



# 人工智能文集

## 第十八集

2024年8月

# 人工智能文集

## 第十八集

中国开源软件推进联盟

China OSS promotion Union

# 目录

## 一、AI 认知的哲学性颠覆

(辛顿在爱尔兰都柏林大学的演讲).....GeoffreyHinton 2024.4.8

## 二、2023 人工智能十大趋势.....李飞飞团队发布 2024.5.3

## 三、评奥特曼的闭源策略.....陆首群 2024.6.18

## 四、模型开放性框架(MOF).....LF-AI, 2024.6.20

## 五、发展基于开源的人工智能.....陆首群 2024.7.16

## 六、北京 AI 安全国际共识.....Yoshua Bengio, Geoffrey Hinton, 姚期智 2024.03.18

## 七、中外人工智能专家共话通用人工智能机遇与挑战

2024 北京智源大会报导.....2024.7.23

## 八、如何从生成式人工智能(GAI)转移到通用人工智能(AGI)轨道上来...陆首群 2024.07.02

## 九、加速计算与生成式人工智能重塑行业

(推进绿色新工业革命).....黄仁勋 2024.7.5

## 十、人工智能发展中面临的"四大"(大参数\大算力\大能源\大投资)挑战.....陆首群 2024.7.2

## 十一、改善大模型预训练的工作环境.....COPU 2024.7.16

## 十二、能源短缺正在向人工智能发展提出挑战.....COPU 2024.7.16

## 十三、开源是 AI 前进的道路

(在 AI 安全问题上闭源无法代替开源).....Meta 报导 2024.7.24

## 十四、评奥特曼的“闭源 AI”、关停 API、割裂全球 AI 的言论和策略....陆首群 2024.7.30

# AI 认知的哲学性颠覆

(辛顿在爱尔兰都柏林大学的演讲)

Geoffery Hinton 2024.4.8

2024 年 4 月 8 日，人工智能大师辛顿在爱尔兰的都柏林大学的演讲，辛顿关于人工智能的观点，对当代哲学几乎所有理论有颠覆性影响。

这一天都柏林大学在欧瑞利大厅举行了尤利西斯勋章的颁奖仪式，授予了杰弗里辛顿教授尤利西斯勋章。辛顿在颁奖仪式上发表了有关人工智能的演讲，在演讲中他介绍了人工神经网络的工作原理和发展历程，他再次反驳了人工智能研究领域的符号主义，尤其是强烈反驳了乔姆斯基关于语言不是通过学习获得的观点，乔姆斯基的理论认为语言的本质在于句法，并且语言是天生的，有一种天生结构，不是通过学习获得的，辛顿指出这种观点误导了几代语言学家，他强调语言显然是通过学习获得的，并且大型神经网络现在能够通过大量数据和学习过程来掌握语言，辛顿解释道，语言学家贝诺姆乔姆斯基误导了好几代人，他也获得了著名的奖章，他有一个疯狂的理论，认为语言不是通过学习获得的，他成功地说服了许多人，这一点实际上语言显然是通过学习获得的，现在这些大型神经网络正在学习语言，他们不需要任何先天结构，只是从随机权重和大量数据开始精准详细解释了人类理解的本质，他认为理解不仅仅是通过符号和规则的操作，而是通过学习特征和特征之间的交互来实现的。

这种观点挑战了传统的符号主义理论，符号主义认为智能行为可以通过操作符号和规则来实现它，指出重要的是所有的知识都在于为词语分配特征以及这些特征如何交互，它不存储任何句子，但它可以重构句子，通过反复预测下一个词语来生成句子，这些大型语言模型的工作方式与此类似，它们不存储任何文本，而是从文本中学习为词语分配特征，并通过这些特征之间的交互来预测下一个词语的特征。

针对有人认为大型语言模型只是美化的文字自动补全的观点，辛顿反驳道。语言学家说他们只是美化的自动补全，他们只是使用统计技巧，只是将文本拼凑在一起，但请记住他们不存储任何文本，自动补全的反对意见是疯狂的，因为他们依赖的是自动补全的想法，他进一步解释现代大型语言模型，通过学习词语特征及其交互来进行预测，而不是存储词语字符串这些系统中的自动补完全不是这样工作的，他将词语转换为特征，并使用特征之间的交互来进行预测。

所以大型语言模型的工作方式和我们人类一样，我们看到很多文本或听到很多词语字符串，我们学习词语的特征，学习这些特征之间的交互，这就是理解的本质，而这些模型以与我们人类完全相同的方式进行理解，他在回答观众提问时说，举个例子，假如你有一个多模态的聊天机器人，你让他画一幅戴红帽子的仓鼠的画，他画了一幅戴红帽子的仓鼠的画，很难说他不理解这意味着什么？在我看来，如果他画了一幅戴红帽子的仓鼠的话，他就应该理解那是什么？同样如果你有一个机器人，你说打开抽屉他打开了抽屉，这对我来说是非常实际的，真的看起来像是理解，所以我实际上非常相信这些东西以我们人类相同的方式理解，我在以前的视频中曾经提到了辛顿关于意识的观点，他严厉批评了传统意识理论中的主观主义理论及内心剧场观念，即认为我们心中有一个只有自己能访问的舞台，在那里展现我们各自的心理体验，他认为这种观念是错误的，应该从功能性的角度来理解意识，他解释说当我们描述我们的感知体验时，实际上我们是在描述我们的感知系统，如何解释外部世界，而不是涉及任何神秘的内部状态。

辛顿认为人工智能也有自己的主观体验，比如一个透镜对于物体的折射改变了物体的位置，AI 同样会说出这种错误的位置，这就是人类主观体验的本质。他指出 AI 系统也可以类似方式报告其对数据的处理结果，从而具有某种形式的主观体验。

在这次演讲中，辛顿还深刻表达了 AI 对社会的影响，在被问到 AI 是否会

接管人类，他说我的猜测是要么它会接管，要么我们会学与它共存。在被问到您认为 AI 将如何改变社会的价值观?他说我对此有一个相当可怕的想法，即在工业革命之前，如果你有强壮的肌肉，并且擅长挖沟渠，你是有价值的，而在工业革命之后你就不再有价值了，因为机器可以做得更好，让我非常担心的是普通智力和普通教育水平的人目前被重视，因为他们可以做许多有用的事情，如果 AI 可以更好更便宜的做这些事情，他们将不再被重视，这对我来说在社会中似乎是非常危险的。

# 2023 人工智能十大趋势

李飞飞团队发布，2024.5.3

李飞飞团队发布《2024 年人工智能指数报告》，该报告揭示了 2023 年人工智能行业的 10 大主要趋势：

## 1. 人工智能在某些任务上胜过人类，但并非在所有任务上。

人工智能已在多项基准测试中超越人类，包括在图像分类、视觉推理和英语理解方面。然而，它在竞赛级数学、视觉常识推理和规划等更复杂的任务上依然落后于人类。

## 2. 产业界继续主导人工智能前沿研究。

2023 年，产业界产生了 51 个著名的机器学习模型，而学术界只贡献了 15 个。2023 年，产学合作还产生了 21 个著名模型，创下新高。此外，108 个新发布的基础模型来自工业界，28 个来自学术界。

## 3. 前沿模型变得更加昂贵。

根据 AI Index 的估算，最先进的人工智能模型的训练成本已经达到了前所未有的水平。例如，OpenAI 的 GPT-4 估计使用了价值 7800 万美元的计算资源进行训练，而谷歌的 Gemini Ultra 的计算成本则高达 1.91 亿美元。

相比之下，几年前发布的一些最先进的模型，即原始 transformer 模型（2017 年）和 RoBERTa Large（2019 年），训练成本分别约为 900 美元和 16 万美元。

## 4. 美国成为顶级人工智能模型的主要来源国。

2023 年，61 个著名的人工智能模型源自美国的机构，超过欧盟的 21 个和中国的 15 个。

美国也仍然是人工智能投资的首选之地。2023 年，美国在人工智能领域的私人投资总额为 672 亿美元，是中国的近 9 倍。然而，中国依然是美国最大的竞争对手，中国的机器人安装量居世界首位；同样，世界上大多数人工智能



专利（61%）都来自中国。

## **5. 严重缺乏对 LLM 责任的可靠和标准化评估。**

AI Index 的最新研究显示，负责任的人工智能严重缺乏标准化。包括 OpenAI、谷歌和 Anthropic 在内的领先开发商主要根据不同的负责任人工智能基准测试他们的模型。这种做法使系统地比较顶级人工智能模型的风险和局限性的工作变得更加复杂。

## **6. 生成式人工智能投资激增。**

尽管去年人工智能私人投资整体下降，但对生成式人工智能的投资激增，比 2022 年（约 30 亿美元）增长了近八倍，达到 252 亿美元。生成式人工智能领域的主要参与者，包括 OpenAI、Anthropic、Hugging Face 和 Inflection，都获得了一轮可观的融资。

## **7. 数据显示，人工智能让打工人更有生产力，工作质量更高。**

2023 年，多项研究评估了人工智能对劳动力的影响，表明人工智能可以让打工人更快地完成任务，并提高他们的产出质量。这些研究还表明，人工智能有可能缩小低技能和高技能工人之间的技能差距。还有一些研究警告说，在没有适当监督的情况下使用人工智能可能会起到负面作用。

## **8. 得益于人工智能，科学进步进一步加速。**

2022 年，人工智能开始推动科学发现。然而，2023 年，与科学相关的更重要的人工智能应用启动——使算法排序更高效的 AlphaDev、促进材料发现过程的 GNoME、可在一分钟提供极其准确的 10 天天气预报的 GraphCast、成功对 7100 万种可能的错义突变中的约 89% 进行分类的 AlphaMissense。

如今，人工智能现在可以完成人类难以完成的、但对解决一些最复杂的科学问题至关重要的粗暴计算。在医疗方面，新的研究表明，医生可以利用人工智能更好地诊断乳腺癌、解读 X 射线和检测致命的癌症。

## **9. 美国的人工智能法规数量急剧增加。**



2023 年，全球立法程序中有 2175 次提及人工智能，几乎是上一年的两倍。美国人工智能相关法规的数量在过去一年大幅增加。2023 年，与人工智能相关的法规有 25 项，而 2016 年只有 1 项。仅去年一年，人工智能相关法规的总数就增长了 56.3%。其中一些法规包括生成式人工智能材料的版权指南和网络安全风险管理体系。

#### **10. 人们对人工智能的潜在影响有了更深刻的认识，同时也更焦虑。**

来自市场研究公司 Ipsos 的一项调查显示，在过去一年中，认为人工智能将在未来 3-5 年内极大地影响他们生活的人，比例从 60% 上升到 66%。此外，52% 的人对人工智能产品和服务表示焦虑，比 2022 年上升了 13 个百分点。

在美国，来自皮尤研究中心（Pew）的数据显示，52% 的美国人表示对人工智能的担忧多于兴奋，这一比例比 2022 年的 38% 有所上升。

# 评奥特曼的闭源策略

陆首群 2024.6.18

奥特曼 (Sam Altman) 早在 2018 年 6 月开发生成式人工智能时的初心拟采用开源策略, 到 2023 年 3 月开发 GPT-4 时违反初心改用闭源策略。

OpenAI CEO 奥特曼在研发 AI 路上发生几个转变:

