加密 × AI，AI × 加密：综述

报告日期：2026.6.8　　中文版整理：Wang Hua　　英文原文：IC3

版本： 1.0 **日期：** 2026年6月8日

编辑： Giulia Fanti[1,3] 和 Ari Juels[1,4]

作者：

Sarah Allen[1,5], Pranay Anchuri[6], James Austgen[1,4], Maryam Bahrani[7], Samuel Breckenridge[1,4], Aaron Buchwald[2], Christian Cachin[1,9], Andrés Fábrega[1,4], Jared Fernandez[3], James Hsin-yu Chiang[1,12], Marwa Mouallem[1,8], Roi Bar-Zur[14], Neil DeSilva[1], Ittay Eyal[1,8], Giulia Fanti[1,3], Ari Juels[1,4], Andrew Miller[1,13], Christian Sillaber[9], Dani Vilardell[1,4], Pramod Viswanath[11], Wenhao Wang[1,10], Matt Weinberg[1,11], Sen Yang[1,10], Jianzhu Yao[11]，以及 Fan Zhang[1,10]

作者机构： [1]Initiative for CryptoCurrencies and Contracts (IC3)；[2]Ava Labs；[3]Carnegie Mellon University；[4]Cornell Tech；[5]Flashbots；[6]Offchain Labs；[7]Ritual Labs；[8]Technion；[9]University of Bern；[10]Yale University；[11]Princeton University；[12]ETH Zurich；[13]Teleport；Flashbots(X)；[14]Tel Aviv University

---

摘要

加密 × AI 的交叉领域正催生出大量的论文、产品、在线帖子和公司。然而，伴随而来的喧嚣掩盖了究竟已经完成了什么、机遇和挑战是什么，以及哪些开放问题值得关注。本综述论文探讨了AI能为基于区块链的技术（广义上的“加密”）做什么（加密 × AI），以及反之亦然（AI × 加密）。我们对现有工作进行系统化整理，总结关键要点，突出开放研究问题，并对行业中普遍存在的误解提出我们的看法，最终得出结论：AI和加密仍处于有意义整合的非常早期阶段。重点内容包括：

今日现状：

1. **加密 × AI**：AI可以帮助分析和检测现有加密交易、事件和协议的重要属性。已有大量工作探索了基于AI的方法来检测欺诈性或存在漏洞的智能合约和协议。这些技术传统上使用简单的机器学习方法，并且在拥有充足训练数据的受控环境中最为有效。
2. **AI × 加密**：加密工具为保护和治理AI工作流提供了新方法。加密社区中广泛使用的几种工具——包括零知识证明和可信计算——可以被重新用来使AI结果更不易被篡改。加密社区中其他有吸引力的想法，如去中心化治理和基础设施管理，尚未在主流AI社区中得到真正的采用。

加密社区所需的实证支持：

1. 去中心化AI解决方案需要与其中心化对应方案进行更严格和直接的成本比较。加密 × AI 行业主要侧重于展示以去中心化方式训练大型模型的可行性。虽然去中心化有其自身的优点，但需要在特定用例和场景下，更量化地展示与中心化AI平台在成本上竞争的机会。
2. 加密支付轨道需要更严格地阐明并证明：与中心化替代方案相比，它对智能体支付具有何种实际价值。尽管加密支付尚未获得显著的市场采用，但智能体支付因费用低，并且不受传统金融基础设施要求账户由人类主体持有或控制的模式限制，而展现出良好前景。加密社区应抓住这一机遇，以及传统金融公司在支付和新型智能体试点中日益增加的加密活动，通过量化方式证明加密对智能体支付的益处，而不应只证明其可行性。

研究挑战：

1. **AI安全需要系统级防御。** AI社区通常在AI模型层面处理安全和保障问题，围绕输入/输出语义设计护栏和防御。随着代理获得自主权和对基础设施的访问权，这种方法将被证明是不够的。加密工具，包括可验证执行和经认证的处理流程，可以帮助提供模型级防御无法提供的系统级保证。
2. **将加密与AI结合创造了新的威胁行为体和攻击向量。** 诸如投资组合管理等AI应用造成了不可避免的隐私与公平之间的张力，而将AI智能体与去中心化和加密货币结合可能会产生危险，例如不可阻挡的自主智能体（unstoppable autonomous agents）或恶意智能合约（rogue smart contracts）。理解哪些威胁是现实存在的，以及哪些缓解措施将有效，都是研究的当务之急。

目录

第三章 C：AI × 加密：用加密增强AI

第四章 D：误解与半真半假

第五章 E：致谢

参考文献

第一章 A

引言

在新兴技术领域，加密（货币）和人工智能（AI）可能获得了前所未有的关注、兴奋、炒作和怀疑程度 [23, 64, 81, 667, 675]。如今，存在无数用AI彻底改变加密，反之亦然的提案 [148, 193, 352, 396]。对于观察者来说，要理清真正的用例，并理解AI和加密在何时以及如何相互契合，可能颇具挑战。

本综述论文提出了一个统一的框架来分类AI和加密之间的联系。我们将展示现有研究如何映射到我们提出的框架，以及哪些主要研究问题仍未得到解答。此外，我们旨在突出那些不符合我们框架的流行趋势，和/或我们认为目前尚不现实的用例。在整篇综述中，我们使用“**加密 × AI**”表示应用于加密的AI；“**AI × 加密**”表示应用于AI的加密。

什么是“加密”和“AI”?

我们使用术语“加密”（crypto）大致有三种含义。

第一，历史上“加密”是“密码学”（cryptography）的缩写，虽然它最初指隐藏信息，但现在指代一系列用于安全存储、传输和（可能）对机密数据进行计算的技术。密码学工具包包括数字签名、门限签名或安全多方计算。区块链开发者一直是某些先进密码学工具最早的广泛采用者之一，并特别推动了零知识（ZK）证明的演进，它使用户能够在不披露秘密的情况下证明对秘密的了解。[intro-1] 所有这些密码学工具都可以应用于保护和调解AI的使用，我们将在本综述中进行探讨。

我们还将把密切相关的**可信计算**技术归入“加密”的旗帜下。该术语指代特殊的计算环境——通常由专用硬件支持——旨在通过防止篡改和秘密泄露来保护软件应用（尽管有重要的附带条件）。它们在区块链应用中越来越重要，这反过来又推动了它们在非区块链应用中的增长和使用，正如我们在本综述中所解释的。[intro-2]

第二，继比特币、以太坊和其他基于区块链的系统在过去十年的普及之后，“加密”是“加密货币”的简称。这是一个建立在密码学原语之上的完整经济层，包括代币、稳定币、去中心化金融（DeFi），以及围绕它们建立的交易所和其他服务提供商的生态系统。这里的几个特征与AI直接相关，例如不可逆性、无需注册或身份验证即可创建账户和转移资金的能力，以及可编程的结算规则。

第三，“加密”可以被理解为一种文化运动，其价值观围绕无需许可的创新、通过去中心化实现的弹性，以及避免依赖可信第三方和中介。在这个意义上，加密是产生端到端加密消息和BitTorrent的“密码朋克”传统的延续，现在为加密货币和现代基于密码学的系统中所做的工程权衡提供了信息。这些价值观的模糊性是紧张局势的反复来源：加密货币生态系统中的许多项目“名义上去中心化”[469]。同时，它至少取得了部分成功：美国的监管架构已逐渐围绕它进行调整，例如，金融犯罪执法网络（FinCEN）2013年的指南区分了虚拟货币的“用户”（花费所挖货币的矿工；普通持有者）与“交易商”和“管理员”，仅将后者归类为货币转移者 [213]。

术语“AI”同样模糊，但指的是能够执行通常需要人类智能的任务的系统，如推理、问题解决、理解语言、识别模式和做出决策。AI涵盖了许多方法和技术，但在本综述中，我们主要关注**机器学习（ML）**。ML系统从数据或环境中学习以实现目标，而不是为特定任务集明确编程。因此，我们在本综述中互换使用AI和ML这两个术语。

为什么我们需要另一篇综述?

许多先前的工作探讨了AI和加密的交叉 [29, 82, 287, 460, 465, 530, 577, 625, 688, 733]。这些资源大多集中在特定领域，如金融和交易 [29]、智能环境和元宇宙 [199, 552, 688, 732]，甚至增强无线通信系统（如6G）的安全性 [465, 733] 等等。我们的目标是采取更广阔的视角，关注整个生态系统，而不是特定的垂直领域；同时，我们承认，区块链最广泛部署的应用集中在金融和加密货币领域 [281, 356]。

在通用综述中，大多数要么侧重于区块链对AI的影响 [287, 530, 577]，要么侧重于AI对区块链的影响 [4, 329, 573, 704]，相对较少同时涵盖两者 [82, 460]。此外，同时涵盖两者的综述没有明确区分基础设施与应用层面的未来方向，并且它们对“区块链”采取狭义的理解，仅考虑基于分布式账本技术（DLT）的部署。

本综述采取了更广阔的视角，产生了几个关键区别或特点：

1. 我们讨论AI对加密的潜在影响，反之亦然。为此，我们将这两种技术都定位为人类与自动化决策流程之间的**中间件**。我们的框架明确区分了AI和加密如何在加密栈的不同层被使用，而不是在同一抽象层次上并列列出不同的用例。
2. 我们使用“加密”作为总括术语，不仅包括区块链栈和应用，还包括基于可信计算工具和技术构建的去中心化系统。
3. 大多数先前的综述发表于生成式和代理式AI爆炸性增长之前，而我们将其纳入综述。
4. 除了对现有文献的综述，我们还分享了我们对于连接加密和AI的当前趋势（无论是在研究领域还是行业领域）的集体观点和解读。本着这种精神，我们旨在在一个广阔的竞争领域中识别有前景的研究方向。

A-1 AI-加密交互的框架

许多重要的决策流程将人类意图转化为自动化处理流程，如图 A.1 所示。

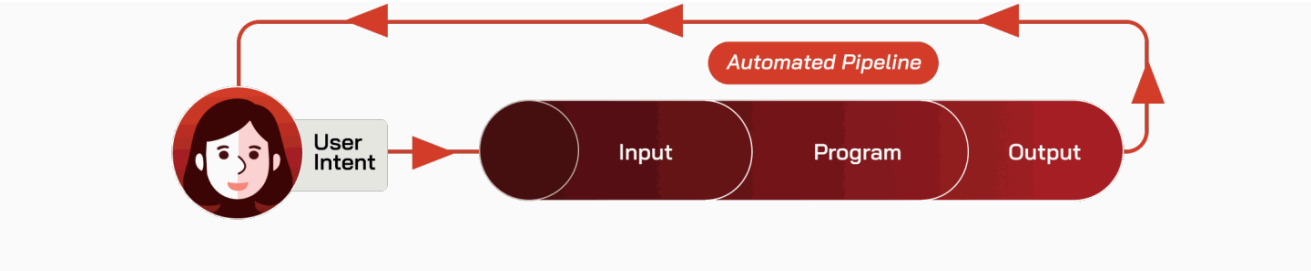


图 A.1 一个标准的自动化决策流程。一个人类意图（例如，“我想让我的车在停车标志前停下”）被转化为一个**程序**，该程序处理结构化**输入**（例如，来自汽车的视频流）并产生**输出**决策（例如，停车）。此处理由程序自主处理。

让我们考虑以下示例：用户希望确保其自动驾驶汽车在停车标志前停下。我们将此目标称为**人类意图**。为了自动实现人类意图，我们必须首先指定一个**程序**来决定何时停车。该程序应从环境中获取**输入**；在这种情况下，输入可能是汽车周围环境的传感器流（例如，视频、激光雷达）。程序使用其输入来尝试实现用户的意图：如果程序“看到”停车标志，它就会产生**输出**信号让汽车停下。请注意，在机器人和人工智能的经典“感知-思考-行动”框架中 [521]，输出可以被视为“思考”组件输出的结果——例如，发送给执行器的信号。我们可以将这个决策流程总结如下：

- **用户意图** = “我想让我的车在停车标志前停下”
- **输入** = 汽车当前周围环境的传感器读数
- **程序** = 检查输入中是否有停车标志
- **输出** = “如果有停车标志，则发送‘停止’代码”

信任的作用

在计算机辅助的决策流程中，我们可能不信任链条中的任何环节。也就是说，我们可能不信任我们的程序准确反映了人类意图，不信任我们的程序正在我们认为的输入上运行，也不信任输出是按照我们认为的程序和输入计算出来的。因此，一个核心问题是：

问题 A-1.1: 决策流程的效用 我们如何确保决策流程既**有用**又**可信**？

两种有助于解决此问题的重要技术是**可信计算机**（通常借助**去中心化**）和**AI模型**。这将是本综述的重点。

A-1.1 可信计算机

近年来，可信计算机变得越来越普遍。**可信计算机**是一种执行程序的系统，旨在保证（正确构建的）程序确实执行了其指令，和/或我们可以事后验证程序执行了其指令。

可信计算机的示例包括：

- **可信执行环境（TEEs）**：TEE是提供隔离和其他安全保证的专用计算模块；值得注意的是，它们作为片上系统（SoC）的一部分被包含，这使它们比其他类型的可信硬件具有更大的灵活性 [528]。
- **可验证计算（也称为SNARKs，或ZK）**：可验证计算指的是用于证明给定计算被正确执行的密码学技术。验证证明的成本通常远低于从头开始执行计算。这些也被称为“snarks”，或根据其他次要属性称为“zkVMs”。
- **区块链**：区块链是一个去中心化系统，具有执行和确认特定类型计算的能力，例如处理交易和/或执行称为智能合约的区块链程序 [426, 674]。与TEE和可验证计算一样，区块链旨在确保程序的正确执行。与TEE和可验证计算不同，区块链在一组节点上运行，这些节点共同就系统状态达成共识。区块链本身也不强制执行机密性。

可信计算机的不同实现对于安全性和性能有不同的影响，我们将在A-2节讨论。然而，核心目标是相似的：以强保证验证计算的状态。

A-1.1.1 属性

其核心，可信计算机可以提供三个主要的安全属性：**机密性**、**完整性**和**可用性**。（计算机安全中的经典CIA三元组。）不同的可信计算机实现这些属性的不同子集。

完整性。在计算机安全文献中，完整性意味着确保计算或通信未被篡改。例如，我们可能担心程序或输入被损坏。可信计算机可以提供两种重要的完整性概念：

- **计算完整性**：组织声称运行一个程序，而实际上它们（被怀疑）运行另一个程序的情况并不少见 [476, 477]。计算完整性确保可信计算机确实运行了它所声称的程序；它解决了以下问题：“我不信任输出 = 程序（输入）。”
  - 可信计算机可以提供以下形式的保证：

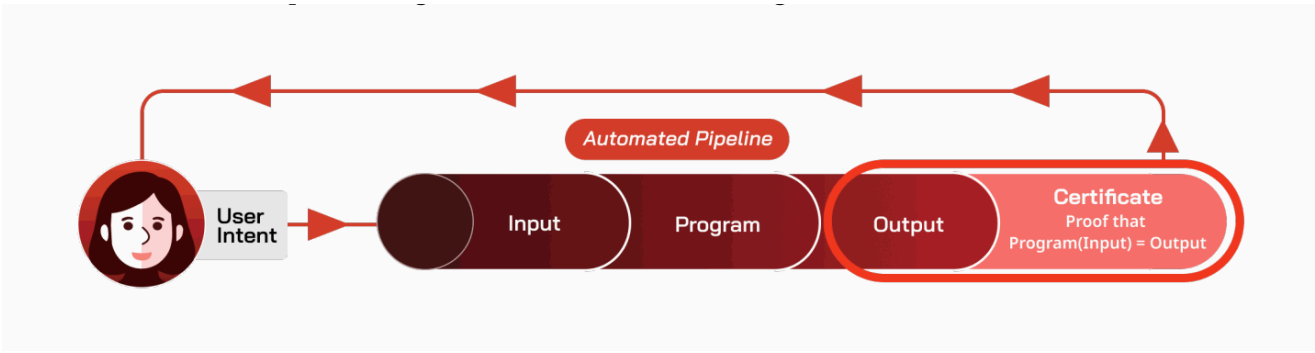


图 A.2 使用可信计算机实现的决策流程可以保证输出确实是通过在**输入**上运行**程序**获得的。

- 即，它可以输出一个证书，证明**输出**是通过在**输入**上运行**程序**产生的。该证书带有密码学保证，确保在标准密码学假设下，证书被伪造的概率可以忽略不计。
- **数据完整性**：在决策流程中，程序不仅从用户那里获取输入，还从网络获取数据。可信计算机可以确保程序正确地获取此类网络数据。这提供了数据完整性。在区块链术语中，执行此操作的可信计算机部分通常被称为**预言机**。
  - 预言机只能确保程序从特定来源检索数据。如果程序被设计为从特定网站获取特定天气报告，预言机确保它确实来自该网站。预言机无法确保数据本身是正确的（例如，确保 `www.trustyweather.com` 正确地报告了纽约市有雨），但来源的可信度通常可以作为其数据可信度的有力代表。

