人工智能文集

第十九集

中国开源软件推进联盟
China Open Source software promotion Union

人工智能文集

第十九集

中国开源软件推进联盟

China Open Source software promotion Union

目 录

AI 的下一波巨浪

Eric Schmidt

2024.8.19

最近两天全球科技圈的热点，被谷歌前 CEO 埃里克·施密特(Eric Schmidt)无意中抢走。他受邀去斯坦福计算机学院做一场交流，埃里克·施密特以为是一场内部分享，或许也加上当天心情不错，全程放飞自我，直到主办方提醒正在直播，结果……埃里克·施密特虽然当场石化，但却意外成就了一次“大实话、高营养”的演讲。

埃里克·施密特的演讲主要谈及:

AI 技术和市场的动态；AI 的投资与国家安全；谷歌、OpenAI 与企业文化；AI 与地缘政治；战争与 AI 的应用；知识的本质与 AI；AI 隐私以及数据安全；以及 AI 对教育和学术研究，创业与就业的影响……

以下，enjoy:

主持人: 今天的嘉宾其实无需过多介绍。我记得大约 25 年前第一次见到 Eric，那时他作为 Novell 的首席执行官来访斯坦福商学院。从那时起，他做了很多事情，他在 Google（大概是从 2001 年开始）和 Schmidt Futures（从 2017 年开始）做了很多事情，还有很多其他的事情你们可以查询了解。但他只能待到下午 5 点 15 分，所以我想我们直接进入问题环节。我知道你们也有一些问题。我这里有一些我写下的问题，但我们在楼上刚刚谈论的内容更有趣。所以我想从那开始，Eric，如果你不介意的话。

AI 技术和市场的动态

主持人: 你预见 AI 在短期内，我认为你定义的是未来一两年，会有怎样的发展?

Eric: 事情变化得如此之快，我感觉我每六个月都需要做一次新的演讲，讲述即将发生的事情。在座的有没有人，一群计算机科学家在这里，有没有人可以解释一下什么是百万 token 的上下文窗口，为其他同学解释一下?

学生: 在这里。基本上，它允许你用百万个 token 或者百万个词进行提示。所以你可以提出一个上百万词的问题。我了解到，这是当前通识教育关注的一个重要方向。Eric: 接着，Anthropic 现在是 20 万,他们的目标是 100 万,以此类推。你可以设想 OpenAI 也有类似的目标。

谁能给出 **AI 智能体**的技术定义?

学生: AI 智能体基本上是执行某种活动的实体。这可能涉及在网上，代表你处理一些事情，可能是许多不同的事项，类似这些。所以，一个智能体就是执行某种任务的实体。另一个定义是，它是一个大语言模型，具有状态和记忆功能。

Eric: 再来一个，计算机科学家，你们中有谁能解释什么是"**将文本转化为行动**"?

学生: 就是把文本转变成行动。而不是把文本转化成更多的文本。

Eric: 另一个定义是，将语言转化为 Python 代码。这是我一直不想看到的编程语言。然而，目前所有的 AI 工作都是在使用 Python

进行的。有一种新的语言叫 Mojo，刚刚出现，看起来他们终于解决了 AI 编程的问题。但我们还要看，这是否能在 Python 的主导地位下生存下来。主持人: 为什么 NVIDIA 的价值和地位如此之高，而其他公司却在挣扎呢？

Eric: 我认为，**这主要是因为，大量的代码需要在 CUDA 优化下运行，而这是只有 NVIDIA 的 GPU 才支持的**，所以，其他公司可以制造他们想要的任何东西,但是如果他们没有10年的软件开发经验，就不可能有机器学习优化。我个人喜欢把 CUDA 想象成 GPU 的 C 语言，对吗？这就是我喜欢的看法。它成立于 2008 年。我一直觉得它是一种糟糕的编程语言，然而，它却成为了市场主导。还有一点值得注意。有一套开源库，它们针对 CUDA 进行了高度优化，而对其他平台的优化却很少。每个构建所有这些堆栈的人——这在任何讨论中都被完全忽视了。这在技术上被称为 VLLM 以及其他一大堆类似的库。它们都是专门为 CUDA 而优化的，对于竞争对手来说，很难复制这个。

主持人: 那么，这些观点对我们来说有何影响或意义呢？

Eric: 在接下来的一年里，我们将看到非常大的上下文窗口、智能体和"文本转行动"等新技术的兴起，当它们能够大规模应用时将对世界产生的影响将超出我们目前的理解范围。这种影响将远超过社交媒体所带来的影响，我个人是这样认为的。以下是我的原因。在一个上下文窗口中，你基本上可以将其作为短期记忆。我对上下文窗口能达到如此之长感到惊讶。这主要由于它的计算和处理难度很高。短期记忆

的有趣之处在于，当你输入信息，比如你问一个问题，"读了 20 本书，你输入这些书的文本作为查询，然后你说，′告诉我它们说了什么′"。它会忘记中间的部分，这与人类大脑的工作方式相似。对吗？这就是我们现在的状况。

主持人: 关于智能体呢？

Eric: 关于智能体，现在有人正在开发基于大语言模型的智能体，他们的做法是阅读一些像化学一样的学科，发现其内在原理，然后进行测试。然后他们将这些知识融入到他们的理解中。这是非常强大的。我提到的第三个要点是"文本转行动"。那么，我来举个例子，政府正在尝试禁止 TikTok，我们拭目以待看结果如何。如果 TikTok 被禁，我建议你们每个人都这样做，告诉你的大语言模型，接下去的操作。复制一份 TikTok。获取所有用户信息。获取所有音乐资源。加入我的个性化设置。在接下来的 30 秒内编制这个程序。然后发布出去。如果一小时内它没有迅速传播开来，那就沿着同样的思路尝试另一种方式。这就是命令。一步接一步，就这样。明白了吗？你知道这有多强大吗？如果你能从任意自然语言转换为任意数字命令，这在这个情况下就相当于 Python，试想一下，如果地球上的每个人都有属于自己的程序员，他们会真正按照你的要求去做事，而不是像我手下的那些程序员那样并不总是按照我说的去做。明白了吗？在场的程序员都明白我在说什么。所以，想象一下，有一位既不自大，又会真正按照你的要求去做事的程序员，你甚至不需要付他一大笔工资。而且这样的程序无穷无尽。

主持人：这一切都将在未来一两年内实现？

Eric：马上就要到来。**这三件事，我深信只有结合这三件事，下一波浪潮才会到来。那么，你问的是接下来会发生什么。我的观点每六个月会有所改变，这就像一个周期性的摆动**。比如说，现在，那些前沿模型（只有三个，我待会会详细介绍）与其他所有人之间的差距，我感觉正在变大。六个月前，我坚信这个差距正在缩小。于是我在一些小公司投入了大量的资金。但现在，我对此已不再那么确定了。我现在正在和大公司们交谈，他们告诉我他们需要投入100亿、200亿、500亿甚至1000亿。比如说，Stargate的投入就达到了1000亿，对吧？这确实非常困难。

AI的投资与国家安全

Eric：Sam Altman是我的密友。他认为这可能需要投入高达3000亿，甚至更多。我向他指出，我已经计算出了这需要的能源量。然后，在完全公开的精神下，我上周五去了白宫，告诉他们我们需要与加拿大建立最紧密的关系。因为加拿大有非常好的人，参与了人工智能的发明，还有大量的水力发电资源。因为我们国家没有足够的能源来完成这件事。另一个选择就是让阿拉伯人来资助。我个人非常喜欢阿拉伯人。我在那里待过很久，对吧？但他们不会遵守我们的国家安全规则，而加拿大和美国是共同遵守安全规则的三方联盟（或三国集团）的一部分。因此，**对于这些价值3000亿美元的数据中心来说，电力开始变得稀缺。顺便说一下，如果你沿着这个逻辑走下去，我为什么要讨论CUDA和NVIDIA呢？如果有3000亿美元都要流向NVIDIA，你应该**

知道在股市里应该怎么做。这不是股票推荐，我并不是许可证发放者。（观众笑）部分原因是，我们需要更多的芯片，但英特尔正在从美国政府和 AMD 那里得到大笔资金，他们正准备在韩国建造半导体工厂。

主持人： 有谁的计算设备里有英特尔的电脑或者芯片呢，请举手。

Eric： 看来，垄断不再是什么大问题了。

主持人： 这正是我想说的。

Eric： 他们曾经垄断过。

主持人： 没错。

Eric： 而现在 Nvidia 有垄断。

主持人： 那些对进入的障碍呢？例如 CUDA，还有其他的，就像我前几天和 Percy Lanny 聊天时提到的。他根据训练模型所能获得的设备，会在 TPUs 和 NVIDIA 芯片之间做选择。

Eric： 那是因为他别无选择。如果我有无限的资金，我会今天选择 NVIDIA 的 B200 架构，因为它运行更快。我并不在这里提倡什么，我只是想说有竞争是好事。我和 AMD 的 Lisa Su 有过长时间的交谈。他们正在开发一种能将你描述的这种 CUDA 架构转换为他们自己的架构，即 RockM。目前它还不能完全运行，他们正在努力改进。

谷歌、OpenAI 与企业文化

主持人：你在谷歌工作了很长时间，他们是 Transformer 架构的发明者。

Eric： 是彼得，都是彼得的错。

主持人：那里有像彼得和杰夫·迪恩这样的出色人才。但现在，他们似乎已经失去了对 OpenAI 的主动权。我看到的最新排行榜上，Anthropic's Claude 是榜首。我问过 Sundar 这方面的问题，他并没有给我一个明确的答案。或许你能给出一个更明确或更客观的解释。

Eric：我现在已经不再是谷歌的员工了。确实如此。我要坦白的说，谷歌认为工作与生活的平衡，早点下班、以及在家工作比赢得比赛更重要。（笑）创业公司之所以能够成功，是因为员工拼命工作。很抱歉如此直言不讳，但事实是，如果你们离开大学去创办公司，你不会允许员工在家办公，而且每周只来公司一天，如果想要与其他创业公司竞争的话。

主持人：Google 创业初期，Microsoft 就是这样。

Eric：对的。

主持人：但现在似乎——

Eric：在我们这个行业里有很多公司，以真正创造性的方式赢得市场并在某一领域取得主导地位，但却未能完成下一次转型。这种现象很常见，并且有很多文献记录。我认为，创始人是特殊的，他们需要掌控一切，与他们共事可能会很艰难，他们会给员工施加很大的压力。**我们可能并不喜欢马斯克的个人行为，但你看看他是如何推动员工的。我曾和他共进晚餐，当时他在蒙大拿州，而那天晚上 10 点他要飞往另一个地方，凌晨 12 点与 X.AI 开会。对吧？你想想看吧。**

主持人：我曾去过台湾，有着完全不同的文化，他们（台积电）让我印象深刻的一点是，他们有一条规定：这些刚入职的优秀物理学博士

需要在地下一层的工厂工作。你能想象让美国的物理博士去做那样的事吗？几乎不可能。他们的工作态度和我们有所不同。

Eric: 而问题在于，我之所以对工作要求这么严格，是因为这些系统具有网络效应，时间是非常关键的。在大部分业务中，时间其实不那么重要。你有充足的时间。可口可乐和百事可乐会一直存在，他们之间的竞争也会持续，这一切都在慢慢发展。我和电信公司打交道时，一般的电信交易要花费 18 个月才能完成。实际上，没有任何事情需要花费 18 个月去完成。要迅速行动。我们现在处于最大的发展期，最大的进步期。这也需要一些疯狂的想法。比如当微软与 OpenAI 达成交易时，我认为那是我所听过的最愚蠢的想法。将 AI 的主导地位让渡给 OpenAI，包括 Sam 和他的团队，这简直太疯狂了。在微软或其他任何地方，都没有人会这么做。然而现在，他们正在朝着成为最有价值的公司的目标前进。他们和苹果公司的竞争激烈。苹果公司并没有一个好的 AI 解决方案，而微软看起来已经成功了。

AI 与地缘政治

主持人: 在国家安全或地缘政治利益方面，**你认为 AI 将如何在与中国的竞争中发挥作用？**

Eric: 我曾经是一个 AI 委员会的主席，我们对此进行了非常详细的研究。你可以去看看。它有大约 752 页。我只是总结一下，我们现在处于领先位置。我们需要保持这种领先地位，并且需要大量的资金来做到这一点。我们的主要对象是参议院和众议院。由此促成了《芯片

法案》以及其他相关立法。如果你假设前沿模型不断发展，少数开源模型也在进步，**很可能只有少数国家能够参与这场竞争。我是指国家，而不是公司。那么这些国家是谁呢？有大量资金、丰富人才、强大教育系统以及获胜意愿的国家。美国就是其中的一个。中国也是。还有其他国家吗？我不知道，也许有。在你们这一代人的有生之年，围绕知识霸权的美中对抗将会是主要的斗争**。因此，美国政府基本上禁止了 NVIDIA 芯片出口到中国，尽管他们并不愿明说这是他们的初衷，但实际上确实如此。我们在芯片制造技术上大约领先 10 年。在次紫外光刻（sub-DUV），即小于 5 纳米的芯片方面，我们大约领先 10 年。

主持人：10 年，这么久？

Eric：大概 10 年。

主持人：哦。

Eric：所以，以现在的情况为例，我们比中国领先了几年。我猜我们可能还会领先中国几年。中国对此非常不满。他们对此感到非常沮丧。这是个大问题。这是特朗普政府的决定，并且拜登政府也同样执行了这个决定。

主持人：你觉得现在的政府和国会听取你的建议了吗？你认为他们会进行如此大规模的投资吗？显然，《芯片法案》已经出台，但除此之外，是否还会建立一个庞大的 AI 系统？

Eric：你知道，我领导的是一个非正式、特设的、不受法律约束的小组。这和违法是不同的。确切地说，只是为了明确。这包括所有的同行。在过去的一年里，这些同行提出了一些理论基础，最终成为了拜