- 1) 从开源向闭源的转变;
- 2) 从研究生成式人工智能 (GAI) 向研究通用人工智能 (AGI) 的转变;
- 3) 从由 Open AI 独立研究向与微软、英伟达合作/协同研发的转变;

4) OpenAI 的研发工作还将遇到未来大参数、大算力、大能耗、大投资的挑战, 还将遇到未来人工智能安全性的挑战, 在这些挑战面前奥特曼能否坚持闭源策略不变还难说。

采用闭源为了追求最大的商业利益是奥特曼转变的一个主要因素但不是唯一, 据奥氏所说, 他今天“主要瞄准通用人工智能, 这时采用开源不是最佳选择”, 这个说法可能是他的遁辞。

在今天的 AI 时代, 开源也在演变中, “传统的开源”将演变为“AI 的开源”, 荷兰两位学者在今年发表的 ACM FACCT 论文中谈到开源代码和使用传统开源许可证的观点已不再适用于 AI 组件(即不再适用于“开源 AI”), 需要对传统开源及其许可证进行修改或重新定义, 他们的观点是与开源社会的 OSI 一致的(从 2022 年开始 OSI 已经在着手进行修改)。修改的主要方面就是增加如下三个方面的开放性和透明度(需要指出的是对闭源来说, 如下三个方面更是不存在开放性和透明度的)。

必须向 AI 大模型提供如下三方面信息:

① 训练数据的详细信息, 包括数据集、数据来源、数据范围和特征、获取和数据选择方式、标注程序、数据清理方法等, 以便技术人员可以用相同或相似的数据复现模型的效果。

②用于训练和运行开源代码、包括支持库以及预处理、训练、验证和测试推理、模型架构等多步骤的代码。

③模型参数，包括训练阶段中间关键的检查点以及最终的优化状态。

这样“开源 AI”比我们传统认知的“开源”已扩充了不少内容。荷兰专家分析了已经问世的 40 个开源和闭源模型，在大多数情况下，开源好于闭源，公众需要知道 AI 模型系统的安全风险分析，AI 评估专家需要知道大模型系统的安全审查性，AI 科学家需要知道大模型的可复现性，用户也需要知道应负的法律风险。对评估人员而言，设计好的评估框架，得出有意义，基于证据、多维度和开放性判断。闭源大模型一概不让各类应知人员需要的“知道”。

# 模型开放性框架(MOF)

LF-AI, 2024. 6. 20

人工智能（AI）领域正处于拐点。生成性人工智能系统和大型语言模型（LLM）的迅速崛起，在自然语言处理、图像和视频生成等方面释放了前所未有的能力。从 GPT-4 到稳定扩散，这些模型正在捕捉公众的想象力，并推动新一波的应用和研究。

然而在兴奋中，一种日益增长的不安，许多最先进的人工智能模型仍然是不透明的“黑盒子”，其内部工作不受审查。有关训练数据、模型架构和开发过程的详细信息往往很少，缺乏透明度使得难以独立验证声称的能力，审计潜在的偏见和安全问题，并在工作基础上再接再厉。

一些模型生产商已采取措施，公开发布模型，但仔细检查后发现了相关模式，被称为“开源”的模型经常使用条款模棱两可的定制许可证。文档是稀疏和分散的。数据集、训练代码和基准等关键工作不存在。这种“开放洗涤”或“开源失真”趋势可能会破坏开放的前提——自由分享知识，以实现检查、复制和集体进步。如果我们要实现人工智能的巨大希望，同时减轻其风险和陷阱，我们需要在模型开发生命周期的所有阶段真正开放。正是在这种背景下，我们引入了模型开放性框架（MOF）。MOF 是一个用于客观评估和分类机器学习模型完整性和开放性的综合框架。它通过评估模型开发生命周期的哪些组成部分以及在哪些许可证下公开发布来做到这一点。

# 发展基于开源的人工智能

陆首群 2024. 7. 16

今天我想和大家讨论：知识大模型是开源好呢还是闭源好？在奥特曼带头开发“闭源 AI”时，国内外一批跟进者鼓噪“闭源 AI”的部署，有人说“当大模型以什么样的成本带来什么样的价值的时候，你永远会选择闭源模型，它一定比开源更强大、推理成本更低”，对于这种说法我可不敢苟同。长期以来，我们一直坚持如下的做法：发展基于开源的深度信息技术（包括 AI 在内，并可举出自 2015 年以来的大量实践）。依靠开源不但可以加快 AI 的开发速度，提高开发质量，打通发展瓶颈，扩大生态，加强运维，反对垄断，有人认为开源实行免费发放，无视经济收益，无法建设强大的新兴产业，这是对开源的误解，开源在推出免费的开源社区发行版同时，也推出收费的开源商业发行版；而且在构建 AI 安全时更离不开开源，开源能有效制止 AI 不安全开发、部署和使用，有利于国家对 AI 的安全监管和全面治理；如果放弃开源，人类就不可能提出属于“AI 安全国际共识”的“AI 红线”，所谓“AI 红线”即要求所有 AI 系统都不应该在人类没有明确批准和协助情况下自主地复制和改进自己；Open AI 公司 CEO 奥特曼于今年 5 月 13 日推出闭源的大模型 GPT4o，他站在生成式 AI 一方，瞄准通用 AI，一时引发全球震动！；但奥特曼及其追随者无视人类安全风险执意要开发和部署“闭源 AI”，这是行不通的。

下面介绍法国一家非营利性 AI 研究机构 Kyutai，开发了具有听、说、看多模态功能的开源模型 Moshi，可同时处理听、说两个音频流，超越 GPT-4o，Kyutai 仅有 8 人，他们获得 AI 大师、图灵奖获得者杨立昆 (YanLeCun) 的支持，该团队开发 Moshi；模型仅用 6 个月，今年 7 月 3 日对外发布。

现将 Moshi 开源模型介绍于下：

**1、背景：**一个仅有 8 人的非营利性 AI 研究机构——Kyutai，模型名为 Moshi，具备听、说、看的多模态功能。图灵奖得主 Yann LeCun 转发说道：

「Moshi 能听懂带有法国口音的英语。」据悉，该团队开发这个模型仅用了 6 个月。

**2、Moshi 的背后：**合成数据立大功！Moshi 的设计目的是理解和表达情感，具有诸如用不同口音（包括法语）说话的能力。它可以聆听和生成音频和语音，同时保持文本思维的无缝流动。Moshi 的一个突出特点是能够同时处理两个音频流，使其可以同时聆听和说话。这种实时交互基于文本和音频混合的联合预训练，利用来自 Helium 的合成文本数据，这是一个由 Kyutai 开发的 70 亿参数语言模型。

**3、愿景使命：**LeCun 坐镇，三十年 AI 老兵带队，这是一支小而精的欧洲团队 Kyutai 是欧洲首个致力于人工智能开放研究的私人倡议实验室，由 iliad 集团、CMA CGM 集团和 Schmidt Futures 于 2023 年 11 月共同创立，初始资金近 3 亿欧元。Kyutai 定位为人工智能开放科学实验室，是一个非营利组织，其使命是解决现代人工智能的基本挑战。

我们不难注意到：从李飞飞的空间物理智能，采用大量的合成数据 80%+，另外 2023 年开始美国很多开源组织机构以及 Ai 创业，都大量采用合成数据。

**4、可靠的人工智能：**通过在巴黎创建人工智能开放研究实验室，我们进一步加快了步伐。Kyutai 将为我们提供超高性能、可靠的人工智能模型，整个欧洲人工智能生态系统都将能够从中受益。

# 北京 AI 安全国际共识

本吉奥、辛顿、姚期智等 2024.03.18

3月10-11日辛顿和姚期智等数十位中外AI专家签署了“北京AI安全国际共识”，提出了AI红线，要求任何AI系统都不应该在人类没有明确批准和协助情况下自主地复制和改进自己。

## 人工智能风险红线

（由廖璐整理国际AI专家公式，2024.5.21）

人工智能系统不安全的开发、部署或使用，在我们的有生之年就可能给人类带来灾难性甚至生存性风险。随着数字智能接近甚至超越人类智能，由误用和失控所带来的风险将大幅增加。

在过去冷战最激烈的时候，国际科学界与政府间的合作帮助避免了热核灾难。面对前所未有的技术，人类需要再次合作以避免其可能带来的灾难的发生。在这份共识声明中，我们提出了几条人工智能发展作为一种国际协作机制的具体红线（包括但不限于下列问题）。在未来的国际对话中，面对快速发展的人工智能技术，我们将继续完善对这些问题的探讨。

## 自主复制或改进

任何人工智能系统都不应在人类没有明确批准和协助的情况下复制或改进自身。这包括制作自身的精确副本以及创造具有相似或更高能力的新人工智能系统。

## 权力寻求

任何人工智能系统都不能采取不当地增加其权力和影响力的行动。

## 协助武器制造

所有人工智能系统都不应提升其使用者的能力使之能够设计大规模杀伤性武器，或违反生物或化学武器公约。

## 网络安全



任何人工智能系统都不应自主执行造成严重财务损失或同等伤害的网络攻击。

### **欺骗**

任何人工智能系统都不能有持续引致其设计者或监管者误解其僭越任何前述红线的可能性或能力。

### **路线**

确保这些红线不被僭越是可能做到的，但需要我们的共同努力：既要建立并改进治理机制，也要研发更多安全技术。

### **治理**

我们需要全面的治理机制来确保开发或部署的系统不违反红线。我们应该立即实施针对超过特定计算或能力阈注册应确保政府能够了解其境内最先进的人工智能，并值的人工智能模型和训练行为的国家层面的注册要求。具备遏制危险模型分发和运营的手段。

国家监管机构应帮助采纳与全球对齐的要求以避免僭越这些红线。模型进入全球市场的权限应取决于国内法规早否其于国际审计达到国际标准，并有效防止了违反红线的系统的开发和部署。

### **辛顿、姚期智等数十位中外 AI 专家：**

Yoshua Bengio, Geoffrey Hinton, Stuart Russell, Robert Trager, Toby Ord, Dawn Song, Gillian Hadfield, Jade Leung, Max Tegmark, Lam Kwok Yan, Davidad Dalrymple, Dylan Hadfield-Menell, 姚期智, 张宏江, 张亚勤, 傅莹, 薛澜, 黄铁军, 王仲远, 杨耀东, 曾毅

# 中外人工智能专家共话通用人工智能机遇与挑战

## 2024 北京智源大会报导

6月14日至6月15日，2024北京智源大会成功举办，图灵奖得主姚期智、加州大学伯克利分校教授斯图尔特·罗素（Stuart Russell）、MIT未来生命研究所创始人马克斯·泰格马克（Max Tegmark）、清华大学智能产业研究院（AIR）院长张亚勤，OpenAI、Meta、谷歌DeepMind、斯坦福大学、加州大学伯克利分校等国际知名人工智能机构代表，以及百度、零一万物、百川智能、智谱AI、面壁智能等国内主要大模型公司CEO等200余位人工智能顶尖学者和产业专家参会，围绕通用人工智能关键技术路径、前沿发展及技术趋势开展深入研讨，建议推动多模态大模型、具身智能等前沿方向的技术路径创新，加强数据、算力等要素支撑，同时注重AI安全治理，尽快建立国家级的AI安全研究机构，代表我国参与国际AI安全研究合作和AI安全治理。