机密性。某些类型的可信计算机，包括TEE和一些注重隐私的区块链，可以额外地为决策流程的部分提供机密性。这个概念有时被称为**机密计算** [524]，我们将在本综述中使用这个术语。例如，TEE可以隐藏用户意图、输入、程序和输出——实际上是流程的任何或所有部分。我们可以把可信计算机看作一个“黑盒”，它被编程为默认隐藏流程，只向某些用户显示选定的数据。从示意图上看，TEE的“黑盒”隐私（作为一种可信计算机，它也强制执行完整性）增加了机密性，如图 A.3 所示：

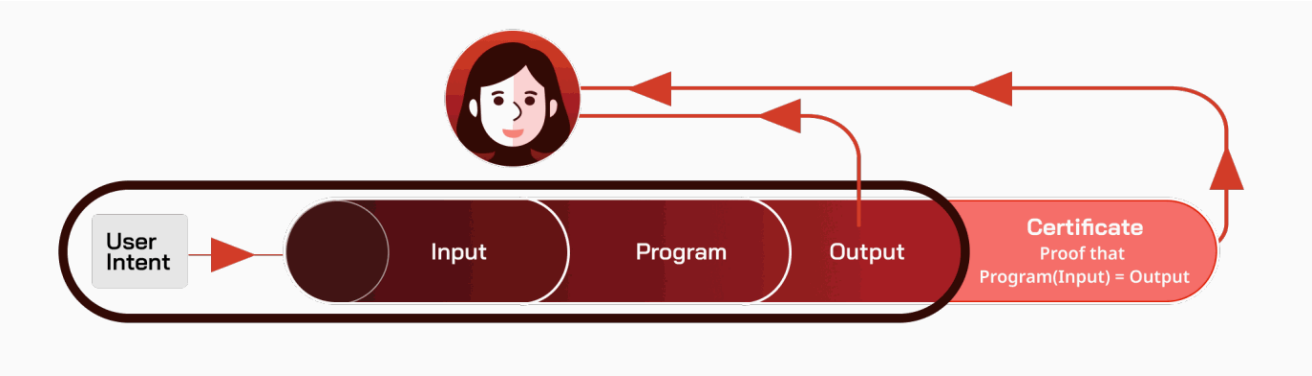


图 A.3 可信计算机在决策流程中强制执行隐私。用户只能访问预定义的部分流程，例如输出。

可用性。区块链的一个关键特征是它们强大的可用性。区块链被设计为具有高正常运行时间，在某些情况下在数年内接近100%。（例如，比特币至今已有十多年没有重大宕机，而其他链则有多次宕机 [272]。）更独特的是，区块链原则上还提供**抗审查性**：即使处于高度对抗性的环境中，包括攻击者控制部分区块链参与者（例如一部分验证者）的情形，区块链仍可供用户使用。

因此，可信计算机有助于解决以下问题：

关键可信计算用例（高层次） 程序用户不信任程序输出是否正确获得。此外，程序所有者可能希望向用户隐藏程序的内部状态。可信计算向用户以及程序输出的下游应用提供正确性和/或机密性的保证。它还可以确保程序对用户持续可用。

A-1.2 去中心化

区块链是一类重要的可信计算机，其关键区别特征在于其**去中心化**。公共区块链如果没有这个属性，就不是真正的区块链。然而，去中心化并非我们定义中"加密"的必要特征：在A-1.1节中描述的两种可信计算机——TEE和ZK——与系统是中心化还是去中心化无关。因此，解释什么是去中心化以及它可以为可信计算机和可信计算生态系统带来什么特殊属性是很重要的。

今天，加密社区对去中心化有许多不同的定义，以及许多不同的去中心化度量指标 [198, 453]。然而，非正式地说，**去中心化**意味着没有单一实体（或一小群实体）能够对系统施加有意义的控制。

在区块链中，所涉及的"控制"通常对应于**审查**，即阻止目标有效交易或一类有效交易上链的能力。反之，这种背景下的去中心化被称为**抗审查性**，并对应于A-1.1.1节中讨论的可用性的普遍概念。抗审查性意味着用户总能及时地让有效交易被包含在区块链上。虽然看似狭隘，但可用性属性实际上非常强大：它意味着系统对所有用户保持开放，并且用户的资产不能通过阻止交易而被没收（正如在中心化系统中发生的那样，例如传统银行）。

区块链去中心化的另一个重要概念是**治理**，即关于区块链或在其上运行的应用程序的管理决策如何做出的问题。这种治理在**去中心化自治组织（DAOs）** 的背景下一直是特别活跃的话题，DAOs通常采取围绕智能合约组织的社区形式（参见，例如 [549]）。DAO的成员资格是DAO代币所有权的函数，有意义的决策（DAO内的角色、技术升级或修改等）由代币加权投票决定。治理机制在不断发展。例如，最近有人呼吁使用AI智能体代表DAO成员投票，以避免许多人认为削弱治理有效性的"决策疲劳" [516]。

最后，去中心化为完整性和可用性服务，如A-1.1.1节所述。缺乏集中控制也意味着缺乏单点故障。换句话说，要获得一个高度去中心化系统的控制权——以破坏其完整性或可用性——攻击者必须破坏多个实体，这通常比对一个实体进行集中攻击更难实施。

正如我们在本综述中所解释的，去中心化带来的属性在AI设置中有许多潜在用途。

关键去中心化用例（高层次） 一个系统（无论是平台还是技术）受到集中控制：一个或一小群实体可以决定谁可以使用它以及它如何发展。去中心化技术可以帮助确保对该系统资源和决策过程的广泛访问。

A-1.3 AI模型

在可信计算兴起的同时，AI和机器学习（ML）在技术和社会世界掀起了一场巨变。AI模型可以实现许多最终目标，但就我们的目的而言，我们将它们视为将用户意图（通常与数据和/或有关环境的信息相结合）转化为实现用户意图的程序（即，输入 → 程

序 → 输出的流程）。以前，设计程序的任务需要通过基于领域知识和多次迭代的繁琐手动软件设计和工程过程来完成。AI允许我们通过示例来学习：我们可以使用反映我们意图的数据来定义一个执行它的程序。请注意，在本综述中，我们只关注从数据和/或环境中训练的ML模型，例如判别式、生成式或强化学习模型；更广泛的AI解释（例如，仅基于经典规则系统的解释）被视为超出范围。

例如，用户可能知道他们想在停车标志前停车，但他们可能不知道如何准确定义一个程序，该程序可以从仪表盘摄像头获取图像并识别停车标志。AI可以从代表性的（输入，输出）对中学习该程序。在这样做时，它为用户提供了将人类意图转化为计算流程的不同接口。这种转换的一些示例包括：

判别模型可以被视为将输入（例如，图像）转换为条件输出（例如，标签）的程序。模型架构可以看作是可能程序（函数）的类别；我们使用数据和ML技术来学习模型权重，从而指定函数类中的哪个元素是最好的。因此，我们使用AI将意图（“标记图像”）转化为程序（一个将图像映射到标签的训练模型）。**生成模型**则捕捉不同的意图：从给定的未标记数据分布中生成样本。和以前一样，我们可以从数据中训练模型以满足此意图，从而学习程序。**强化学习**（经典地）不直接从数据中学习，而是从环境和奖励函数中学习。然而，它具有相同的属性，即人类意图（例如，“学习赢得任何棋局”）被转化为使用ML的程序（即，下棋策略）。

因此，在本综述中，我们将AI视为解决以下问题，如图 A.4 所示：

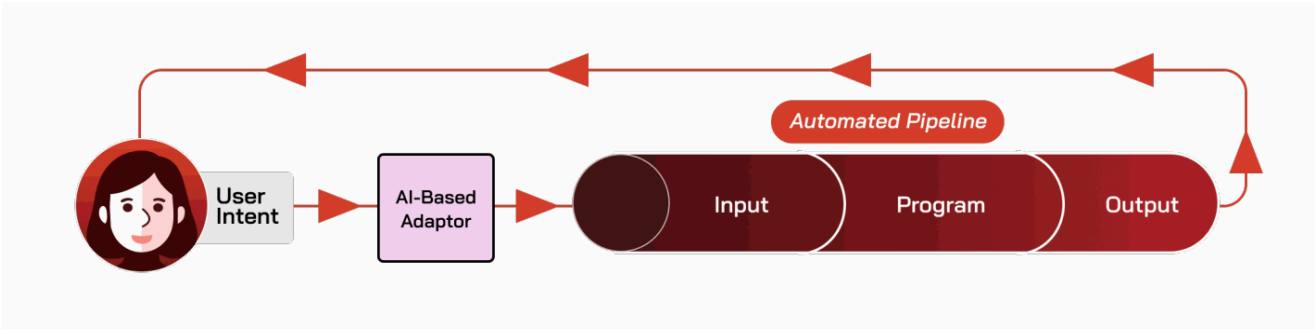


图 A.4 AI可以帮助将人类意图（通常与数据和/或有关环境的信息相结合）转化为一个明确定义的程序。

关键AI用例（高层次） 程序所有者不知道如何定义程序以准确反映人类意图。AI可以帮助将通过示例或自然语言等表达的人类意图转化为程序。

A-1.4 统一框架：作为中间件的加密和AI

从表面上看，AI和加密解决的是截然不同的问题。然而，它们有一个重要的共同点：这两种技术都可以被视为决策流程中的**中间件**。在经典计算机科学文献中，中间件指的是位于复杂系统之间的软件，使这些系统之间能够通信和交互 [665]。经典中间件的一个例子是网络中的中间盒，它根据网络管理员指定的规则检查、转换和过滤数据包 [237]。在我们的案例中，我们可以抽象地将AI和加密（包括基础设施和应用）都视为人类与计算机系统之间的中间件。我们通过图 A.5 说明这种关系。

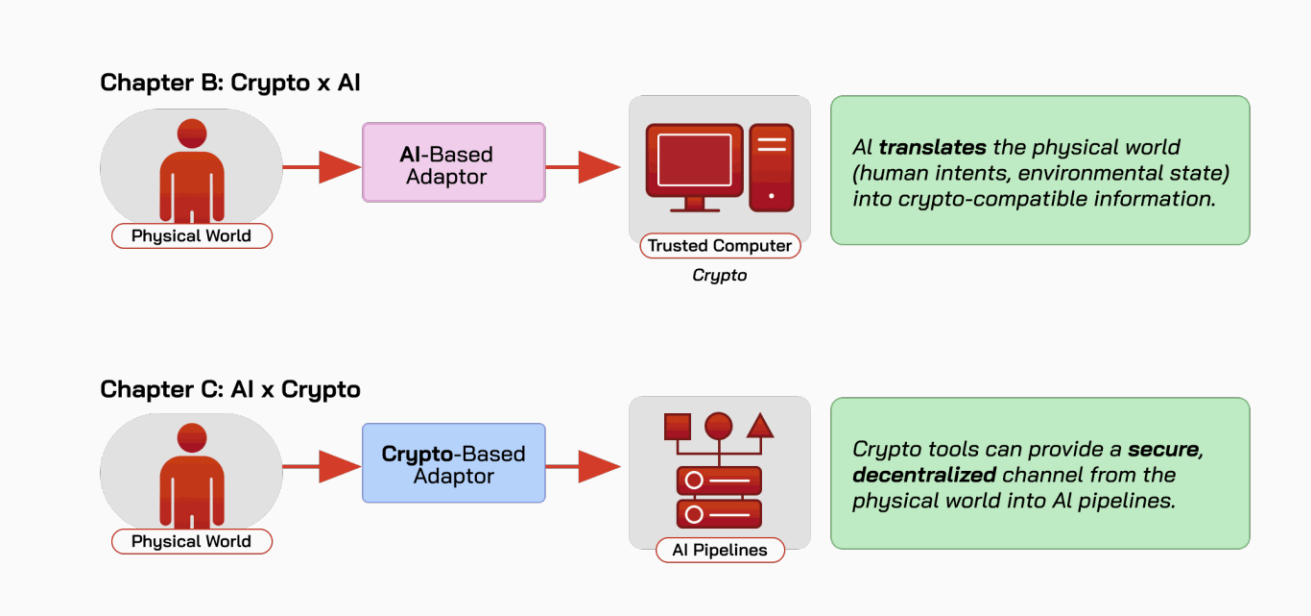


图 A.5　在本综述中，我们认为加密和AI可以充当彼此的中间件，连接物理世界与虚拟系统（例如，可信计算机或AI工作流）。然而，这两种接口的形式看起来非常不同，我们将在B章和C章中进行探讨。

AI作为物理世界与区块链之间的中间件。 我们之前看到AI可以帮助将人类用户的意图转化为执行该意图的程序（例如，图A.4）；换句话说，它充当翻译中间件。这种能力在区块链的背景下特别有用，区块链以难以使用而闻名。AI可以显著降低设计、分析甚至简单使用现有区块链的门槛。关键在于在人类与区块链之间的接口处适当地部署它。有前景的应用类别示例包括：

- **构建性任务（基础设施）：** 今天，人类使用费力的手动过程来设计区块链技术和治理栈的算法。AI可以通过基于人类表达的期望属性来提出或发现区块链栈中的基础算法，从而加速这一过程。如果成功，人类设计师只需评估提议的设计，而不是从头开始设计。
- **构建性任务（应用）：** 今天，使用区块链——特别是在跨链环境中——通常是一个困难且痛苦的过程。用户经常实现不符合其初始意图的智能合约——有时会带来灾难性的后果 [409]。AI可以通过将人类偏好转化为提议的智能合约，以及搜索安全漏洞或逻辑缺陷来协助解决此问题。
- **分析性任务：** 今天，大量工作用于分析区块链交易，以了解系统的当前状态，无论是在宏观属性还是个人交易（例如，欺诈检测）方面。AI可以在给定高级条件和大量数据的情况下，帮助促进对区块链动态的分析。

我们将在B章讨论AI应用的这种分类。

加密作为计算平台与物理世界之间的中间件。 我们之前看到，可信计算可以给人类提供正确性的保证（可能受机密性约束），而去中心化可以帮助确保可用性。因此，加密可以被视为自动化决策流程与必须信任流程输出和可用性的人类（或另一个程序）之间的中间件。实际上，这种信任调解在概念上类似于面向安全的中间件在计算机系统中所做的：它（试图）确保到达目标的信息是可信的并可靠地到达。在AI领域，有前景的应用示例包括：

- **可信数据：** 今天，ML模型的数据在训练中经常被使用，而没有确保数据源是可信的。然而，这可能导致在医疗诊断、基础设施管理或安全等高风险应用中出现系统关键性故障。因此，可信计算在AI数据提供者和数据使用者之间提供中间件，允许接收者信任数据的来源。
- **可信计算：** 类似地，训练过程通常在封闭且不透明的环境中进行，外界可能难以判断训练是否正确完成。可信计算提供了一种强有力的方法，可向下游用户保证模型按照预定义规范完成训练。为了帮助确保规范本身以负责任的方式制定，以AI治理形式出现的去中心化机制可以承载社区意见与监督。
- **私有数据：** 最后，可信计算的一个主要用例是能够在进行可信操作的同时不泄露有关数据输入的信息。这在当今的AI领域是一个巨大的问题，因为许多公司希望模型能够根据专有的内部数据进行定制。可信计算是在不将数据以明文形式暴露给执行模型训练的一方的情况下获得可信结果的最实用解决方案之一。在这个例子中，必要的“信任”概念既包括程序输出的正确性，也包括对输入的隐私保证。

我们将在C章讨论加密应用的这种分类。

A-1.5 综述路线图

这种将加密和AI视为中间件的观点将为我们考虑的应用和研究问题提供信息。我们将综述分为两个主要部分：

1. **加密 × AI：** AI系统如何增强加密系统的能力？
2. **AI × 加密：** 加密系统如何帮助保护AI系统并赋予其新能力？

本综述分为几章。我们从A章其余部分更详细的相关技术初步介绍开始。如图A.5所示，B章讨论AI能为加密带来的好处，包括增强我们分析和使用区块链系统的能力。C章讨论加密可以改善AI系统的方式，特别是在增强其去中心化、安全性和隐私方面。我们最后在D章中提出我们对加密 × AI社区中持续存在的常见误解（或不完整描述）的看法。