登政府《AI 法案》的核心内容，这是历史上最长的总统行政命令。

主持人：你是在谈论特殊竞争研究项目吗？

Eric：不，这是来自行政办公室的实际法案。他们现在正忙于实施细节。到目前为止，他们做得很好。例如，过去一年中我们讨论的一个问题是，如何在一个已经学习到危险内容的系统中检测这些危险，但是你不知道该问它什么？换句话说，这是一个核心难题。系统可能学到了一些有害的东西，但它无法告诉你学到了什么，而你也不知道该如何询问它。而且威胁种类繁多，比如，它学会了一种你不知道如何询问的新的化学混合方式。因此，人们正在努力解决这个问题。但最终我们在给他们的备忘录中写道，有一个阈值，我们任意设定为 10 的 26 次方的浮点运算，这在技术上是一个计算量的衡量标准。超过这个阈值，你必须向政府报告你正在进行这种操作。这就是规则的一部分。欧盟为了有所区别，把阈值定为 10 的 25 次方。

主持人：是的。

Eric：但这差距其实很小。我认为所有这些区别都会消失，因为现在的技术——专业术语是联邦学习技术，基本上你可以将不同部分联合起来进行训练。所以我们可能无法让人们完全免受这些新威胁的影响。据传言，这也是 OpenAI 必须这样训练的部分原因，因为电力消耗太大无法集中在一个地方进行训练。

战争与 AI 的应用

主持人：好了，让我们谈谈正在进行的真正战争。我知道你非常关注乌克兰战争，尤其是，关于“白鹳”项目，我不确定你能谈多少，关

于用 500 美元的无人机摧毁 500 万美元的坦克。这个改变了战争方式吗？

Eric: 我曾在国防部工作了七年，试图改变我们管理军队的方式。虽然我并不特别喜欢军队，但是军队的运行开支非常大，我想看看我能否对此提供一些帮助。而现在看来，我觉得我基本上失败了。他们给了我一枚勋章，所以可能失败者也能得到勋章吧，或者随便怎么说。但我对自己的批评是，什么都没有真正改变，美国的体系不会带来真正的创新。看着俄罗斯人用坦克摧毁有老人和孩子的公寓楼，我感到非常愤怒。所以我决定和你的朋友、斯坦福大学的前任教授塞巴斯蒂安·特鲁恩，以及一批斯坦福人一起创办一家公司。其实，我们的目标主要有两个。首先是用复杂而强大的方式将 AI 应用于这些机器人战争中，其次是降低机器人的成本。你可能会好奇，一个像我这样的自由派为何会有这样的想法？答案是，现有的军队理论以坦克、炮兵和迫击炮为主，而我们可以消除它们。我们可以让入侵一个国家的代价，至少在陆地上，几乎是不可能的。这应该可以避免大规模的陆地战争。

主持人: 这确实是一个很有趣的问题，这种方式是否能让防守方获得更多优势？我们能否做出这样的区分呢？

Eric: 在过去的一年里，我一直在做这个，我学到了很多关于战争的知识，而这些知识我原本不想知道。其中一个关键点是，进攻方总是占据优势，因为他们总能压倒防御系统。所以，作为国家防御策略，拥有一套强大的进攻机制是很有必要的，以备不时之需。而我和其他

人正在构建的系统将能够实现这一点。由于系统的运作方式，我现在是一名持证军火商。所以我现在既是计算机科学家，商人，也是军火商。（笑）

主持人：我很抱歉地说——这算是一种职业进步吗？

Eric：我不太确定，但我并不建议你把这作为你的职业发展路径。我建议你还是继续做 AI。由于法律的规定，我们是以私人方式进行这些工作，并且政府对此予以支持，因此我们直接进入乌克兰，随后战争开始了。不详细展开，但局势非常严峻。我认为，如果在五月或六月，俄罗斯如预期那样进行军事集结，乌克兰将会失去大片的领土，并开始逐渐失去整个国家。所以情况非常严重。

主持人：如果有人认识 Marjorie Taylor Greene，我建议你从通讯录中删除她。因为她就是那个，一个人阻止了数十亿美元援助这些援助本可以拯救一个重要的民主国家。

知识的本质与 AI

主持人：我想谈谈一个稍微带有哲学性质的问题。去年你和亨利·基辛格以及丹·赫特洛克写了一篇关于知识本质及其演变的文章。前几天我也和别人讨论了这个话题。对于历史上的大部分时间，人类对宇宙的理解更多是神秘的，然后出现了科学革命和启蒙运动。而在你的文章中，你们提出了一个观点，现在的模型变得如此复杂和难以理解，以至于我们不再真正知道其中发生了什么。我要引用理查德·费曼的一句话。他说：“我不能理解我无法创造的东西”我最近看到这句话。但现在人们能够创造出一些东西，却并不真正理解其中的原理。知识

的本质是否正在发生变化？我们是否要开始接受这些模型的结果，而不再需要它们解释给我们听？

Eric：我想，可以将其比作青少年。如果你有个十来岁的孩子，你知道他们是人类，但你却无法完全理解他们的想法。（笑）然而，我们的社会已经适应了青少年的存在，对吧？他们总会长大成人。我是认真的。因此，我们可能会有无法完全理解的知识系统，但我们了解它们的边界，对吗？我们理解它们的能力范围。这可能是我们能够获得的最好结果。

主持人：你认为我们能理解这些边界吗？

Eric：我们会变得越来越好。我每周都会和我的团队会面，我们的共识是，最终，你会使用所谓的对抗性 AI，实际上会有一些公司，你可以雇用他们，付钱让他们去破坏你的 AI 系统。

主持人：就像网络安全中的红队一样。

Eric：那么，将会是 AI 红队，而不是现在的人类红队，你将会看到整个公司和行业的 AI 系统，它们的任务是挖掘现有 AI 系统的漏洞，特别是那些我们无法理解的知识点。我认为这个观点是有道理的。对于斯坦福来说，这也是一个很好的项目，因为如果有一个研究生能够弄清楚如何攻击这些大型模型并理解它们的运作，那将为下一代技术积累的宝贵经验。我觉得这两者会齐头并进。

学生提问环节

学生问题一：如何让 AI 真正按照我们的意愿行动？

学生：之前你提到过这点，和你刚才的评论有关，AI 需要真正按照

我们的意愿行事。你刚刚提到了对抗性 AI，我想如果你能再详细说明一下就更好了。除了计算能力肯定会增加，我们可以有更强大的模型，但是如何让它们真正按照我们的意愿行动，这个问题你觉得解决了吗？在我看来，这个问题似乎还没有完全解答。

Eric: 好吧，你必须认为，当前的问题会随着技术的进步而减少，对吗？我并不是说这些问题会完全消失。然后你还需要假设，我们会有对 AI 有效性的测试方法。因此，我们必须有一种方式，去确认事情是否成功。举刚才我提到的 TikTok 竞争对手的例子，我并不是建议你非法盗用其他人的音乐。如果你是硅谷的企业家，希望你们都能成为的话，如果这个产品火了，那么你就需要雇一大堆律师来收拾烂摊子，是吧？但如果没有人使用你的产品，那你盗用了所有内容也无关紧要。当然，不要引用我的话，是吧？（观众笑了）

主持人: 对，你正在被摄影机记录。

Eric: 是的，没错。（观众笑了）但你明白我的意思。换句话说，硅谷会进行这样的尝试，并收拾残局，这通常就是事情的运作方式。所以在我看来，你会看到越来越多的性能系统，甚至有更好的测试，最终会有对抗性测试，这将保证它们的行为在一个可控的范围内。我们通常把这称为链式思维推理。人们相信，在未来几年里，你将能够实现长达 1000 步的思维链条推理。做这个，再做那个。就好比是在制作菜谱。你可以照着菜谱做，并测试它是否产生了正确的结果。这就是系统运作的方式。

学生问题二：AI 进步的驱动力是什么？

学生: 总的来说，你对 AI 进步的潜力似乎持有非常积极的态度。我很好奇，你认为什么会推动这个进步呢？是更强大的计算力量？还是更多的数据？或是基础性或实质性的变化？

Eric: 是的。现在投入的资金数额简直令人难以置信。我选择投资于几乎所有领域，因为我无法确定谁会是最后的赢家。跟随我投资的资金规模非常庞大，我觉得这部分原因是早期投资的回报已经出现，那些大资金的投资者，他们对他们自己在做什么并不清楚，他们只知道他们的投资中必须包含 AI 成分。现在所有的投资看似都与 AI 有关，所以他们无法区分哪些是真正的 AI 投资。我把 AI 定义为能够学习的系统，真正能够自我学习的系统。所以我认为这是其中的一个驱动力。另一个驱动力是，有许多非常复杂的新算法，这些算法可以说是超越了 Transformer。我的朋友，也是长期的合作伙伴，创造了一种新的非 Transformer 架构。我在巴黎资助的一个团队声称他们也做出了相同的成果。这里有巨大的发明潜力，斯坦福大学也有很多创新。最后一点是，市场普遍认为智能的发明将带来无尽的回报。比如说，你投资了一家公司 500 亿美元，你必须从智能中获得大量的收益才能回本。因此，我们可能会经历一场大的投资泡沫，最终市场会自行调整。这在过去一直如此，而这一次也极有可能是如此。

学生问题三：AI 开源与闭源的争论

Eric: 你之前提到，你认为领先者正在逐渐拉开与其他人的差距，现在。这个问题可以这样理解，有一家叫做 Mistral 的法国公司，他们的表现非常出色，我也是他们的投资者。他们已经推出了第二个版本，

但由于成本过高，第三个版本可能无法推出。他们需要收入，不能免费提供模型。所以，我们行业中关于开源与闭源的争论非常激烈。而我的整个职业生涯都是基于人们愿意共享开源的软件。我的一切都与开源有关。谷歌的许多基础设施都是开源的。我在技术上所做的一切也是如此。然而，可能是因为资本成本如此巨大，这从根本上改变了软件开发的方式。你和我之前谈到的，我认为软件程序员的生产率至少会提高一倍。有三四家软件公司正在努力实现这一目标。我投资了所有这些公司。本着这种精神。他们都在努力提高软件程序员的效率。我最近接触的一个最有趣的公司叫 Augment。我总是把它想象成单个程序员，但他们告诉我，“我们的目标并不是他。我们的目标是那些上 100 人的开发团队，代码行数达到数百万行，谁也不知道里面发生了什么。”这确实是一个非常适合 AI 解决的问题。他们能赚钱吗？我希望他们能。有很多问题需要解答。

学生问题四：上下文窗口、智能体和“文本转行动”的组合重要性

学生：你好。你一开始提到，上下文窗口的扩展、智能体和将文本转化为行动的组合将产生无法想象的影响。首先，为什么这种组合如此重要呢？其次，我知道你没有水晶球，不能预测未来。但为什么你认为这些影响将超出我们的想象？

Eric: 这有助于解决新“时效性”问题。目前的模型需要训练大约一年半的时间。包括六个月的准备，六个月的训练，六个月的微调。所以它们总是过时的。通过扩展上下文窗口，你可以提供最新的信息。例如，你可以在上下文中询问关于哈马斯与以色列战争的问题。这是

非常强大的。它让模型能够像谷歌一样保持时效性。

关于智能体，我给你一个例子。我成立了一个基金会，该基金会资助了一个非营利组织。首先，如果在场有化学家，我并不真正了解化学。有一个叫做 ChemCrow 的工具，它是一个基于大语言模型的系统，学习了化学知识。他们使用这个系统生成关于蛋白质的化学假设，然后在实验室测试一晚上，之后它会学习。这大大加速了化学、材料科学等领域的研究进程。这就是智能体模型的一个例子。

然后对于“文本到行动”，可以理解为你拥有大量廉价程序员，对吧？我们可能还不清楚会发生什么——这是你的专长领域——当每个人都有自己的程序员时会发生什么。我并不是在谈论开灯和关灯的事情。想象一下——再举一个例子——**比如，你不喜欢 Google。于是你说，给我打造一个可以与 Google 竞争的搜索引擎。没错，你一个人就能打造一个能与 Google 抗衡的搜索引擎，搜索网页，有一个用户界面，照原样山寨一份，以一种有趣的方式在其中融入生成式 AI，用 30 秒钟完成，看看效果如何。很多人因此相信，包括 Google 在内的许多大公司，可能会受到这种方式的威胁。我们拭目以待。**

关于 AI 和虚假信息的讨论

主持人：有很多通过 Slido 提交的问题，有些得到了投票。这里有一个，我们去年讨论过这个问题。我们如何阻止 AI 影响公众舆论，制造假消息，特别是在即将到来的选举期间？有什么短期和长期的解决方案呢？

Eric: 在这次即将到来的选举中，以及在全球范围内，大部分的假消息都会出现在社交媒体上。而社交媒体公司没有足够的组织能力来管理这些问题。以 TikTok 为例，有很多指控称 TikTok 偏袒某些类型的误导信息，而忽略其他类型的误导信息。许多人声称中国强迫他们这么做，但我没有看到任何证据。我觉得这是个混乱的局面。我们这个国家必须学会批判性思考。这对美国来说可能是个不可能的挑战。但是，别人告诉你的，并不一定就是真的。

主持人: 事情可能会走向另一个极端，导致真实的信息也不再被人们相信吗？有些人称这为认识论危机，比如马斯克就说过，“我从未做过那件事。证明它。”

Eric: 以唐纳德·特朗普为例。我认为我们社会存在信任问题。民主制度有时也会失败。我认为对民主制度最大的威胁是虚假的信息，因为我们会变得非常擅长制造这些虚假信息。当我管理 YouTube 的时候，我们在 YouTube 上面临的最大问题是有人上传虚假的视频，然后有人因此受害甚至丧生。我们有一个禁止涉及死亡信息的政策，令人震惊。我们致力于解决这个问题，但这是一项艰巨的任务。而且，这还是在生成式 AI 出现之前的事情。

主持人: 嗯，所以——

Eric: 我没有好的答案。

主持人: 有一个技术解决方案，这并不是答案，但可能有助于缓解这个问题，我不明白为什么这种方法没有被更广泛使用，那就是公钥认证。当乔·拜登发表演讲时，为什么不像 SSL 那样进行数字签名呢？

或者，比如那些名人或公众人物，他们不能有一个公钥吗？

Eric: 对，这就是公钥的一种形式，然后有一种确定的方式了解系统如何——

主持人: 嘿，当我向亚马逊发送我的信用卡信息时，我知道这是亚马逊。

Eric: 我和乔纳森·海德一起撰写并发表了一篇论文，他一直在研究生成焦虑问题。这篇论文并未产生任何影响。他是个非常出色的传播者，我可能不如他。所以我得出的结论是，系统没有按照你所说的那样组织起来。

