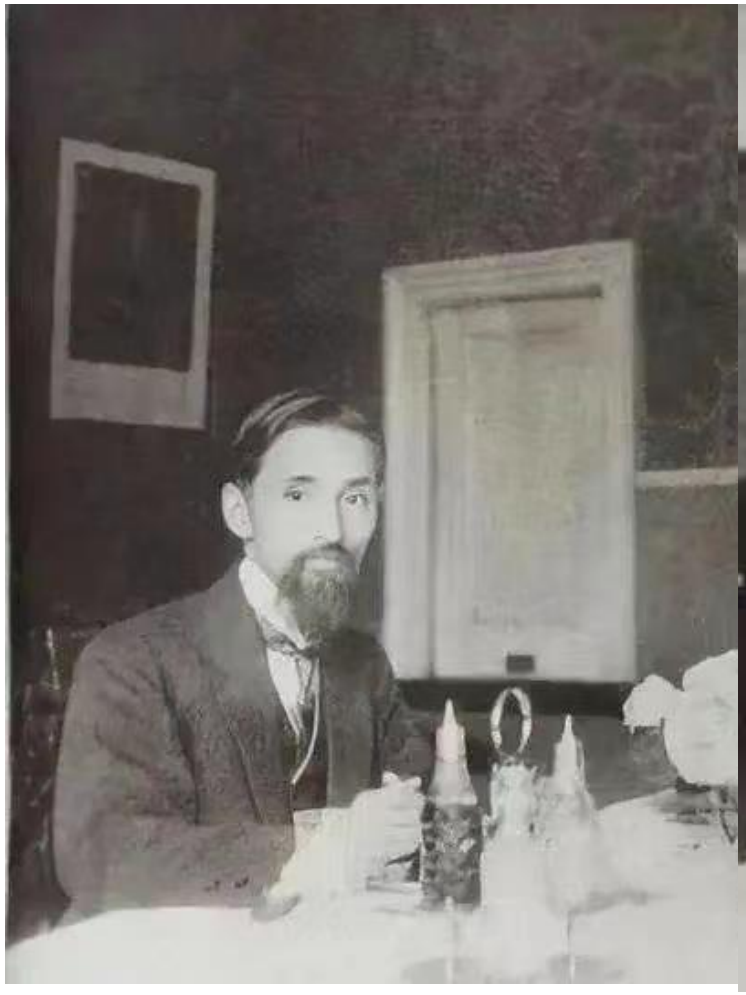
- 从图形界面进入对话界面
- 从桌面计算进入空间计算
- 从真伪难分进入人机莫辨
- 从零成本分发到零成本生产