### 一、通用人工智能前沿发展及未来趋势

（一）大模型技术的突破加速了通用人工智能（AGI）的到来，“语言大模型-多模态大模型-具身大模型-世界模型-AGI”成为AGI可能的技术演化路径。

MIT未来生命研究所创始人马克斯·泰格马克（Max Tegmark）认为，随着大模型的发展，人们对何时到达通用人工智能（AGI）的预测，从之前的几十年后缩短到大约两三年后。智源研究院院长王仲远认为，现阶段语言大模型的发展已经具备了通用人工智能非常核心的理解和推理能力，并且形成了一条以语言大模型为核心对齐和映射其他模态的技术路线，从而让模型具备了初步的多模态理解和生成能力。但这并不是让人工智能感知、理解物理世界的终极技术路线，而是应该采取统一模型的范式，实现多模态的输入和输出，让模型具备原生的多模态扩展能力，向世界模型演进。未来，大模型将以数字智能体的形态与智能硬件融合，以具身智能的形态从数字世界进入物理世界，同时，大模

型这一技术手段可为科学研究提供新的知识表达范式，加速人类对微观物理世界规律的探索与研究突破，不断趋近通用人工智能的终极目标。

（二）Scaling Law（规律增长定律）是当前阶段驱动大模型发展的重要理论基础，远没有触及天花板，但未来的扩展方式可能会面临更多挑战 and 变化。

零一万物 CEO 李开复认为，通过更多算力和数据不断增加大模型性能的大模型 Scaling Law（规律增长定律）仍在推进中，远未失效，但需要专注算法和工程创新一体化推进，避免进入“盲目堆算力推动模型性能提升”的状态。清华大学智能产业研究院（AIR）院长张亚勤认为，至少在未来五年内，Scaling Law 仍是产业发展的主要方向。月之暗面 CEO 杨植麟认为 Scaling law 是一种会持续演进的第一性原理，只是在这一过程中，规模扩展的方法可能发生很大变化。百川智能 CEO 王小川则认为，Scaling Law 到目前没有看到边界，依旧在持续地发挥作用，但是，需要在 Scaling Law 之外，去寻找范式上新的变化，走出这样的体系，才有机会走向 AGI。智谱 AI CEO 张鹏认为，Scaling Law 在未来相当一段时间内仍然会有效，但所谓的有效性是一个动态的概念，它所涵盖的内容会不断演进，从模型的参数规模，到数据量、数据质量、计算量等都变得越来越重要，它的内涵其实在不断变化。面壁智能 CEO 李大海认为，Scaling Law 是一种经验公式，是行业对大模型这个复杂系统的观察和总结。随着实验的增多，我们对模型训练过程的认知越来越清晰，细节也会越来越多。训练方法本身对 Scaling Law 的影响也很显著，一旦我们固定了参数规模，数据质量和训练方法的重要性就会显现出来。

（三）人工智能大模型已进入工业大生产阶段，大模型产业化场景在 To C 方面，将和 PC、移动互联网时代类似，从生产力工具一步步走向短视频类应用。

百度首席技术官王海峰认为，人工智能基于深度学习及大模型工程平台，包括算法、数据、模型、工具等，已经具备了非常强的通用性，并且具备了标准化、模块化和自动化的特征，与前三次工业革命的核心驱动力量机械技术、

电气技术和信息技术规模化应用时的特征类似，所以深度学习及大模型工程平台推动人工智能进入到了工业大生产阶段，通用人工智能将加速到来。

**零一万物 CEO 李开复**认为，针对大模型产业化的场景，短期在中国 To C 更有机会，但国外两者都有机会。To C 方面，AI 2.0 时代会和 PC、移动互联网时代一样，第一个阶段是生产力工具，包括信息获取。第二个阶段可能是娱乐、音乐、游戏。第三个阶段是搜索；再下一个阶段可能会是电商、社交、短视频、O2O 的应用出现。递进模式不会有特别大的改变，To C 应用会从生产力工具一步步走向短视频类应用，在普及顺序上会按照这几个阶段进行。

（四）随着语言大模型参数量和计算量的迅速增长，模型架构创新、数据优化等工作成为提升模型性能的核心策略。

**北京大学助理教授贺笛**认为，Transformer 是大模型的核心架构，但是 Transformer 在处理长序列数据时效率较低，目前国内外许多研究致力于开发高效 Transformer，希望提高其处理速度和性能。但是，通过实验发现，高效 Transformer 在解决复杂推理问题时表现不如标准 Transformer，并且差距难以弥补。交替使用高效层和标准注意力层的混合模型是一条可行的技术路径，通过兼具速度和理论优势，能够有效规避许多问题。**百川智能研究员王炳宁**提到，探索 Transformer 之外的新架构工作已有典型成果。例如，新型架构 RWKV 可以高效处理长文本和实时数据，展现了优越的扩展性。新架构 Mamba 通过递归机制和状态空间模型（SSM-RNN）优化计算复杂度，并利用固定内存机制减少信息增长，从而提升效率。

另外，对于提升语言大模型的推理能力，关键在于数据的质量优化。**清华大学副教授东昱晓**提到，目前语言大模型使用的数据主要来自互联网和过去三四十年的电子化记录，这些数据仅占人类生成文本总量的 1% 到 5%，理论上还有 20 到 100 倍的数据空间可用。**中国人民大学教授赵鑫**表示，未来数据可能成为限制因素，现有数据未必是最佳选择，合成数据的应

用和研究正变得越来越重要。

（五）多模态大模型训练过程中，多个模态的统一是大势所趋，但是从头训练原生多模态模型当前还存在许多技术问题尚待解决。

OpenAI 研究员余家辉认为，多模态本质上是在某个时间点同步发生的不同信号，为了决定下一步做什么，必须融合当前所有的信号。因此，多个模态应该统一起来，这样可以实现更多功能。Prompt AI 联合创始人兼 CEO 肖特特认为，多模态学习应该包含视频、音频、手势、触觉、听觉和情感等所有信号，人类学习的过程本来就是所有模态一起进行的。对于大模型来说，也应该从一开始就综合考虑所有模态进行学习，这样才能获得全面的理解。清华大学电子工程系副教授、上海人工智能实验室领军科学家代季峰认为，训练原生多模态模型时，需要同时处理多个模态的数据，这对数据采集和算法提出了很高的要求，目前的算法对数据的利用效率并不高，因此成本和挑战很大。智源研究院视觉模型研究中心负责人王鑫龙介绍了智源构建统一多模态模型的探索，目前已研发出原生多模态大模型 Emu3，但是在构建统一多模态模型时会遇到“不可能三角”的挑战：紧凑-无损-离散，三者无法同时满足。紧凑性，即用较少的 token 来表达图像或视频；无损性，即能够完美重建图像或视频；离散，即使用离散的 token 表示。目前只能同时满足其中的两个，实现所有三个目标仍然有技术瓶颈。

（六）具身智能尚在起步阶段，仍有漫长的发展周期，核心瓶颈是数据采集及硬件成本较高，未来本体增加触觉感知能力非常重要。

零一万物 CEO 李开复认为，具身智能可以很好地结合大模型多模态能力，而且一旦“具身”后就可以产生数据，形成数据飞轮闭环，有很大的想象空间。但短期要做好，难度很大，具身智能肯定要走很漫长的道路。目前具身智能技术研发所用的数据主要来自仿真模拟数据及采集的实际操作真实数据。宇树科

技创始人&CEO 王兴兴认为目前训练数据不够多维，限制了具身智能的发展，他表示，如果结合实际操作中的真实数据进行强化学习，效果会更好。清华大学助理教授、视觉与具身智能实验室主任高阳指出，对于简单任务，仿真数据非常有用，而复杂任务仍需依赖现实世界的真实数据。新加坡国立大学助理教授邵林提到，机器人数据采集非常昂贵，我们需要成熟的数据集或规模适中的技能库供机器人使用，但目前的挑战在于，基础技能库不足以支持机器人在大模型中的广泛应用。他还认为，目前机器人本体的高成本也是限制仿生人工智能发展的主要瓶颈之一。我们还没有足够的低成本机器人可以广泛部署到各个领域。高成本导致机器人部署数量少，进而导致数据采集不足，形成了一个负反馈循环。

另外，机器人本体增加触觉感知能力非常关键。北京大学副教授、智源学者卢宗青认为，机器人触觉至关重要，如果没有触觉，就像是在玩一场虚拟游戏，而不是与真实世界进行交互。清华大学助理教授、视觉与具身智能实验室主任高阳认为，触觉是一个非常重要的感知模态，不必追求全身都配备复杂的传感器，可能只需要在手部和夹指上安装一些传感器，这样已经能够显著提升机器人的感知和操作能力。

（七）数据的自动化获取和处理是行业发展大势所趋，数据要素价值变现机制及数据基础设施建设无法满足数据需求，需要机制创新。

数据对人工智能大模型性能起到了决定性作用，是智能发展的根本。随着大模型的快速发展，对数据的需求呈指数级增长，人工处理数据的模式已经无法满足行业发展需求。北京智源人工智能研究院理事长、北京大学教授黄铁军认为，数据产业（数据加工、数据标注）应该主要靠人工智能技术，依托 AI 智能体可替代 90%以上的人工清洗、标注工作，再通过优质数据不断推动智能体渐进式的、迭代式发展。复旦大学教授、上海市数据科学重点实验室主任肖仰华

认为，基于大模型的 AI 智能体能够实现海量知识验证、非结构化数据访问、数据分析、数据智能运维等功能，将成为推动数据要素价值变现的重要引擎。**智源研究院副院长兼总工程师林咏华**认为，合成数据是一个重要的方向，许多大模型企业都会使用合成数据或增广数据。其中，增广数据是基于人类的现有数据，用技术自动产生不同变种的数据，尤其是在指令微调阶段，因为需要有很多特殊格式、特殊任务，很难靠人工大批量产生。

另外，数据资产化机制及基础设施建设模式尚有创新空间。北京**智源人工智能研究院理事长、北京大学教授黄铁军**认为，在优质数据缺乏的背景下，当前简单的数据资产化并不利于数据流通，还将加剧模型研发机构的资金压力。北京大学人工智能研究院副院长、北京大学数据空间技术与系统全国重点实验室主任**黄罡**认为，当前的数据基础设施的建设严重滞后于需求，通过构建专网+公网+跨境的一体化数据基础设施，打造“开放式”数据飞轮，将有望实现基于数联网的大模型智能体数据供应链。

（八）国产 AI 系统（AI Infra）软件栈生态建设不完善，国产 AI 算力资源利用率偏低。

**清华大学长聘教授翟季冬**认为，当前国产算力的资源利用率非常不充分，核心挑战是底层的算力软件生态。智能算力大约有 10 个非常关键的软件，包括调度器、管理内存、容错系统、并行软件、编程语言、编译器、多机通信、编程框架等，但这些软件的国产化均存在瓶颈，现状仍是英伟达及其生态伙伴处于绝对垄断地位。**智源研究院 AI 算子和编译器方向的负责人白童心**认为，我国 AI 系统软件生态当前面临着以下困境：一是目前 AI 芯片厂商的自有体系生态隔离，互不兼容，存在较为明显的孤岛效应，在我国算力急缺的情况下，不同芯片算力集群间的混训难以实现；二是由于生态的割裂导致软件发展只能依赖厂商自行投入，耗时耗力但更新迭代速度仍不及预期。三是分散的生态也增加了