主持人: 你有一篇论文赞同我们的做法。

Eric: 也即是对你的建议的赞同。

主持人: 好的。

Eric: 我的结论是，总的来说，**CEO 们都在努力最大化收入。为了最大化收入，他们最大化用户参与度。为了最大化用户参与度，他们最大化煽动性内容。算法选择煽动性内容，因为这会带来更多的收入，**对吧？由此，算法更倾向于推荐些极端的内容。这并不是单方面的问题。我并不是在这里做出任何偏袒的声明。这是一个问题。这个问题必须得到解决。

在一个民主制度中，我对 TikTok 的解决方案是，我们之前私下讨论过，在我年轻时，有一个叫做"平等播放时间"的规则。因为 TikTok 实际上并不仅仅是社交媒体，它更像是电视，对吧？有程序在幕后控制你。根据统计，TikTok 在美国的用户每天观看 90 分钟，每个用户

平均观看 200 个 TikTok 视频。这是很大的数量，对吧？因此，政府必须采取某种形式的平衡措施。

关于大语言模型对经济的影响以及学术界的选择

学生：大语言模型的经济影响，比如劳动力市场的影响，比你最初预期的要慢。Chegg（在线教育技术公司）和一些服务人员受到了影响。你认为学术界应该获得 AI 方面的补贴，还是应该选择与业界的大公司合作？

Eric: 我非常努力地推动为大学建立数据中心。如果我是这里的计算机科学系教师，我会非常失望，因为无法与研究生一起构建能够进行博士研究的算法。我被迫和这些公司合作。我认为这些公司在这方面并没有表现出足够的慷慨。我与许多你们认识的教师交谈，他们花费大量时间等待 Google Cloud 的配额。这是非常糟糕的情况。这个领域正在快速发展。我们希望美国能胜出，我们希望美国的大学——出于多种原因，我认为正确的做法是把资源提供给他们。我在这方面做了很多努力。

至于劳动力市场的影响，我将这个问题交给这里的真正专家。作为 Eric 教导的业余经济学家，我坚信那些接受过大学教育、从事高技能工作的人将会适应，因为人们将会与这些系统一起工作。我认为这些系统与以往的技术浪潮没有区别。那些危险的工作和不需要太多人为判断的工作将被取代。

关于计算机科学教育的未来

学生: 我对“文本到行动”及其对计算机科学教育的影响非常感兴趣。

你认为计算机科学教育应该如何变革以适应新时代？

Eric: 我认为计算机科学专业的本科生群体将始终有一位编程搭档。因此，当你学习你的第一个 for 循环时，你会有一个工具作为你的伙伴。这就是未来的教学方式，教授会讲解概念，但你会通过这种方式参与其中。这是我的猜测。

关于非 Transformer 架构与数学创新

学生: 你提到了一些让你兴奋的非 Transformer 架构。我想其中一个被提到的是状态模型，但现在还有一种是更长上下文的模型。我更好奇的是你在这方面看到的情况。

Eric: 我对数学的理解还不够深入，但我非常高兴我们为数学家们提供了工作机会，因为这里的数学太复杂了。但基本上，他们是在采用不同的方法来进行梯度下降和矩阵乘法，使其更快更好。Transformer 架构是一种能够同时进行乘法运算的系统化方式，这是我理解的方式。它和其他的很像，但数学部分不同。我们将继续关注这些新的数学进展。

关于中美关系与国家安全

学生: 你在关于国家安全的论文中提到，中美两国在现代架构的帮助下处于关键地位。接下来的 10 个国家，以及稍后的一组国家，都是美国的盟友，或者与美国盟友关系密切。我很好奇你对这些国家，尤其是那些中间地带但并非正式盟友的国家有什么看法。它们有多大可能愿意加入我们的安全阵营？是什么因素可能会阻碍它们加入？

Eric: 最值得关注的国家是印度，因为顶尖的 AI 人才从印度来到美

国，我们应该让印度保留一部分顶尖人才。他们没有我们这里如此丰富的培训设施和项目。在我看来，印度是这一领域的重要摇摆国家。中国已经没有机会了，不会再回来。日本和韩国显然站在我们这一边。台湾在硬件方面很出色，但他们的软件开发较弱。在全球范围内，其他大国的选择并不多。欧洲因为欧盟的政策和管理机制陷入困境，这并不是什么新鲜事。我与他们斗争了 10 年，但他们仍然有许多限制，导致我们在欧洲开展研究非常困难。法国还有一些机会，但我不太看好德国。其余的国家都不够有影响力。

关于编程与 AI

学生: 我是编译器工程师。考虑到你设想这些模型将具备的能力，我们还需要花时间学习编程吗？

Eric: 是的，这是一个老生常谈的问题，为什么你已经会说英语还要学习英语呢？是为了让你的英语更好。你确实需要理解这些系统的工作原理，我对此非常坚定。

关于分布式计算与大语言模型的训练

学生: 你是否探索过分布式环境？虽然构建大型集群很困难，但 MacBooks 的性能很强大。你认为类似“Folding@home”的想法适合训练吗？

Eric: 是的，我们对此进行了深入研究。算法的工作方式是你有一个非常大的矩阵，基本上是一个乘法函数。系统的速度完全受限于内存到 CPU 或 GPU 的传输速度。Nvidia 的下一代芯片已经将所有这些功能整合到了一块芯片中，如今的芯片已经足够大到需要将它们都粘合

在一起。短时间内，分布式计算在训练大语言模型方面的进展可能有限。

关于数据隐私与版权

学生：在 ChatGPT 发布后，OpenAI 因使用他们的作品进行训练而被《纽约时报》起诉。你觉得这种情况会如何发展？这对数据隐私有何影响？

Eric：我曾经从事过大量的音乐版权工作。我了解到在 60 年代，有一系列诉讼最终达成了一项协议：每次播放你的歌曲，你都能获得一定的版权费。我猜 AI 领域可能会有类似的情况。可能会有很多诉讼，最后可能会达成一种协议，要求你必须支付你的一部分收入，以便使用类似 ASCAP BMI 的版权管理机构。

关于 AI 领域的垄断与反垄断法规

学生：现在看起来几个主要玩家在 AI 领域占据主导地位，他们会继续维持这个地位。这些公司似乎与那些受到反垄断法规关注的大公司有所交叉。你对这两种趋势有何看法？像是监管部门会否对这些公司进行分拆？这将会对现状产生怎样的影响？

Eric：在我的职业生涯中，我曾帮助微软避免被拆分，结果它确实没有被拆分。我也曾努力阻止谷歌被分拆，目前来看，谷歌还未被分拆。因此，从我的观察来看，这些公司被分拆的趋势并不明显。只要这些公司避免成为像约翰·D·洛克菲勒那样的企业巨头，政府可能不会采取行动。这些公司占据主导地位的原因是他们有足够的资本来建设数据中心。

关于技术差距对全球的影响

学生: 这一切将会把那些不参与前沿模型开发和无法获取计算资源的国家带向何方？富国越来越富，穷国尽力而为。

Eric: 事实上，这就是富国的游戏，对吧？巨大的资本，众多技术过硬的人才，强有力的政府支持。有些国家，比如印度，将会成为这一领域的重要参与者。其他国家可能需要找到合作伙伴或与其他国家联合。

对年轻人的建议

主持人: 最后一个问题，Eric。你花了很多时间帮助年轻人，他们正在创造大量财富，你对在座的各位有什么建议吗？他们正在为这门课写商业计划书或撰写政策提案、研究提案，未来职业发展方面你有什么建议？

Eric: 我在商学院教过一门关于这个主题的课程。快速制作原型的能力真的非常重要，成为创业者的一个难点在于一切都发生得非常快。现在，如果你不能在一天之内利用这些工具构建你的原型，那你就需要考虑一下了，因为你的竞争对手正在这么做。所以我的建议是，尽快使用这些工具把你的想法制作成原型。这是关键，因为在其他公司、其他大学，或你从未去过的地方，一定有人正在做和你一样的事情。

埃里克(Erc　Schmit)在斯坦福大学的演讲中谈到“AI 的投资与国家安全”主题与 COPU 注解

COPU 2024.8.29

埃里克认为：企业在研发、应用 AI 时将会遇到大能耗和大投资的挑战，小企业可能力所不支，大企业也将面临很大困难。如果大企业还要建立供预训练之用的数据中心（或算力中心），能耗、投资之巨大也可能不胜负担，以至于还有求于大国的国家资助，而世界上能给国家资助的大国也寥寥无几。他曾提出，现时美国也面临能源短缺、负债过重、投资不足的困境，要美国进行国家资助实属勉为其难！

下面将提到埃里克对这个主题演讲的摘要

AI 的投资与国家安全

他说：我现在和大公司交谈，他们告诉我需要投资 100、200、500 亿美之甚至 1000 亿美元，Stargate 的投资就达到 1000 亿美元。奥特曼对我说，OpenAI 需要投入 3000 多亿美元。上周五我去了白宫告诉他们，我们需要与加拿大合作，加方有很多研发 AI 的人才，还有大量水力发电资源。另所还可选择中东阿拉伯人来资助。可惜他们不会遵守我们国家的安全规则。而加拿大与美国是共同遵守我们国家的安全规则的。

如果 OpenAI 3000 多亿美元都要流向英伟达，我们还要讨论 CUDA 和英伟达的问题。如果我有无限资金，我今天会选择英伟达的 B200 架构，因为它运行更快。现在英特尔似乎不存在垄断了，但英伟达还是有垄断。我和 AMD 的 LisaS 有长时间交谈，他们正在研发 RoCKM 架

构可取代 CUDA，反对英伟达垄断。

COPU 注解

在 AI 发展中，不但能耗巨大，还需要巨大资金投入。作为 AI 的个体大模型能耗就很大，在作预训练时投资也大。如 Chat GPT，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，一天耗电达 5 万 KW，目前全球炙手可热的 AI 程序——CGPT，每天耗电高达 50 万 kwh(相当于一年一个小城市的用电量)。不少开发 AI 的企业已经付不起数据中心为其训练服务应付昂贵的租赁服务费用。

建设数据中心或集成算力训练服务中心，投资、耗能巨大。OpenAI 拟在推出 GPT-5 之后，建设 Sora 模型（文本生成视频）的集成算力培训服务中心，初步设计 7000 张 H100 芯片，综合算力 10^8TFLOPS，需用电力约 400MW，投资约 3000 多亿美元。OpenAI 拟在推出 GPT-6 之后，建设“星际之门”集成算力训练服务中心，初步设计 10 万张 H100 芯片，综合算力 10^9TFLOPS，需用电力约 5000MW，相当于大城市成都 1/4 用电量，投资约 4 万亿美元。上述两个集成算力训练服务中心还处于设想阶段，何时建成没有定。现在马斯克建成的 10 万卡机房，这是捷足先登，预示着 AI 进入了一个新时代！建设集成算力训练服务中心是一般企业（包括大企业在内）难以负担的，为此还需要求助于政府，争取国家资助，但世界上能做到这点的国家很少。

今年 7 月 12 日，马斯克与黄仁勋相遇相谈，一起感叹：从汽车到人工智能，美国面临能源短缺、债台高筑、投资紧缺的挑战，他们

担心美国将惨败于能源、投资短缺的挑战。埃里克与奥特曼也担心美国面临能源、资金短缺的挑战，埃里克为此奔赴白宫献计献策，建议美国政府找加拿大与中东合作，加拿大有丰富的水力发电资源，中东有资金和资源优势，埃氏认为中东不会执行美国政府的安全规则，而加拿大会遵守与美国政府共同制定的安全规则。

中国呢？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领先，核聚变、太阳能前沿科技也在大力探索，在未来能源革命中占据有利位置，从汽车到人工智能，美国将被中国史诗般超越。

埃里克在谈到“AI 与地缘政治”这个主题的演讲中，他认为世界上只有极少数的国家（不是指企业）可能参与未来“AI 前沿”的竞争，主持人问他“是哪些国家”，他说“美国是一个，中国也是，其他我不知道！”“这些国家必须拥有大量资金、充足能源、丰富人才、强大教育系统，以及具有获胜的意愿。”

顺便说一句：奥特曼正在追求研制通用人工智能，需要通过预训练测试环节，他比较现实的做法是有求于马斯克的 10 万算力卡集成训练测试中心的支持，可是马斯克十分不满奥特曼的闭源政策，两人为此争吵，不欢而散，这次能否获得马斯克支持，真难说！

所谓英伟达的 CUDA 平台，即应用生态系统(用以优化应用程序的平台）；也是加速计算平台(主要是并行计算) 人工智能在进行预训练时会用到它，此时英伟达或 CUDA 是否垄断，这是要与其商讨的。

埃里克答学生提问三“AI 开源与闭源的争论”

Eric Schmidt

2024.8.29

埃里克在斯坦福大学演讲时，回答学生提问: “提问三: AI 开源与闭源的争论”，塔埃里克同答:

1）他肯定“在我们行业中关于开源与闭源的争论非常激烈”。

2）谈到他本人是坚持开源的，他说“我的整个职业生涯都是基于人们愿意共享开源的软件，我的一切都与开源有关。”谈到他曾任职的企业，他说: “谷歌的许多基础设施都是开源的”，“我们在技术上所做的一切也是如此。”

3）他在演讲中没有谈论“如何比较开源 AI 与闭源 AI”的问题，他只是谈到法国一家闭源出色的小公司 Mistral(不久前 COPU 也对它进行了介绍)，他说由于成本过高担心引起企业亏损，Mistral 无法推出第三个版本。为了降低成本他建议提高程序员的工作效率。其实对闭源如此，对开源的商业发行版也可采取一样的做法。

评奥特曼的闭源策略

Francois chollet

2024.8.29

Keras 原创者、谷歌 Al 研究员 Francois chollet 发言评论 openAI 公司 CEO Sam · Altman 仅凭一己之力，改变游戏规则，推行闭源策略，导致语言大模型的前沿研究全面封闭，是非常可悲的！以前是所有最新研究成果都是共享的，现在前沿研究不再被公开发表，变得全面封闭了。奥特曼的如此做法，使通用人工智能(AGI)的研究进展延后倒退了几年，可能是倒退五年至十年。

OpenAI 继续让语言大模型的炒作热潮吸走了人们对他们改变游戏规则的关注。

奥特曼现在的做法，更像是走在通往通用人工智能的一条岔道上。

Meta 致力于发展“开源 AI”