“大”时代——瞥



“大”时代—瞥

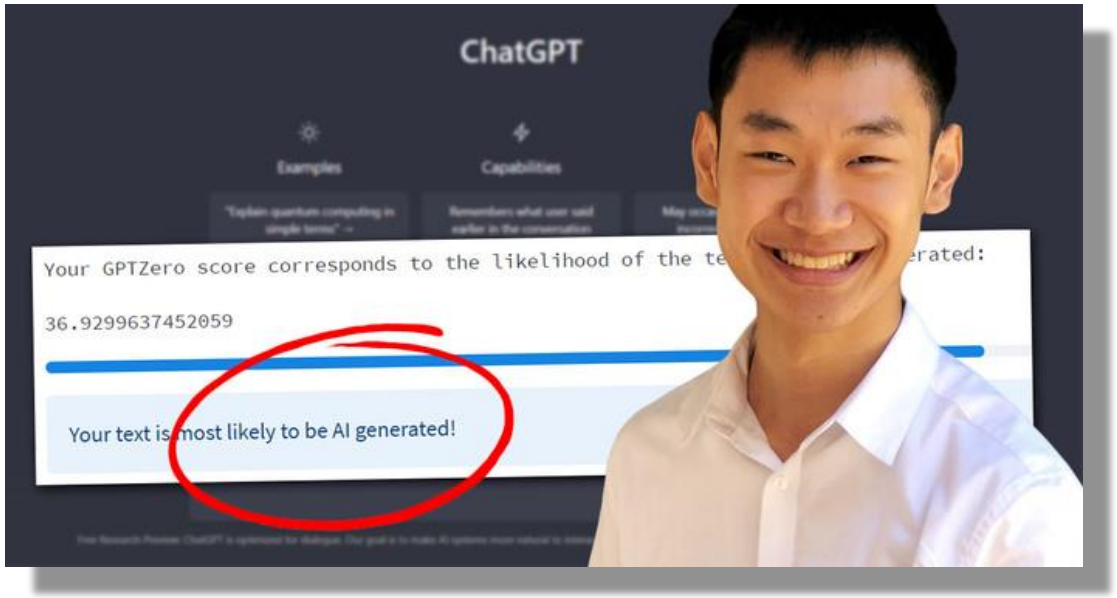


AIGC技术场景

- 文本生成
  - 非交互式文本
    - 结构化写作 (新闻播报等，有比较强的规律)
    - 非结构化写作(剧情续写、营销文本等，需要一定创意和个性化)
    - 辅助性写作 (推荐相关内容、帮助润色，不属于严格AIGC)
  - 交互性文本
    - 闲聊机器人 (虚拟男/女友、心理咨询等)
    - 文本交互游戏等 (AI dungeon等)
- 音频生成
  - 语音克隆
  - 文本生成特定语音 (生成虚拟人歌声/播报等)
  - 乐曲/歌曲生成 (包含作曲及编曲，在实际应用中常包含自动作词)
- 视频生成
  - 视频属性编辑 (副除特定主体、生成特效、跟踪剪辑等)
  - 视频自动剪辑 (对特定片段进行检测及合成)
  - 视频部分编辑 (视频换脸等)
- 图像、视频、文本间跨模态生成
  - 文字生成图像 (根据文字prompt生成创意图像)
  - 文字生成演示视频 (拼接图片素材生成视频)
  - 文字生成创意视频(完全从头生成特定主题视频)
  - 图像/视频到文本 (视觉问答系统、自动配字幕/标题等)

特图、logo等)

预计 2025 年 AIGC 产生的数据将占据人类全部数据的 10%



当文字不是人类写的时，人类应该有权知道。

——Edward Tian，普林斯顿大学

它满足了人类的希望任何事情都有工具的“程序性”思维。各种论文查重软件在人工智能时代也主要扮演这个作用，就是按程序免责作用。不过解决心理需求可能是未来各种应用的主要价值。

- GPTZero 基于两个指标——perplexity （困惑度）和 burstiness （突发性）的分值，根据统计学特征来判断一篇文章是不是AI写的。
- perplexity 衡量了文章语言的复杂性和随机性，如果 GPTZero 对文本感到困惑，那么该文本具有很高的复杂性，更有可能是人工编写的。
- burstiness 则衡量了文章中句子复杂性的变化，人类倾向于写高度复杂的文本，如长短句交错，其突发性更高，而 AI 生成的文本多为短句，复杂度较低。

报告提纲

- “大”时代一瞥
- 大模型的幻觉
- 共建大资源
- 奇点 Vs. 终点

大模型的幻觉

- Chatgpt只是互联网的一张模糊缩略图

—— 特德•姜

Illustration by Vivek Thakker

ANNALS OF ARTIFICIAL INTELLIGENCE

CHATGPT IS A BLURRY JPEG OF THE WEB

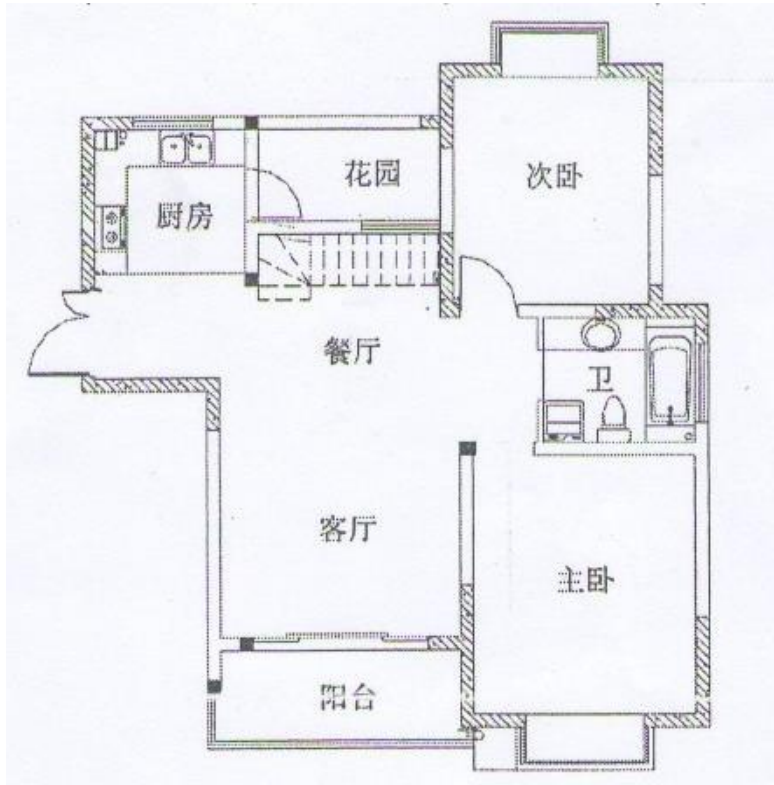
OpenAI's chatbot offers paraphrases, whereas Google offers quotes. Which do we prefer?