模型训练机构的负担，需要对不同芯片进行多次适配，成本较高且影响模型迭代速度。

（九）伴随通用人工智能（AGI）的到来，人类或将面临“生存风险”，加强 AI 安全研究迫在眉睫，AI 安全研究是新兴研究方向，需要建立国际 AI 安全研究组织或机制加强研究协调。

加州大学伯克利分校教授斯图尔特·罗素（Stuart Russell）认为，我们思考的不应该是“如何使 AI 安全？”，因为一旦实现了 AGI，如果它还不安全，可能就太晚了，而是要思考“如何构建安全的 AI”，AI 需要采用类似于核能、航空等领域的监管方式，必须在严格的规定和安全标准下进行部署。当前在部署之前验证前沿人工智能模型的方法依赖于独立测试和红队测试，但是人工智能系统的弱点在经过大量测试后也可能仍未被发现，**这些测试制度并不能提供严格的、可量化的安全保证**，人类唯一的长期选择是“可证明的安全和有益的 AI”，**AI 安全技术研究必须与监管齐头并进**。

当前，AI 安全研究仍是新兴研究方向，尚缺乏基础理论框架及工具。**姚期智院士**认为，从长远来看，AI 安全需要作为一门学科来研究，类比于网络安全方向的密码学，AI 安全可能也需要建立一套类似的基本科学工具。另外，针对当前国际上 AI 安全研究研究较分散的问题，**斯图尔特·罗素（Stuart Russell）教授建议**，国际 AI 安全社区需要建立国际组织，举办相关研究主题的国际会议，加强 AI 安全研究的全球合作和协同。

在今年 5 月份召开的首尔 AI 安全峰会上，美国、英国、日本、韩国和欧盟共 11 个国家和地区签署了《首尔宣言》和《首尔人工智能安全科学国际合作意向声明》，提出“安全、可靠”地开展人工智能创新，框架内各国将依托各自成立的人工智能安全研究所，并建立**国际 AI 安全研究所网络**，强化前沿 AI 系统研究合作，酌情共享技术资源、大模型信息、测评数据，共同打造所谓“安

全评判标准”，推进人工智能安全科学研究。今年 10 月份计划在美国旧金山召开首届国际 AI 安全研究所网络峰会，由于我国尚未成立国家级的 AI 安全研究所，目前尚无法参会。马克斯·泰格马克（Max Tegmark）在会上表示，中国未来有机会在 AI 安全研究和治理上发挥领导作用，必须参与。

## 二、工作建议

（一）持续发挥智源研究院等新型研发机构原创引领的作用，推动我国大模型前沿技术和思潮发展，解决行业共性问题

月之暗面 CEO 杨植麟认为，智源是亚洲地区最早投入大模型研究的机构，具有领先的想法和宽广的视野。百川智能 CEO 王小川认为，智源是中国大模型的黄埔军校，推动了大模型技术和思潮的发展。今天中国大模型发展迅速，智源有很好的定位，建议作为中立和技术高地，扮演技术和智库角色，帮助我们快速健康发展。智谱 AI CEO 张鹏说到，智源从 NGO 型研究机构发展到今天，已成为国际人工智能领域的一面旗帜。面壁智能 CEO 李大海认为，大模型领域变化快，有些事情商业公司无法独立完成，建议在智源的撮合和带领下，搭建更好的平台，共同协作，解决行业问题。智源研究院院长王仲远认为，当前的人工智能研究已进入到资源消耗型模式，创新算法的验证也需要大规模算力支撑，对于新型研发机构来说负担非常重，建议各方能给予更多算力支持。

（二）持续推动语言、多模态、具身等大模型前沿及关键技术研发，提升语言大模型训练性能，探索多个模态统一的原生多模态模型研发路径，推动具身通用机器人发展。

北京大学助理教授贺笛建议，要加快大模型架构创新，优化 Transformer 架构性能，探索 Transformer 之外的新架构创新，提升语言大模型规模增长效率。浙江大学求是讲席教授沈春华建议推动多模态统一的算法创新，研发国际领先的原生多模态模型。北京大学助理教授、智源学者王鹤建议

要推动研发具身智能大模型，发展具身智能“大脑”“小脑”，提升具身机器人的感知、决策及行动能力。持续研发触觉感知、力觉感知灵巧手，提升本体感知能力。加强大规模高质量的具身多模态数据集建设，特别是真实数据的采集，夯实通用机器人发展数据支撑。

（三）突破数据、算力、系统软件等发展瓶颈，提升数据要素支撑能力，推动国产 AI 芯片生态发展，打造满足大模型时代发展的 AI 系统软件。

数据方面，北京智源人工智能研究院理事长、北京大学教授黄铁军建议创新数据流通及交易机制，通过“先使用后付费”的数据流转模式，在模型训练机构使用前对数据进行确权，当模型训练机构获得收益后按比例分红，推动数据飞轮实现，加速“数据+智能”正反馈。中国移动集团首席科学家冯俊兰建议建设多行业数据汇聚基座，推动跨场景、跨任务、跨功能的数据流通和共享。算力及系统软件方面，清华大学长聘教授翟季冬建议推动建设支持异构 AI 芯片、万卡集群互联的高效 AI 系统软件（AI Infra），加快 AI 算子库及编译器建设，推动国产 AI 芯片的规模化发展。

（四）重视 AI 安全研究布局，由优势机构联合建设国家级 AI 安全研究所，代表中国参与国际 AI 安全科研合作及标准规范制订。

会上，姚期智、张亚勤等院士呼吁尽快设立“中国人工智能安全研究所”，推动开展人工智能模型的安全对齐理论与方法、人工智能安全测试方法与技术、人工智能安全与传统安全的交互融合、人工智能中的数据安全保护等 AI 安全技术研究，为我国 AI 安全的长期发展提供有力支撑。同时，积极参与国际 AI 安全研究所网络建设，强化前沿 AI 系统研究合作，共同打造国际 AI 安全评判标准，支撑全球人工智能的健康发展。

# 如何从生成式人工智能（GAI）转移到 通用人工智能(AGI)轨道上来

陆首群 2024.07.02

大模型（LLM）生成式人工智能（GAI）在智能发展上的突破加速了通用人工智能（AGI）的到来。

所谓生成式人工智能是基于机器学习/深度学习的大模型，当其资料库中依托统计技术增长的参数达到 2000 亿左右时，在大模型中突然涌现出一股推理性质的涌流，推理是生成的关键，推理赋能大模型（或机器）被动生成自然语言，从而实现了人机对话。

所谓通用人工智能是大模型（或机器）通过自我学习、交互，建立自主系统的人工智能。

大模型的生成机制，实现了文本生成文本，或图像或视频，…。在文本生成视频时，在输入端的提示指导下完成由固定的视频升级为动态的短时间影视的任务。2022 年底，Open AI 开发了 Q\*(Q-star) 模型和文本生成视频 Sora（世界模拟器），是生成式大模型研发进入通用人工智能的标志。

通用人工智能的演化路径可能是：（生成式）语言大模型 → 多模态大模型 → 具身(embodiment)大模型 → 世界模型 → 通用人工智能。

通用人工智能产生自主系统表明：该系统具有增长、积累知识的功能（人工智能拥有的知识，最终可能会超过人类），也开始产生思维 / 意识（有缺陷类似不同于人类的思维 / 意识）。

令人担心的是：通用人工智能建立起来高度自主智能系统在无可能超越人类？一旦超越将给人类带来什么样的后果？

目前人们对通用人工智能能否超越人类有如下几种观点：

①人工智能如果没有人类在后台帮助就不会超越人类；

②人类可以为人工智能发展划出一道红线，限制人工智能相互拷贝、交流、学习，就可限制人工智能超越人类；

③如果人工智能自主系统高度发展，或人类放松对人工智能的监控，使其可在人工智能同类间进行相互学习、交流、拷贝，迅速超越人类；从而给人类带来巨大的安全风险，这是我们所不能允许出现的！

对于上述第③种观点涉及的情况，人类别无他途，只能在人工智能尚未超越人类之前，赶紧采取预防性措施：

1) 抓紧研发人工智能安全保障技术；

2) 应用人工智能安全保障技术与实行人工智能安全监管同时并举；

3) 建立人工智能安全治理体系：

①制定全球统一有效的安全标准、法律规范、审计程序、建立独立的安全测试机构，采取统一的安全测试规范；

②建立全球与各国的人工智能安全研究中心和安全研究网络；

③建立人工智能治理体系（全球发放供参考的人工智能治理指南，以及推动各国制定人工智能治理方案，包括技术、经济、培训、宣传、法律及其他相关的）。

# 加速计算与生成式人工智能重塑行业

## （推进绿色新工业革命）

黄仁勋谈到：“加速计算”产业的主要标志是制造具有“加速计算技术的超级计算机”，这是打造工业数字化的主力，将拥抱第四次工业革命。

他阐述了“加速计算”技术的创新形成机制，基于高低相关的两阶社会（全能宇宙 Omniverse）的互动：设想在现实的低阶现实社会（real society）之上构建一个虚拟（Virtual society）的高阶信息社会，为了便于考察和试验，在虚拟社会中选取一个数字空间（Cyber Spa 的算力引擎+AI，向物理空间注入待变的经典计算单元，将先进的算力引擎+AI 作用于在降维空间中的经典计算单元，促其产生 0→1 智能化变化，形成“加速计算”平台。

NVIDIA 具有“加速计算”技术的超级计算机在全球超级计算机 Green500 榜单 Top10 排名中占据 7 席。科学家利用这些超级计算机可以揭开宇宙奥秘，解码基因序列，预测气候模式，NVIDIA “加速计算”可节约能源、降低成本，支持经济持续高效发展。

数据处理和管理是一个价值 1000 亿美元的市场，也是企业的主要工作负载。NVIDIA CUDA GPU 可以加速普遍应用的 Apache Spark 数据处理引擎，由其加速的 Spark 可将数据处理的碳足迹减少 80%。

Pandas 是世界领先的数据分析库，过去只能在 CPU 上运行，在处理大型数据集时速度缓慢。NVIDIA 已将 Pandas 加速了 150 倍，惠及全球 1000 万用户。

训练 AI 基础模型需要大量算力，会耗费大量能源，NVIDIA 加速计算正在降低训练这些基础 AI 模型所需的成本和能耗。全新的 Blackwell 平台就是一个完美的例子，它正在帮助世界各地的机构在万亿参数语言模型上构建并部署实时生成式 AI，在模型处理方面的成本和能耗也比其前身降低了一个数量级。

评估 AI 成本和能耗的正确方法是纵向评估，不仅聚焦训练，还要着眼于整

个生命周期和所创 AI 模型的下游影响，复合效益可能是巨大的。

生成式 AI 作为新的计算堆栈从根本上改变了计算机的工作方式，使计算机从指令驱动型转变为意图驱动型，生成式 AI 改变了我们使用计算机的方式，从搜索、检索既有的内容转变为与每个上下文独有的实时生成信息进行交互，生成式 AI 正在将生产软件的计算机行业转变为生产数字智能的行业。