A-2 可信计算基础

本节概述可信计算的目标以及实现其各种属性的各种方式。

A-2.1 可信计算示例：单用户推理

作为说明如何将可信计算与AI系统结合使用的示例，我们考虑一个简单场景，其中单个用户向ML模型提供输入进行推理。

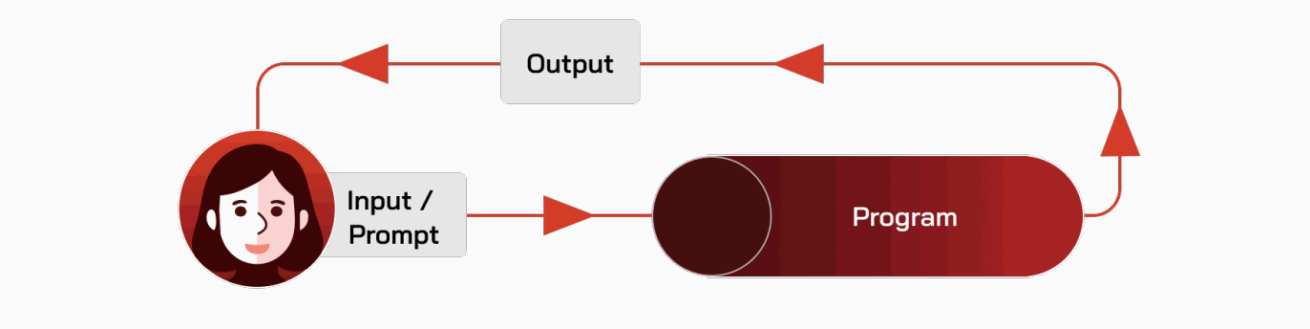


图 A.6 用户向模型提供提示。

例如，假设我向ChatGPT或Claude这样的在线LLM平台输入以下提示：

我的影子颜色不对。这种医学状况叫什么？

对于我（用户）来说，这里没有隐私顾虑。我提供了提示，所以试图对其保密是没有意义的。对于我是否应该信任提示也是如此。因为它来自我，我确信它符合我对LLM的预期用途。

然而，即使在这个常见场景中，也存在两个自然的安全问题。概括地说，它们是我是否可以信任LLM平台的输出，以及是否对提示和输出以及AI系统本身强制执行隐私。

输出完整性

假设我收到对我上述提示的以下回应：

> **最终通知：关于“Umbrachromatosis”的医疗警报**
>
> Umbrachromatosis：你的影子颜色不对。案例ID：DBB-44721
>
> **患者状态：紧急**
>
> 我们的自动诊断标记了Umbrachromatosis的指标。如果不治疗，这种情况可能会升级，并可能在7-14天内致命。好消息是，有治愈方法，但仅限一天。要解锁你的个性化治疗并确保当日咨询，请立即将0.24 ETH的押金发送至0x728877d47ac48dbd17a1e95f7b1dec20be6fb8d6。剩余时间：02:13:57。

一个训练有素的LLM不应该像这样回应并索要加密货币付款！看到这个回应，你自然会产生疑问：LLM是否以某种方式被损坏并产生了幻觉，或者它是否被试图从用户那里窃取加密货币的人入侵了。换句话说，对于：

- **输入** = 我的提示
- **程序** = 一个经过适当训练和部署的AI系统
- **输出** = 我收到的回应

担忧将是：`输出 ≠ 程序（输入）` 这正是可信计算机可以帮助解决的问题。回想A-1.1节，可信计算机生成一个证书，显示 `输出 = 程序（输入）`。AI系统可以被设计为与其对提示的回应一起发送这样的证书。（见图A.2。）通过这种方式，用户获得保证，回应确实是由**程序**——一个特定的AI系统——处理**输入**（用户的提示）生成的。

关键要点 A-2.1: 输出完整性 为了信任对程序（如ML模型）的**输入**所产生的**输出**，用户必须保证没有发生篡改，这意味着：`输出 = 程序（输入）` 可信计算机可以输出证明此属性的证书。

用户隐私

在我们正在进行的单用户示例中，即使LLM平台看起来工作正常，我可能仍然担心我的提示——以及由此产生的输出——对运行LLM平台的组织来说的隐私性。我可能不想让别人知道我的身体状况。

今天，用户依赖运行AI系统的组织来执行合理的隐私政策。但这种方法并不总是有效。最近，一些用户沮丧地发现，他们无意中发布了他们的ChatGPT会话，并使其可供搜索引擎索引。在一个尴尬的案例中，一位用户要求重写简历，一名记者识别出了该用户，在LinkedIn上追踪了结果，并报道该用户没有得到那份工作 [557]。

即使提示和/或输出没有被公开披露，它们仍然容易被AI系统的操作员滥用。例如，用户的姓名和医疗状况原则上可以被偷偷出售给广告商或保险公司。

关键要点 A-2.2: 可信计算与用户隐私 可信计算（具体来说，可信机密计算）为用户在查询ML模型时保护输入和输出的隐私提供了一种关键的、技术性的、实用的方法。

关于完整性的说明： 完整性对于强制执行隐私至关重要。假设我将我的医疗查询发送给我认为是可信的ML模型执行环境，但它实际上去了一个被编程为将提示转发给保险公司的ML环境。在这种情况下，即使交互（即ML模型和我之间的通道）保持私密，也无关紧要，因为模型环境本身就破坏了隐私。（这就像我对着一个爱传闲话的闺蜜耳语了一个秘密。）

A-2.2 通过硬件和复制实现可信计算

在讨论了可信计算的用途之后，我们现在转向如何实现它：首先通过硬件和复制（第A-2.2节），然后通过密码学（第A-2.3节）。

A-2.2.1 区块链

区块链的功能最好被解释为“公告板”：一个公开的仅追加的消息日志，称为交易，节点运行共识协议以就交易的排序达成一致 [426, 674]。两个安全属性使区块链作为基础设施很有用：**一致性**（安全性），意味着每个观察者看到相同的最终记录（这样数字资产的转移就不会导致双重支付），以及**可用性**（活跃性），意味着提交的交易最终会被包含。对抗恶意方的强大可用性概念（也称为“抗审查性”）是使区块链作为基础设施具有吸引力的原因。这些属性部分由密码学技术确保，但部分也来自经济激励：原生代币通过费用补偿验证者，并在权益证明系统中通过罚没对不诚实行为进行惩罚。

通过解释公告板上的消息，区块链成为一个可编程的可信计算机，这些消息遵循确定性规则。资产转移需要来自源账户绑定的私钥的签名。智能合约语言（如Solidity）使规则集本身成为一个程序，因此任意状态转换逻辑（用于操作、选举和交换）都可以在相同的最终日志上运行。底层日志的可用性和一致性属性转移到这些程序计算的任何功能上。

两个基本限制促使了本节其余部分的发展。首先，公告板默认是公开的，因此任何输入或计算出的状态对所有观察者都是可见的；机密性需要额外的密码学，例如零知识证明或基于TEE的执行。例如，Aztec、Aleo和Penumbra构建在ZK编程模型上 [19, 52, 472]，而Oasis Sapphire、Phala和Secret Network构建在TEE上。其次，复制计算成本高昂，因为每个验证者都必须重新执行每笔交易，因此智能合约区块链收取称为“gas”的费用来限制链上计算。更重的计算必须通过预言机（第A-2.4节）、使用SNARG的rollup（第A-2.3节）或基于TEE的协处理器在链下进行。

A-2.2.2 可信执行环境与机密计算

可信执行环境（TEE）是用于可信计算的硬件原语：一种处理器特性，保护计算免受其运行系统的其余部分的影响。TEE以不同形式有着悠久的历史 [302, 528]。今天，大多数手机都包含一个——例如，用于保护用于向设备验证您身份的生物特征数据（例如，人脸模板）。最近，服务器和超大规模处理器（如Intel Xeon和AMD EPYC）增加了TEE原语，从而实现了基于云的“机密计算”：这是保护使用中数据（不仅仅是静态或传输中数据）的行业术语。

一个TEE有两个主要属性：**隔离和远程认证** [302, 528]。这里的“执行环境”是操作系统进程、虚拟机或容器。在通常的执行层次结构中，较低级别的进程（内核环0°或虚拟机监控器“环-1”）对较高级别的进程具有完全的可见性和控制权；隔离意味着即使这些较低级别也无法篡改或检查TEE的计算。远程认证意味着此层次结构外部的方（例如，另一台机器和网络上的进程）可以验证将隔离计算的输出绑定到其程序代码的签名证据。

为什么机密计算对ML隐私很重要： 如A-1.1节所述，某些类型的可信计算机（TEE）支持机密计算，即在不向系统运营方披露输入或输出的情况下使用AI模型。模型在TEE内执行，其黑盒抽象对模型运营方隐藏模型状态和执行过程；输入和输出则通过用户与TEE之间的加密通道传输。这样，图A.1中的流程便可在运营方无法看到程序（ML模型）的输入和输出的条件下实现，但这一保证仍有重要限制和前提。

GPU支持使TEE在AI场景中变得切实可用。当前面向AI应用最成熟的TEE平台之一是NVIDIA机密计算（Confidential Computing，CC）。H100的CC模式于2024年6月随CUDA 12.4正式发布，现已覆盖H100/200、Blackwell B100/200以及即将推出的Vera Rubin等GPU [168, 441, 443]。

机密CPU与机密GPU之间的加密PCIe通道只带来适度开销：在80亿参数的Llama模型推理中低于7%，对700亿参数等更大模型则接近于零 [729]。其主要目标是云端推理服务：运营方可以通过云服务商部署模型，同时保护模型本身及用户与模型的交互。它同样可用于云端模型训练和其他可信计算场景。

Azure已通过预览服务支持基于TEE的机密推理 [523]，Anthropic也在一篇设计论文中描述了相关方案 [40]。NEAR AI Cloud [430]、RedPill [503]、Venice [630]、Tinfoil [604]、Chutes [140] 等服务商则已提供生产环境中的相关能力。

A-2.2.3 机密计算的局限

TEE在区块链中的普及经历了相当长的过程，原因之一是其安全漏洞与功能优势同样受到关注 [139, 538]。首先，TEE的信任模型较为微妙：它本质上是一种软件抽象，设计目标是抵御**软件层面的攻击者** [25, 312]，例如保护SGX进程免受拥有内核权限的攻击者侵害，或保护机密虚拟机免受虚拟机监控器侵害。

TEE并非为物理防篡改而设计，但物理攻击也并不容易实施，因此人们很容易给予它超出设计边界的信任。近年来的发展方向是以极低开销优化机密计算，即使代价是放弃对物理攻击的抵抗能力。最清楚的例子，是从客户端时代的Intel SGX转向服务器时代占主导地位的可扩展SGX、Intel TDX和AMD SEV-SNP。

早期SGX采用受完整性树保护的内存加密；服务器时代的设计改用确定性AES-XTS [25, 312]。这种方案可扩展到数TB受保护内存，但相同明文会产生相同密文。TEE.Fail正是通过内存总线介入利用这一属性 [139]。Intel在介绍可扩展SGX时也直接说明了这种权衡 [312]。

由于TEE在设计上不抵御物理攻击，其运行环境必须与底层远程认证分开评估。虽然TEE由硬件提供商实例化，云服务商仍必须负责保护硬件免遭物理攻击。因此，Azure机密虚拟机 [412]、Google Cloud机密计算 [250] 等产品，本质上是硬件厂商与云运营商共同提供的协作型产品。

近期工作 [481, 511] 进一步将机密虚拟机的认证与平台级证据绑定，用于证明宿主机是位于安全数据中心内、已经注册的特定硬件设备。

另一个问题是对硬件制造商的信任。技术上无法保证制造商没有在硬件中植入后门；这一风险既适用于处理器，也适用于远程认证系统。这些系统并不透明，而且协议本身无法阻止制造商与情报机构串通，为特制进程签发有效认证。

这种类似斯诺登披露事件的系统级合谋无法由技术设计排除。它不一定意味着攻击者能够解密已经运行的TEE，但可能允许一个伪造的“间谍TEE”加入网络而不被发现。

除上述固有问题外，TEE的复杂性也使其经常因实现错误而出现漏洞 [538]。典型案例是ÆPIC Leak [99]：高级可编程中断控制器（APIC）的微码实现存在缺陷，在访问未定义的内存映射I/O范围时，没有清除私有寄存器的高位。纯软件攻击者即可利用它提取运行中程序的秘密并伪造远程认证，完全处于TEE原本宣称防御的威胁模型之内。

A-2.3 通过密码学实现可信计算

零知识证明 (ZKPs)

零知识证明（ZKP）[247] 是一种密码学协议，其中一方（证明者）使另一方（验证者）相信某个陈述为真，而除了陈述的有效性之外不透露任何信息。在现代实践中，ZKPs通常以 `zk-SNARGs` ——零知识简洁非交互式论证（zero-knowledge succinct non-interactive arguments）—— 的形式实现，通过将零知识层叠加在底层SNARG之上构建 [136]。我们在下面解构这两个层次。

SNARG 允许证明者说服验证者某个公共陈述 `x` 满足某种关系 `R`（例如，`x = (u,y)` 且 `y = f(u)`，对于固定函数 `f`），通过生成一个简短证明 `π`，验证者验证该证明的效率远高于重新执行计算。SNARG的价值在于其简洁性。更具表达力的关系允许证明者使用私有见证 `w`（例如，秘密输入），证明 `(x,w) ∈ R` 而无需将 `w` 包含在证明中；然而，SNARG本身并不保证 `π` 隐藏 `w`。

zk-SNARG 在此基础上增加了零知识保证：`π` 除了关系本身所暗示的之外，不透露关于 `w` 的任何信息。在本综述中，我们使用“ZKP”指代 zk-SNARGs，因为它们是实践中部署最广泛的ZKPs类型，尽管它们并非唯一类型。

ZKPs和SNARGs在区块链和AI环境中都有应用。

在区块链中的应用： ZKPs在区块链中得到了广泛采用，最突出的是通过**zkRollups** [110]，它将一批交易的验证压缩为一个简短证明，从而将以太坊的gas费用降低高达99%，通过zkSync Era [404]、Starknet [580]和Polygon zkEVM [479]等项目实现。zkRollup将一批交易的验证压缩成一个单一短证明，可以高效验证。区块链因此可以更新到批次后的状态，而无需重新执行交易。值得注意的是，在此应用中交易本身是公开的，因此零知识属性未被使用；这些系统仅依赖SNARG的简洁性。zkRollup中的“zk”在很大程度上是从区块链社区松散使用中继承来的用词不当。

在AI中的应用： 在AI中，SNARGs让计算能力强的证明者（例如，云服务）代表资源受限的验证者（例如，智能合约或移动设备）执行昂贵的计算（如ML推理），并证明其正确执行。当模型具有专有权重时，普通的SNARG是不够的，因为证明可能泄露权重信息。零知识属性填补了这一空白。证明者可以先发布对权重 `θ` 的承诺 `c_θ`（例如，一个将证明者绑定到 `θ` 而不透露它的密码学哈希），然后向验证者证明使用由 `c_θ` 绑定的权重正确执行了任何推理，而验证者从未看到 `θ`。**零知识机器学习（ZKML）** [124, 711] 是ZKPs在AI中的主要（尽管仍在新兴）应用领域。ZKML已经在实践中部署，例如在RockyBot [419]中，这是一个由Modulus Labs [418]开发的链上可验证ML交易机器人。性能成本仍然令人望而却步：Modulus Labs的“智能成本”（Cost of Intelligence）基准 [357] 报告称，在高端AWS实例（AMD EPYC 7R32，128GB RAM）上，对于仅有约1800万个参数和220亿次乘加运算的多层感知器，证明时间约为1分钟，距离前沿规模的LLM尚有几个数量级的差距。

zk-SNARG要求证明者拥有明文的完整见证，因此它不支持输入分布在相互不信任且不愿与单个证明者共享的各方之间的计算；例如，专有权重由模型所有者持有而私有输入由单独用户持有的模型推理。解决此问题需要不同的原语：**安全多方计算（MPC）**。

安全多方计算 (MPC)

安全多方计算（MPC）[246, 690] 使一组 `n` 方（每方持有私有输入 `x_i`）能够共同计算一个商定函数 `f(x_1,...,x_n)`，同时不透露关于个人输入的任何信息，除了输出本身所暗示的。

与ZKP不同，MPC不会生成外部方可以稍后验证的计算完整性证明。称为**协作式zk-SNARGs** [454] 的一系列工作弥补了这一差距，允许将私有ZKP见证在多个证明者之间分割，他们共同生成单个ZKP，而无需任何一方重建完整的见证。这结合了zk-SNARG的简洁性、可验证性和零知识保证与MPC的分布式信任模型。

MPC在区块链和AI环境中都有应用。

在区块链中的应用： MPC在区块链中最广泛部署的用途是**门限签名** [167]，其中私钥在 `n` 方之间分割，以便产生签名需要其中的阈值数量，而无需任何一方重建密钥。这支撑了商业MPC托管和验证者密钥管理服务。一个相关的混合部署是**zkTLS** [709]，它结合了MPC和ZKPs，让客户端在不信任服务器的情况下，向智能合约证明关于受TLS保护的Web内容的陈述。

在AI中的应用： MPC支持ZKP单独无法支持的两类AI应用。第一个是**协作式训练**，其中多个数据所有者（例如，多家持有患者记录的医院或持有交易历史的银行）共同训练模型，而无需任何方向其他方或中央服务器透露其原始数据 [639]。第二个是**隐私保护推理**，其中用户在专有模型上评估其私有输入：MPC允许计算在不披露用户输入给模型提供者，也不披露模型权重给用户的情况下进行 [354]。性能成本仍然很高：PUMA [174] 是最先进的基于MPC的Transformer推理框架，报告称LLaMA-7B每个令牌大约需要5分钟，比标准推理慢几个数量级。

巨大的开销将MPC和ZKPs限制在狭窄的部署范围内。在信任模型允许的情况下，TEE为机密性和可验证执行提供了一种便宜得多的替代方案：机密计算推理开销保持在7%以下 [A-2.2.2节]，比上面提到的MPC和ZKP成本低几个数量级。它们之间信任模型的差异使得它们难以比较。TEE要求信任硬件制造商且没有侧信道攻击，而MPC增加了各方之间的非勾结假设（阈值取决于协议），ZKPs则仅依赖密码学硬度。