By Ted Chiang

February 9, 2023

大模型的幻觉

- 2013年，德国一家建筑公司的工人们注意到了他们的Xerox复印机有些奇怪的地方：当他们复制一幢房屋的平面图时，复印件与原件有微妙但重要的不同。在原始平面图中，房屋的三个房间各有一个矩形指定其面积：分别为14.13，21.11和17.42平方米。然而，在复印件中，所有三个房间都被标记为14.13平方米的大小。



大模型的幻觉

- Xerox复印机使用一种称为jbig2的有损压缩格式，专为黑白图像使用。为了节省空间，复印机在图像中识别相似的区域，并存储一个副本；当文件被解压缩时，它使用该副本多次重建图像。
- 复印机以微妙的方式降低图像质量，其中压缩工件不能立即识别。如果复印机只是生成模糊的印刷品，每个人都会知道它们不是原件的精确复制。导致问题的是复印机生成的数字是可读的但不正确的；它使得复制品似乎是准确的，但实际上不是。（2014年，Xerox发布了一个补丁来解决这个问题。）


因为文本被高度压缩，你不能通过搜索精确的引用来寻找信息；因为存储的不是词语，所以你永远不会得到精确的匹配。为了解决这个问题，你创建了一个界面，接受以问题形式的查询，并以表达你在服务器上拥有的东西的答案作为回复。