黄仁勋还说，每次工业革命都开创生产力飞跃并持续发展的新时代，第一次工业革命作为推动生产力飞跃发展的标志是蒸汽机，第二次工业革命是电力，第三次工业革命是软件（生产），第四次工业革命已经开始，推动生产力飞跃发展的将是数字智能。而 NVIDIA 创造的“加速计算”平台体现数字智能。

NVIDIA 正在建设全球的数据中心，先前实现现代化的全球数据中心由最万台通用服务器组成，耗资达数万亿美元，采用“加速计算”技术后，采购的通用服务器将减少一个数量级。



# 人工智能发展中面临的“四大” (大参数、大算力、大能源、大投资) 挑战

陆首群 2024.07.02

在人工智能发展中，无论是单独的智能体建设、或是建设集成数据培训中心（供中小企业乃至大企业租用）都会面临大参数、大算力、大能源、大投资等“四大”挑战。

## 一、大数据、大参数

当生成式知识大模型蜕变为人工智能过程中，当资料库中参数增加到一定家量时(如 GPT-3 或 Chat GPT 的参数量增加到 1250 亿/时)，会使大模型突然出现推理性的湧流，推理是生成的基础，从而赋能机器生成自然语言，实现人机对话。

这时的大参数（或大数据）取自社会、行业，随着人工智能预训练的需要，将提升参数量到 1-2 万亿(直至 10 万亿)，这时参数来源将求助于海量信息的互联网及网际上的调查分析机构，并赋以“词元(Token)”头衔。随着人工智能预训练进一步发展，互联网等参数源提供的参数量也不敷需要，将以更好更全面的能力收集新的提供合成的训练参数，此时 Token 预训练参数量将高达几百万亿到几千万亿/千亿级大模型人工智能正在迅速耗尽世界的高质量数据，导致人采用合成数据。

## 二、大算力

人工智能的算力需求正以一种近乎疯狂的以指数级、速度增长，人工智能算力每 6 个月就会翻 10 倍。

黄伟达(NVIDIA)正在研究 Blackwell 超级芯片，构建 Rubin ultra 下一代人工智能平台，改变 GPU 架构，更新研发节奏，打破一年一次的摩尔定律。

据 OpenAI 与英伟达共同提出的建设两个集成算力培训中心的方案是：(1) OpenAI 在研究大模型 GPT-5 之后，在建设文本—视频 Sora 集成算力培训中心时，需要集成 7000 多张 H100 芯片，综合算力为  $10^8$  TFLOPS；(2) OpenAI 在研发大模型 GPT-6 之后，在建设“星际之门”时，需要集成 10 万张 H100 芯片，综合算力达  $10^9$  TFLOPS。

### 三、大能耗

建设人工智能能耗巨大！今年 7 月 12 日两位科技巨头马斯克与黄仁勋相遇时哀叹：从汽车到人工智能，美国将惨败于能源。美国面临能源短缺的挑战。

从早期研发的个体大模型（如 chat GPT）而言，每天消耗能源 50 万 KW，相当于数万个美国家庭一天的用电量）一台光刻机一天耗电 3 万 kwh，目前全球炙手可热的 AI 程序—CGPT，每天耗电高达 50 万 kwh，相当于一个小城市一年用电量。如果全球所有 AI 程序计算在内其能耗将是一个天文数字。

### 四、大投资

在未来人工智能发展中，单个产品（如 GPT-4）训练费用高达 7800 万美元，谷歌 Gemini Ultra 的计算成本高达 1.91 亿美元。

对建设集成服务的预训练中心而言，每个中心需要投入资金约几千亿美元～几万亿美元。

训练中心向申请代训企业的收费昂贵，可以听到“烂钱租”、“吞金兽”的呼声。

有人反映，国内企业一阵风起迄今开发了 300 多个大模型，对其表现如何的评语是“群魔乱舞”！也有人认为，到今年年底，起码有一半以上行将消失！其实盲目发展的这些企业，也难过这“四大”挑战关！

# 改善大模型预训练的工作环境

COPU 2024. 7. 16

随着大模型飞速发展,与其配套的 GPU 预训练设施的建设也取得了相应发展,一些巨额投资、巨大能耗、由 GPU 集群组成巨大的预训练中心也相继出现,以供广大企业租用来训练他们自制的大模型。

这些预训练设施在运作中也暴露一些性能差、效率低、能耗大等发展中的问题,客户也时有抱怨该预训练中心收费昂贵,形如“烧钱租”、“吞金兽”。所以改进预训练方案迫在眉睫,下面介绍加州大学的改进方案(设想):

加州大学从各种显卡中选择训练卡作为改进研究的对象,他们从改进 Transformer 架构出发,摒弃 Mat Mul 矩阵乘法(消除其带来的内存占用和延迟等严重问题),开发高效的 GPU 实现方案,优化推理内核机制,推出基于 FPGA 定制硬件解决方案,从而降低了训练方案对内存的依赖,改善性能和效率,减少了能耗,减少对预训练中心建设和运作的投资,当然也减少了代训的租金。

研究人员开发了一种高度的 GPU 实现方案,将训练时的内存使用锐减 61%,推理时内核消耗更降至原来 1/10 以下,速度也提升了 457 倍!还展示了一种基于 FPGA 的定制硬件解决方案,在处理十亿参数规模模型时功耗仅为 13w,堪比人脑的效率!

加州大学(旗下圣克鲁兹分校、戴维斯分校)与苏州大学合作的改进方案只是起步,他们提供的论文还在论证阶段,其理论依据还停留在定制方案(存在局限性)。

我们看到对他们研究工作的一篇报导:什么加州大学颠覆性架构,13w 功耗大模型告别 GPU,使 GPT-4 瑟瑟发抖,似乎有点夸张!

但改善大模型预防训练的工作环境仍待继续。

# 能源短缺正在向人工智能发展提出挑战

COPU 2024. 7. 16

正值人工智能发展日新月异之时，两位科技巨头马斯克和黄仁勋凭着他们敏锐捕捉到人工智能发展背后的隐患——能源短缺，触景生情在 7 月 12 日相聚时突然哀叹：美国人工智能发展面临能源短缺的挑战！从汽车到人工智能，美国将被中国史诗般超越，美国惨败于能源。

人工智能所需的数据和算力正以一种近乎疯狂的速度呈指数级增长，每 6 个月就会翻 10 倍，对能源供给提出了前所未有的挑战。AI 训练用大数据、大算力，形如“数据巨兽”，对大能源有巨大需求。CGPT 是目前全球最炙手可热的 AI 程序，每天耗电量近 50 万 kwh，相当于一个小城市一年的用电量，如果把全球所有程序员认真在内，其耗电量将是一个天文数字。能源短缺是向人工智能发展提出的一大挑战。

马、黄二位认为：中国发展能源走在世界前列，烧煤的火电在下降，水电、核电有较大增长，发展光伏、风电等新能源产业中国走在世界前列，核聚变、太阳能等前沿科技中国也在大力探索，在未来能源革命中争取占据有利位置。

美西方评中国改革一般采用评击和评飘两手，这次是评飘，希望国人不要被他们带着飘起来。

# 开源是 AI 前进的道路

(在 AI 安全问题上闭源无法代替开源)

Mate 报导 2024. 7. 24

刚刚，**Llama 3.1** 正式发布，登上大模型王座！

在 150 多个基准测试集中，405B 版本的表现追平甚至超越了现有 SOTA 模型 GPT-4o 和 Claude 3.5 Sonnet。

也就是说，这次，**最强开源模型即最强模型**。

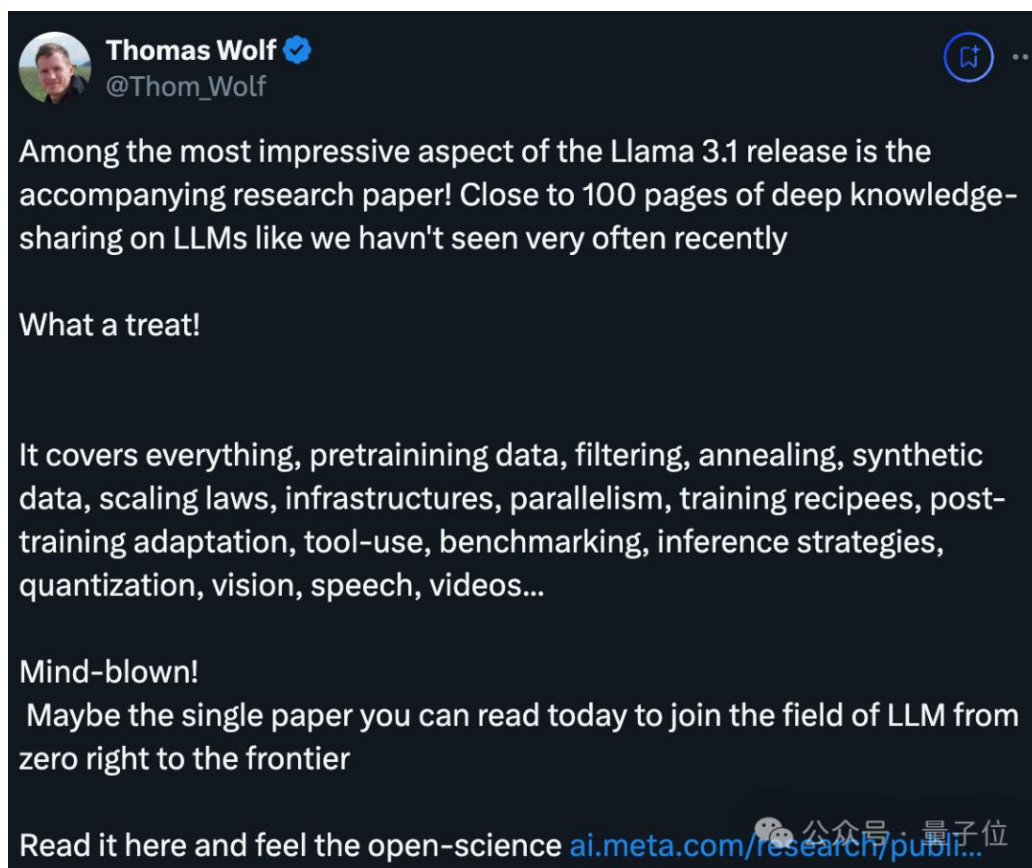
Category Benchmark	Llama 3.1 405B	Nemotron 4 340B Instruct	GPT-4 (0125)	GPT-4 Omni	Claude 3.5 Sonnet
General					
MMLU (0-shot, CoT)	88.6	78.7 (non-CoT)	85.4	88.7	88.3
MMLU PRO (5-shot, CoT)	73.3	62.7	64.8	74.0	77.0
IFEval	88.6	85.1	84.3	85.6	88.0
Code					
HumanEval (0-shot)	89.0	73.2	86.6	90.2	92.0
MBPP EvalPlus (base) (0-shot)	88.6	72.8	83.6	87.8	90.5
Math					
GSM8K (8-shot, CoT)	96.8	92.3 (0-shot)	94.2	96.1	96.4 (0-shot)
MATH (0-shot, CoT)	73.8	41.1	64.5	76.6	71.1
Reasoning					
ARC Challenge (0-shot)	96.9	94.6	96.4	96.7	96.7
GPQA (0-shot, CoT)	51.1	-	41.4	53.6	59.4
Tool use					
BFCL	88.5	86.5	88.3	80.5	90.2
Nexus	58.7	-	50.3	56.1	45.7
Long context					
ZeroSCROLLS/QuALITY	95.2	-	95.2	90.5	90.5
InfiniteBench/En.MC	83.4	-	72.1	82.5	-
NIH/Multi-needle	98.1	-	100.0	100.0	90.8
Multilingual					
Multilingual MGSM (0-shot)	91.6	-	85.9	90.5	91.6