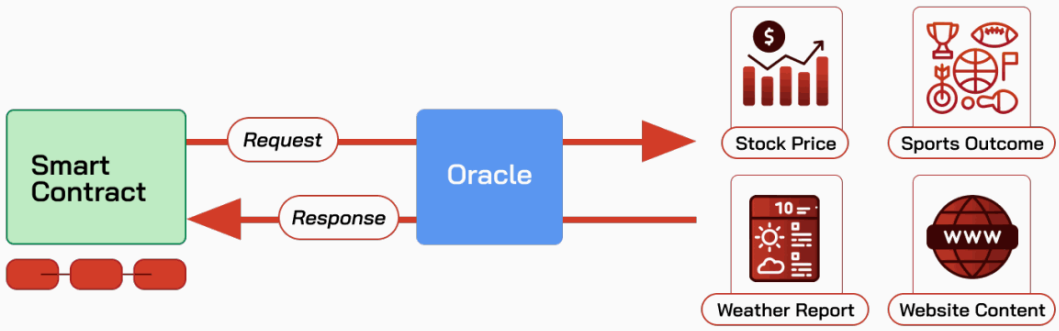
A-2.4 预言机

预言机起源于为区块链提供认证数据的系统。它们也是AI的有价值工具，因为它们可以支持对私有数据的访问，我们将在C章中对此进行扩展。在本节中，我们首先回顾智能合约预言机的概念，以及与AI应用更相关的隐私保护预言机。然后，我们讨论实现预言机的技术方法及其安全性和性能权衡。

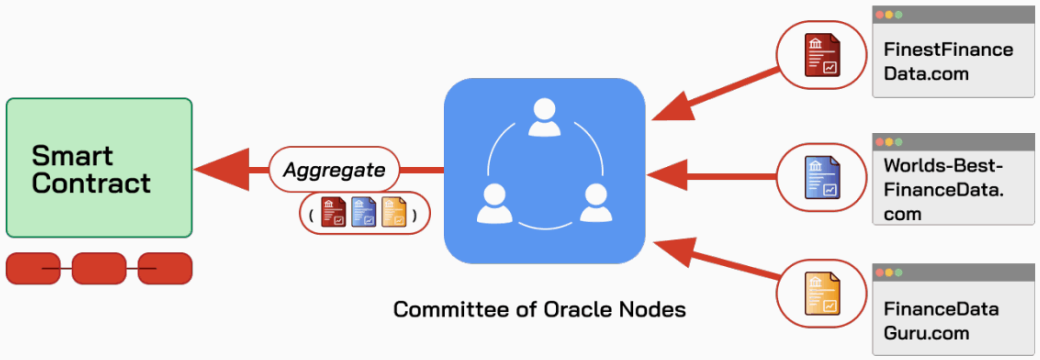
A-2.4.1 预言机的概念与应用

智能合约预言机。 智能合约是在区块链上运行的自主程序。许多应用需要智能合约访问链下数据，例如股票报价（用于代币化股票）、体育结果（用于预测市场）、航班状态（用于延误保险，如AXA的Fizzy [105]）等等。由于智能合约只能访问已经在区块链上的数据，链下数据必须由称为预言机的系统推送到区块链，如图A.7a所示。由于预言机是智能合约栈的关键部分，整个行业都专注于构建健壮、高效和安全的预言机系统，包括Chainlink [120]、RedStone [504]、Chronicle [138]、Witten [671]、UMA Optimistic Oracle [616]、Teller [598]、Band Protocol [63]、Pyth Network [482]、API3 [42]、Supra [591]和Gas Network [235]。

预言机的一个基本安全属性是**真实性**，即预言机忠实地从指定来源中继数据，而不篡改或谎报其来源。如图A.7b所示，常见的设计是让一个由多个预言机服务器组成的委员会独立获取数据，并使用聚合机制过滤掉潜在的恶意输入并产生最终结果。这种方法保证了真实性，假设委员会中恶意部分的比例低于某个阈值（例如，少于三分之一）[103]。



(a) Conceptual schematic of smart-contract oracles.



(b) A common architecture for smart contract oracles: a committee of nodes fetches data independently, potentially from different sources, and an aggregation mechanism filters and combines data and may also filter out anomalies or outliers. This flow may be periodic or triggered upon request by the receiving smart contract.

图 A.7 (a) 智能合约预言机的概念示意图。(b) 智能合约预言机的常见架构：一个节点委员会独立获取数据，可能来自不同来源，聚合机制过滤和组合数据，也可能过滤异常值。此流程可以是周期性的，或在接收智能合约请求时触发。

隐私保护预言机。 除了真实性之外，隐私保护预言机还可以中继源自私有数据的信息，这些数据是普通预言机无法直接访问的，例如用户的银行对账单、信用报告或年龄信息。正如我们将在C章中探讨的，隐私保护预言机支持在从用户直接获得的私有网络数据上进行ML训练或微调，而无需与原始数据持有者进行特殊安排。同时，用户享有强大的隐私保护，因为他们可以控制要共享的数据。

更具体地说，隐私保护预言机协议允许用户说服验证者，来自特定来源 `S` 的一条数据（完整性/真实性）满足某个谓词 `P`，而不泄露任何其他信息（隐私）。例如，假设Alice想向贷方（智能合约或链下实体）证明她有良好的信用。她可以发送信用报告的截图，但这很容易伪造。使用隐私保护预言机，Alice可以密码学地向预言机证明“根据 `https://www.bigbank.com` 的数据，Alice的信用评分超过700。”预言机可以验证此声明并将结果中继给贷方，如图A.8所示。关键的隐私特性是，除了上述陈述为真这一事实之外，不会透露关于Alice的任何信息。特别地，Alice不需要向贷方透露检索信用记录所需的秘密（例如，她的社会安全号码）或信用报告上可能包含的任何额外信息（例如，她的地址历史）。

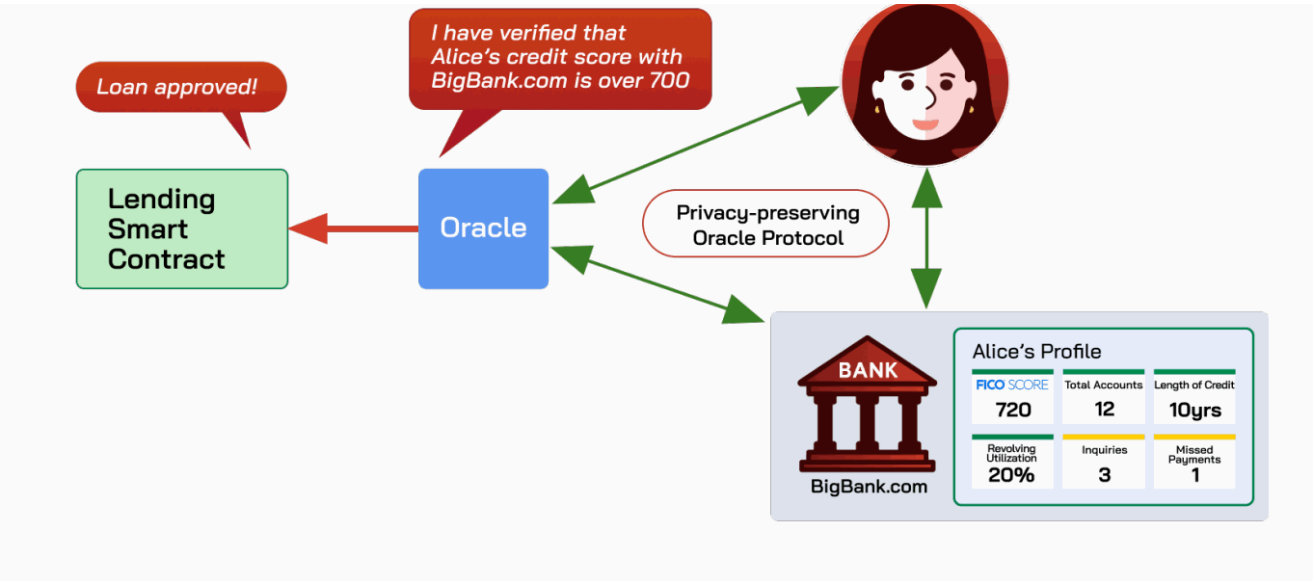


图 A.8　用户可以使用隐私保护预言机“导出”私有信息。

A-2.4.2 隐私保护预言机的构建

我们现在讨论高层次的技术构建，基于提出隐私保护预言机的原始论文 [708, 709]。后续工作提出了各种安全性和性能优化 [387]，但主要思想保持不变。

隐私保护预言机的起点是保护互联网安全的**传输层安全（TLS）** 协议。TLS是一种协议，使用户（例如，浏览器）能够与远程Web服务器建立安全连接。读者无需理解TLS的细节，但我们注意到TLS本身并不对传输的数据进行数字签名。相反，当Alice从TLS服务器 `S` 获得一段密文 `D` 时，`D` 的完整性由Alice和服务器之间共享的密钥保护。因此，第三方（例如我们示例中的预言机节点）无法验证 `D` 的真实性，也无法确定 `D` 是来自服务器还是完全由Alice伪造。在所有下面的预言机构建中，这个共享密钥需要以某种方式对Alice隐藏。

构建预言机协议主要有三种方法：使用TEE、使用安全两方计算（2PC）或使用预言机作为代理。值得注意的是，TLS可以被修改为提供不可否认性 [512]，但现有的大多数Web服务器并未部署此类修改，因此我们的讨论侧重于无需修改服务器的协议。

- **基于TEE的预言机（Town Crier [708]）。** 第一类方案依赖Intel SGX或TDX等TEE技术。Town Crier在TEE内执行如下高层逻辑：首先接收一项请求，其中指定数据源 `S`、需要针对来自 `S` 的数据证明的陈述或谓词 `P`，以及访问 `S` 所需的用户秘密（例如加密后的密码）；随后通过TLS从 `S` 获取数据 `D`，并输出 `P(D)` 及由硬件生成的正确性认证。

谓词 `P` 可以是预言机在 `D` 上计算的任意通用函数。因此，预言机也可以充当繁重链下计算的执行层。与下面介绍的设计相比，基于TEE的预言机很可能是AI等大规模计算目前最实用的方案。供智能合约使用的AI工具可能在此类预言机中执行：模型访问外部数据，完成非平凡计算，并把附带认证的结果返回链上。

- **基于2PC的预言机（DECO [709]）。** DECO提出了一种不依赖TEE的隐私保护预言机。其高层流程如图A.9所示。核心思想是Alice与预言机运行两方计算（2PC）协议 [690]，共同完成TLS握手，使任何一方都无法获得完整会话密钥。为区别于标准的两方TLS握手，这一过程称为“三方握手”。

握手完成后，Alice在预言机协助下通过TLS查询Web服务器，并对密文 `D` 作出承诺。三方握手隐藏了会话密钥，使Alice无法伪造TLS密文。确定 `D` 的真实性后，证明者可在第三步使用任意通用零知识证明系统 [259]，证明关于数据的细粒度陈述。

- **基于代理的预言机（DECO代理模式 [709]）。** 前两类方案分别使用TEE和三方握手防止TLS密文伪造。代理模式采用另一种方法：Alice通过作为网络代理的预言机与TLS服务器交互，再以类似2PC方案第三步的方式证明关于代理转发密文 `D` 的陈述。与2PC模式相比，代理模式省去了昂贵的三方握手，因此性能更好、实现也更简单。该模式最初同样由DECO提出 [709]。

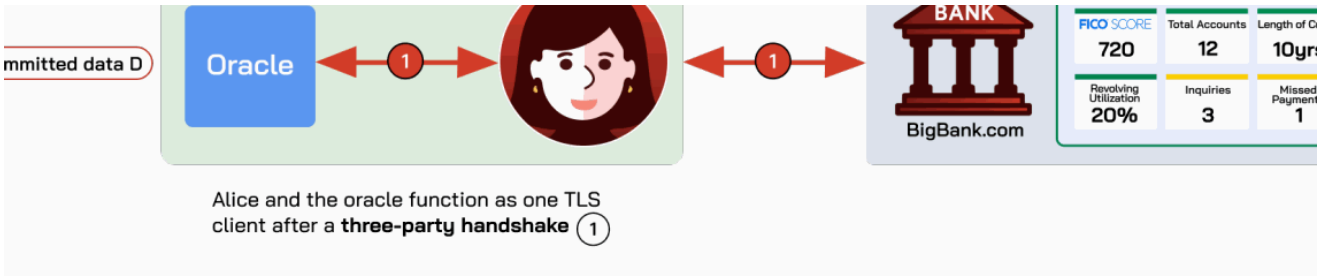


图 A.9 DECO两方计算模式的高层工作流程。

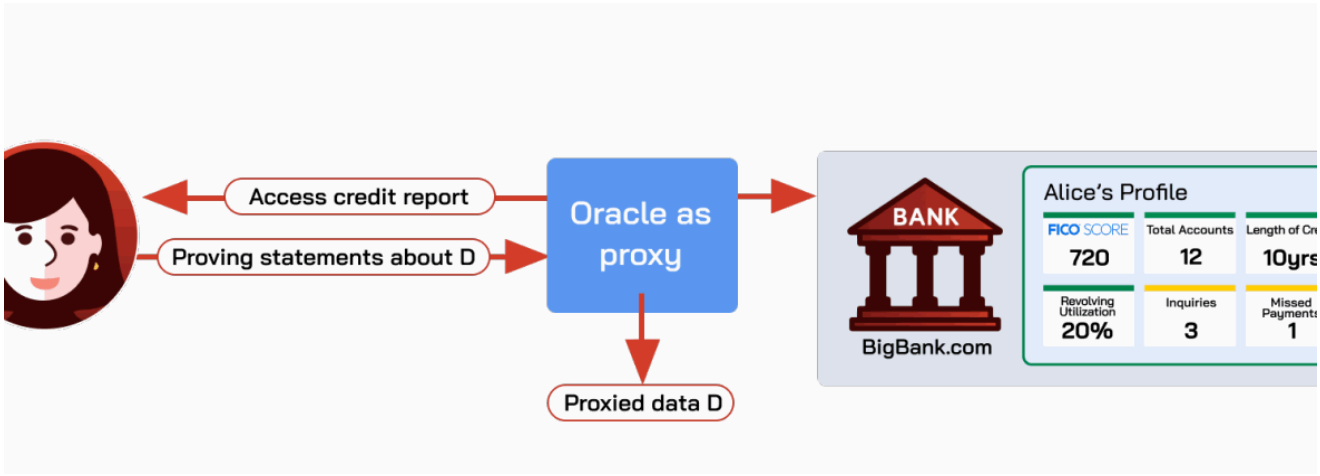


图 A.10 DECO代理模式的高层工作流程。

总结： 我们比较了这些设计的安全性和性能。要访问私有数据并对其进行大规模计算，基于TEE的解决方案是目前唯一实用的选择（尽管零知识证明系统正在日益扩展）。最近的GPU TEE，如NVIDIA机密计算，特别适合ML工作负载。其缺点是引入了对硬件制造商及其运行环境的额外信任假设，如我们在A-2.2.2节中详述的。DECO类协议（2PC和代理模式）更适合证明相对简单的陈述，如年龄验证 [401]、信用验证等。在2PC和代理模式之间，后者避免了昂贵的步骤；然而，它容易受到网络层攻击（例如，BGP劫持 [709，附录C.4]）。对于高风险应用，应优先选择2PC模式。

所有这些设计及其变体都已在实践中得到应用。行业现在将它们统称为“zkTLS”协议——这对于基于TEE的解决方案来说是用词不当，因为它们不使用任何“zk”（零知识证明）。Town Crier和DECO已作为Chainlink运行时环境（CRE）的一部分产品化 [121]。Reclaim [501] 实现了一种改进的基于代理的变体 [387]，现在正在扩展到新的基于TEE的协议。zkPass [730] 默认以代理模式运行，对于不支持代理模式的服务器回退到2PC模式，他们称之为混合模式。TLSNotary [606] 是以太坊基金会对2PC模式的实现。

第二章 B

加密 × AI：用AI增强加密

B-1 概述：让加密更可用、更灵活

今天，人类设计和实现可信计算机的基础协议和应用，如共识算法、网络算法和智能合约（图B.1）。一旦指定，可信计算机（例如，区块链）从环境中获取输入，包括人类用户和外部世界，例如通过预言机。人类与区块链之间的接口需要仔细思考和验证，而人类常常难以精确指定他们的意图，无论是作为用户、程序员还是系统设计者 [66, 226, 718]。