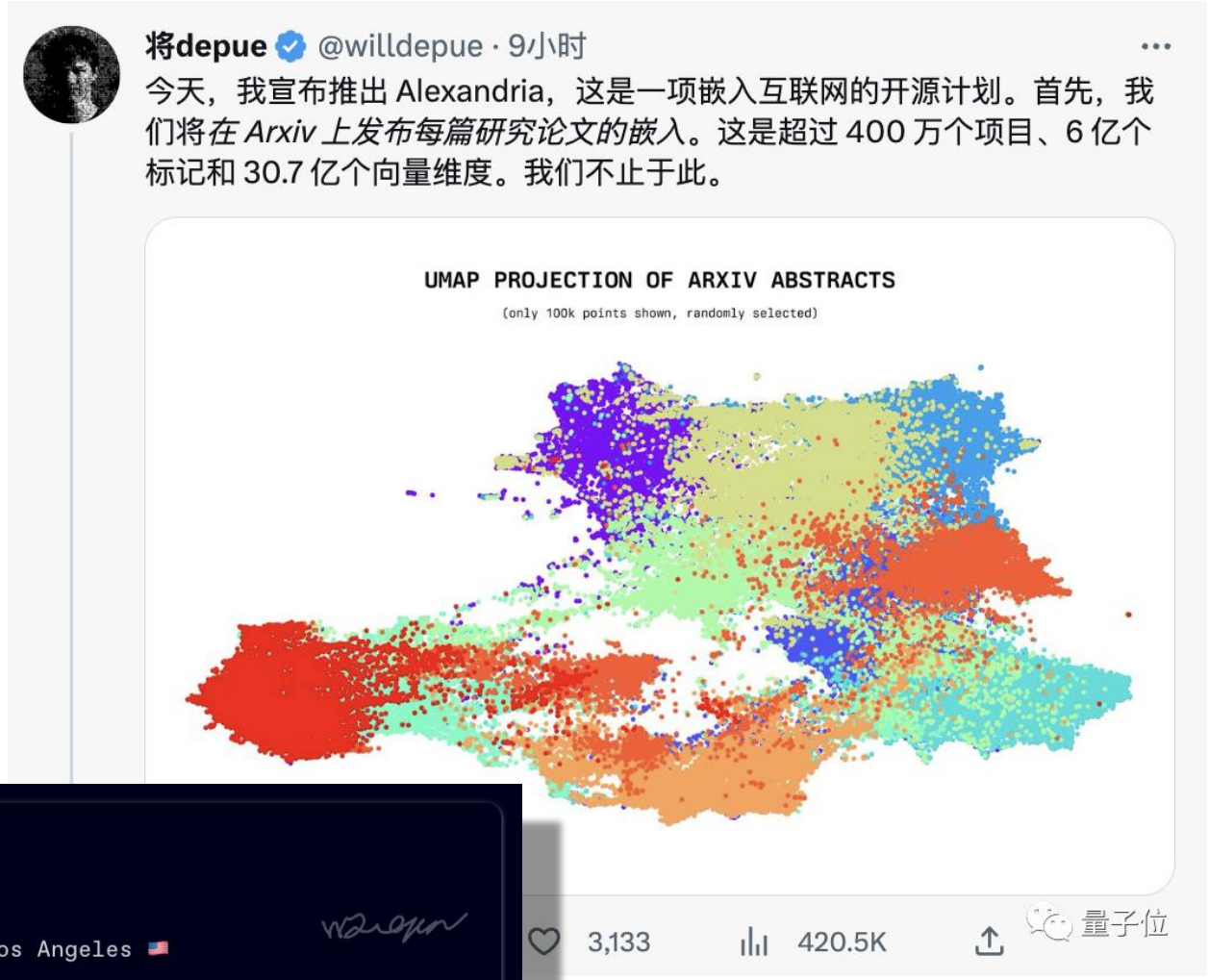
大模型的幻觉

- 把ChatGPT想象成所有网页文本的模糊的jpeg。它保留了网上大量的信息，就像高分辨率图像的jpeg保留了大量信息一样，但是，如果你正在寻找一系列比特，你将永远找不到它；您将永远得到一个近似值。但是，因为近似值以语法文本的形式呈现，ChatGPT擅长创建语法文本，它通常是可以接受的。您仍在查看模糊的jpeg，但模糊发生的方式不会使图片整体看起来更模糊。
- 对有损压缩的类比可以帮助我们理解像我这样的大型语言模型的局限性和潜在的不准确性。就像有损压缩算法会在重建的图像中引入错误和伪影一样，大型语言模型在回答问题时也会产生无意义的或不准确的回答。
- 像我这样的大型语言模型是为了生成连贯的文本，而不一定是准确或可信的信息。

https://mp.weixin.qq.com/s/OxzFnga8pSov2VWTM5Aycw

大模型的幻觉

- 5月26日，最新爆火开源计划**亚历山大**将Arxiv上所有论文转成Token，加起来不过14.1GB而已。
- 他们想要将整个互联网变成**Tokens**，换言之全都转化成ChatGPT等大模型理解这个世界的方式。



大模型的幻觉

- 大模型面临拐点
- “压缩”造成“模糊”
- AIGC 带来语料污染
- 中文资源稀缺形成思维定式

报告提纲

- “大”时代一瞥
- 大模型的幻觉
- 共建大资源
- 奇点 Vs. 终点

共建大资源

- 何为大资源？
  - 广而博 资源容量巨大
  - 约而精 学科知识精深
  - 跨媒体 媒体类型齐全
  - 多机构 GLAM and More……

共建大资源

“人工智能刚刚破解了人类文明的操作系统。在历史上的每一种人类文化中，操作系统始终是语言。一开始就有了话语。我们用语言创造神话和法律，创造神和金钱，创造艺术和科学，创造友谊和国家。”

“在几年内，AI可能会消化掉整个人类文化，消化掉我们几千年来创造的一切，然后开始大量产出新的文化创作，新的文化物品。我们要记住，我们人类从未真正直接接触到现实，我们总是被文化包围，我们总是通过文化的棱镜来体验现实。”