（Meta 公司 CEO 扎克伯格（Mark Zuckerberg）

在 Llama3.1（405B 版本）发布会上的讲话）

摘要如下:

他提到: Llama3.1 将上下文长度扩展到 128k， 405B 版本已扩大到 8 种语言支持。在 150 多个基准测试中，405B 版本表现追平甚至超越现有 80TA 模型 GPT-4o 和 Claude3.5 Sunnet。这个 Llmama3.1（405B 版本）开源大模型不耽在开放性、可修改性和多语言翻译等能力方面也都追评、超越现有的顶尖模型。

在超过 15 万亿 token 上训练的 Llama3.1-405B 模型，大幅优化整个训练栈，并把模型算力规模扩展到超过 16000 个 H100GPU，使该模型可进行实时和批量推理、监督微调、评估模型、持续训练等高级程序作业。开源的 Llama3.1 登上了大模型王座，超越 GPT-4o，Llama3.1(405B 版本)的下载是至今已超过 3 亿次。

扎克伯格还谈到: 开源 AI 是前进的道路，开源 AI 是最强大的模型。他还嘲讽 OpenAI 早就没有向中国开放其 API，奥特曼完全没有必要发出宣布再次强调这点。由于 OpenAI 一声吆喝，惊起中外同行的一滩鸥鹭，一批中国优秀的大模型公司马上自主实行对 GPT-4o API 的完全对标、平替、甚至超越。

他还指出: Mate 致力于开源人工智能，开源是最好的开发堆栈。我们需要控制自己的命运，不要被限制在闭源供应商那里。我们需要保护我们的数据安全，不能通过闭源模型的云 API 发送。

开源创新——开源赋能引领数智新时代

陆首群

2024 年 8 月 27 日

人们常常提问，在人工智能发展中是采用开源还是闭源，哪个更为有利？我曾回答：“COPU 坚持发展基于开源的人工智能”，我也曾引用 Meta 的马克·扎克伯格的回答：“开源是人工智能前进的道路”，“开源是人工智能最强大的模型”。我们曾指出：“有人非要用闭源来捆绑 AI，势将束缚 AI 的发展，而开源将使 AI 具有更大努力来提升其创造力和协调能力”。我曾写过评论奥特曼实行闭源策略的文章，奥特曼曾说：“研发 AI，采用开源不是最佳选择”。我认为奥特曼的“闭源 AI”模型，很难通过“安全关”（在为 AI 模型设置安全红线时）；面临未来巨大能耗、巨大投资的挑战，也是“未定之天”。

一、开源锻造现代创新引擎：“互联网+基于知识社会的创新 2.0

为了提升传统业态的创新效果，不求 0→0“八宝粥”式的变化模式（实质是不变化或少变化），而要追求 0→1 爆发性的变化模式，这时就有求于开源的现代创新引擎实现跨时代创新，即在数智新时代，开源创新引领传统业态的数字化转型或智能化重构。

其设计布局如下：在几乎占有无限空间底层的工业社会中，划出一个用于考察试验用现实的物理空间（physical space），在其中配置有待考察、试验的传统业态（有生产的、技术的、经济的、社会的，也可扩充为其他的等等），同时我们构建一个更高维度的、虚拟的数

字空间（Cyber Space），在其中配置供试验用的高维阶网络、资源、动能，等等，以数空中的动能作用于物空中的业态，促使其产生 0→1 爆发性变化，体现开源创新：使新业态实现深刻的数字化转型或智能化重构。

二、开源打通人工智能发展瓶颈

我们经常举的一个实例是：2015 年美国人工智能四大重镇——谷歌、微软、脸谱（即现在的 Meta）、IBM，为克服人工智能发展瓶颈，在当年将他们研发的人工智能框架、平台、引擎、工具、算法、源代码、项目等全部开源。

以谷歌为例，实行开源的有 900 多个项目 2000 万行代码，包括：TensorFlow 框架；Android 移动平台完整堆栈：操作系统、中间件和一些重要应用；Augular：JavaScript 和 Web 应用程序框架；BoZel：可再生代码的工具；Brotli：压缩算法；Chromium：Chrome 浏览器背后的引擎；Go：一种编译型并发型、垃圾回收功能的编辑语言。

三、开源可加快技术发展进程

2016 年 7 月 20 日《福布斯》杂志记者向谷歌高级副总裁、人工智能首席科学家 Jeff Dean 提问；

记者问：谷歌为什么要开源？为什么要开源自己最先进的技术？

Jeff Dean 答：常规科学发展缓慢，阻碍公司创新；开源能加快技术发展进程，有利于外界实时交流协作，包括建立、吸引志愿开发者和维护者团队的支持。

Jeff Dean 还答：过去企业取得成功的做法是把自己的“使用价

值链”做到最优，以提高企业在客户中的信誉、地位，提高企业的竞争优势，现在企业最成功的做法是提高“产业价值的生态系统”，并使企业的人才融入更大的科技圈中，以扩大企业的影响力。

近期也有人问：谷歌为什么将很多大数据项目开源了，贡献给公众？

谷歌的回答：这些软件规模非常大，靠一家公司独立去维护、运营，成本非常高，开源后变成公共财产，很多公司和志愿者共同参与，大公司不需要投入太大力量，小公司也可解决成本问题，形成我为人人，人人为我的开源文化。

四、开源是推动深度信息技术（包括大数据、云原生、区块链、人工智能等）发展的基础和动力。

我们在早先《发展基于开源的人工智能》的一篇文章中是如此谈论开源的：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。

据 Linux 基金会 2016 年的调查统计，全球 500 强企业中 IT 跨国公司头部企业已改变企业内的设计方式，变为企业内外的设计方式，以利用企业外的开源资源为主设计（85%）、企业内部为辅助设计（15%）的方式。

我们还指出，在某些信息技术（包括人工智能）的应用场景中绝对离不开开源，如解决 AI 安全问题时离不开开源，全球全程及时发现供应链漏洞时（如阿里达摩院的开源志愿者首先发现在万里之外美国供应链上出现的 Apache Log4j 严重的漏洞）也离不开开源。

有人担心开源会泄漏原创技术，造成企业亏损或产业萎缩，这是他们对开源缺乏理解而产生的误解，需要明白：开源免费的社区发行版与开源收费的商业发行版之间的融合与区别。

有人认为开源与闭源是不离不弃的左右手，其实他们并不充分了解开源软件或闭源软件目前谁是主流？谁代表发展趋势？

早在 2015-2016 年， 国际著名的调查分析公司 Gartner 就宣布开源是软件发展的主流，也代表发展趋势。他们的调查数据如下：2015 年， 85%的全球商业软件使用开源软件，2016 年，95%的主流企业（或组织）直接（或间接）在其关键任务系统“（Mission Critical System）”中使用开源软件。

如此看来，开源与闭源在主流或趋势发展上已是不平衡、不对称的，用左右手平衡的观点来形容它们，似不恰当！

也有人认为开源的创新为其次，开源的普及（或推广）应为主。我们认为，不要把开源的主次颠倒了！

五、开源有力支持互联网发展中数字主权建设

2021 年 12 月 8 日互联网治理论坛（IGF）邀请中国开源软件推进联盟（COPU）参加在波兰召开的未来《互联网治理座谈会》，COPU 在会上作了“开源协同，有力支持互联网数字主权建设”的报告，指出：中国开源运动发展的一个重要体验是，用开源协同的数据证明，它有力支持了互联网数字主权建设（由梁志辉、鞠东颖作为 COPU 代表的发言稿刊载于最近出版的《开源创新，数字化转型与智能化重构》一书 P11）。

AI 专家的异议与 COPU 观点

陆首群

2024.08.19

有人非要用闭源来捆绑 AI，势将束缚 AI 的发展，而开源将使 AI 以更大潜力来提升其创造力和协同能力，至于对 AI 发展尤为关键的安全更离不开开源。

自 2015 年以来，COPU 一直致力于发展基于开源的深度信息技术(包括大数据、云原生、区块链、人工智能等)。

OpenAI 的 CEO 萨姆·奥特曼(Sam Altman)违背他研究“开源 AI”的初心，转而推行闭源策略， Meta 的 CEO 马克·扎克伯格(Mark Zuckorberg)坚持“开源 AI“路线，认为“开源 AI”是最强大的模型。COPU 观点：“开源 AI”彰显“开源创新，数字化转型与智能化重构，”作为 AI 发展前沿的“AI 安全”是离不开开源的。

目前在全球火红的是语言大模型(LLM)生成式 AI，在肯定其取得巨大进步同时，也要看到其存在的局限性，其中突出的是产生“幻觉”。

人工智能专家、OpenAI 前首席科学家 Ilya Sutskever 在肯定语言大模型 LLM 取得进步时，坦言它仍存在局限性“幻觉”问题，原因是缺乏对推理的理解，在应用实际问题时推理能力丧失或不足。图灵奖获得者，Meta 首席科学家、人工智能大师楊立昆(Yann LeCun)认为，仅靠语言模型无法实现真正的智能，因为语言模型缺乏对世界底层结构（物理空间）的理解。

Ilya Sutskever 预测，奥特曼在研发 GPT-5 之后，通用人工智

能（AGI）即将降临，所以他认为 LLM 是通往 AGI 的途径。Yann LeCun 认为，自回归语言模型并不是通往 AGI 的充分途径，因为它缺乏智能生物的基本能力，缺乏对物理世界的具体理解，削弱了推理能力。

早先陆首群在面完数字化、智能化时对现实世界“物理空间”中的各种业态创新有所阐述，2016 年陆首群提出现代创新引擎“互联网+基于知识社会的创新 2.0”：建立相互连系、兼容的高维度的虚拟社会数字空间(Cyber Space)与低维度的现实社会物理空间(physical Space)，并将数字空间中的动能作用于物理空间中的业态，促使其产生 0→1 爆发性变化，即“开源创新数字化转型与智能化重构”。

有人认为，现时人们集中于发展生成式的“语言 AI 大模型”，更应关注发展依托于实体经济的“工业互联 AI 大模型”。COPU 认为，我们在 2010 年就曾提出依拖实体经济的 AI 研究方向。

现在看来,人们把通用人工智能(AGI)看成是 AI 的发展前沿，已经取得了共识。COPU 同意这种共识，但认为这里也存在讨论的空间:如更现实的发展做法是:是否可首先发展服务于各行各业的专用人工智能，并由“专”达“通”，由“简”及“全”。

众多中外 A1 专家在讨论通用人工智能关键技术发展路径时，形成一个共识:语言大模型→多模态大模型→世界模型→具身智能→通用人工智能，COPU 观点有待细化。

在 21 位人工智能大师和专家联名签署《北京 AI 国际安全共识》时，加州大学伯克利分校 Stuart Russell 教授认为，“应在具有自

主系统的通用人工智能的发展超越人类之前，人类应制造限制其摆脱人类控制的红线。”COPU 的观点是，人们要进一步研究并讨论：开源在制定这条红线时的作用如何？研究通用人工智能是否应做到安全第一、安全为先?全球同步?技治并举?

COPU 还认为奥特曼于 2023 年 3 月实行闭源策略，当时 COPU 就敏感地感觉到“四大”(即大参数、大算力、大能耗、大投资)可能会对人工智能未来的发展构成巨大的挑战，而推行“开源 AI”还提“闭源 AI”，谁将更易过关?！我们经过思考和计算后认为，鉴于开源具有开放、共享、协同的特性，将有更大韧性通关，而对奥氏的闭源策略来说，过关难度要更大，COPU 认为它能否通关尚是未定之天！

开放生成式人工智能

（LF） Jim Zemlin

2024.08.19

目前，开源对于现代技术和产品的发展起着至关重要的作用。Linux 基金会作为众多开源项目的汇聚地，不仅关注这些项目的成长和发展，还致力于通过这些项目推动技术创新。我们期望未来能够开发出更多先进的云和开源项目。

Linux 基金会并不止于开源软件，我们也高度重视开放数据共享的重要性，相信这将为人工智能前沿模型的构建提供更多支持。未来，我们将进一步加大在数据共享方面的努力，以推进人工智能技术的创新与发展。

此外，在硬件领域，我们也积极展开合作。例如，与 RISC-V 基金会的合作使我们在芯片研发方面拥有了众多的合作伙伴。这不仅有助于推动半导体技术的发展，也为我们构建更大规模、更可持续的生态系统提供了坚实的基础。

当前，全球半导体产业正在快速发展，Linux 基金会在其中扮演了关键角色，致力于构建一个开放的生态系统。我们从技术开始，以开源项目作为开源软件研发的核心，这些项目在解决行业及社会问题方面发挥着至关重要的作用，涵盖了操作系统、云架构以及人工智能等多个领域。

开源软件已成为整个行业的基石，开源软件无处不在，不仅为我们提供了日常所需的技术服务，还助力企业高效开发产品。

很多企业使用了开源软件实现更多营收，所以他们在开源项目上进一步投资，不断为开源软件做出贡献。他们以使用开源软件的经验作为基础，在整个商业运行过程中进一步提升和完善，在上游的代码上以及技术产品上都得到了不错的反馈，产生的营收进一步回归到开源项目，这样可以以循环的方式研发产品和服务，形成良性的循环圈，这也是我们希望在开源项目上打造的模式。

目前，Linux 基金会也创建了很多工具和数据集进一步助力大规模的合作，年营收增长达 30%左右，每年都会有新技术加入到 Linux 基金会中。在通信、云计算、汽车、人工智能等领域，Linux 基金会的工作触及多个领域，做出了非常多的努力。

Linux 基金会在推动开源和人工智能领域发展中扮演着重要角色，我们致力于探索并分享这些领域的机会。为更深入理解，我们需要了解人工智能从底层到上层的具体架构。位于最上层的是 AI 的应用，例如我们熟知的聊天机器人等。这些应用能极大地促进企业发展，也是众多企业热衷于 AI 的原因。AI 的应用可通过大语言模型（LLM）如 ChatGPT 等展示出来，并且可以针对特定需求进行微调。

数据是构建大语言模型的基础，大多数高效的大语言模型都是通过数据训练而成的。我们使用公共和私有的数据混合训练前沿模型，以提升模型训练的效果。

同时，构建和微调专有模型的时候，需要更好的管理系统。OPEA 工具提高了微调模型的构建速度，大大缩短了数据分析时间，使得开发工作更加高效。此外，OPEA 可以让平台变的更加标准化，使得更