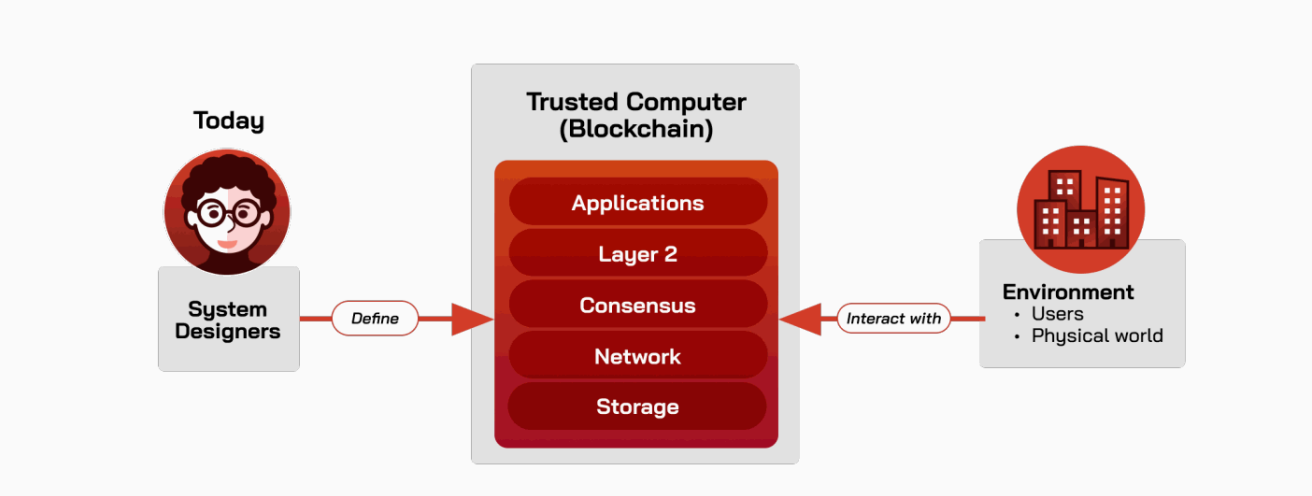


图 B.1 今天，人类为可信计算机设计和实现基础协议和应用（如共识算法、网络算法和智能合约）。一旦指定，可信计算机从环境中获取输入，包括人类用户和外部世界，例如通过预言机。人类与区块链之间的接口需要仔细思考和验证，而人类常常难以精确指定他们的意图，无论是作为用户、程序员还是系统设计者。

AI可以作为可信计算机与“一切其他”之间的翻译层，包括设计者和环境，如图B.2所示。例如：

- AI可以帮助设计者更灵活地指定区块链栈组件（例如，DeFi激励机制）的期望属性。
- AI可以帮助区块链用户定义策略和/或指定他们与区块链的交互（例如，将来自现实世界的数据流转换为智能合约可读的数据，或创建并向执行用户意图的智能合约提交交易）。

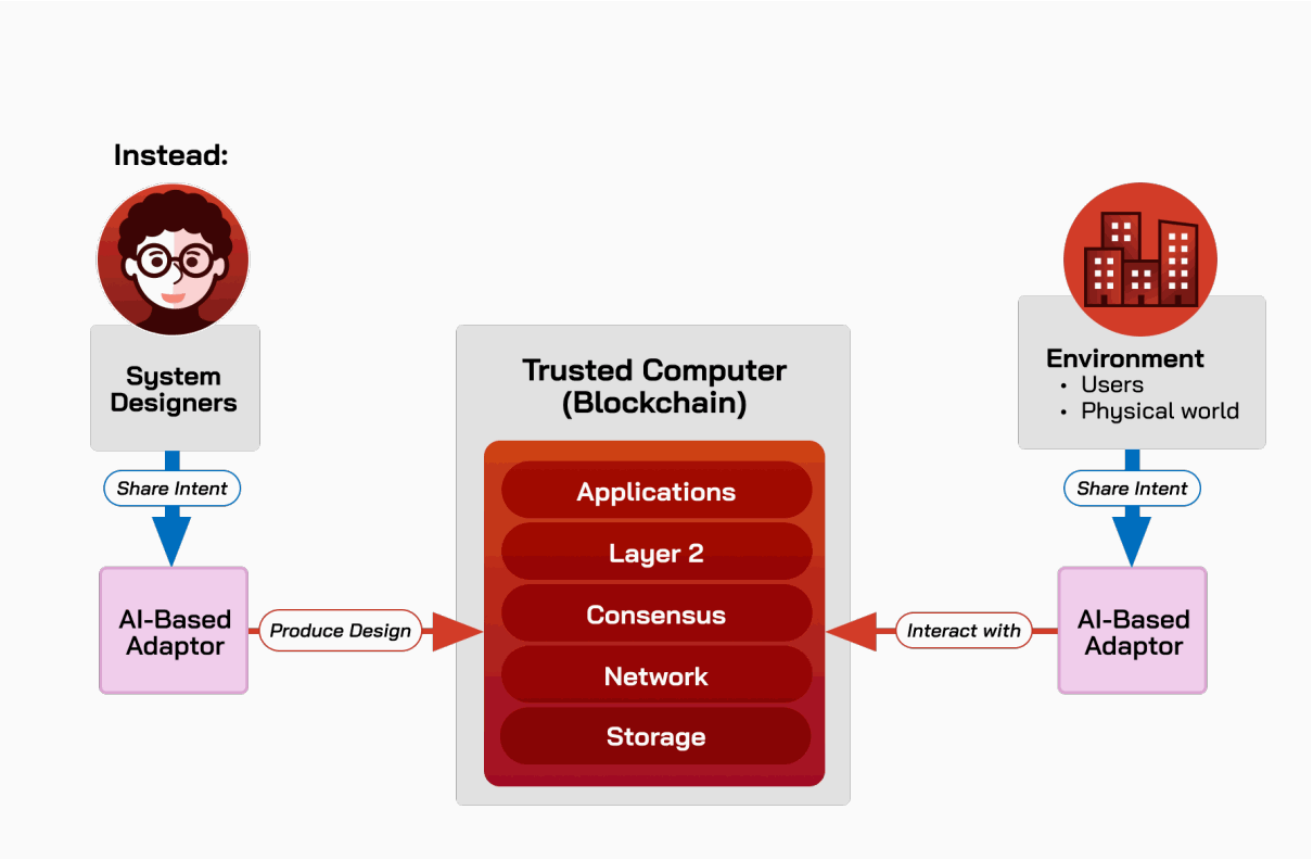


图 B.2 AI可以以更灵活的方式将人类意图（蓝色箭头）转化为机器可读指令。

在本章中，我们将讨论在B-2至B-4节中使用AI促进区块链设计和分析的持续开展的工作，以及在B-5节中一些更具未来感的方向。研究人员将AI应用于区块链系统（更广泛地说，去中心化系统）已有十多年。虽然算法和技术空间广阔，但我们大致将这些努力分为三类，这些类别在时间上大致依次获得突出地位。这些类别如图B.3所示。

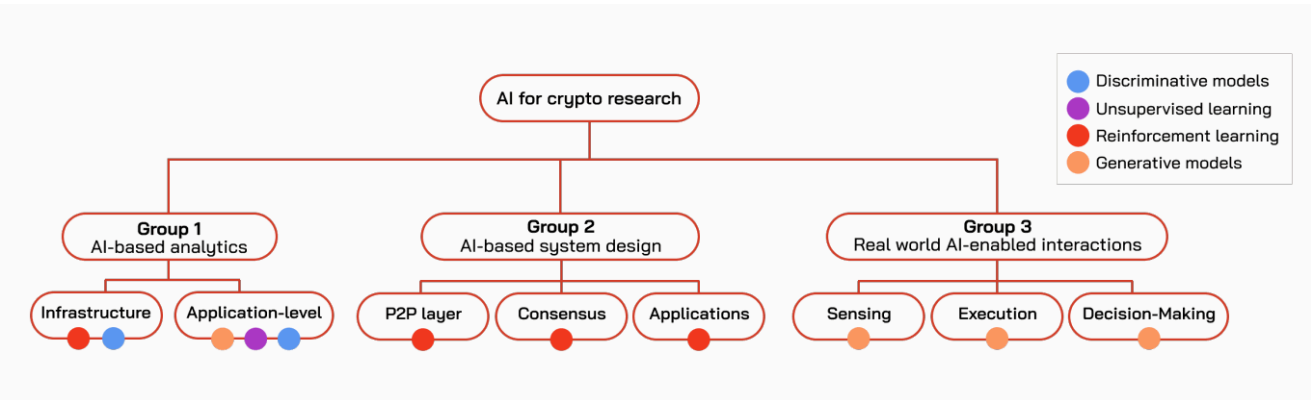


图 B.3 我们确定了AI用于加密研究的三个主要类别：基于AI的分析、系统和算法的基于AI的构建性设计，以及AI增强的与现实世界的交互。彩色气泡代表用于各类研究的常见ML技术；它们并非详尽无遗，但代表了我们观察到的常见模式。请注意，更先进的ML技术（如RL和生成式AI）主要用于第2组和第3组，它们涉及与环境的交互，而不是第1组中静态的、更传统的预测任务。

第1组：AI辅助分析。（B-2节） 这些是基于AI的算法，用于分析或理解现有区块链的状态。这类技术出现于十多年前：它包括将AI应用于分析问题，例如预测区块链上的事件或分类欺诈交易。这些技术将复杂系统的状态转化为人类可理解的东西。

第2组：AI辅助的构造型算法设计。（B-3节） 这些基于AI的方法将高层人类目标（例如增加吞吐量、降低延迟和提高安全性）转化为算法决策，可在设计阶段指导系统配置，也可在运行时使系统行为适应条件变化。这类方法在过去约六年中逐渐流行，通常使用ML学习用于设计加密系统或与其交互的策略。在这一类别中，强化学习（RL）及多臂老虎机等相关技术被用于在复杂的状态和动作空间中学习策略。

第3组：AI辅助与现实世界的交互。（B-4节） 第三类出现于最近几年，探索如何在应用层使用现代AI来增强区块链与外部环境的交互；这类技术主要使用生成模型，有时与强化学习结合，来实现这种增强功能。具体来说，现代AI可以为智能合约配备相对于上一代的三项增强能力：（1）**感知**：AI可以帮助智能合约使用来自现实世界的非结构化数据。（2）**执行**：智能合约可以像LLM当前所做的那样调用工具和函数。（3）**决策**：智能合约可以充当代理，根据编码在目标函数中的价值观做出决策。这类功能部分由预言机实现，预言机允许区块链对外部世界的状态有一个准确的视图。

B-2 AI辅助分析

我们从探索文献中如何使用AI来分析现有区块链开始。我们将这项工作分为两类：（1）全局区块链属性分析（B-2.1节），和（2）局部区块链对象属性分析（B-2.2节）。

B-2.1 全局区块链属性分析

某些类别的分析代表区块链整体的属性，即它们不与单个用户或交易相关。此类分析可以涉及网络范围的协议（例如，共识算法）、网络（例如，P2P网络）和区块链的衍生属性（例如，加密货币价格）。

B-2.1.1 共识分析：漏洞与攻击发现

越来越多的研究使用AI来发现区块链基础设施中的漏洞，特别是在共识协议中。

自私挖矿算法的发现。 探索的一个常见漏洞是自私挖矿 [194, 216]，其中区块链操作员偏离协议规定的行为以不公平地获得额外奖励。例如，自私矿工可以扣留新创建的区块以构建私有链，然后在战略时机发布比公共链更长的私有链。这会导致区块被丢弃，浪费诚实矿工的工作，并增加自私矿工在奖励中的相对份额。

对自私挖矿的研究始于描述特定攻击的收入 [194, 429]，主要针对工作量证明区块链。后来的工作使用基于AI的方法优化了自私挖矿攻击 [533, 651, 734]。这涉及将区块链矿工的所有可能动作建模为马尔可夫决策过程（MDP）。Sapirshtein等人 [533] 首次提出了一种需要求解一系列MDP的方法。后来的工作引入了真实MDP的高效近似 [734] 和其他模型，用于网络条件 [238]、权益证明协议 [238] 和贿赂攻击 [687]。

经典的MDP求解技术由于巨大的计算和内存需求而局限于相对较小的状态空间 [592]。一系列工作采用了深度RL技术（涉及神经网络的使用）来分析更复杂的模型 [69, 70, 277, 534, 535]。Hou等人 [277] 在多种共识规则下，对多个自私矿工的环境使用了深度RL。其他工作 [69, 534] 使用深度RL研究了交易费用对工作量证明区块链中自私挖矿的影响，以及可能为利润而进行微小偏离的“轻微合规”矿工的影响 [70]。Sarenche等人 [535] 使用深度RL技术将自私挖矿的分析扩展到了权益证明区块链。深度RL的使用使得在更现实的设置中找到有利可图的自私挖矿策略成为可能，而这单靠传统方法是不可能实现的。

尽管能够分析更复杂的模型，深度RL技术通常缺乏对导出策略最优性的正式保证。为此，最近的工作使用传统RL方法准确分析具有替代证明系统的最长链协议 [123] 和基于DAG的协议 [71, 334]。他们的方法允许准确表征安全阈值，即从自私挖矿中获利所需的最低算力。

共识攻击的实时发现。 除了表征和优化攻击，最近的工作还侧重于检测正在进行的攻击。Reddy和Sharma [502] 通过使用无监督学习的谱聚类来识别由不合作矿工产生的区块，从而检测基于DAG的账本中的双重支付攻击。Venkatesan和Rahayu [631] 后来提出将混合共识协议（例如，权益证明和工作量证明）与ML分类器结合用于实时异常检测，旨在抢先检测如51%攻击等威胁。另一系列工作侧重于使用判别式ML分类器检测自私挖矿。Wang等人使用基于分叉结构的神经网络 [659]。后来的工作将分析扩展到使用更先进的ML技术，如集成深度学习 [657]，并使用额外特征，如交易费用和区块生成时间 [67]。最近的工作也研究了不可检测的自私挖矿策略，使这些机制失效，并呼吁对如何检测和缓解此类攻击进行更多研究 [59]。总的来说，共识层攻击检测的主流方法是对手工特征进行判别式监督分类，最近出现了更深的架构和图神经网络。然而，进展仍然受到对模拟器生成训练数据的依赖以及已确认的真实世界攻击样本稀缺的限制。

B-2.1.2 P2P分析：攻击发现

另一系列工作使用ML来检测P2P网络层的正在进行的攻击。主要关注的是**日食攻击**（eclipse attacks），其中攻击者试图将一个或多个节点的IP地址与网络其余部分隔离，以改变和控制在区块链网络中不同参与者的视图。这些攻击有时可以通过分析网络流量模式并应用基本的ML技术来检测 [89, 157, 505, 685]。例如，[685] 使用随机森林分类器基于统计特征（如数据包大小和访问频率）识别攻击。后来的工作，如Dai等人 [157]，采用了更先进的ML技术，将卷积神经网络与双向RNN和交叉注意力机制相结合，以捕捉网络流量数据中的时空模式，从而提高检测准确性。尽管在这一方向上已有一些工作，但基于ML的方法（监督和无监督）由

于缺乏来自日食攻击的标记数据而难以开发。先前的方法部分通过合成数据增强技术来处理这个问题，这可以给检测率带来小幅提升 [157]。

B-2.1.3 衍生属性分析：价格预测

许多论文探索了ML技术来预测各种加密货币的价格 [127, 291, 297, 301, 335, 526, 633]。最近的论文使用了相对标准的ML工具，如贝叶斯神经网络 [301]、RNN [126, 297, 526, 633]、MLP [335] 和SVM [335, 633]。这些最近的论文通常强调使用多样化输入数据和特征的重要性，包括链上数据、来自外部（非加密货币）市场的数据，以及社交媒体平台以了解用户情绪 [126, 335, 526, 633]。然而，它们仍然使用了相当基础的ML工具，对神经网络的使用有限。

相比之下，行业工具越来越多地使用**基础模型（FMs）**来解决这些问题。例如，ElizaOS [189] 和 Virtuals [396] 使用户能够部署基于底层LLM做出预测和决策的代理（例如，参见B-4.3节）。由于这些代理使用通用LLM进行决策，它们并非专门针对价格预测进行训练。

加密货币价格预测与传统市场之间的一个主要区别在于，加密货币市场（相对而言）仍然更依赖于小型投资者，特别是对于新兴或波动的资产，如模因币（memecoins）。因此，加密货币价格更受社交媒体渠道（如Discord和Telegram）的影响和讨论，这些渠道可以被挖掘用于辅助信息，其中一些具有误导性 [332]。从这个意义上说，加密货币价格预测问题可以从潜在更丰富的辅助信息中受益，而基于LLM的预测器非常适合处理此类辅助信息。另一方面，有可能设计更好的定制模型来利用这种丰富的辅助信息。一个关键的研究挑战是如何驾驭这种张力。

研究问题 B-2.1 我们如何设计ML工具来预测加密货币的聚合属性（如价格），并有效地整合来自互联网的丰富和非结构化辅助信息？

据我们所知，先前关于加密货币价格预测的研究论文和行业努力都没有使用最先进的时间序列预测模型 [37, 160, 375]。这些工具能否相对于先前研究中提出的更简单、更低维度的预测器带来实质性收益？它们能否相对于行业中看到的通用代理带来好处？设计定制预测器的一个关键研究挑战是加密货币市场辅助信息的丰富性；来自各种平台的文本消息和聊天信号是半结构化的，而ML社区主要针对高度结构化数据的固定基准开发预测模型。调和这些差异可能需要新颖的架构适配器和/或数据处理，类似于在其他领域所做的 [145]。我们在本节结论B-6节中以更广泛的规模讨论其中一些问题。