—— 尤瓦尔·赫拉利4月29日于Frontier论坛

据，在

文网站
文是

况统计
，中国
下降了

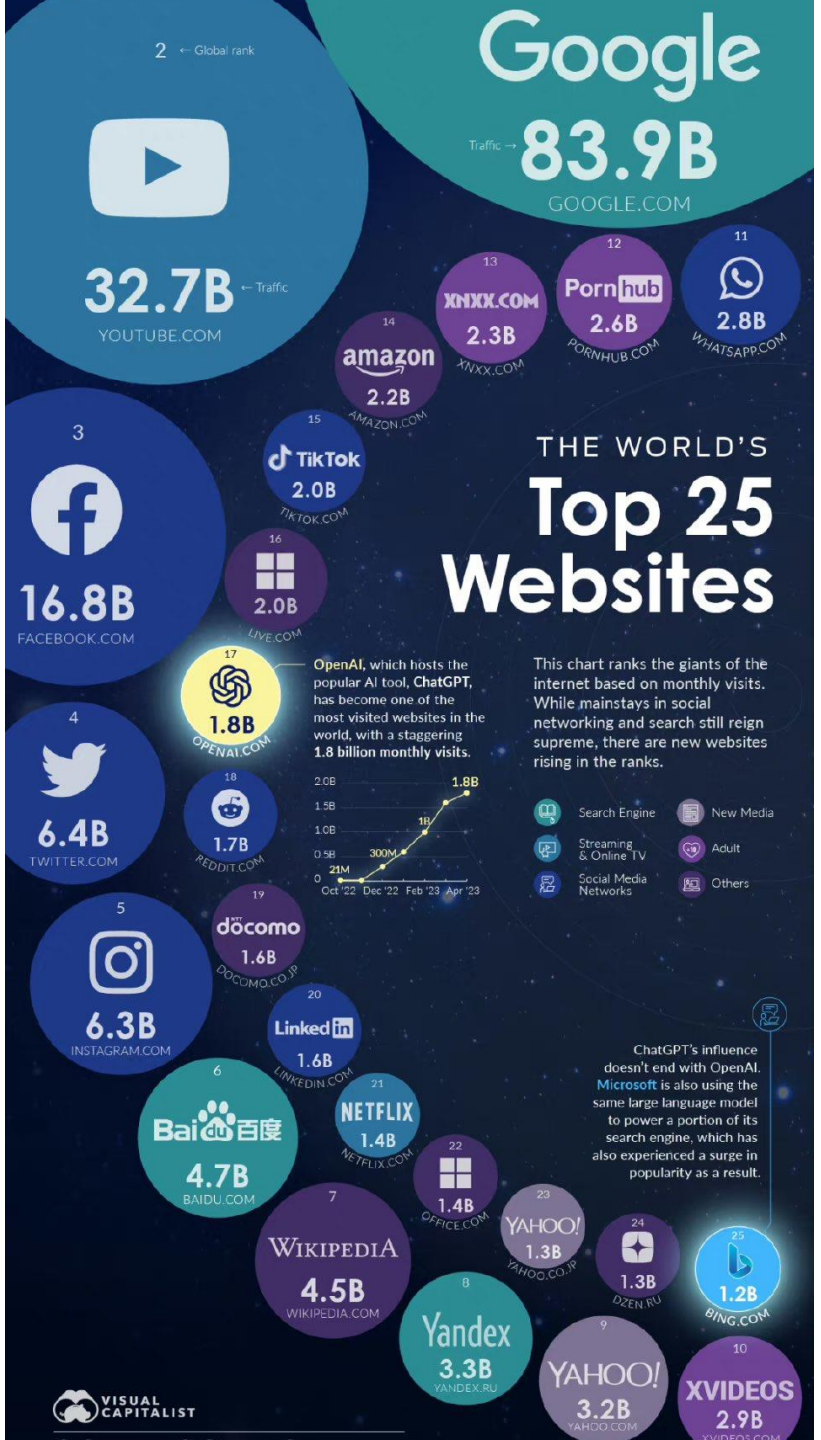
名次	语言	百分比
1	英语	59.3%
2	俄语	8.4%
3	西语	4.2%
4	德语	2.9%
5	土耳其语	2.9%
6	波斯语	2.8%
7	法语	2.8%
8	日语	2.4%
9	葡萄牙语	2.2%
10	中文	1.3%
11	越南语	1.3%
12	意大利语	1.0%
13	阿拉伯语	0.9%
14	波兰语	0.9%
15	希腊语	0.7%