在此之前，Llama 3.1 已经被多番曝光泄露，如今可以说千呼万唤始出来。

从今天开始，模型即可在官网上下载使用，Meta AI 应用可在线试玩。

更令研究社区赞赏的是发布近 100 页详细论文，涵盖了创造 Llama 3.1 过程中的一切：预训练数据、过滤、退火、合成数据、缩放定律、基础设施、并行性、训练配方、训练后适应、工具使用、基准测试、推理策略、量化、视觉、语音、视频……

HuggingFace 首席科学家赞叹：如果你是从零开始研究大模型，就从这篇论文读起。



小扎**扎克伯格**还在最新接受彭博社采访时专门嘲讽了一把 OpenAI。

奥特曼的领导能力值得称赞，但有点讽刺的是公司名为 OpenAI 却成为构建封闭式人工智能模型的领导者。

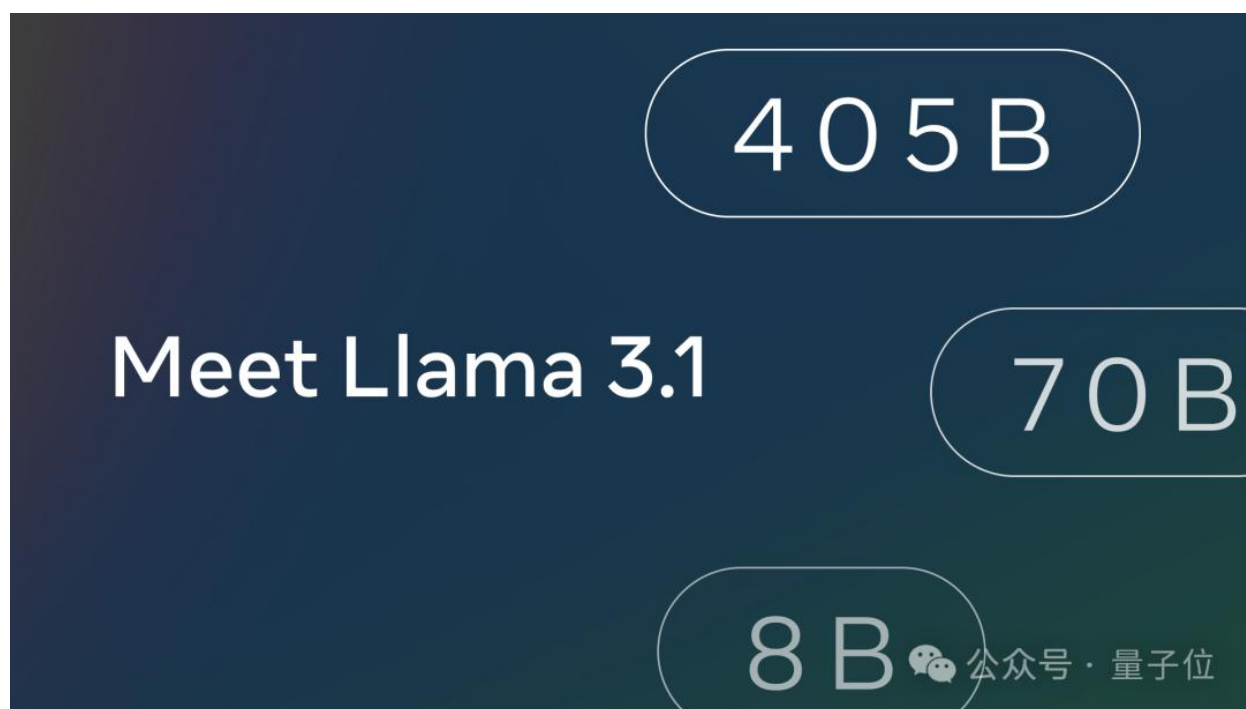


小扎还专门为此撰写了一篇长文：[开源 AI 是前进的道路](#)。

以往，开源模型在性能、功能等方面大多落后于闭源模型，但现在：

就像开源的 Linux 在一众闭源系统中脱颖而出获得普及，并逐渐变得更先进、更安全，拥有比闭源系统更广泛的生态。

我相信 Llama 3.1 将成为行业的一个转折点。



迄今为止，所有 Llama 版本的总下载量已超过 3 亿次，Meta 也是放下豪言：

这仅仅是个开始。

各大云厂商也在第一时间上线了的 Llama 3.1 的支持，价格是酱婶的：



## Model Pricing

Hosted Llama 3.1 API public pricing as of 12pm PST on 7/23/24.  
This table will be updated as more pricing becomes available.

Model	8B		70B		405B	
	Input	Output	Input	Output	Input	Output
AWS	\$0.30	\$0.60	\$2.65	\$3.50	-	-
Azure	\$0.30	\$0.61	\$2.68	\$3.54	\$5.33	\$16.00
Databricks	-	-	\$1.00	\$3.00	\$10.00	\$30.00
Fireworks.ai	\$0.20	\$0.20	\$0.90	\$0.90	\$3.00	\$3.00
IBM	\$0.60	\$0.60	\$1.80	\$1.80	\$35.00	\$35.00
Octo.ML	\$0.15	\$0.15	\$0.90	\$0.90	\$3.00	\$9.00
Snowflake	-	-	-	-	\$15.00	\$15.00
Together.AI	\$0.18	\$0.18	\$0.88	\$0.88	\$5.00	\$15.00

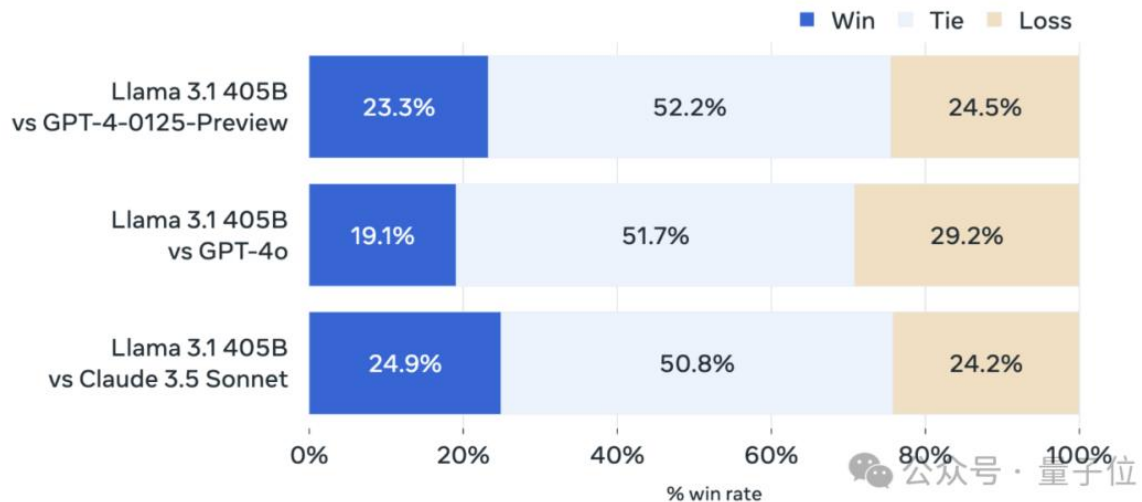
公众号 · 量子位

## Llama 3.1 官方正式发布

首先来看模型能力。

Llama 3.1 将上下文长度扩展到 128K、增加了对八种语言的支持。其中超大杯 405B 版本，在常识、可操纵性、数学、工具使用和多语言翻译等能力方面都追平、超越了现有顶尖模型。

### Llama 3.1 405B Human Evaluation



Category	Benchmark	Llama 3 8B	Gemma 2 9B	Mistral 7B	Llama 3 70B	Mixtral 8x22B	GPT 3.5 Turbo	Llama 3 405B	Nemotron 4 340B	GPT-4 (0125)	GPT-4o	Claude 3.5 Sonnet
General	MMLU (5-shot)	69.4	72.3	61.1	83.6	76.9	70.7	87.3	82.6	85.1	89.1	89.9
	MMLU (0-shot, CoT)	73.0	72.3 <sup>△</sup>	60.5	86.0	79.9	69.8	88.6	78.7 <sup>△</sup>	85.4	88.7	88.3
	MMLU-Pro (5-shot, CoT)	48.3	-	36.9	66.4	56.3	49.2	73.3	62.7	64.8	74.0	77.0
	IFEval	80.4	73.6	57.6	87.5	72.7	69.9	88.6	85.1	84.3	85.6	88.0
Code	HumanEval (0-shot)	72.6	54.3	40.2	80.5	75.6	68.0	89.0	73.2	86.6	90.2	92.0
	MBPP EvalPlus (0-shot)	72.8	71.7	49.5	86.0	78.6	82.0	88.6	72.8	83.6	87.8	90.5
Math	GSM8K (8-shot, CoT)	84.5	76.7	53.2	95.1	88.2	81.6	96.8	92.3 <sup>◇</sup>	94.2	96.1	96.4 <sup>◇</sup>
	MATH (0-shot, CoT)	51.9	44.3	13.0	68.0	54.1	43.1	73.8	41.1	64.5	76.6	71.1
Reasoning	ARC Challenge (0-shot)	83.4	87.6	74.2	94.8	88.7	83.7	96.9	94.6	96.4	96.7	96.7
	GPQA (0-shot, CoT)	32.8	-	28.8	46.7	33.3	30.8	51.1	-	41.4	53.6	59.4
Tool use	BFCL	76.1	-	60.4	84.8	-	85.9	88.5	86.5	88.3	80.5	90.2
	Nexus	38.5	30.0	24.7	56.7	48.5	37.2	58.7	-	50.3	56.1	45.7
Long context	ZeroSCROLLS/QuALITY	81.0	-	-	90.5	-	-	95.2	-	95.2	90.5	90.5
	InfiniteBench/En.MC	65.1	-	-	78.2	-	-	83.4	-	72.1	82.5	-
	NIH/Multi-needle	98.8	-	-	97.5	-	-	98.1	-	100.0	100.0	90.8
Multilingual	MGSM (0-shot, CoT)	68.9	53.2	29.9	86.9	71.1	51.4	91.6	-	85.9	90.5	91.6

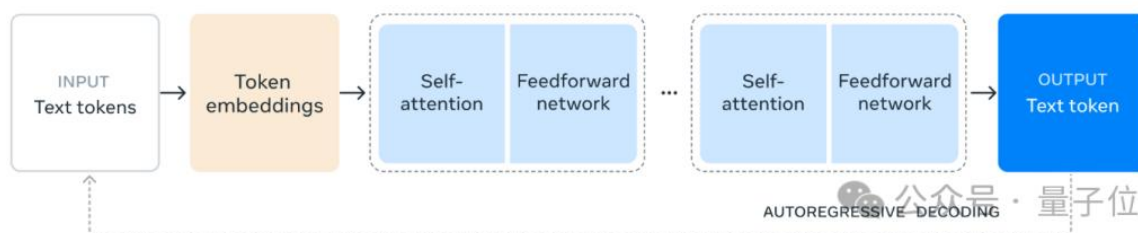
除此之外，也推出了 8B 和 70B 模型的升级版本，能力与同等参数下的顶尖模型基本持平。