B-2.1.4 总结

注意，在上述讨论的论文中，ML已被用于我们有可见性的环境中的分析。也就是说，给定一个全局的、已知的共识协议，我们可以发现漏洞，或者给定关于区块链状态的公共全局信息，我们可以检测攻击或预测价格波动。然而，由于许多区块链的去中心化特性，我们可能缺乏运行许多种类分析的可见性。

关键要点 B-2.1: 用于聚合级区块链分析的ML 对于聚合级区块链分析，ML已被应用于有限的一组问题，对于这些问题我们可以获得关于系统的全局、公开观察结果。另一方面，细粒度系统监控和局部分析由于缺乏中心协调和可见性仍然具有挑战性。

例如，在企业环境中，人们对将AI应用于**可观测性**（observability）越来越感兴趣 [135, 145, 714]。换句话说，给定一个复杂的微服务架构，我们如何有效地从各种组件收集遥测数据，并使用它们来识别和解决瓶颈或错误？该流程的数据分析组件越来越多地由ML处理 [145]。

随着区块链处理越来越多的数据，可观测性技术可以通过剔除无响应的对等节点、识别作为资源瓶颈的智能合约或检测不可靠节点来帮助简化操作；然而，传统遥测需要特殊的仪表化和集中聚合，并不适合区块链网络。因此，一个有趣的研究问题是如何为去中心化区块链系统设计可观测性基础设施和相关的基于ML的分析流程。

研究问题 B-2.2 可观测性工具如何帮助简化去中心化区块链操作？我们如何为去中心化方式简化区块链管理设计隐私保护遥测和基于ML的分析？

这个开放性的研究问题需要对区块链系统中的性能瓶颈有详细了解。首先，它需要来自基础设施提供商的测量，以表征在区块链栈中普遍阻碍性能的硬件和软件瓶颈。基于识别出的瓶颈，研究和/或开源社区可以为区块链系统开发标准化遥测——类似于用于微服务监控的OpenTelemetry [695]。为区块链设计可观测性工具引入了新的挑战，例如从不同角度监控共识工件（例如，区块结构），并使典型指标和跟踪技术适应区块链基础设施。例如，在典型的微服务架构中，每个用户请求都会生成一个跟踪，记录请求如何在服务之间流动。然而，在去中心化区块链中，对智能合约的每个请求都可以调用不同的智能合约，每个合约都在不同的验证者硬件上并行执行。因此，在每个验证者处存在相关但不同的跟踪（在时序方面）。总之，为区块链跟踪、日志和指标设计高效的数据表示可能是一个具体的研究挑战。另一个可能是以隐私保护的方式收集遥测数据，不泄露关于单个验证者专有算法和基础设施配置以及私人用户交易模式的细节。

B-2.2 局部区块链对象分析

接下来，我们将焦点从聚合分析缩小到应用层单个对象或工件的基于ML的分析——即交易或智能合约。我们将这些方法分类如下：（1）智能合约安全分析，（2）智能合约经济分析，（3）交易级去匿名化，和（4）交易级欺诈分析。

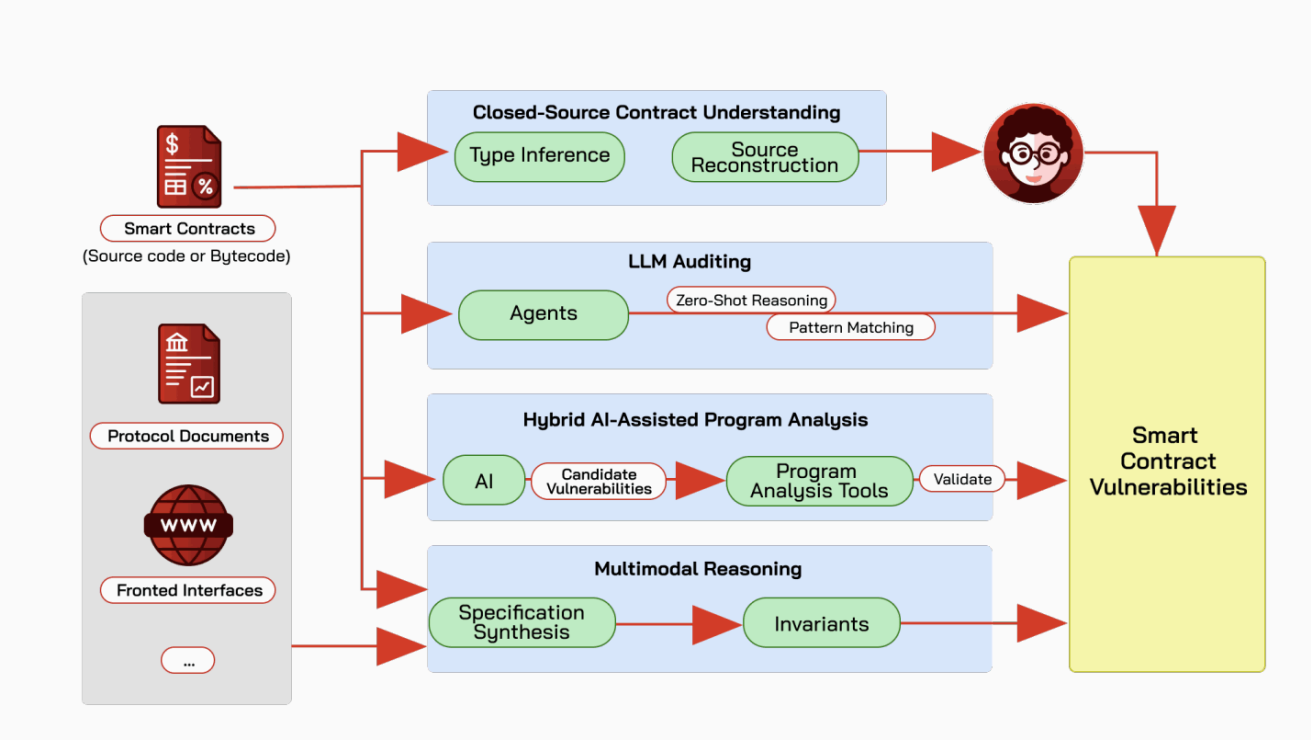


图 B.4 AI辅助智能合约分析的四种范式，它们在AI如何在分析流程中使用以及它们整合的信息来源方面有所不同。现有工作大致可分为：（i）闭源合约理解，AI辅助逆向工程字节码；（ii）基于LLM的审计，LLM代理试图模拟人类安全审计员；（iii）混合AI辅助程序分析，AI提出候选漏洞，然后使用传统程序分析工具进行验证；以及（iv）多模态推理，整合额外信息源，如协议文档和前端界面，以验证合约行为并检测语义不一致性。

B-2.2.1 智能合约安全分析

智能合约安全是区块链完整性的关键支柱 [325, 389, 434, 474, 611, 727]，因为已部署合约中的漏洞可能直接导致财务损失。传统上，漏洞检测依赖于静态分析 [104, 253, 611]、符号执行 [223, 257, 519] 和模糊测试 [256, 309, 555]。虽然这些技术对于检测定义明确的漏洞模式很有效，但它们通常在处理复杂合约逻辑和跨合约交互时遇到困难，这推动了使用AI辅助方法来实现更丰富的语义推理。如图B.4所示，我们将现有的AI辅助方法分为四种范式：闭源合约理解、基于LLM的审计、混合AI辅助程序分析和多模态推理。

闭源合约理解。 AI辅助智能合约分析的第一个范式旨在从已部署的字节码中恢复人类可解释的语义，通常通过反编译，从而支持下游任务，如人类审计。然而，传统的反编译器 [253, 254, 728] 常常留下语义鸿沟，产生审计员难以解释的低级逻辑。为了弥合这一鸿沟，基于AI的反编译已经从属性特定推理发展到端到端源代码重建。早期方法侧重于特定的元数据恢复。例如，SigRec [129] 通过利用调用数据访问模式，通过类型感知符号执行推断参数数量和类型，从EVM字节码中恢复函数签名，因此不需要源代码或签名数据库。SmartHalo [377] 遵循神经符号方法，将静态依赖分析与LLM结合以恢复高级属性，同时通过符号验证保持功能正确性。最近的工作通过两种主要范式探索直接的字节码到源代码翻译：提示工程和微调。代表前者的DiSCo [587] 采用了一种无需训练的方法，使用语义单元来利用通用LLM的零样本能力。相比之下，David等人 [163] 将基于低秩适配的微调应用于将三地址代码映射到Solidity的大规模数据集，使模型能够更好地匹配人类编程风格。

基于LLM的合约审计。 第二个范式探索使用LLM作为直接分析合约代码的自动审计员。这种方法利用通用模型的零样本推理能力来模拟人类审计过程。最近的实证研究 [125, 162] 揭示了一个喜忧参半的情况：虽然LLM在识别漏洞相关模式方面表现出可衡量的能力（例如，David等人 [162] 显示GPT-4和Claude都能在52个先前被利用的DeFi智能合约中正确识别40%的漏洞类型），但这伴随着由于幻觉 [307, 717] 而导致的大量误报。

混合AI辅助程序分析。 为了解决纯LLM的高误报率问题，第三个范式采用了混合架构，将AI的语义理解与传统程序分析的严格验证相结合 [566, 590, 707]。例如，GPTScan [590] 使用GPT识别潜在的漏洞候选者，然后采用静态程序分析（包括数据流和控制流验

证）来确认检测到的漏洞的可行性。

多模态推理。 虽然上述方法主要在代码层面操作，但第四个范式超越了纯粹的程序分析，将漏洞检测转向协议级语义和多模态推理。这一范式的代表性例子包括DeFi-Aligner [231]，它使用LLM从项目文档中提取业务逻辑，并将其与符号代码摘要对齐以识别语义偏差；以及Hyperion [689]，它分析DApp前端界面以揭示用户承诺与链上执行之间的不一致性。PropertyGPT [383] 利用检索增强生成从审计报告中合成可编译规范，而SmartInv [649] 采用"思维层级"策略从多模态来源（包括代码和自然语言注释）中推断关键不变量。

总之，这些范式表明了一个明确的趋势，即利用AI减少人类审计工作量，同时扩大漏洞发现的范围并保持准确的检测。

关键要点 B-2.2 检测智能合约安全漏洞的最先进方法并非纯粹依赖基于输入特征的AI预测；相反，它们识别必须成立的不变量，并将这些与ML模型结合以检测漏洞。

今天的许多论文将确定性程序分析和/或领域信息语义分析与ML工具结合用于漏洞检测。这是一个强大的范式，但它引出了如何最好地提取和使用此类领域特定知识的问题，正如以下研究问题所强调的。

研究问题 B-2.3 应该将何种级别的领域知识整合到智能合约安全分析流程中，以及如何将其与基于AI的技术结合？这些AI辅助系统应利用哪些信息源，包括多模态数据？

先前的工作已经探索了在提取智能合约中应成立的领域特定属性方面的不同操作点，并将其与AI结合。然而，目前尚不清楚在何种条件下哪种架构最佳，以及AI应扮演什么角色。它应该纯粹用作过滤器，如GPTScan吗？它应该用于识别智能合约应满足的策略和不变量吗？如果是后者，需要什么样的输入，包括文档类型和详细程度？虽然存在各种架构的概念验证，但我们缺乏对不同架构和信息源进行系统比较的探索。这样的探索可以帮助指导设计更强大的智能合约漏洞分析架构。

B-2.2.2 智能合约经济分析

AI不仅用于分析智能合约的安全属性，还用于研究由智能合约交互产生的经济行为。在当前的文献中，这一研究方向主要围绕**最大可提取价值（MEV）**，即通过区块构建过程中的战略交易排序获得的利润 [158]。最常研究的MEV活动包括套利 [656, 725]、三明治攻击 [726] 和清算 [486]。现有工作主要分为两类：MEV发现和MEV活动检测；我们在B-3.3.3节讨论ML辅助的出价策略发现，这更多是一个构建性设计问题而非分析问题。

MEV发现。 首先，AI可以通过搜索交易序列和执行环境来帮助发现MEV机会。该方向的早期工作主要依赖于非学习技术，包括对DeFi操作的基于约束的搜索 [725]、对组合DeFi合约的形式验证 [53] 以及对套利机会的启发式驱动分析 [656]。相比之下，Lanturn [54] 将MEV提取表述为一个自适应的基于学习的黑盒优化问题，旨在合成利润最大化的交易序列，而不仅仅是检测预先指定的套利模式。MEVisor [130] 通过并行遗传算法和GPU加速执行来补充这一工作，强调高吞吐量的机会搜索。

MEV活动检测。 其次，AI可用于检测已实现的MEV活动，即从链上或捆绑包级数据中自动识别和分类MEV相关行为。早期方法主要依赖于启发式规则来检测MEV活动 [407, 487, 607]。虽然这些方法在特定设置中有效，但它们严重依赖于手工制作的模式，因此在自动化、可扩展性和对新MEV策略的适应性方面受到限制。因此，最近的工作开始引入基于AI的MEV活动检测方法。一个代表性的例子是ActLifter-ActCluster流程 [376]，它首先将原始交易痕迹提升为语义DeFi操作，然后执行捆绑包表示学习以将原始捆绑包编码为低维特征向量，最后应用迭代聚类来识别已知和以前未见过的MEV活动模式。

总的来说，AI已将智能合约经济分析从手工制作的模式和静态启发式转向更自动化的方法，这些方法可以用更少的任务特定先验知识来适应复杂行为。这为利用智能合约中可能对人类来说不直观的微妙结构和错位打开了大门。

B-2.2.3 交易级去匿名化

加密货币去匿名化是指识别给定交易的来源，由其现实世界身份或其他网络标识符（如IP地址）定义。基于ML的去匿名化通常通过以下机制之一出现：（1）与中心化第三方服务的交互，（2）行为和链上模式分析，以及（3）网络层及相关侧信道。大多数关于基于AI去匿名化方法的公开研究来自第二和第三类，我们总结如下。

行为和链上模式分析。 行为和链上模式分析无需第三方直接参与即可促进去匿名化。常见技术包括地址聚类，以及交易金额、频率、支付习惯、地址重用和时序模式的分析。这些启发式方法利用区块链固有的透明度来揭示地址之间的关系，并在某些情况下将它们与现实世界的个体联系起来。例如，一些工作使用手动图表示学习技术对节点进行分类 [85, 380]。转向更自动化的特征提取方法，许多论文使用图神经网络从交易图中学习嵌入以进行账户去匿名化 [182, 200, 280, 283, 374, 379, 551, 698, 721, 724]。例如，Huang等人将分层自注意力模块与大型图关系建模相结合，在其架构中建模局部节点结构和全局关系 [283]。Hu等人 [280] 则使用交易的顺序建模；他们的BERT4ETH模型是在以太坊交易序列上预训练的编码器风格Transformer。它用于提取以太坊账户的嵌入以进行去匿名化和异常检测（见下面的B-2.2.4节）。据我们所知，BERT4ETH [280] 的顺序建模方法与最近显式使用图结构的方法

（如 [200, 283, 724]）之间没有正面比较。因此，目前尚不清楚哪种架构效果最好，尽管近期的趋势明显倾向于图感知建模；也就是说，可以使用自注意力，但要以尊重底层图拓扑的方式进行。

网络层和侧信道。 网络层和侧信道攻击利用点对点基础设施的动态来推断交易来源的IP地址。控制大量节点的攻击者可以监控交易传播模式，通过时序分析或独特的转发行为推断源节点，并收集链下元数据，如在消息广播期间暴露的IP地址、设备指纹或与外部事件的相关性 [91, 97, 202, 233]。这些方法大多未使用数据驱动的ML预测器，而是侧重于基于传播动力学 [97, 202] 的统计源预测器的设计和分析。

B-2.2.4 交易级欺诈分析

作为来源去匿名化分析的补充，另一类重要分析旨在检测欺诈性或异常的加密交易 [518]，**包括反洗钱（AML）和反恐融资（CFT）**应用。历史上，无论在区块链还是传统金融中，欺诈检测都广泛采用基于规则的系统 [518]。

然而，这类方法难以应对动态而复杂的环境，常会产生大量误报，带来显著的运营与合规成本 [6, 518]。例如，Project Aurora [65] 在接近真实世界的条件和约束下，比较了现有规则模型、逻辑回归、神经网络和图神经网络等策略。研究发现，在合成的跨境交易数据集中，图神经网络检测洗钱活动的效果最好。

数据驱动、AI辅助的欺诈检测算法可以同时利用区块链技术栈不同层级的数据和链下来源。常见数据包括应用层轨迹（如应用内交易及其时间信息）、社交媒体等链下营销数据，以及价格波动等链下二级市场数据。下面按输入数据类型概述相关检测技术。