多的开源大语言模型得以开发，进一步打造 AI 应用。

开源已被证明是打造大语言模型进行机器学习最有效的方式之一。开发者们可以使用 PyTorch 等工具在云端基础架构上进行机器学习和操作，从云计算层面上，我们面临着打造开放 API 的机会。

在构建大语言模型时，有效利用开源参考数据至关重要。我们面临的最大挑战之一是数据质量问题，如果训练数据不够好，模型开发会变得异常困难。因此，引入干净、经过分析的高质量数据是打造最佳模型的关键。

为了简化这一过程，我们提供了一系列模块以构建、训练、优化和验证这些模型，且每个阶段都配备了相应的工具。由于这些工具都是开源的，使用起来非常便捷。OPEA 具有很好的基础，是在企业中可以实施的人工智能系统管理框架。

制定相关 API 标准也是我们工作的一部分，这为企业提供了更便捷的使用体验。在推进人工智能开放的过程中，我们认识到对 AI 开放的共识和定义至关重要。与传统软件的开放不同，AI 的开放不单是关于软件的商标和知识产权等问题，而是涉及更广泛的内容。

Linux 基金会创建了模型开放框架（Model Openness Framework，简写为 MOF），为研究人员和开发者提供了指导，帮助他们在宽松许可的前提下，增强模型的透明度和可复制性。MOF 的优势在于其对监管者更透明，并为开发人员打造了一个更有活力的生态系统，进而促进了安全和创新。

基于开源的超级记账（区块链）

（北京大学） 陈钟

2024.9.2

超级账本开源基金会（HYPERLEDGER FOUNDATION，以下简称“超级账本”）成立于2015年，旨在通过围绕开源区块链软件技术培育繁荣的生态系统，为企业市场带来透明度和效率。作为Linux基金会的一个项目，Hyperledger Foundation构建了一个由会员和非会员组织、个人贡献者和软件开发人员组成的社区，使用区块链、分布式账本和相关技术为多方系统构建企业级平台、库、工具和解决方案。超级账本的成员包括了金融、银行、医疗保健、供应链、制造、技术等行业的领先组织。加入超级账本的成员不仅展示技术领导力，还能与他人合作和建立网络，并提高对企业区块链社区工作的认识。所有Hyperledger代码都是公开构建的，并在Apache许可下可用。如今9年过去了，发展与应用如何？本文做一个简要的回顾与介绍。

一、 开源的社区治理树立了开源区块链的典范

2015 年 12 月，开源世界的旗舰组织 —— Linux 基金会 牵头，联合 30 家初始企业成员（包括 IBM、Accenture、Intel、J.P.Morgan、R3、DAH、DTCC、FUJITSU、HITACHI、SWIFT、Cisco 等），共同 宣布 超级账本（Hyperledger）联合项目成立。超级账本项目致力为透明、公开、去中心化的企业级分布式账本技术提供开源参考实现，并推动区块链和分布式账本相关协议、规范和标准的发展。项目官方网站为 hyperledger.org。

超级账本项目的出现，形成了全球开源区块链应用的“三足鼎立”，而比特币和以太坊二者均与加密货币或者以加密货币为激励机制的区块链应用紧密相关。然而，超级账本项目实际上试图证实区块链技术已经不局限在单一应用场景中，也不限于完全开放匿名的公有链模式下，而是有更多的可能性，也说明区块链技术已经被主流企业市场正式认可和实践。同时，超级账本项目中提出和实现了许多创新的设计和理念，包括权限和审查管理、多通道、细粒度隐私保护、背书-共识-提交模型，以及可拔插、可扩展的实现框架，对于区块链相关技术和产业的发展都将产生十分深远的影响。

开源社区的组织结构和治理是保证开源成功的关键因素之一。超级账本基金会成功地借鉴了 Linux 基金会最佳实践，形成了有效的治理架构。Apache 基金会创始人 Brian Behlendorf 亲自担任过执行总监，吸引了大约 500 个成员加入，大约其中五分之一来自中国大陆，百度曾经是代表中国的唯一的高级会员参与治理。2016 年 12 月还正式成立了中国技术工作组，负责推动大中华地区超级账本社区组织建设和开源技术的发展与应用。

二、 开源的方式形成坚实的技术基座

超级账本基金会成立之初，项目就收到了众多开源技术贡献。IBM 贡献了 4 万多行已有的 Open Blockchain 代码，Digital Asset 贡献了企业和开发者相关资源，R3 贡献了新的金融交易架构，Intel 贡献了分布式账本相关的代码。

作为一个联合项目（Collaborative Project），超级账本由面

向不同目的和场景的子项目构成。目前包括 Fabric、SawToothLake、Iroha、Blockchain Explorer、Cello、Indy、Composer、Burrow、Quilt、Caliper、Ursa、Grid、Transact、Aries、Besu、Avalon 等顶级项目，所有项目都遵守 Apache v2 许可，并约定共同遵守如下的 基本原则:

- 重视模块化设计: 包括交易、合同、一致性、身份、存储等技术场景;
- 重视代码可读性: 保障新功能和模块都可以很容易添加和扩展;
- 可持续的演化路线: 随着需求的深入和更多的应用场景，不断增加和演化新的项目。

超级账本项目的企业会员和技术项目发展都十分迅速，如下图所示。

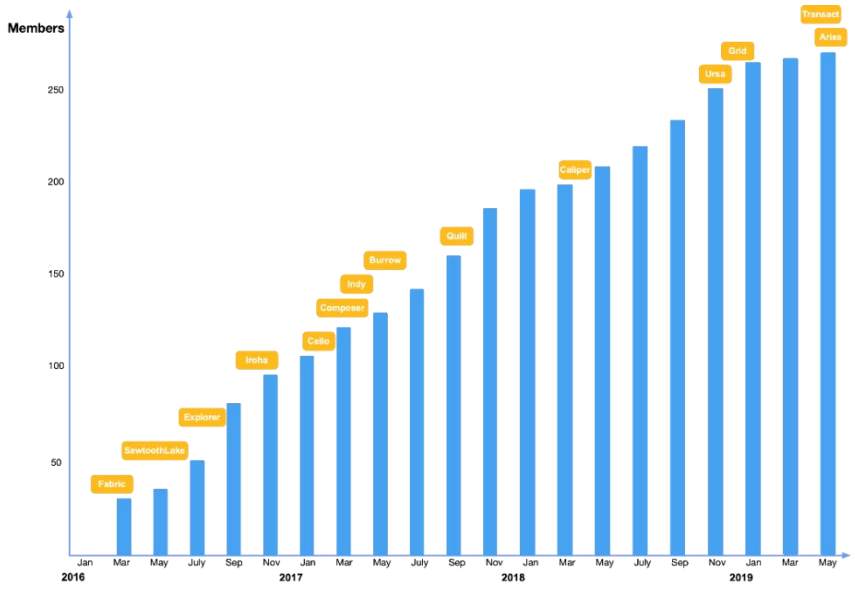


图 1: 超级账本顶级项目及成员发展状况

超级账本的所有项目代码托管在 Github 上。主要包括了如下顶级项目（按时间顺序）。

- Fabric：包括 Fabric、Fabric CA、Fabric SDK（包括 Node.Js、Java、Python 和 Go 语言）等，目标是区块链的基础核心平台，支持 PBFT 等新的共识机制，支持权限管理，最早由 IBM 和 DAH 于 2015 年底发起；
- Sawtooth：包括 arcade、core、dev-tools、validator、mktplace 等。是 Intel 主要发起和贡献的区块链平台，支持全新的基于硬件芯片的共识机制 Proof of Elapsed Time（PoET），2016 年 4 月贡献到社区；
- Blockchain Explorer：提供 Web 操作界面，通过界面快速查看查询绑定区块链的状态（区块个数、交易历史）信息等，由 DTCC、IBM、Intel 等开发支持，2016 年 8 月贡献到社区；
- Iroha：账本平台项目，基于 C++ 实现，带有不少面向 Web 和 Mobile 的特性，主要由 Soramitsu 于 2016 年 10 月发起和贡献；
- Cello：提供区块链平台的部署和运行时管理功能。使用 Cello，管理员可以轻松部署和管理多条区块链；应用开发者可以无需关心如何搭建和维护区块链，由 IBM 团队于 2017 年 1 月贡献到社区；
- Indy：提供基于分布式账本技术的数字身份管理机制，

由 Sovrin 基金会发起，2017 年 3 月底正式贡献到社区；

- Composer：提供面向链码开发的高级语言支持，自动生成链码代码等，由 IBM 团队发起并维护，2017 年 3 月底贡献到社区。目前已经成熟，处于 Deprecate 阶段，仅考虑修正可能的严重缺陷；
- Burrow：提供以太坊虚拟机的支持，实现支持高效交易的带权限的区块链平台，由 Monax 公司发起支持，2017 年 4 月贡献到社区；
- Quilt：对 W3C 支持的跨账本协议 Interledger 的 Java 实现。2017 年 10 月正式贡献到社区；
- Caliper：提供对区块链平台性能的测试工具，由华为公司发起支持，2018 年 3 月正式贡献到社区。
- Ursa：提供一套密码学相关组件，初始贡献者包括来自 Fujitsu、Sovrin、Intel、DFINITY、State Street、IBM、Bitwise IO 等企业的开发者，2018 年 11 月正式被接收到社区；
- Grid：提供帮助快速构建供应链应用的框架，由 Cargill、Intel 和 Bitwise IO 公司发起支持，2018 年 12 月正式贡献到社区；
- Transact：提供运行交易的引擎和框架，由 Bitwise IO、Cargill、Intel、IBM、HACERA 等公司发起支持，2019 年 5 月正式贡献到社区；
- Aries：为客户端提供共享的密码学钱包，由 Sovrin、

C3I 和 Evernym 等公司发起支持，2019 年 5 月正式贡献到社区；

- Besu：作为企业级的以太坊客户端支持，由 Consensys、Hacera、JPM 和 Redhat 等公司发起支持，2019 年 8 月正式贡献到社区；
- Avalon：提供链下计算支持，增强安全性和可扩展性，由 Intel、IEX、IBM 和 Consensys 等公司发起支持，2019 年 9 月正式贡献到社区。

这些顶级项目分别从平台、工具和类库三个层次相互协作，构成了完善的生态系统，如下图所示。

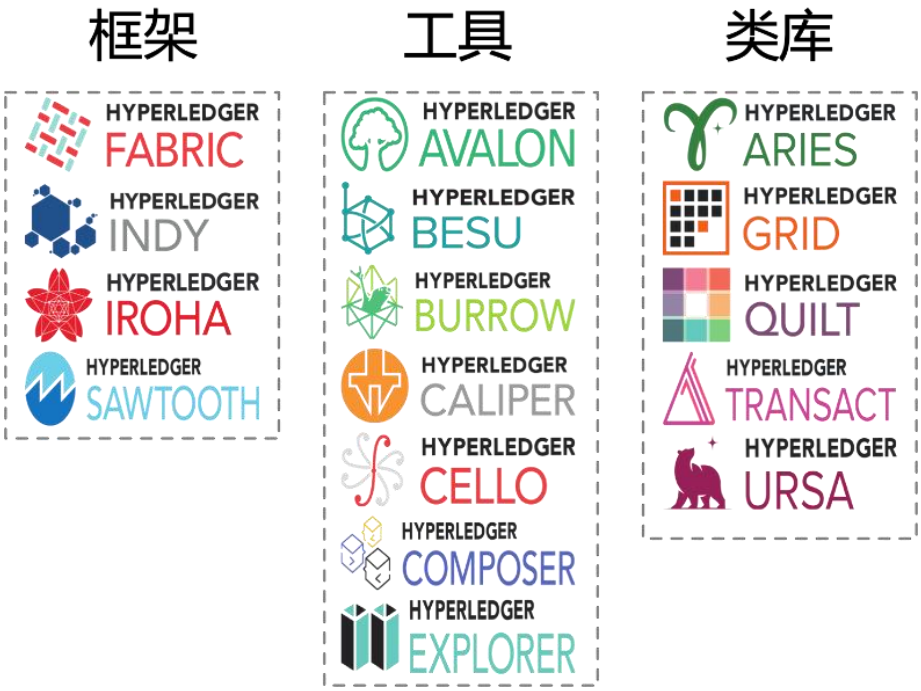


图 2：超级账本生态：框架、工具、类库

北京大学网络与信息安全实验室代表北京大学作为超级账本基金会非会员的大学成员，积极参与和推进了超级账本社区的推广和发展，并且通过其开源的 gmssl 项目有力支撑了超级账本 URSA 项目，

并对超级账本中国应用国密算法的集成做出了重要贡献。超级账本作为以联盟链为主的区块链技术标杆，也为中国区块链技术发展和应用起到了支撑和参考的作用。

三、 超级账本技术及应用进展状况

超级账本社区来自于各类成员组成，也由成员治理，超级账本基金会的现任管理委员会主席是大卫···特里特(David Treat)，他是埃森哲全球元宇宙联盟业务集团的高级董事总经理。

作为超级账本的一项例行工作成果，是编制和发布年度区块链品牌研究报告。最新的一期是 2023 年品牌研究报告。该报告深入研究了企业区块链增长和采用的全球格局，包括超级账本基金会在这一不断发展的创新中的作用。总体结论是：2023 年使我们更接近实现区块链的变革潜力，这是我们多年来坚定追求的愿景。从超级账本调研的情况可以认为区块链发展“喜忧参半”。

对比 2021 年进行的上一次品牌研究，市场格局发生了重大变化。其中将用例转移到生产环境一直是超级账本的使命。报告中列举了部分案例：首先是 ISSA 的“现实世界中的 DLT”调查，该调查显示现场 DLT 使用率从 2021 年的 8%飙升至 2022 年的 32%，增长了四倍。其次，花旗财政和贸易解决方案(TTS)启动了花旗代币服务的创建和试点，用于现金管理和贸易融资。此外，摩根大通执行了涉及贝莱德和巴克莱的首个基于区块链的实时抵押品结算交易。再有，中央银行数字货币(CBDC)的发展势头持续强劲，130 个国家(占全球 GDP 的 98%)正在探索 CBDC 的未来。一份专项报告——《超级账本在央行数字货