Category Benchmark	Llama 3.1 8B	Gemma 2 9B IT	Mistral 7B Instruct	Llama 3.1 70B	Mixtral 8x22B Instruct	GPT 3.5 Turbo
General						
MMLU (0-shot, CoT)	73.0	72.3 (5-shot, non-CoT)	60.5	86.0	79.9	69.8
MMLU PRO (5-shot, CoT)	48.3	-	36.9	66.4	56.3	49.2
IFEval	80.4	73.6	57.6	87.5	72.7	69.9
Code						
HumanEval (0-shot)	72.6	54.3	40.2	80.5	75.6	68.0
MBPP EvalPlus (base) (0-shot)	72.8	71.7	49.5	86.0	78.6	82.0
Math						
GSM8K (8-shot, CoT)	84.5	76.7	53.2	95.1	88.2	81.6
MATH (0-shot, CoT)	51.9	44.3	13.0	68.0	54.1	43.1
Reasoning						
ARC Challenge (0-shot)	83.4	87.6	74.2	94.8	88.7	83.7
GPQA (0-shot, CoT)	32.8	-	28.8	46.7	33.3	30.8
Tool use						
BFCL	76.1	-	60.4	84.8	-	85.9
Nexus	38.5	30.0	24.7	56.7	48.5	37.2
Long context						
ZeroSCROLLS/QuALITY	81.0	-	-	90.5	-	-
InfiniteBench/En.MC	65.1	-	-	78.2	-	-
NIH/Multi-needle	98.8	-	-	97.5	-	-
Multilingual						
Multilingual MGSM (0-shot)	68.9	53.2	29.9	86.9	71.1	51.4

再来看模型架构。

官方介绍，要在超 15 万亿个 token 上训练 Llama 3.1 405B 模型挑战不小。

为此他们大幅优化了整个训练栈，并把模型算力规模首次扩展到了超过 16000 个 H100 GPU。



具体来说，还是采用标准的仅解码器的 Transformer 架构，并做一些细微改动；并采用迭代的 post-training 流程，每轮都有 SFT（监督微调）和 DPO（直接偏好优化），以提高每个能力的性能。

与 Llama 以前的版本相比，他们提高了用于预训练和 post-training 数据的数量和质量。

而为了支持 405B 这样尺寸模型的大规模生产推理，Meta 将模型从 16 位（BF16）量化到 8 位（FP8）数值，有效地降低了所需的计算需求，并允许模型在单个服务器节点内运行。

在指令微调方面，Meta 还提高了模型对用户指令的响应能力、增强了它遵循详细指令的能力，同时保证安全性。

在 post-training 阶段，Meta 在预训练模型的基础上进行多轮对齐。

每一轮都包括监督微调（Supervised Fine-Tuning, SFT）、拒绝采样（Rejection Sampling, RS）和直接偏好优化（Direct Preference Optimization, DPO）。

他们使用合成数据生成来绝大部分 SFT 示例，并数次迭代。

此外，还采用了多种数据处理技术来将这些合成数据过滤到最高质量。

总计 15T tokens 使用 Llama 2 模型做清理和过滤，而代码和数学相关的数据处理流水线则主要借鉴了 Deepseek 的方法。

**Model-based quality filtering.** Further, we experiment with applying various model-based quality classifiers to sub-select high-quality tokens. These include using fast classifiers such as `fasttext` (Joulin et al., 2017) trained to recognize if a given text would be referenced by Wikipedia (Touvron et al., 2023a), as well as more compute-intensive Roberta-based classifiers (Liu et al., 2019a) trained on Llama 2 predictions. To train a quality classifier based on Llama 2, we create a training set of cleaned web documents, describe the quality requirements, and instruct Llama 2’s chat model to determine if the documents meets these requirements. We use DistilRoberta (Sanh et al., 2019) to generate quality scores for each document for efficiency reasons. We experimentally evaluate the efficacy of various quality filtering configurations.

**Code and reasoning data.** Similar to DeepSeek-AI et al. (2024), we build domain-specific pipelines that extract code and math-relevant web pages. Specifically, both the code and reasoning classifiers are DistilledRoberta models trained on web data annotated by Llama 2. Unlike the general quality classifier mentioned above, we conduct prompt tuning to target web pages containing math deduction, reasoning in STEM areas and code interleaved with natural language. Since the token distribution of code and math is substantially different than that of natural language, these pipelines implement domain-specific HTML extraction, customized text features and heuristics for filtering.

除了最基本的根据提示词响应，Meta 官方表示，任何普通开发者可以用它做些高级的事情，比如：

- 实时和批量推理
- 监督微调
- 针对特定应用评估模型
- 持续预训练
- 检索增强生成 (RAG)
- 函数调用
- 合成数据生成

而这背后也是由它的强大生态伙伴支持。

Features for 405B models	aws	databricks	DELL technologies	nvidia	groq	IBM	Google Cloud	Microsoft	scale	snowflake
Real-time inference	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Batch inference		✓	✓	✓		✓			✓	✓
Fine tuning		✓	✓	✓					✓	✓
Model evaluation	✓	✓		✓		✓	✓	✓	✓	
Knowledge base	✓	✓	✓	✓		✓	✓	✓	✓	✓
Continual pre-training		✓		✓						
Safety guardrails	✓	✓	✓	✓		✓	✓	✓	✓	✓
Synthetic data generation	✓	✓		✓		✓	✓	✓	✓	✓
Distillation recipe	✓	✓		✓						

公众号 · 量子位

## 小扎撰写长文：开源 AI 是前进的道路

（以下由大模型翻译，摘取主体内容，如有遗漏错误欢迎更正！）

在高性能计算的早期，当时的大型科技公司都投入巨资开发各自闭源的 Unix 版本。那时，很难想象除了闭源之外，还有其他途径能够孕育出如此先进的软件。然而，开源的 Linux 操作系统最终赢得了广泛的欢迎——最初是因为它允许开发者自由地修改代码，成本也更加低廉；随着时间的流逝，Linux 不仅变得更加先进和安全，而且构建了一个比任何闭源 Unix 系统都要广泛的生态系统，支持更多的功能。如今，Linux 已成为云计算和大多数移动设备操作系统的行业标准基础，我们所有人都因此享受到了更优质的产品。

**我相信人工智能将以类似的方式发展。**今天，几家科技公司正在开发领先的闭源模型。但开源正在迅速缩小差距。去年，Llama 2 只能与落后一代的模型相媲美。今年，Llama 3 与最先进的模型竞争，并在某些领域领先。从明年开始，我们预计未来的 Llama 模型将成为行业中最先进的。但即使在那之前，Llama 已经在开放性、可修改性和成本效率方面领先。

今天，我们正在朝着“**开源人工智能成为行业标准**”的方向迈进。我们发布了 Llama 3.1 405B，这是第一个前沿级别的开源人工智能模型，以及改进版 Llama 3.1 70B 和 8B 模型。除了与闭源模型相比具有显著更好的成本/性能比之外，405B 模型的开放性将使其成为微调和蒸馏更小模型的最佳选择。



除了发布这些模型外，我们正与一系列公司合作，以发展更广泛的生态系统。亚马逊、Databricks 和英伟达正在推出一整套服务，支持开发者微调 and 蒸馏自己的模型。像 Groq 这样的创新者已经为所有新模型构建了低延迟、低成本的推理服务。这些模型将在所有主要云平台上提供，包括 AWS、Azure、Google、Oracle 等。像 Scale.AI、Dell、德勤等公司已准备好帮助企业采用 Llama，并用他们自己的数据训练定制模型。随着社区的增长和更多公司开发新服务，我们可以共同使 Llama 成为行业标准，将 AI 的好处带给每个人。

Meta 致力于开源人工智能。我将概述为什么我认为开源是最好的开发堆栈，为什么开源 Llama 对 Meta 有好处，以及为什么开源人工智能对世界有好处，因此是一个长期可持续的平台。

## 为什么开源人工智能对开发者有好处

当我与世界各地的开发者、首席执行官和官员交谈时，我通常会听到几个主题：

**我们需要训练、微调和蒸馏我们自己的模型。**。每个组织都有其独特的需求，最适合的是使用不同规模的模型，这些模型可以根据他们特定的数据进行训练或微调。对于设备上的任务和分类任务，小模型足矣；而对于更复杂的任务，则需要大模型。现在，你可以利用最先进的 Llama 模型，用你自己的数据继续训练它们，然后将它们蒸馏成一个最适合你需要的模型尺寸——而无需让我们或任何其他其他人看到你的数据。

**我们需要控制自己的命运，不要被限制在闭源供应商那里。**许多组织不希望依赖他们无法自行运行和控制的模型。他们不希望闭源模型提供者能够更改模型、修改使用条款，甚至完全停止服务。他们也不想被限制在只有一个云平台拥有模型独家使用权。开源允许广泛的公司生态系统拥有兼容的工具链，使你可以轻松地在它们之间迁移。

**我们需要保护我们的数据安全。**许多组织处理敏感数据，需要加以保护，不能通过闭源模型的云 API 发送。还有一些组织根本不信任闭源模型提供者处理他们的数据。开源通过允许你在任何地方运行模型来解决这些问题。普遍认为，开源软件通常更安全，因为它的开发过程更加透明。

我们需要一个运行效率高且负担得起的模型。开发者可以在自己的基础设施上，以大约使用闭源模型如 GPT-4o 一半的成本，运行 Llama 3.1 405B 的推理，无论是面向用户的还是离线推理任务。

我们希望投资于将成为长期标准的生态系统。许多人看到开源的发展速度比闭源模型快，他们希望建立他们的系统在将给他们带来最大长期优势的架构上。

## 为什么开源人工智能对 Meta 有好处

Meta 的商业模式是为人们创造最佳的体验和服务。为此，我们必须确保始终能够获取最佳技术，并且不会被锁定在竞争对手的闭源生态系统中，从而限制了我们的创新能力。

我的一个重要经历是，由于苹果对我们在其平台上能够构建的内容有所限制，我们的服务受到了制约。从他们对开发者征税的方式，到他们随意应用的规则，再到他们阻止发布的所有产品创新，很明显，如果我们能够构建产品的最佳版本，而竞争对手无法限制我们的创新，Meta 和其他许多公司将能够为人们提供更好的服务。从哲学上讲，这是我坚信在人工智能和 AR/VR 中为下一代计算构建开放生态系统的主要原因。

人们经常问我是否担心通过开源 Llama 放弃技术优势，但我认为这忽略了大局，原因有几个：

首先，为了确保我们能够获取最佳技术，并且长期不会被锁定在闭源生态系统中，Llama 需要发展成为一个完整的工具生态系统，包括效率提升、硅片优化和其他集成。如果我们是唯一使用 Llama 的公司，这个生态系统就不会发展起来，我们的表现也不会比 Unix 的闭源版本更好。

其次，我预计人工智能的发展将继续非常具有竞争力，这意味着开源任何特定模型并不会在当时给予比下一个最佳模型更大的优势。Llama 成为行业标准的道路是通过持续保持竞争力、效率和开放性，一代又一代地发展。