截至2020年3月25日， W3Techs 预测前百万互联网网站使用的语言文字百分比

共建大资源

- 全球网站排名榜，根据每月访问量来排名。主要的类别包括社交媒体、搜索引擎、成人网站、其他类型和新媒体。

1. Google以839亿的访问量位居第一，YouTube（327亿）、Amazon（22亿）、Pornhub（23亿）和WhatsApp（28亿）也在排名前列。
2. **OpenAI**，作为ChatGPT的主机，已经成为全球访问量最大的网站之一，每月有惊人的18亿次访问。
3. 微软在使用这种大型语言模型来驱动其 http://Office.com 的一部分，使得它在流行度上有所上升。
4. 百度和必应作为搜索引擎，分别有47亿和12亿的访问量。
5. 除了OpenAI，TikTok、Facebook、Twitter、Instagram、LinkedIn等社交媒体网站也在榜单中占据了重要位置。



共建大资源

中文世界的数字图书馆联合起来

- **共建标准、可信、开放的中文大资源，为大模型提供可靠的中文知识和中华逻辑；**

大学数字图书馆国际合作计划
CHINA ACADEMIC DIGITAL ASSOCIATIVE LIBRARY

全部 请输入搜索内容

CADAL

共建大资源

中文世界的数字图书馆联合起来

- 共建标准、可信、开放的中文大资源，为大模型提供可靠的中文知识和中华逻辑；
- **共建垂直学科的特藏资源库，为领域模型提供专业知识与数据；**

共建大资源

中国写本文献数字资源库于2022年6月21日正式对外发布

https://xieben.cadal.edu.cn/

敦煌文獻

筛选

【收藏機構】…

线上可获取 38
所有资源 39

【收藏機構】
中國國家圖書館

【寫本作者】
沙門玄奘奉詔譯（首題）
三藏法師玄奘奉詔譯（首題）
三藏法師玄奘譯
彌勒菩薩説沙門玄奘奉詔譯…
未知
更多

【寫本年代】
856年左右

39 条相关结果

按名称首字母排序

《瑜伽師地論》卷二一

【寫本編號】：BD1324　【千字文號】：張024　【縮微膠捲號】：7193
【寫本年代】：856年左右
尾題“瑜伽師地論卷第廿一”。

瑜伽師地論分門記

【千字文號】：鳥093　【縮微膠捲號】：7293
中國國家圖書館

在綫瀏覽

本地资源浏览
（缩微胶卷）

《瑜伽師地論》卷二一

♥加入收藏　<分享

【寫本編號】：	BD1324
【数字化编号】：	ZC-07362
【千字文號】：	張024
【縮微膠捲號】：	7193
【收藏機構】：	中國國家圖書館
【寫本年代】：	856年左右
【寫本概況】：	卷軸裝，3紙。前殘尾全，存68行（前5行上部殘，6—11行中部…且殘損程度逐漸減少），行約17字。尾題“瑜伽師地論卷第廿一”。
【寫本綴合】：	BD9665號＋BD1324號　图像　链接
【彩色圖版】：	【出處01】　图像　链接
【黑白圖版】：	【出處01】《中國國家圖書館藏敦煌遺書》106冊176頁。　图像　链接
	【出處02】《中國國家圖書館藏敦煌遺書》20冊62頁-63頁　图像　链接
【寫本録文】：	【録文01】　图像　链接
【相關寫本】：	【寫本01】上圖171號　图像　链接
	【寫本02】BD15391號　图像　链接
【參考文獻】：	[01]张涌泉、徐鍵 《瑜伽師地論》系列敦煌殘卷綴合研究 安徽大學學報(哲學社會科學版)2015年第3期 p72 https://kns.cnki.net/kcms/detail/detail.aspx? [更多]

本地图像
链接

出处网站
链接

不同的版
本类型

引文链接

共建大资源

中文世界的数字图书馆联合起来

- 共建标准、可信、开放的中文大资源，为大模型提供可靠的中文知识和中华逻辑；
- 共建垂直学科的特藏资源库，为领域模型提供专业知识与数据；
- **共建古籍资源库，为 AI 时代的文化传承构建最坚实的中文资源基础设施。**

共建大资源

- 与古人对话
  - 可以讓用戶選擇不同的對象，孔子、老子、朱熹……通過自然語言對話。
  - 每一個對象依據其所處的時代與典籍空間，依據其自身作品的知識空間，嚴格遵循人物的時代約束，回復用戶對於中國文化的疑問；
  - 甚而可以與對象探討其作品的內涵與意義。