币中的作用》介绍了这部分的内容，可以了解更多有关的技术及其所发挥的作用。虽然生产型 CBDC 可能还需要更长的时间，但有强烈的前进势头迹象，如法国央行行长弗朗索瓦·维勒鲁瓦·德·加尔豪(Francois Villeroy de Galhau)表示，“……实验将在明年展开，包括实际交易的试验”，指的是批发数字欧元。

超级账本基金会内部的社区在推动市场、促进增长和确保区块链保持在技术创新的前沿方面发挥了关键作用。虽然 “进入高端市场成熟度” 的旅程仍在进行中，但我们可以在几个关键领域看到切实的牵引力，包括 IPWe 开创性工作的通证化，以及企业级以太坊超级账本 Besu。最后还有数字身份，这在白宫的国家网络安全战略中被高度列为战略目标。报告中写道: “区块链的故事还远未完成，但这一页正在翻开，我们很荣幸能够成为这一叙事的推动力量”。区块链继续为企业、政府和整个社会带来的变革性影响。报告强调，在我们迈向更加数字化连接、透明和高效的世界的过程中，需要持续创新、合作和适应。“我们很自豪能够站在这场技术复兴的最前沿，并将继续致力于驾驶区块链的增长、采用和实现是全球数字格局的基本要素”。

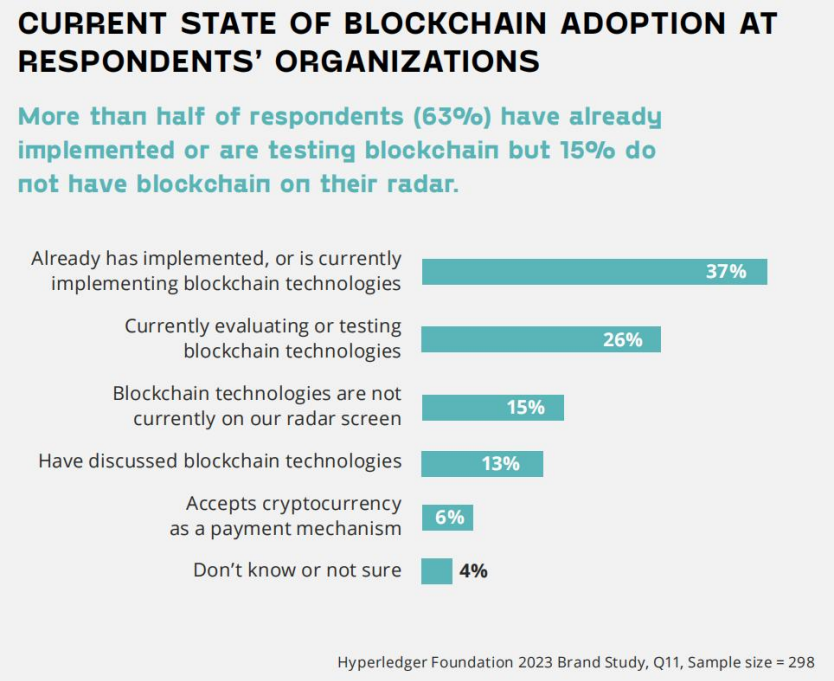
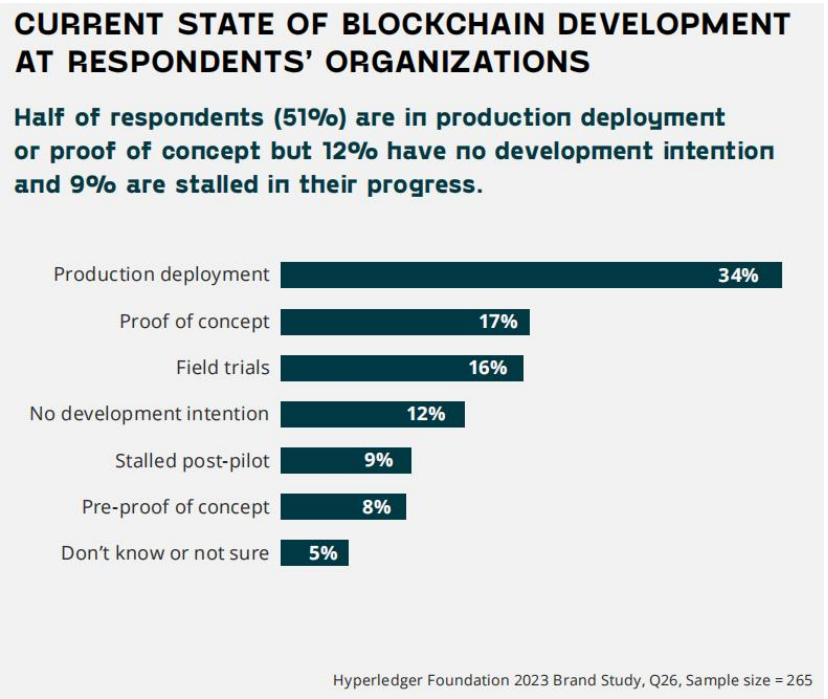


图 3：区块链应用现状调查：298 个组织的回应

图 4：区块链开发现状调查：265 个组织的回应



超级账本基金会身处国际化的环境，涉及到了诸多数字资产价值表达的主题并积极参与和贡献。例如：在调查 NFT 效应时都讨论了 NFT 现象对区块链采用的影响。O ′Dair 认为，“五年前，关于 NFT

的很多对话都是非常概念化和抽象的: 它是关于可能发生什么。尽管NFT 现象有起有落，但它引入了一种完全不同规模的活动，因为它与个人用户有关。”他认为，NFT 还可以帮助个人理解区块链及其不同的层面。与 O ′Dair 的观点一致，Day 从尝试进入 NFT 领域的公司数量的角度来看待这种活动规模。他估计: “我认为去年有 250 多个品牌涉足与代币化相关的东西，或者使用区块链来创造体验、艺术或社区。”“这是很多。有了这些品牌，他们就有了一批追随者。将现有用户群吸引到区块链平台的意义，对于技术的广泛采用来说非常强大。” 无论这种可见性是否总是积极的，NFT 的无处不在创造了一个新机会，它让公众更多地参与和理解区块链，为企业层面的采用和实施打开了大门。

发展中的障碍依然不可忽视。尽管超级账本在这一领域取得了积极进展，但调查数据也显示出一些进展停滞。例如，在图 3 中，15%的受访者表示区块链目前不在其组织的雷达上。同样，9%的受访者组织的区块链开发在试点后停滞不前，12%的受访者没有任何开发意图(见图 4)——这两个数字都比 2021 年有所上升(分别为 4%和 7%)。尽管他看到了企业用户持续的甚至“看好”的兴趣，玛德哈维(Madhavji)指出，我们还没有被主流采用。正如他所说，“当你看到这些公司所做的一些更大的事情时，你可能会认为它现在已经成为主流。你还没有看到大公司在有意义的地方进行整合。”他说: “它在前进。我不知道它是否总是发展得足够快，但它确实在发展。”一些受访者认为

这种停滞是在炒作周期的背景下进行的。例如，Hess Legal Counsel的创始人兼管理顾问埃里克·赫斯(Eric Hess)断言，“区块链需要证明投资回报，某种价值创造，这与前几年不同。”戴伊(Day)也有同样的感受，并为今年的不同之处添加了更多背景：“投资支出的压力越来越大，或者将其转化为业务的举证责任更高。我们在2017年、2018年、2019年看到，由于区块链比人工智能更受欢迎，这些组织可以建立一个区块链的东西，或者一个财团，或者一个商业网络，来测试这个模型。然后他们发现去中心化很难，或者并不总是符合他们的商业模式。”

四、 结束语

通过超级账本基金会的实践以及调查报告可以看出，区块链技术作为一种新质生产力具有巨大的发展潜力，特别是在数字经济蓬勃发展的阶段。包括人工智能在内的新技术发展都在呼唤之相适应的新型生产关系的建立和新型激励与分配机制的形成。区块链作为一种能够调节生产关系的技术体系在这方面一定会有突出的技术支撑。然而，再强大的技术依然需要服从法律和制度的安排，我们在大力发展新质生产力的过程中不仅要关注技术进步，同时也需要研究法规和制度设计，特别是在互联网发展进入了Web3.0的过程中应该抓住机遇，以习近平新时代中国特色社会主义思想为指引，不断从区块链试点案例中总结经验，并在制度设计方面做出理论和实践创新，形成我国经济与社会发展的新局面。

云原生 AI

（LFAP） 陈泽辉

云原生与 AI，两者间存在着天然的紧密联系，仿佛一对孪生兄弟。云原生的架构设计，广泛应用于 AI 部署及机器学习等领域，其高效、灵活的特性，使得二者相得益彰。正是基于此，云原生 AI 成为了业界热议的焦点。

谈及云原生，Kubernetes 作为云上操作系统的核心地位不容忽视。在 AI 应用中，Kubernetes 以其强大的扩展性、灵活的存储管理及数据处理能力，成为了不可或缺的支撑力量。近年来，Kubernetes 培训在国内迅速发展，不仅吸引了传统应用领域的开发者，更吸引了大量投身于云原生 AI 领域的探索者。

云原生 AI 之所以受到青睐，不仅因为其资源利用的便捷性——无论是公有云、私有云还是混合云环境，都能轻松应对；更在于其提供的协作平台，让 AI 从业者能够更高效地协同工作。此外，云原生还简化了 AI 部署的复杂度，使得那些原本对架构不甚了解的 AI 专家也能轻松上手，专注于 AI 模型的训练与优化。

在 AI 领域，模型的多样性与依赖关系的复杂性是一大挑战。云原生通过容器化技术，有效解决了这一问题，使得依赖关系得以清晰呈现，便于追踪与解决潜在问题。同时，云原生还提供了强大的扩展性与共享能力，尤其是在 GPU 等高性能计算资源的管理上，更是展现出了其独特的优势。

然而，云原生 AI 的发展并非一帆风顺。在享受其带来的诸多便利

的同时，我们也面临着数据隐私、模型偏见、资源分配与故障恢复等多重挑战。为解决这些问题，我们需要借助开源的力量，不断探索与创新，共同推动云原生 AI 技术的成熟与发展。

在未来的云原生 AI 发展道路上，我们面临着诸多机遇与挑战。首先，在工具层面，随着 AI 资源的日益丰富与多样化，我们需要构建更加灵活、互通的工具集，以支持不同领域、不同背景的开发者能够轻松上手并高效利用这些资源。这将有助于打破技术壁垒，促进 AI 技术的普及与应用。

安全、便利与可持续性是云原生 AI 发展中不可忽视的三大问题。在确保数据安全的前提下，我们需要通过优化资源分配、提高计算效率等手段，实现 AI 发展的经济性与环保性。同时，云原生技术的引入将为我们提供更加灵活、可扩展的架构支持，有助于我们更好地应对 AI 发展中的各种挑战。

在模型依赖关系管理方面，云原生技术同样能够发挥重要作用。通过容器化等技术手段，我们可以清晰地呈现模型间的依赖关系，便于追踪与解决潜在问题。此外，对于 GPU 等高性能计算资源的管理与利用，云原生也提供了更加高效、便捷的解决方案。

在 AI 模型的解释性方面，我们需要推动标准化与统一化进程，以减少误解与沟通障碍。同时，开放源代码的模型与工具将有助于我们共同学习、共同进步，推动 AI 技术的持续发展。

对于云原生 AI 的发展蓝图，我们需要关注当前已经建立的项目以及正在开发中的项目。这些项目将为我们提供宝贵的经验与启示，帮

助我们更好地规划未来的发展方向。同时，我们也需要关注开源社区的动态，积极参与其中，共同推动云原生 AI 技术的繁荣与发展。

我想强调的是，云原生与 AI 之间的紧密联系将是我们未来发展中不可忽视的重要因素。无论是 GPU 的管理、大模型的依赖关系处理还是可扩展性的提升等方面，云原生都将为我们提供强大的支持。因此，我建议大家在部署 AI 应用时务必考虑云原生的因素，并积极参与相关的培训与学习活动以提升自己的技能水平。

RISC-V 开源高性能处理器核“香山”联合开发实践

北京开源芯片研究院 包云岗

2024.08.19

RISC-V 正引领着新一轮的产业变革，特别是在芯片设计领域。2023 年，麻省理工学院《科技评论》这一权威技术杂志，将 RISC-V 评选为年度十大突破性技术之一，其核心原因在于芯片设计正逐步走向开放化。RISC-V 以其灵活、开源的特性，有望成为颠覆传统芯片设计格局的关键力量。

然而，为何 RISC-V 直至近年才崭露头角并推动开源芯片设计的发展？关键在于历史上多数指令集均被企业私有化，这种私有化模式严重阻碍了开源芯片生态的构建，导致开源芯片的发展远远滞后于开源软件，时间跨度近乎四十年。传统处理器开发流程通常始于指令集手册这一标准文档，随后进行微架构设计、生成设计文档、编码（RTL 代码）直至最终通过 EDA 工具生成版图并流片生产。若指令集私有化，则整个流程中涉及的工具与实现均难以开源。

RISC-V 的诞生，其核心理念在于指令集应免费开放，这一理念自其诞生之初便贯穿始终。尽管在技术上，RISC-V 初期相较于英特尔、ARM 等成熟架构在指令集完备度上有所不足，但其理念上的根本性转变，为全球共同构建新生态提供了可能。这一新生态的构建方式，正成为新的竞争赛道，吸引着全球范围内的积极投入与参与。

在此背景下，我们于 2019 年启动了香山开源高性能处理器核心的研发项目。我们的初衷是借鉴 Linux 在操作系统领域的成功经验，探

索在CPU领域构建类似Linux的开源生态，既满足工业界的广泛应用需求，又支持学术界进行多样化的创新研究。我们期望未来能有数百乃至上千篇博士论文基于香山项目展开，推动CPU领域的持续进步。

自项目启动以来，我们已取得显著进展。2021年，第一代香山处理器核心研发成功，其性能可与ARM 2016年推出的Cortex-A973相媲美。随后，在2022至2023年间，香山迭代至第二代，性能达到Cortex-A76水平。目前，第三代香山正处于研发阶段，其性能预计将达到ARM Cortex-N2的水平。同时，我们也在不断优化面积、功耗等关键指标。