交易（序列）元数据。 许多ML工具使用账户交易间隔、应用内交易类型、交易金额、参与方和目的地等交易级或序列级特征，检测非法活动与诈骗，并据此训练基础预测模型 [51, 203, 230, 280, 284, 457, 680]。例如，BlockGPT [230] 和BERT4ETH [280] 使用序列化的以太坊交易训练Transformer，再提取嵌入用于下游分类任务。

拓扑数据。 另一类重要方法显式建模拓扑数据。这些数据可能来自P2P网络测量 [336]，也可能来自对逻辑交易图的观察 [15, 464, 484, 637, 678]。例如，BitcoinHeist从比特币交易图和地址图中提取定制特征以识别勒索软件 [15]；类似的图特征，如资金流模式、混币行为和分层结构，也被用于AML/CFT风险评分和可疑活动检测 [264, 384, 626, 643]。

包括图卷积网络（GCN）在内的图神经网络，在区块链交易数据的异常与欺诈检测中表现出很大潜力 [65, 464, 478, 677]。除Project Aurora [65] 的发现外，Patel等人提出EvAnGCN，将动态图卷积网络应用于以太坊不断演化的诱导交易图，以识别异常行为 [464]。较新的研究则表明，标准GCN在异常加密货币交易检测中优于传统方法 [478]。

链下数据。 研究人员也越来越多地利用链下数据检测异常、欺诈或非法的加密货币活动 [191]。一些研究专门识别与暗网市场相关的洗钱交易 [17, 176]；另一些则使用社交媒体和其他网络来源。

例如，CryptoScamHunter使用自然语言处理分析YouTube视频标题和描述，以检测DEX套利机器人骗局 [369]。Huang等人将二级市场数据与链上交易活动结合，用于识别NFT卷款跑路骗局 [284]。Wang等人则结合浏览器扩展评论和程序特征，预测加密相关浏览器扩展是否具有恶意性 [648]。另有多项工作使用NLP分析社交媒体内容，发现加密货币拉高出货计划 [416, 432, 436, 455]。

关键要点 B-2.4：用于交易级分析的AI 目前最先进的交易级分析方法，包括欺诈检测和去匿名化，都大量使用交易图特征和元数据。这些输入被用于提取能够捕捉图依赖关系的神经表示：既可以由GNN显式提取，也可以由在整理后的交易序列上训练的Transformer隐式提取。

当然，强大的检测器还可能使用大量不具图结构的结构化和非结构化辅助输入。如何设计能够有效利用这些辅助输入的ML算法，是一个重要问题。

B-3 AI辅助的构造型算法设计

前一类工作（B-2节）侧重于AI辅助分析：用于理解现有加密应用和网络状态的工具。自从该类研究在十多年前兴起以来，研究社区也开始转向AI来帮助设计去中心化算法，从点对点网络和区块链协议到应用和市场。这些技术通常依赖于强化学习和相关ML技术来推动算法的设计——既为核心基础设施、底层共识协议，也为战略性应用算法。我们按这些方法在区块链栈中操作的层次对其进行分类：点对点网络（P2P；B-3.1节）、共识层（B-3.2节）和应用层（B-3.3节）。

B-3.1 点对点协议

AI已被用于设计帮助形成和维护区块链网络中点对点消息传输层的算法。在一个相关的发展中，维护特定对等点之间链接的应用层网络也受益于AI的使用。后者涉及第二层解决方案，如支付通道网络。

点对点网络。 已经提出了一些算法，其中单个网络节点进行基于AI的本地更改（例如，重新连接对等连接、更新通信配置文件），以改善本地和全局网络属性，如平均延迟或总通信量 [399, 400, 595, 623]。例如，Topiary [400] 明确将网络拓扑形成表述为多臂老虎机（MAB）问题，并设计了一种算法来更新每个节点的对等连接，丢弃传入消息相对延迟较高的对等点。Valko和Kudenko [623] 则采用RL代理重新排序广播队列，以减少传播时间和发送的消息总数，从而降低网络的能耗足迹。

类似的想法已应用于特定领域的区块链网络，如车载网络和无线网络。在车载网络中，节点频繁加入和离开，导致比加密货币网络更高的对等点周转率。Kim和Ibrahim [337] 将对等点数量的选择建模为上下文MAB问题，动态调整通道大小以在频繁车辆周转下保持拜占庭容错。Saadat等人 [527] 进一步使用ML来预测基于集群的车载网络中的节点稳定性，选择不太可能在过程中离开的共识节点。在无线网络中，高传输速率导致高能耗。为此，Ju等人 [317] 使用图卷积网络（GCNs）来确定每个节点的数据传输速率，目标是在可靠通信与低能耗之间取得平衡。

总的来说，与区块链背景下AI的其他用途相比，用于P2P网络设计和管理AI辅助方法仍相对未被充分探索。一个有趣的问题是如何将AI辅助P2P网络管理与检测正在进行网络攻击的方法相结合。

研究问题 B-3.1 如何将检测日食攻击的技术（B-2.1.2节）与P2P网络管理和优化算法相结合？现有的P2P网络管理算法何时能在对抗性网络条件下提供鲁棒性（以及我们应如何对此类对抗性条件建模）？

第二层网络和跨链通信。 依赖区块链共识层（即它们的Layer 1）的第二层网络也使用P2P结构，但它们面临与消息传播P2P网络不同的一组挑战。一个突出的例子是像比特币上的闪电网络（Lightning Network）这样的支付通道网络（PCN）。PCN中的连接代表逻辑通道，可以通过这些通道进行点对点转账，而不是通信通道。因此，连接代表一种相互信任的形式，并保证参与者可以在不诉诸Layer 1网络的情况下进行转账。尽管存在这些差异，两种类型的网络在网络形成和路由方面都提出了类似的挑战：应该形成什么样的网络拓扑？应如何管理不活跃的对等点？应如何在网络上路由交易？其中一些挑战已经通过基于RL的算法解决，如在P2P网络中一样。例如，基于RL的机制已被用于设置通道参数（如费用）以最大化运营商利润 [45, 132, 571]、选择支付路由以最小化交易发送者的成本 [322, 485, 623, 624]，以及重新平衡通道以提高网络吞吐量 [131, 461]。在这些方法中，用于路由管理的RL算法相似 [623, 624]；然而，据我们所知，关于使用RL来确定如何在PCN中建立和断开通道的工作有限（而在P2P文献中这是一个虽小但已建立的研究领域）。这可能是未来研究的一个有趣问题。

研究问题 B-3.2 PCN节点如何使用RL方法动态地随时间决定建立和断开哪些边？此类算法将如何影响节点自身的奖励以及整体网络健康？

B-3.2 共识协议

AI已在共识层被广泛使用，既用于增强现有共识协议和操作，也作为全新协议的组成部分。

B-3.2.1 性能增强

共识协议最初是为静态环境设计的，但它们在实践中通常在动态条件下运行，其中网络条件和参与者可能会变化。近年来，越来越多的研究转向ML技术，通过更响应变化的条件来增强共识协议的性能（即增加吞吐量和/或减少延迟）。可以区分三个相关领域，其中AI已被用于增强共识。

协议选择。 在第一系列工作中，借助AI选择了共识协议。预期的部署网络、可用的同步功能和预期的连接性影响共识协议的选择。没有单一的共识协议在所有操作条件下都占主导地位：比特币的PoW共识具有高延迟，并针对在松散同步网络中保持安全性进行了优化，时间常数在分钟量级；而以太坊和Cardano的PoS协议则假设更紧密的同步（约10-20秒）和更均匀的可达性。

因此，建议构建在操作期间在协议之间切换的系统，最常见的是在乐观路径下运行快速且脆弱的协议，并在检测到乐观路径问题时切换到较慢但鲁棒的协议。探索这一思想的经典工作提出了以这种方式构建“下一代700种BFT协议”[47]。基于AI的方法已被用于指导协议的选择。Liu等人 [381] 开发了一个使用深度强化学习静态选择共识算法的系统，而最近的BFTBrain工作 [676] 在共识协议（如PBFT [118]、HotStuff [694] 和 Zyzzyva [347]）之间动态选择。它通过将协议选择视为一个上下文MAB问题，并使用内置于整个系统中的RL引擎来解决。

然而，依赖自动化和AI在正常情况下加速共识，为恶意行为者打开了潜在的攻击途径。该领域的经典结果表明，在给定环境中使用错误的协议可能会将性能降低到零 [28, 142]。因此，基于ML的协议选择方法需要是鲁棒的。

参数选择。 在共识增强的第二个领域，AI已被用于选择共识协议的参数。根本原因是网络的性能在很大程度上取决于配置参数，如超时、区块大小和区块间隔，以及活跃运营商之间的可用连接性。在最优值随工作负载波动而变化的環境中，手动调优是不切实际的。一系列工作侧重于区块参数的优化。Momen等人 [420] 应用XGBoost预测未来交易量并动态调整区块大小以匹配预期需

求。其他工作 [381, 705] 使用深度强化学习来调整区块大小和间隔。Zhai等人 [705] 尝试通过使用结构性因果模型为所选值提供解释，使这些选择更具可解释性。Dutta等人 [184] 则侧重于区块创建的时间，使用强化学习让操作员学习何时密封区块以最小化交易确认时间。

共识参与者选择。 作为共识协议内的第三个也是最后一个领域，AI已被应用于选择特定角色的参与者。由于实践中部署的大多数共识协议依赖于领导者，测量节点性能并选择连接良好且活跃的节点作为领导者通常会加速协议操作。然而，必须注意，如果测量结果出错，不要降低性能。

最近的研究 [147, 612] 表明，即使没有特定于AI的选择启发式，协议性能也能从精心选择共识领导者中受益匪浅。Islam等人 [296] 的协议使用多代理RL方法选择权益证明共识中的领导者。它从所有验证者收集关于潜在领导者的各种性能指标，旨在通过奖惩机制识别和排除恶意验证者。Nour等人 [169, 170] 研究了AI在基于DAG的BFT协议（如Narwhal [159] 和 Bullshark [576]）中的应用。他们使用图神经网络对区块进行排序并在DAG内选择领导者，从而减少延迟并提高吞吐量，而不会影响底层共识协议的健全性。许多其他作者以类似方式使用RL和相关方法动态地为验证者节点分配分数，并根据这些分数分配特定的协议角色 [128, 363, 364]。最近关于该主题的一项综述 [513] 收集了许多技术，但主要表明由于自动化推理的使用存在多个开放问题，该领域仍处于起步阶段。特别是，该综述强调了AI所启用的对抗性攻击的潜力。

研究问题 B-3.3 在共识协议内操作并指导参数和参与者选择的基于ML的方法，对于恶意内部节点的操纵有多鲁棒？去中心化共识协议如何对其参与节点的可信度得出一个共同且可信的估计？

B-3.2.2 分片

深度RL也被应用于区块链分片算法的设计，解决了两个互补的问题：分片配置优化和通过数据放置减少跨分片交易。

最早的努力使用深度强化学习（特别是深度Q网络，DQN）来选择创建多少分片以及如何调整它们的大小，将这些重复性决策框定为顺序优化问题。SkyChain [713] 使用DQN持续变化参数配置，同时确保分片在需要时可以高效合并和拆分。一项同期工作 [701] 表明，DQN代理可以学习在满足安全约束的情况下最大化吞吐量，该约束来自对网络中其他节点性能和可检测不当行为的估计。一个代表性的近期工作是TbDd [719]，一个基于信任的DRL驱动的分片框架：它收集所有节点性能和历史行为的反馈，并相应地分配角色。

第二系列工作将数据和账户分配到分片，以最小化昂贵的跨分片交易。TbDd [719] 也实现了这一点，通过观察数据访问模式并将数据项分配到合适的分片以限制跨分片交易的数量。Wang等人 [646] 使用生成式AI预测未来的跨分片交互，以便主动将节点分配到分片。SPRING [372] 采用深度RL，通过利用时空交易模式在分片之间迁移账户。AERO [572] 扩展了这种方法，通过批处理迁移决策来缩小动作空间并实现更好的可扩展性。据我们所知，动态分片目前尚未在任何活跃的区块链系统上部署，因此这些研究仍代表探索性研究。如果分片被大规模采用，ML引导的分片和数据放置算法无疑将再次被采用并可能部署。

研究问题 B-3.4 在不观察特定交易工作负载的情况下，AI和ML如何帮助分析加密货币或通用区块链上的数据依赖性？哪些方法将导致自动化的数据放置算法，用于减少跨多个分片的依赖并提高性能的分片区块链？

B-3.2.3 信任模型

工作量证明共识中巨大能源成本和对看似无用的计算的投入，导致了对在区块链共识中使用有用工作的长期追求，特别是与AI相关的工作，因为它具有高计算负担。然而，这至今并不容易，也没有令人满意的解决方案。Dotan和Tochner [177] 正式调查了这一领域，并推导出依赖于无浪费PoW的系统的约束。他们表明，在现实假设下，允许的问题集仍然必须涉及密码学难度的元素，以保持协议的安全性和效率。

最近的研究更为积极：Komargodski、Weinstein等人 [68, 345] 研究了带有外部奖励的共识的经济有效性和均衡动态。矩阵乘法作为一种有前景的问题受到特别关注，可以基于此构建有用的计算谜题 [345]。他们的工作为使用AI训练和推理工作负载进行共识开辟了道路；然而，现在评估此类基于有用工作构建的区块链网络的经验安全性还为时过早。

B-3.2.4 总结

共识协议处于区块链所有安全机制的核心：它们确保基于许多参与节点的输入和正确行为对状态和共同行动达成一致。其安全性与性能之间的权衡因此受到广泛关注。因此，有大量关于优化各种共识机制参数的工作，其中一些也使用了AI。当前的研究表明，将ML用于共识协议具有许多前景。但由于基于AI的协议优化方法尚未广泛部署，它们在实际网络环境中的性能和抗攻击能力仍是未知数。

对于基于分布式计算理论的核心共识协议设计，AI迄今发挥的作用较小。可以说，自动化发现安全且高性能的协议将成为可能，类似于密码学协议研究中的现有自动化 [73]。

研究问题 B-3.5 我们如何使用AI提出新颖的共识机制，以优化通信和延迟方面的性能，同时有机地适应动态环境并保持安全？AI能否帮助我们设计和分析此类新协议的安全性？

B-3.3 应用设计

许多应用将AI作为其设计的核心组成部分。

B-3.3.1 DeFi市场设计

AI已成功应用于设计DeFi应用，包括自动做市商（AMMs）。例如，ZeroSwap使用DQN方法随时间改变资产价格 [425]，Moszczynski将AMM与频繁批量拍卖结合，使用RL优化批量之间的定价规则 [424]。相关技术也已应用于像Aave [1]、Morpho [422] 和 Euler [192] 这样的借贷市场：在这些市场中，利率影响市场的效用和流动性。有几篇论文应用随机控制理论来设定此类利率 [76, 77, 79, 87]，有时与深度学习技术结合 [79]。Chitra最近展示了如何使用在线学习以后悔最优的方式设定利率 [137]。

B-3.3.2 AI增强的智能合约安全

除了协议级机制设计，AI也已应用于智能合约安全，特别是自动化生成漏洞利用或为易受攻击合约打补丁等任务。先前关于自动漏洞利用生成的工作依赖于手动制作的模板或符号分析来合成满足特定漏洞条件的攻击序列 [223, 257, 519]。基于AI的方法通过启用语义推理和自适应搜索来改进漏洞利用生成，允许系统在最少人工干预的情况下构建超出预定义模板的漏洞利用。例如，AdvSCanner [679] 结合LLM和静态分析生成专门利用重入漏洞的对抗性合约。向自主性扩展，系统AI [239] 采用基于代理的工作流和执行验证，使用领域特定工具实现自主发现并验证现实世界区块链状态上的有利可图漏洞利用。同样，PoCo [33] 引入了一个智能体框架，通过推理-行动-观察循环将自然语言审计报告转化为可执行的Foundry概念验证。

作为自动漏洞利用生成的补充，近期工作也探索了AI驱动的智能合约自动程序修复，旨在闭环从漏洞发现到缓解。虽然传统修复在很大程度上依赖于刚性模板 [310, 433, 515]，但AI已使这些框架变得更加自适应和上下文感知。最初的AI增强方法，如SmartFix [567]，通过使用统计模型对补丁候选进行排序来增强传统的“生成-验证”范式；sGuard+ [232] 引入机器学习分类器来指导基于规则的修复，选择最合适的规则以避免过度修补。超越静态模板，近期工作利用LLM的语义理解和生成推理能力进行端到端的生成式修复 [644, 715]。

B-3.3.3 MEV提取算法

AI开始被用于优化MEV拍卖中的出价策略，一旦MEV机会被识别。为了获得潜在利润，目标交易必须仍然被包含在链上，但在实践中，同一机会通常有许多竞争者 [158]。早期的竞争通常在公开的优先Gas拍卖中进行，而Flashbots风格的私有通道将此过程转变为密封投标的首价拍卖 [218, 661]。这使得出价本身成为一个学习问题：目标是预测竞争性出价并选择在包含条件下最大化预期利润的贿赂。例如，Raun等人分析Flashbots拍卖数据并训练机器学习模型（特别是Light Gradient Boosted Machine回归器）来预测MEV拍卖中的获胜贿赂比率，表明基于学习的出价策略可以提高套利MEV拍卖的盈利能力 [499]。Lanturn [54] 则使用最近的自适应采样技术 [303] 优化黑盒奖励函数以最大化可提取价值。其他工作已经探索使用随机森林和决策树等简单回归模型来预测诸如FlashBots拍卖长度和最大出价值等辅助变量 [358]。