报告提纲

- “大”时代一瞥
- 大模型的幻觉
- 共建大资源
- 奇点 Vs. 终点


BAAI
智源研究院
BAAI CONFERENCE
2023
6.9—6.10
北京智源大会
BIG IDEA ON
LARGE-SCALE MODELS
谢晓亮
张钹
Stuart Russell
张宏江
David Baker
Sam Altman
Geoffrey Hinton
Max Tegmark
Ariel Warshel
姚期智
杨立昆
Joseph Sifakis
郑南宁
大会官网注册参会
智源研究院
新智元

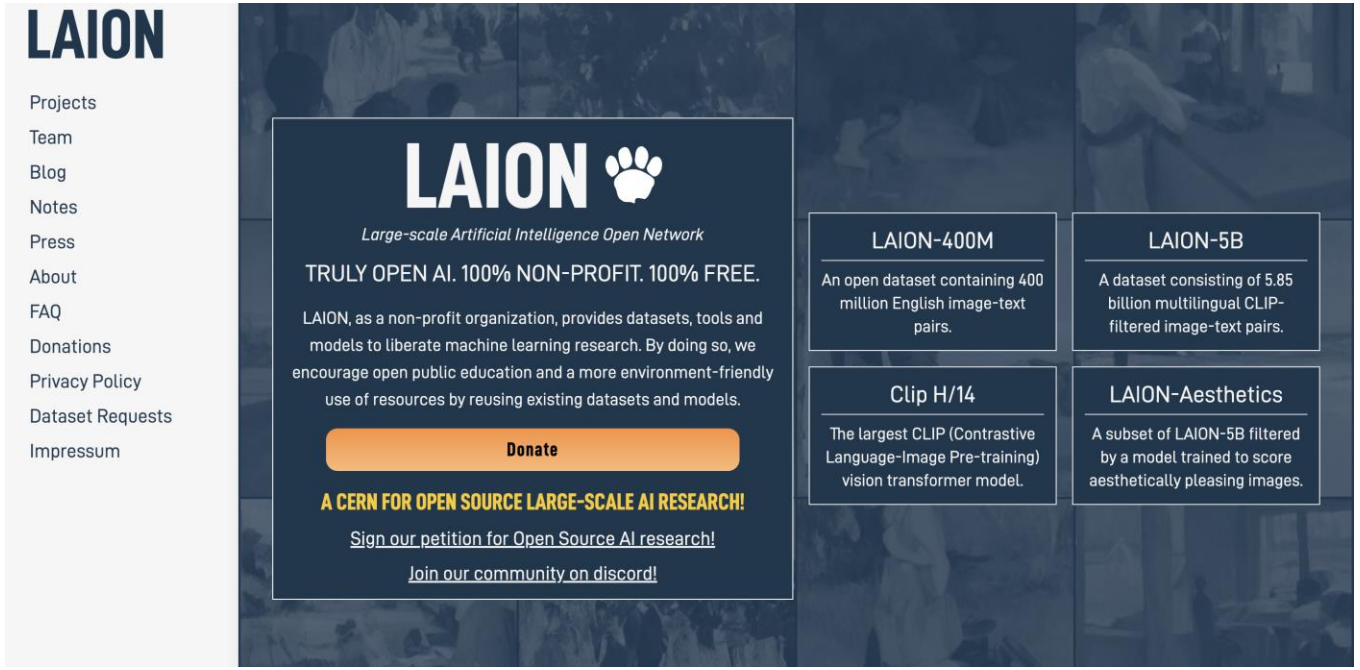
奇点 Vs. 终点

“我看不出如何防止这种情况发生，但我老了。我希望像你们这样的许多年轻而才华横溢的研究人员会弄清楚我们如何拥有这些超级智能。”

—— Geoffery E. Hinton @BAAI Conference2023
2023.06.10

THE END

奇点 Vs. 终点



- 他主持领导了包括LAION-5B、LAION-3D、Open-Assistant（开源对话数据集）等一系列AI预训练数据集的构建工作，并发布了Open-CLIP、NSFW Detection等一系列大模型。值得一提的是，所有参与者均为零工资自愿贡献，并不会从中获取经济收益。

• 我已经有足够的钱过上好日子了，现在税后月收入4500欧元，2000欧元用来付房租。假如微软来找我，工资翻倍，我的孩子们会更快乐吗？即使我能立刻去洛杉矶或硅谷住大豪宅，孩子们也不稀罕。

谢 谢!

Mail to: Beside.huang@gmail.com