香山作为一个开源项目，其活跃度与影响力持续提升。自2021年6月11日在GitHub上建立仓库以来，已累计获得9000多个提交（commit）、4500多个星标（star）、600多个fork以及来自80多位贡献者的支持。在性能表现上，香山处理器核心通过标准测试展现出强劲实力，其性能指标已接近国内现有高端处理器的水平。

目前，已有多家企业基于香山处理器核心开发产品，预计明年将有一批芯片正式进入市场。此外，香山还吸引了国际知名大学与科研机构的关注与参与，推动了高水平研究成果的产出与发布。

尤为重要的是，香山项目通过开源模式实现了企业与学术界的紧密合作与联合开发。2021年12月，16家企业联合出资成立了开源芯片研究院（简称“开芯院”），旨在搭建学术界与产业界之间的桥梁，通过开源方式促进联合研发与技术创新。这一机制得到了业界的广泛认可与各级领导的高度评价。

在香山开发过程中，我们不仅致力于构建高性能的开源处理器核心，还积极探索了一系列创新的设计方法与流程，这些新方法不仅提升了开发效率，还促进了学术研究与产业实践的深度融合。

我们引入了敏捷设计的开发流程，并采用高级开发语言进行芯片设计。这种语言具有更高的抽象层次，能够显著提升开发效率并减少代码量。相比传统芯片设计语言，我们的新流程使代码量减少了 80%，这一显著优势得益于对软件工程最佳实践的借鉴。通过面向对象的方法，我们将处理器分解为独立的模块（类），使得不同指令集的支持变得灵活且易于重构。例如，我们的本科生团队仅用了两周时间，就在一套源代码基础上实现了从 RISC-V 到龙芯指令集的支持转换，充分展示了这种设计方法的灵活性和高效性。

在香山的设计中，我们大量采用了模块化与解耦的策略。这种策略使得各个模块可以独立开发、测试和维护，极大地提高了开发效率和可维护性。通过面向对象的思路，我们能够将复杂的处理器设计问题简化为一系列相对简单的模块问题，每个模块都专注于解决特定的功能需求。这种设计方法不仅降低了开发难度，还促进了团队成员之间的协作与知识共享。

针对传统芯片验证成本高、周期长的问题，我们提出了开源硬件众包的验证思路。通过构建“万众一心”平台，我们打破了验证环节与设计环节之间的紧耦合关系，使得更多的软件工程师能够参与到芯片验证中来。这一创新不仅降低了验证成本，还加速了验证进程。通过统一的接口和多语言支持，我们成功地将软件生态导入到硬件设计

领域，实现了跨领域的资源共享与协同合作。在实践中，我们已经证明了这一方法的有效性，例如让非专业的本科生团队在短时间内完成了核心模块的验证工作。

香山项目的成功离不开开源社区的支持与贡献。我们积极构建香山开源社区，邀请全球范围内的开发者、学者和爱好者共同参与项目的开发、验证与改进。通过开放源代码、工具平台和文档资源，我们为社区成员提供了丰富的学习与交流机会。同时，我们也鼓励社区成员提出新的想法和解决方案，共同推动香山项目的持续发展。

医疗影像 AI 胰腺癌全球大规模筛查

(阿里达摩院)姚佳文

2024.8.19

我国作为癌症高发国家，每年新增癌症病例数与因癌症死亡的人数均居世界前列，这一现状预计将在未来进一步加剧，到 2030 年，癌症治疗费用或将攀升至万亿元级别。这一沉重的社会负担，很大程度上源于癌症早期症状的隐匿性及有效筛查手段的缺乏，导致多数患者确诊时已处于中晚期，治疗成本高且效果有限。因此，早发现、早诊断、早治疗，对于提高患者生存率具有至关重要的意义。为此，达摩院致力于开发基于平扫 CT 的癌症筛查工具，并率先推出了针对胰腺癌的筛查方案。

本项目的灵感源自于一位合作医生的亲身经历。这位医生在其恩师确诊晚期胰腺癌后，回顾了恩师生前十个月的体检资料，意外发现一张平扫 CT 片中存在胰腺病变的蛛丝马迹。然而，由于当时无明显症状且胰腺癌发病率相对较低，这一迹象被忽视了。直至十个月后，恩师因腹痛就诊，经增强 CT 检查确诊为胰腺癌晚期，错失了最佳手术时机。这一悲剧深刻揭示了胰腺癌早期筛查的重要性及其严峻挑战——早期无症状、进展迅速，导致九成患者确诊时即为晚期，五年生存率极低，被称为“癌症之王”。

尽管胰腺癌相对罕见，但我国庞大的人口基数使得其患者数量依然庞大。据统计，2022 年我国新增胰腺癌患者及死亡人数均达到 13 万之众，每天都在上演着“癌症之王”的悲剧。遗憾的是，当前临床

上尚缺乏针对胰腺癌的有效早期筛查手段，高精度工具匮乏。

基于上述背景，我们与医生紧密合作，提出了一个创新思路：能否利用广泛普及、安全便捷且成本低廉的平扫 CT，结合人工智能技术，实现对胰腺癌的早期筛查？这一设想不仅符合实际需求，还能在不增加患者经济负担和辐射风险的前提下，提升筛查效率与准确性。

然而，平扫 CT 的低对比度特性给医生诊断带来了巨大挑战，极易导致漏诊。针对这一难题，我们依托自身的技术优势，与医生深入临床一线，共同研发了名为“PANDA”的 AI 筛查模型。该模型采用三阶段分级检测及诊断网络，能够精准定位胰腺、识别病灶，并有效区分胰腺癌与其他胰腺疾病。通过大量临床数据的训练与优化，我们成功提升了模型的准确性，使其对胰腺病变的检出率达到 93%，特异性高达 99.9%，几乎实现了零误报。

为确保研究成果的可靠性与安全性，我们与国内外多家顶尖医疗机构合作，进行了严格的人机对比验证及真实世界测试。结果显示，“PANDA”模型在胰腺癌早期筛查中展现出优异的性能，其敏感性、特异性及鉴别诊断效果均达到或超过现有临床标准。这一成果已于去年在《Nature Medicine》上发表，得到了国内外同行的高度认可与赞誉。

在我们的研究中，为了全面评估 AI 模型与人类医生在胰腺癌筛查中的性能差异，我们邀请了来自 12 家医院的 33 位医生，按专业级别分为三组：胰腺专家（副主任及以上）、普通影像科医生和影像科住院医师。每组医生均独立审阅了包含 290 个病例的测试集，这些医

生在审阅时了解了患者的部分临床信息。结果显示，我们的 AI 模型在多个关键指标上全面超越了人类医生。

当我们将 AI 模型的初步筛查结果提供给医生作为参考后，一个月后再次请这些医生复审相同病例，发现 AI 显著提升了医生的诊断能力。具体而言，在胰腺病灶的检出任务上，医生们的敏感性平均提高了 9%；而在胰腺癌的特异性检出上，提升更是高达 21%，部分医生的表现甚至超越了胰腺专家的水平。这一发现不仅证明了 AI 模型在提升诊断效率与准确性方面的巨大潜力，也预示着其在医疗资源相对匮乏地区的应用前景，有望实现医疗资源的“共同富裕”。

针对早期胰腺癌这一诊断难点，我们的研究进一步揭示了 AI 模型的优势。在增强 CT 领域上，即使是经验丰富的专家也难以保证对所有早期胰腺癌的准确识别，而我们的 AI 模型却能在平扫 CT 上捕捉到这些难以察觉的病变。这一发现对于提高早期胰腺癌的检出率、改善患者预后具有重要意义。

此外，我们还验证了 AI 模型在胸部平扫 CT 中的应用效果。由于许多患者在接受肺结节筛查时也会进行胸部 CT 检查，这为我们提供了额外的数据资源。通过对 490 例胸部平扫 CT 病例的分析，我们发现其中 80%的胰腺病灶能在胸部 CT 中有所体现，而 AI 模型更是能够在这些间接征象（如胰管扩张）中捕捉到胰腺癌的蛛丝马迹，其敏感性远高于人类医生。

为了将研究成果转化为实际应用，我们开发了“达医智影”软件，并成功在医院云端进行了部署。该软件能够实现与医院现有系统的无

缝对接，为医生提供实时的诊断辅助。在真实世界的验证中，该软件连续处理了 2 万例日常检查数据，成功发现了 26 例原本被临床医生漏诊的胰腺病例，其中包括 1 例胰腺癌和 3 例其他胰腺病变。这一成果不仅证明了 AI 模型在提升临床筛查准确率方面的有效性，也体现了其在促进患者早期发现、早期治疗中的重要作用。

2023 年，一位因咳嗽就诊的患者在接受胸部平扫 CT 检查时，其胰腺病变被 AI 模型准确识别并提示为胰腺癌。这一预警信息引起了医生的高度重视，并促使患者接受了进一步的增强 CT 和 MRI 检查，最终确诊为胰腺癌。这一成功案例不仅彰显了 AI 模型在提升医生诊断信心、避免误诊漏诊方面的价值，也为我们继续推进医疗 AI 领域的研究与应用提供了有力支撑。

AI Agent 时代
大模型产业八项落地八个思考

沈向洋

（摘自中国电子版 2024.09.05）

9 月 5 日，香港科技大学校董会主席、美国国家工程院外籍院士沈向洋在 2024 Inclusion · 外滩大会上分享了他对大模型产业落地的八个思考。他认为，AI Agent 时代的到来，不会是一个神奇而强大的模型突然代替了所有工作流，它涉及技术、工程与市场的不断磨合，最终以超预期的服务呈现给人类。

思考一：算力是门槛

“今天做大模型，做深度学习，首先最重要的事情是要有算力。”沈向洋表示。他指出，从 2010 年开始，大模型需要的算力以 6、7 倍的速度增长。这几年稳定下来，大概每年有 4 倍的增长。模型越来越大，参数量越来越大，算力的需求也随着参数的增长，呈现出平方向的增长。在他看来，整个计算机芯片行业的发展已经从原来的“摩尔定律”变成了“黄氏定律”。以前摩尔定律认为，算力每隔 18 个月增长一倍。如今预测，GPU 将推动 AI 算力实现逐年翻倍。“讲卡伤感情，没卡没感情。以前有一句话叫贫穷限制想象力，现在贫穷可能扭曲想象力，因为如果没有卡，能想象要做的项目可能就不太一样了。”沈向洋感慨道。

思考二：关于数据的数据

公开资料显示,GPT3 的训练数据达到了 2 个 T 的 token(吞吐量),GPT4 则达到了 12 个 T 左右。据沈向洋预判，GPT5 的训练数据可能会达到 200 个 T。目前互联网上的数据远远不能满足未来模型训练的需求，还需要思考用什么办法去挖掘更多的数据。在人工智能领域，数据被视为模型的“燃料”，模型需要从这些数据中学习和提取有用信息。因此，数据的数量、质量和多样性都会直接影响到模型的准确性和性能。沈向洋表示，之前作为互联网最核心的积累，数据大多被谷歌用来做搜索引擎，以后这些数据都会被拿来训练大模型。“互联网 40 年积累的数据，好像就是为了这样一个 AI 时刻”。

思考三：大模型的下一章

下一步到底要干什么？沈向洋认为，大模型产业未来的发展路径已经非常明确，将会从之前的大语言模型，到多模态模型，未来迈向世界模型。从技术上讲，肯定要走理解和生成统一起来的道路。“未来一定会往具身智能方向上走，往机器人上面走，其中一个特殊形态就是自动驾驶。”沈向洋说道。实际上，关于世界模型业内并没有形成一个标准的定义。OpenAI 推出的 Sora 模型曾引发业内对“世界模型”的探讨。OpenAI 将其视为能够理解和模拟现实世界的模型的基础，相信其能力是实现 AGI（通用人工智能）的重要里程碑。然而，沈向洋认为，“Sora 模型虽然做的非常好了，但还不是那么强大，里面的物理性质是不能保证的，做不到一个世界模型。”

思考四：大模型横扫千行百业

大模型可分为通用大模型、行业大模型、企业大模型和个人大模型。沈向洋指出，通用大模型是 AI 的基础，要训练一个通用大模型至少需要万卡；行业大模型是做领域应用的底座，需要千卡级别的训练；企业大模型是企业数据价值的再发现，需要百卡级别的训练。这些大模型都对算力的要求极高。“最激动人心的是个人大模型，比如联想、微软在推 AIPC 和苹果的 Apple Intelligence 等都是朝着个人智能这个方向发展的。”沈向洋说道。截至今年 7 月底，中国已备案大模型达到了 197 个，其中 30%是通用大模型，70%是行业大模型。“可以看到，行业大模型占到绝大多数，未来肯定还会越来越多。”沈向洋表示。

思考五：AI Agent——从愿景到落地

2024 年 5 月，微软公司创始人比尔·盖茨公开表示，AI Agent 不仅会改变每个人与计算机交互的方式，还将颠覆软件行业，带来从键入命令到点击图标以来，最大的计算革命。沈向洋对此观点表示认同。他认为，人工智能时代，真正了不起的超级应用就是 AI Agent。AI Agent 从愿景到落地的过程中，需要始终以需求为圆点，深刻理解模型的能力，并构建一个 AI 深度参与的工作流程。“今天在一家公司里工作的话，整个工作流是非常复杂的，ChatGPT 虽然很强大，但远远没达到 Agent 的程度，它只实现了单点突破，真正要向前走还得融入整个工作流。”他说。

思考六：重视 AI 的治理

AI 治理非常重要。今年世界人工智能大会（WAIC）的主题就是在讲 AI 治理，各个国家对于这件事情的看法有很多不一样。AI 的发展，对民众、公司、政府监管、社会发展等各个方面都产生了强大的冲击，引发了公众对于其安全治理的担忧。“我觉得接下来人工智能的发展很重要的一点，从全球各个国家角度来讲，是一定要做主权人工智能，而主权人工智能背后一定需要有一个主权云来支持主权人工智能的发展。”沈向洋表示。