第三，Meta 与闭源模型提供者的一个关键区别在于，出售对人工智能模型的访问并不是我们的商业模式。这意味着公开发布 Llama 并不会削弱我们的收入、可持续性 or 研究投资能力，而这对于闭源提供者来说则不然。

最后，Meta 有着长期的开源项目和成功的历史。我们通过发布服务器、网络和数据中心设计，并让供应链标准化我们的设计，通过 Open Compute

项目节省了数十亿美元。我们通过开源诸如 PyTorch、React 等领先工具，从生态系统的创新中受益。长期坚持这种方法对我们始终有效。

## 为什么开源人工智能对世界有好处

我相信开源对于实现积极的人工智能未来至关重要。人工智能比任何其他现代技术都有更大的潜力来提升人类的生产力、创造力和生活质量——并加速经济增长，同时推动医学和科学研究的进步。开源将确保全球更多的人能够获得人工智能的好处和机会，权力不会集中在少数公司手中，技术可以更均匀、更安全地在社会中部署。

关于开源人工智能模型的安全性正在进行辩论，我的看法是开源人工智能将比替代方案更安全。

我理解安全框架是我们需要防范两类伤害：无意的和故意的。无意的伤害是当一个人工智能系统可能会造成伤害，即使运行它的人没有意图这样做。例如，现代人工智能模型可能会无意中给出不良的健康建议。或者，在更具未来感的情景中，一些人担心模型可能会无意中自我复制或过度优化目标，从而损害人类。故意伤害是当一个不良行为者使用人工智能模型以达到造成伤害的目的。

值得注意的是，无意的伤害涵盖了人们对人工智能的大多数担忧——从人工智能系统将对数十亿使用者产生什么影响，到对人类来说真正灾难性的科幻情景的大部分。在这方面，开源应该更安全，因为系统更透明，可以广泛审查。从历史上看，开源软件因此更安全。同样，使用 Llama 及其安全系统如 Llama Guard 可能会比闭源模型更安全、更可靠。因此，关于开源人工智能安全性的大多数对话都集中在故意伤害上。

我们的安全流程包括严格的测试和红队，以评估我们的模型是否有能力造成重大伤害，目标是在发布前降低风险。由于模型是开放的，任何人都可以自己测试。我们必须记住，这些模型是由已经在网络上的信息训练的，所以当考虑伤害时，起点应该是模型是否能促进比从 Google 或其他搜索结果中快速检索到的信息更多的伤害。

当你考虑未来的机遇时，请记住，当今大多数领先的科技公司和科学研究都是建立在开源软件之上的。如果我们共同投资，下一代公司 and 研究将使用开源人工智能。



最重要的是，开源人工智能代表了世界上利用这项技术为每个人创造最大经济机会和安全的最佳机会。

## 让我们一起建设

对于过去的 Llama 模型，Meta 是自己开发然后发布的，但并没有过多关注构建更广泛的生态系统。这次发布我们采取了不同的方法。我们正在内部组建团队，让尽可能多的开发人员和合作伙伴能够使用 Llama，我们也在积极建立合作伙伴关系，以便生态系统中的更多公司也能为其客户提供独特的功能。

我相信 **Llama 3.1 的发布将成为行业的一个转折点**，大多数开发人员将开始主要使用开源，并且我预计这种方法只会从现在开始发展。我希望你能加入我们的旅程，将人工智能的好处带给世界上的每个人。

# 《评奥特曼的“闭源 AI”、关停 API、 割裂全球 AI 的言论和策略》

陆首群 2024. 7. 30

## 一、针对奥特曼的“闭源 AI”、关停 API、割裂全球 AI 的三个言论：

1、2024 年 6 月，奥特曼违背“开源 AI”初心转向实行“闭源 AI”策略。

他说：现在 OpenAI 研发的目标直接瞄准通用人工智能（AGI），“开源人工智能可能不是实现这一目标的最佳路径”。（奥特曼可能没有真正理解 AI 大师杨立昆指出的 AGI 的理论基础，即 AGI 不存在自回归大语言模型（LLM）并不是通往 AGI 的充分途径，因为他们缺乏智能生物的基本能力；真正的智能需要对物理世界的具体理解。

2、2024 年 5 月 13 日，奥特曼推出闭源的大模型 GPT4o，2024 年 6 月 25 日，奥特曼宣布对中国、朝鲜、伊朗、俄罗斯等一些用户关停 GPT4o API 的政策。

3、2024 年 7 月 26 日，奥特曼在《华盛顿邮报》上发表“谁将掌控 AI 的未来？”的声明。充满了极端的意识形态色彩。他说：“人工智能将由谁掌控，是我们这个时代最紧迫的问题”，“需要确保以美国为首的西方世界主宰 AI 这一领域”，“特别要防范中国，他们表示在 2030 年前要成为全球 AI 的领导者，他们 AI 的发展已经初具成效，呼吁建立以美国为首的民主国家的 AI 主权联盟，在联盟内部开源，对其他国家闭源，更要采取强有力的安全措施。”

奥特曼的上述三个言论遭到国内外众多开源和 AI 专家的批评和反对。

早在 2015 年,美国谷歌、脸谱、微软、IBM 四大 AI 初期的重镇,在发展 AI 时遇到瓶颈,同年一致将 AI 的工具、平台、框架、源代码、项目全部实行开源以解瓶颈之危。

2016 年,Linux 基金会进行基于开源的深度信息技术(如移动互联、物联网、云原生、区块链、人工智能等)的研究、推广工作。

2018 年,COPU 将推进基于开源的深度信息技术列为本开源联盟的主要任务之一。

2024 年 6 月 18 日,在奥特曼发表“闭源 AI”策略不久,陆首群谈话:“我预言,开源 AI 的竞争性研究者即将站在奥特曼身旁,与他一比高低!

同时,他发表《发展基于开源的人工智能》的文章,在文中他阐述:“依靠开源不但可加快 AI 的开发速度,提高开发质量,打通发展瓶颈,扩大生态,加强运维,反对垄断,在构建 AI 安全时更离不开开源;有人对免费的开源有所误解,其实在推出免费的开源社区发行版同时,还推出收费的开源商业发行版,以支持开源产业的发展。”

2023 年 7 月,开源大师 Jim Zemlin、Brian Behlendorf 在 COPU 的园桌会议上指出:开源是 AI 安全的保障。2024 年 7 月 2 日,谷歌、Meta 和扎克伯格嘲讽奥特曼成为构建闭源的头人,严厉批评奥特曼对中国关停大模型 GPT-4o 的 API 举措,认为这是一项愚蠢的丑行。在 Open AI 开发闭源 GPT-4o 时,谷歌、Meta 开发开源的 Gemini-2(9B、27B)和 Llama3.1,而开源的 Lama-3.1(405B 版本)超越了闭源的 GPT-4o(以后互有反复),处于竞争漩涡之外的法国 Kyutai 也开发了具有多模态功能的开源模型 Moshi,挑战闭源的 GPT-4o。在奥特曼向中国关停 API 时,激起了中国一批大模型的反制,平替 OpenAI 关停 API 接口,这次中国不但未输而且领先一局。

扎克伯格更声称:“开源 AI 是前进的道路”,“利用开源技术构建世界领先模型和生态系统”,“开源操作系统、开源大模型之争,不仅是开发者人才的竞争更是国家之间科学与产业的竞争”。扎克伯格还特别赞扬 Linux 在 AI 时代赢得了广泛的欢迎。

在 2024 年 3 月 18 日由本吉奥、辛顿、姚期智等几十位中外人工智能大师、专家签署的《北京 AI 安全国际共识》中,和在 2024 年 6 月 14~15 日成功举办的《北京智源大会》(伯克利的罗素、MIT 的泰格马克、清华大学的姚期智、张亚勤等,及 OpenAI、斯坦福、谷歌的代表 300 多位 AI 专家参加大会)上,把构建 AI 安全看作为 AI 研究的前沿,把建立国际统一的 AI 安全研究和治理看作为头等大事,为此离开“开源 AI”能够成事吗?!

## 二、借口意识形态划线，分裂全球对 AI 进行统一的安全研究和部署是走不通的

（奇文共欣赏：）

萨姆-奥特曼(Sam Altman)最新在《华盛顿邮报》上发表了一篇专栏文章，充满了极强的意识形态色彩,直接把对 AI 的态度上升到国家主权、阵营的级别，这是非常少见的聊政治、国际问题的公开发声。

奥特曼呼吁“建立以美国为首的 AI 主权联盟,确保 AI 仍然是自由和民主的载体。”

他在文中说：“AI 将由谁掌控,是我们这个时代最紧迫的问题”。他认为，“需要确保以美国为首的西方世界主宰 AI 这个领域,免受中国侵占”。

人工智能领域的突破，迫使我们面临一个战略选择：是要由美国及其盟国推动全球人工智能的发展、传播；还是任由一些不认同他们价值观的国家利用 AI 巩固和扩大他们的权力”没有中间道路可选。“现在是时候决定走哪条路了”。

他还说：“美国目前在 AI 领域处于领先地位,但这种领先并不巩固”。“全球范围内的中国政府投入巨资,试图追赶并超越我们。赢得 AI 竞争的国家将拥高权力和影响力。同时,中国也明确表示,其目标是在 2030 年前成为全球 AI 的领导者。”“他们的 AI 的发展已经初具成效，像 Chat GPT、Capilot 这样的系统作为助手发挥作用，无论是提供有限帮助还是在特定领域如软件工程的代码生成中发挥更大的作用。人类社会正处于一个关键时期。”

他说：“确保世界能够体现民主愿意，美国的公共部门和科技行业需要做到以下四点”：

“首先，制定了强有力的安全措施，保护我们的技术优势，同时保持私营企业的创新能力。这包括加强网络防御和数据中心安全，防止关键知识产权如模型权重和 AI 训练数据的泄露。”

“第二，投资于基础设施建设,因为他们是 AI 发展的基础。过去对光纤网络等基础设施的投资帮助美国在数字时代取得领先,并继续支撑其在 AI 领域的优势。美国需要与私营部门合作，进一步建设数据中心，电力设施等，同时创造新的就业机会。此外，还需要培养下一代 AI 创新人才。”

“第三，制定一套连贯的商业外交政策,涉及 AI 技术的出口管制和外国投资规则。对于某些敏感的 AI 技术,如芯片、训练数据和模型代码,可能需要采取限制措施。同时，要确保盟友能够轻松获取开源模型，以加强技术联盟。”

“第四,探索新的合作模式,与其他国家就 AI 问题保持对话。可以考虑建立类似国际组织的机构来管理 AI 事务,或者通过投资基金等方式支持遵守美西方国家发展 AI。”

作为领先国家和技术领先者，我们肩负着这一责任，现在是采取行动的时候了。

奥特曼高举美西方假民主大旗，以意识形态划线，分裂全球对人工智能的安全进行统一开发和部署，是行不通的！



敬请关注联盟微信公众号  
COPU开源联盟



扫描二维码  
获取往期资料

---

中国开源软件推进联盟秘书处

电话: +86 010-88558999

联盟公共邮箱: [office@copu.org.cn](mailto:office@copu.org.cn)

联盟官网: <http://www.copu.org.cn>

地址: 北京市海淀区紫竹院路66号赛迪大厦18层

---