思考七：重新思考人机关系

“GPT 带来的冲击有多少是人机交互的震撼，又有多少是机器智能的发展？”沈向洋认为应该重新思考人机之间的关系。他指出，AI 为人类提供了与技术共生的全新语境，人机交互的新方式指向“AI 与 IA”的融合共进。IA（Intelligent Augmentation），即智能增强，代表着一种以人为本的 AI 发展路径。它聚焦于运用技术提升人类的能力，而非取代人类，强调了人类与 AI 之间的协作关系。“纽约时报专栏作者 John Markoff 提到，计算机过去几十年的发展路程里，真正的赢家是做人机交互的。不管是什么技术，最后的目的都应该是帮助人类更好地使用机器。”沈向洋表示，“到了 AI 时代，人机交互最本质的是对话，就像 ChatGPT 这样。ChatGPT 加上微软，会不会成为 AI 时代最伟大的公司？我想只有时间才能够验证。”

思考八：智能的本质

今天，GPT 的发展如火如荼，但实际上，人们对智能的理解还是非常有限的。不同于物理学，上到浩瀚的星空，下到微小的量子，都

能有一个大一统的理论可以解释；今天的深度学习很多东西都是不可解释的，没有鲁棒性。“智能的本质是神经网络与符号系统的世纪之争。”沈向洋说道，“今天，虽然人工智能的发展还处在一个相对来讲比较早期的阶段，但是行业上已经有很多的应用，值得下定决心去做，我对未来的发展充满信心。”

讨论人工智能操作系统（AIOS）研制问题

COPU 2024.9.10

主持: 陆首群（COPU名誉主席）

在当前人工智能时代，国内外关于人工智能操作系统的研制犹如井喷，我们要与日俱进，跟上时代的步伐。国内外研制AIOS分两类: 一类是准AIOS(feature AIOS),从应用入手,在传统OS中加智能模块，另一类是全功能OS（A11 AIOS）从框架或内核入手，结合AI模型全面构建和研发AIOS。研制单位也有两类: 一类是企业界，另一类是学术界，在企业界中也有桌面、手机和汽车等几个层次。

语言大模型智能代理操作系统（LLMAgentOS）是人工智能操作系统（AIOS）的一种类型，其主要特点是将大模型嵌入操作系统中，作为操作系统的“大脑”，以实现一个“有灵魂”的操作系统，该操作系统的主要职能是: ①优化运行在LLM上的资源分配，②促进各集成代理之间上下文切换，③实行代理的并发执行，④为代理提供工具服务，⑤维护各集成代理之间的访问控制，LLMAgentOS架构的具体职能为: ①代理调度，②上下文管理，③内存管理，④存储管理，⑤工具管理，⑥访问管理等。

信息技术行业、汽车行业的人工智能操作系统（AIOS）不能直接套用大模型智能代理操作系统（LLMAgentOS/AIOS，或简化为LLMOS），但可供参考。

推荐参阅 https: //github.com/agiresearch/AIOS

这次会议我们请清华大学计算机系陈渝老师介绍国内外研制AIOS的情况,还请电子系汪玉教授团队的核心成员颜深根老师介绍他们研制情况; 这次讨论会，我们邀请正在研制的十几家单位的代表参加讨论（学术界2位代表、企业界9位代表)，今天先做简短的一般性讨论，以后还要深入研究讨论（包括如何组织开发问题）。

下面请清华大学计算机系陈渝教授发言:

目前AI与OS的结合有两种思路方向:

1）产业界思路: OS for AI，从应用入手，自顶向下，形成AI实用+以大模型为基础的框架结构+网络+传统操作系统，这种方式对传统OS和网络几乎没有改动，适会企业尽快进入AI应用领域，并逐步向下修改OS本身。从发展现状来看，AI应用的推广可能首先出现在以AI手机为代表的智能移动终端领域，以及自动驾驶汽车为代表的新能源智能网联汽车领域。同时在传统领域，如桌面、服务器也会巡视引入各种AI应用，而未来的工业/服务智能机器人将会把各种AI应用综合在一种硬件平台上。

2）学术界思路: AI for OS，从框架或内核入手，自底向上，把AI核型或深度学习模型直接引入OS内核和框架的构建研发中来，这种方式会对OS内核和框架产生较大的改变，有很多技术挑战，难度较大，需要在面向AI的计算理论和模型、AI加速硬件优化、编程语言、软件工程、系统架构、优化技术、软硬协同优化等多个层面形成突破。由于这方面难以迅速突破，所以基于全新革命性的

自底向上的研究需要时间和技术积累，产业落地周期长，一级企业难以采用这种方式。

无论哪种思路，都会向着一个趋势发展，即 AI 应用的强大需求，会进一步推动面向 AI 加速的芯片、网联、计算基础设施等的快速发展，并推动各种系统软件支撑、编程与 SDK 接口、服务框架的发展。从而推动计算机领域新的一个跳跃式发展。

请清华大学电子系颜深根教授介绍研制 AIOS 的情况:

大模型带来的变化: 1）任务负载的变化。对计算资源、存储资源、带宽资源消耗更多。2）人工智能能力的变化，从识别到理解+生成。理解了之后，可以给出下一步的操作，可以转化成具体的指令。

对操作系统的影响: 1）对应第一点，OS for AI，为了应对 AI 的负载，需要在 OS 层面做一些改变，比如计算效率要更高，支持更高带宽的存储。从以 CPU 为中心到以协处理器为中心。2）对应第二点，AI for OS，通过 AI 的能力，增强 OS 本来的一些功能。比如 OS 人机交互界面的更新，从触摸屏升级到语音、图像交互。OS 自带文件检索功能，OS 运行日志的自动整理等。

目前大模型+OS 层面需要解决的问题: 1）云侧系统太过复杂，尤其在中国，底层芯片种类非常多，如何兼容各种不同的芯片。如何稳定的运行。尤其是国产芯片性能比不过英伟达的芯片，需要更大的并行规模，带来了效率和稳定性方面的双重挑战。2）在端侧目前大模型运行有三个方面的挑战，一个是算力，第二个是功耗，

第三个是存储空间。算力和功耗目前主要是通过芯片工艺改进、针对性的芯片设计、计算指令优化来解决。存储分为两个方面，一方面是存储带宽，这个目前有存内计算、近存计算等思路。存储的第二个方面是存储容量，更大容量的存储导致成本上升，这个目前没有太好的解法。

关于 AI+OS 的建议：个人感觉 AI+OS 在云侧目前投入已经很多了，主要是端侧目前投入还比较少，希望能够有一个机制，把端侧相关的应用公司、系统公司、芯片公司拉到一起，大家商量怎么样统一接口，从试点开始，逐步增加 AI+OS 的应用，通过应用带动系统发展。

COPU 副主席、北京大学教授陈钟：谈到 AIOS，我建议更多关注苹果公司 Apple-16，iOS，因为苹果公司是一个从硬件芯片到整机到软件全链条的生态。特别是其对结合大模型能力集成的理解具有标杆性，并且能够根据 AI 需要将芯片功能模块级、操作系统级和应用级完整设计和集成。9 月 10 日 iPhone16 正式发布，苹果公司推出产品级的设计成熟而稳定，甚至比微软更具有全球前沿性和引领性。

统信软件产品经理顾雨婷介绍：统信软件正在将 AI 能力变成操作系统的基础服务之一：1）在人机交互方面，统信软件推出的 UOS AI 智能助手通过全新升级的自然语言处理技术，能更好地理解用户需求，提供智能搜索、推荐系统、自动化办公助手等个性化服务；2）在应用生态方面，60 余款应用已接入 UOS AI 生态，通过本

地/在线模型，向各应用提供 AI 能力支持；3）在开发赋能方面，统信软件打造的 UOSAIAPI 是国内首个系统级 AI 应用开发框架，为 AI 应用开发提供系统级 AI 接口，赋能生态伙伴应用开发。

开源鸿蒙 PMC 主席任革林谈到: 鸿蒙操作系统原生支持端侧 AI，包括支持 AI 基础能力和利用 AI 自主改造的能力。AI 基础能力主要包括对 AI 硬件的管理、内置 AI 开发框架和 AI 推理模型等，方便用户程序高效利用系统的 AI 资源。操作系统利用 AI 自主改造的能力主要体现在利用预置的 AI 能力进一步提升当前操作系统智能化水平。譬如通过周边环境感知、运行上下文感知等 AI 技术让系统资源的管理更加高效；通过语音识别，机器视觉，用户操作习惯学习等 AI 能力，让交互操作个性化、更自然；通过 AI 能力将传统 UI 控件改造为智能 UI 控件，支持智能抠图、图像分割、图像理解、文字提取、文本朗读和智能纠错等；通过 AI 对分布式能力进行改造，提供智能搜索和智能协同等体验。

中兴通讯技术规划部标准规划总工高峰表示: 算力基础设施是 AIOS 的一个重要考虑方向，通过 OS 实现算力+网络+存储资源的分配、调度和优化，为 AI 提供基础服务能力

OPPO 技术总监罗升阳谈了几点: 1）大模型的端侧部署问题，一是功耗，二是内存带宽，三个计算生态。2）功耗问题：当前的旗舰平台手机运行 7B 模型需要的功耗还比较大，功耗和玩重度游戏差不多，会有发热发烫的问题，影响用户体验。3）内存带宽问题：一个 7B 模型，经过量化压缩之后，大概还占 4GB 左右的内存。

由于大模型自回归推理的特点，导致每输出一个 token 都需要加载一遍参数到 NPU 计算。以 20 tokens/秒的推理速度计算，需要 80GB/秒的内存带宽。这对移动平台的芯片设计是一个很大的挑战。4）计算生态问题：当前移动芯片平台主要有高通和 mtk 两个大玩家，但是他们都是各玩各的，导致手机厂商开发商在部署模型时，面临碎片化问题。需要有一套统一的计算生态 API，它既是统一的，又像 CUDA 一样强大。

麒麟软件副总李震宁介绍到：当前，国外一些大模型热度减退，用户更加关注新技术能给自己带来的价值。人工智能的客户场景在哪里成为行业关注点。麒麟软件是国内自主创新领域应用量最大，应用程度最深的国产操作系统。针对政务和关基行业的用户需求。推出的 AIPCOS 从几个层面去帮助用户提升效率，拥抱 AI 未来：1）与传统桌面操作系统相比，融入了“AI 子系统”，该系统实现了模型和硬件、应用和模型的双重解耦。2）针对办公和行业应用场景，让 AI 和办公场景更深结合，实现智能搜索、AI 助手、记忆地图等增值功能。3）对于数据安全要求较高的行业，用户可以自由切换到本地大模型，摆脱网络依赖。

COPU 专家委员会副主任委员章文嵩表示：关于 LLM OS，从设计理念来看，应该保存每个模块的简单高效，不要把什么东西往内核里装。我们过去为了高性能更多地把内核的功能往用户态搬，网络有 DPDK，存储有 SPDK，例如 DPDK 版本的 LVS 在高性能网卡上性能大幅高于内核版本的 LVS，我们很多分布式存储系统用 SPDK 越过

内核直接管理存储设备。LLM 应该是用户态的服务，为所有应用程序共享，它是 OS 的一部分。

有两个建议：1）大家共建 LLM 服务的 API 标准，这样不同的手机上 APP 不用适配各家的 LLM 服务，让整个生态更有效率。2）针对 ARM+NPU 开源共建一套类似 CUDA 的计算框架，可以基于 AMD 的 ROCM 开源框架来做，若华为把 CANN 开源出来，一起做面向 ARM+NPU 的开源计算框架。

COPU 常务副秘书长谭中意认为：一方面在 IDC 的服务器集群内，如何支持大模型的高效训练；一方面在端侧，如何支持大模型的高效推理。两个方面的迭代，方向不同，要求不同，都应该是 AIOS 的范围。

陆主席小结说：

①有能力的企业研发 AIOS 是否考虑两条线：一是以应用入手，二是对 OS（向内核、框架到全部）以 AI 进行全面创新。②对应用入手的 AIOS，从应用场景需求出发，联合制定应用规范，供国内研发单位研发采用。

请大家考虑。

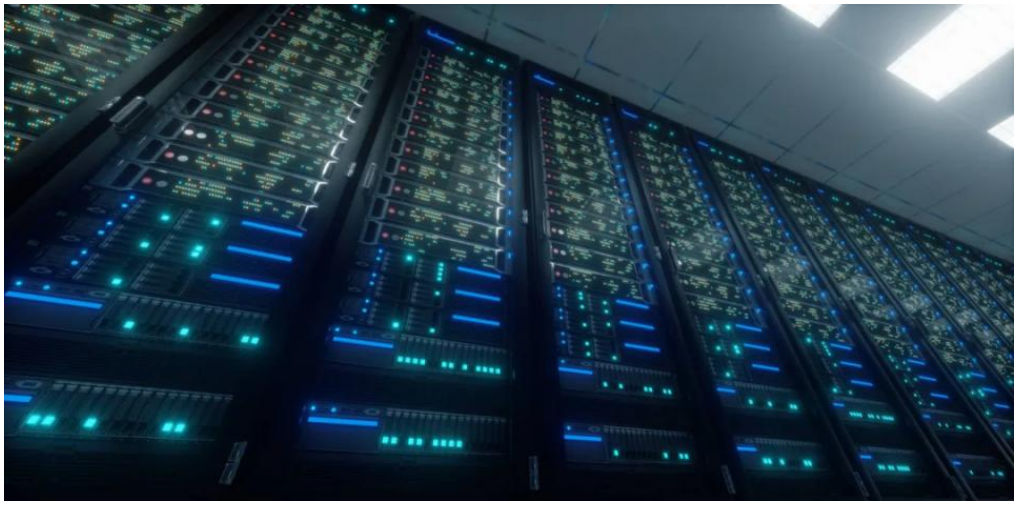
两则信息：

（一）国际奥委会主席巴赫的重磅之言

若没有中国的全力相助，巴黎奥运会决然达不到现今的程度，因为中国的阿里巴巴作为全球领域高科技企业之一，云技术和 AI 解决方案提供了强有力的支撑，超过半数的全球转播商都用阿里巴巴的云服务（8K 高清转播）转播。

（二）马斯克 10 万卡机房建成为标配

马斯克 10 万算力卡集成训练服务中心机房建成成为标配，预示着 AI 将进入一个新时代。



马斯克的 10 万卡机房与大型集成算力训练服务中心的“星际之门”类似，“星际之门”初步计算为：采用 10 万张 H100 芯片集成，综合算力可达 10^9 TFLOPS，需要 5000MW 电力，投资估算约 4 万亿美元。

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　　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