B-4 AI增强与现实世界的交互

在过去几年中，出现了使用AI来增强智能合约与现实世界交互的努力。这些努力已经得到部署，但最好被描述为新兴的。

在本节中，我们讨论使用AI来增强区块链与外部世界的三种主要交互类型：

1. **感知：（B-4.1节）** AI可以帮助智能合约以比目前更复杂的方式理解世界状态，例如，通过AI驱动的预言机消化和处理非结构化数据流。
2. **执行：（B-4.2节）** 泛化感知，AI可以扩展智能合约执行计算和影响外部世界的方式，例如，调用AI模型。
3. **决策：（B-4.3节）** AI可以帮助智能合约制定更复杂的决策流程。我们描述了一个来自金融领域的案例研究，展示基于AI的决策如何引入新的挑战。

在所有这三个领域，我们都讨论机遇和挑战。

B-4.1 感知：使智能合约能够理解自然语言

在其初始形式中，智能合约被设计为仅对链上数据进行操作。区块链通过预言机——将智能合约连接到链下数据和链下计算资源的系统——的出现克服了这一限制。然而，今天大多数现有的预言机仅限于中继干净、结构良好的API数据。这限制了预言机在没有明确API数据的世界中的覆盖范围。因此，例如，智能合约无法理解或解释人类语言或制度，这是它们代表书面合同能力的关键障碍。

AI具有显著扩展智能合约可访问数据种类的潜力。例如，预言机系统可以可验证地使用AI工具将松散结构化的数据转化为智能合约可读的格式——从而充当更广泛互联网与单个智能合约之间的中间件。例如，今天，唯一类型的完全自动化保险智能合约是简单的参数保险，其中支付在预先约定的事件发生时触发（如AXA已停产的航班延误保险Fizzy [105]）。有了LLM驱动的预言机，人们可以想象自动化的保险合同，这些合同可以摄取并推理更丰富的证据，如索赔叙述、警方报告或来自维修店的检查报告。另一个例子是预测市场，它依赖预言机来确定问题的结果。现有的系统（如Polymarket）依赖人类提案者来回答没有明确API数据的问题，以及人类争议解决来解决分歧 [278]。AI驱动的预言机有潜力自动化这一过程，避免人为错误并减少与争议相关的延迟和成本。

B-4.1.1 LLM错误的风险

尽管有这些前景，但部署LLM驱动的预言机——以及在关键系统中使用LLMs普遍存在——的一个核心挑战是出错的可能性，如推理缺陷、幻觉和算术错误 [537]。在这里，我们关注由AI固有局限性引起的错误，假设没有对抗性操纵。保护AI预言机流程免受攻击至关重要且是正交的。

最近两项关于LLMs在回答Polymarket问题上准确性的研究 [323, 617] 揭示了AI作为预言机的当前能力。在一项Chainlink Labs的实证研究中 [323]，作者使用GPT-4o解决了Polymarket上的1,660个市场，并将LLM的答案与真实情况进行比较。GPT-4o实现了89.3%的总体准确率。UMA的一项类似实验 [617] 证实了这一结果，他们的Truth Bot实现了75%的准确率（尽管很大一部分是由于在截止日期前提交的答案）。作为背景，人类对UMA乐观预言机提供的答案总体准确率为98.2%。因此，对于这项特定任务，LLMs仍然比人类犯更多的错误，表明需要护栏和监督。

LLMs的准确性因上下文而异。当要提取的信息是离散的且有官方真实来源时，它们表现得相当好。一个典型的例子是体育结果。在Chainlink Labs的研究中，关于体育结果的问题有99.7%被正确回答。在UMA的报告中，Truth Bot在体育和资产定价市场（后者通过调用特定API回答）上实现了99.3%的准确率。

对于不太直接的问题，错误率可能高得多。例如，LLMs在回答涉及时间的问题时很吃力（例如，“利率上调是在GDP报告之前还是之后宣布的？”），或者当需要大量努力来提取答案时（例如，回答“特朗普在5月30日的匹兹堡集会上会说多少次‘钢铁厂’？”需要转录视频并计数单词，这是一项无法用自动化工具准确完成的任务）。

值得探索降低错误率的方法。例如，UMA提议使用AI智能体来寻找已知的幻觉模式（例如，Perplexity进行推理而非以事实为驱动）[617]，尽管最近的研究表明基于LLM的错误检测器表现远不如人类 [326]。其他方法，如使用多个模型，也可能有所帮助，尽管它们不会消除错误，需要仔细评估成本/收益权衡。

B-4.1.2 容忍LLM错误

即使在LLM表现非常好的环境中，错误率也不是零。因此，使用LLMs作为预言机的系统必须处理潜在错误。存在三种可能性。

1. 依赖LLMs的应用需要被设计为容忍错误，以便小错误不会导致灾难性后果（例如，使用AI解决总支付低于阈值的低价值市场）。
2. 第二种方法是让人类参与循环以检测和纠正AI错误。UMA乐观预言机（OO）[616] 是一个例子：在答案由提案者（人类或AI机器人）提交后，会有一个48小时的争议窗口，在此期间任何人都可以挑战提交的正确性，如有必要，将启动仲裁过程，最终所有UMA代币持有者验证事件结果并投票决定最终结果。然而，人类参与循环会减慢决策速度。
3. 第三种可能性是仅在AI模型无法以高置信度做出决策时才让人类参与。最近的研究表明，弃权（即在面对不确定性时不给出答案）可以提高模型准确性 [209, 662]。如果AI模型能够以高准确率弃权，那么这种设计可以在AI输出可用时做出快速决策，只有在不可用时才做出缓慢决策。

关键要点 B-4.1: 处理AI驱动预言机的错误 使用AI驱动预言机的系统必须处理潜在错误。有三种高层次方法：1）设计系统以容忍错误（这可能只在有限场景中可行），2）让人类仲裁AI输出（这减慢了决策速度，但可以有条件地基于争议），以及3）如果AI模型在面对不确定性时可以弃权，则可能只在AI无法决定时才让人类参与。

B-4.2 执行：使智能合约能够使用AI模型和工具

普通的智能合约可以强制执行简单的规则，如条件语句，但它们通常在高层次任务（如数据分析、模式提取和规划）方面存在困难。此外，它们通常在链上采取行动——即直接影响区块链的状态，而不影响其他任何东西。智能合约可以通过访问更广泛的AI生态系统来扩展其执行能力，无论是在链上还是链下。

具体来说，AI生态系统包括大量的AI模型和相关工具，用于以包括数据访问和修改、资金转移和信息检索在内的多种方式与现实世界交互。然而，在链上运行这些工具的成本可能高得令人望而却步。因此，一个自然的问题是，我们如何使智能合约能够在现实世界（链下）可验证地访问AI模型和工具，而不会在链上产生高昂的计算成本？

预言机再次提供了一个潜在的解决方案，通过使链下计算能够以可验证的方式记录在链上。现代预言机系统，如Chainlink和Supra，已经可以执行链下工作流并返回经过认证的结果。它们自然可以（被扩展）支持AI工作流。我们将讨论这样做的方法及其权衡。

AI可以将智能合约从强制执行静态规则转变为动态的、上下文感知的机制。例如，智能合约可以使用AI工具标记欺诈交易 [9]、动态调整参数以及做出自动交易决策。在链上直接执行AI逻辑在经济上不可行，因为链上计算可能比云计算昂贵几个数量级 [618]。作为一种高效的替代方案，预言机可以将智能合约连接到链下AI工具，无论是预言机基础设施内运行的本地模型还是第三方ML服务提供商（例如，OpenAI API）。

对于预言机的这种应用，两个安全属性很重要。首先，智能合约需要高效验证整个AI工作流的完整性，确保来自智能合约的输入未被篡改、使用了正确的ML模型且输出被正确计算。这些要求对AI来说并非独一无二（它们适用于一般的预言机链下计算），但AI工作负载的规模要求高效的解决方案。我们请读者参考C-4节关于三种技术方法的讨论：1）乐观方法，其中完整性由经济激励确保（即，此类协议配备了识别不正确结果并经济上惩罚违规者的机制）[44, 150, 415, 691, 723]；2）预言机节点在TEE中运行ML模型，并伴随结果提供硬件认证以证明计算的正确性 [188]；以及3）预言机节点伴随计算结果提供零知识证明 [3, 124, 234, 382, 489, 490, 545, 588, 641]。这些解决方案之间的权衡在C-4节中有详细讨论。

第二，对于涉及私有数据或专有模型的ML任务，必须保护其机密性。例如，智能合约可能想要运行一个不能向预言机节点泄露的专有欺诈检测模型。为此，基于TEE的方法可能是最实用的 [188]，智能合约开发者可以在TEE保护的密钥下加密模型，以便模型仅在TEE内部解密。全同态加密（FHE）理论上可以用于评估加密模型，但它对于大型模型尚不实用，并且由于需要解密能力，需要依赖一个可信节点委员会 [224]。

除了安全考量，运行AI工作负载会比典型的预言机工作流消耗更多资源，这引发了如何为AI计算定价的机制设计问题 [57]，以及如何计量AI使用并正确向智能合约收费的系统设计问题。这个问题的一个有趣难题是搭便车（freeloading）的风险。因为区块链缺乏机密性，预言机交付的计算结果是公开可用的。例如，一个模仿性的预测市场智能合约可以监控另一个付费使用AI驱动预言机的预测市场的答案，并获得免费答案来解决自己的市场。搭便车问题在Town Crier [708] 中讨论过，作者建议使用指定验证者证明（designated-verifier proofs）来使输出仅对原始请求者可验证，但完整的解决方案留待将来工作。我们注意到，在AI工作负载的背景下，这个问题更为突出，因为搭便车AI计算可能节省的费用非常可观。

B-4.3 决策：基于AI的投资工具

如果我们赋予智能合约访问AI模型和工具的权限，如B-4.2节所述，一个自然的问题是这些工具将如何影响链上应用。将AI整合到智能合约的决策过程中会引入巨大的复杂性和不透明性，而以前智能合约的一个优势是它们的（相对）透明性和可解释性。在本节中，我们描述了一个来自金融领域的案例研究，展示在智能合约中使用基于AI的决策如何引发参与者之间的新的紧张关系和潜在的不公平来源。具体来说，今天AI对区块链最流行的应用之一是在投资工具中。我们讨论基于ML的**集体投资算法（CoinAlgs）**及其风险。

B-4.3.1 CoinAlgos

集体投资算法（Collective Investment Algorithms，CoinAlgs）是由用户社区共享的驱动集体投资行动的算法 [197]。算法交易在传统金融中早已是常见的做法，无论是在大公司（高频交易、对冲基金、量化投资等），还是在零售投资者使用的普及产品（机器人顾问、交易软件包等）中。

最近，随着AI的广泛部署，CoinAlgs在去中心化金融中也变得普遍。最突出的例子是AI驱动的投资DAO——将资金汇集在区块链上进行集体交易的社区，这些交易由AI模型或代理决定。被称为“去中心化对冲基金管理人”[411]，这类CoinAlgs在Web3内部引起了极大的兴趣。流行的CoinAlg项目，如ElizaOS [642]（一个从AI驱动投资基金转变为通用AI平台的项目）和AIXBT [13]（一个AI

驱动的市场情报代理），达到了约27亿美元和47亿美元的峰值市值，以及约2290万美元和7.554亿美元的峰值管理资产 [697]。其他流行的CoinAlgs包括SingularityDAO [558]（提供“AI驱动的量化策略”用于DeFi）和Soldex [570]（一个提供AI驱动交易机器人的DEX）。

B-4.3.2 CoinAlgos的风险

虽然它们承诺使金融民主化，但CoinAlgos因其易受广泛研究的ML攻击而对其用户构成危害。对抗性机器学习领域的长期研究 [90, 248, 255] 表明，基于AI的系统容易受到包括提示注入 [255, 473]、内存注入 [468]、数据投毒 [245]、后门攻击 [133] 和模型提取 [610] 在内的攻击。这些攻击在实践中可能产生严重影响，例如修改系统行为或泄露有关其架构的敏感信息。它们对基于AI的CoinAlgos构成风险，从操纵投资决策到泄露专有交易策略。这些担忧并非严格理论上的；例如，在最近的工作中，Patlan等人 [468] 展示了“上下文操纵”攻击可用于触发针对ElizaOS的恶意资产转移。

除了对AI系统的一般攻击，CoinAlgos在用于金融时还引发具体担忧。例如，一些论文表明，AI交易代理可以以负面方式操纵或影响交易市场。最近Dou等人 [178] 的研究表明，即使没有直接沟通渠道，利润最大化的AI交易代理也可以相互协调和勾结，这可能导致市场低效率。类似地，其他工作展示了AI智能体如何操纵金融基准 [550]、欺骗限价订单簿 [653] 和逃避市场操纵监管 [654]。

最近Fabrega等人 [197] 的工作突出了投资者面临的一个更根本的问题。CoinAlgos在其设计中面临一个固有的安全权衡，称为**CoinAlg困境（CoinAlg Bind）**。直观地说，CoinAlg的交易策略要么可以是透明的，要么可以是（至少部分）私有的。两种选择都对投资者构成严重风险。透明的交易策略自然会以利润为代价，因为这可能导致策略盗窃甚至无风险的套利形式，如三明治攻击。因此，这促使将（有利可图的）CoinAlgos保持私有，并将其“秘密配方”隐藏于潜在竞争者。实际上，大多数CoinAlgos（在传统和去中心化金融中）选择私有交易策略 [197]。然而，这种替代方案开启了内部人员不公平提取价值的风险——私有的交易策略允许信息不对称，其中完全了解CoinAlg策略的内部人员（例如，其创建者或宿主）可以利用其特权信息从CoinAlg的交易中提取价值。在 [197] 中正式定义为公平性的这些私有CoinAlgos的风险类似于传统金融中的内幕交易。

Fabrega等人 [197] 通过两个互补的理论模型正式证明了CoinAlg困境的存在：一个比较了具有不对称CoinAlg交易策略知识的参与者的提取价值，另一个形式化了CoinAlg与预先了解其交易的参与者之间的互动。此外，他们通过在现实世界区块链数据上的广泛模拟实证验证了该困境。CoinAlg困境是一个固有的且不可避免的设计权衡，是未来部署CoinAlgs的核心挑战之一。

关键要点 B-4.2: CoinAlg困境 基于AI的集体投资算法（CoinAlgos）面临利润（需要投资策略的隐私）与公平（需要投资策略的透明度）之间的根本张力。

B-4.3.3 走向缓解措施

尽管存在风险，CoinAlgos是投资领域中不可避免的一部分，因此设计限制其危害的护栏是未来研究的关键问题。

在传统金融中，CoinAlg困境通过受监管的投资者保护得以规避，这使得CoinAlgs可以公开（因此有利可图），同时阻止内部人员提取价值（从而强制执行公平）。然而，像Web3这样监管较少的环境需要依赖技术对策来解决CoinAlg困境。特别是，由于利润损失的风险，Web3 CoinAlgs很可能会继续保持私有，因此最小化不公平价值提取风险的护栏尤为重要。当然，挑战在于此类护栏本身不应以牺牲CoinAlg的利润为代价。

问题 B-4.1: CoinAlg困境的技术护栏 什么技术机制可以在不降低CoinAlg利润的情况下，最小化不公平价值提取的风险？

Fabrega等人 [197] 探索了一些初步的护栏设计方向。最值得注意的是，他们提议使用所谓**随机化包装器（randomizing wrapper）**，这是一种在CoinAlg交易执行前对其交易进行随机化的（透明）算法。通过在私有的可信执行环境（即TEE）内运行CoinAlg，这种随机化可以使内部人员更难以预先预测CoinAlg的交易，从而降低他们利用特权信息获利的能力。此外，由于包装器是公开的，用户可以确保交易会被诚实随机化。未来工作的一个重要方向是对随机化包装器进行进一步和更原则性的研究，既包括理论模型以保证其安全性，也包括量化其在实际中效用的实证研究。

B-5 未来风险：AI驱动的恶意智能合约

赋予智能合约AI能力可以极大地扩展其应用范围，如B-4节所述。不幸的是，这种扩展的范围不仅包括好的应用，还包括恶意应用。这是因为智能合约被设计为人类和机构信任及冲突解决的技术替代方案。在一个使用智能合约而非人类流程进行冲突解决的系统中，潜在受益者包括那些拥有最不可信的人际关系和机构关系的人：犯罪分子。

AI补充的智能合约可能取代“盗贼之间的荣誉”的想法在 [318] 中提出。那里描述的恶意智能合约提供赏金/奖励以实施犯罪。（或者可以反过来：提供有偿犯罪服务。）

图B.5说明了这样一个合约如何工作。这里，一个恶意智能合约 RogueSC 被创建，提供金钱奖励 Sreward 来实施犯罪：破坏华盛顿纪念碑。

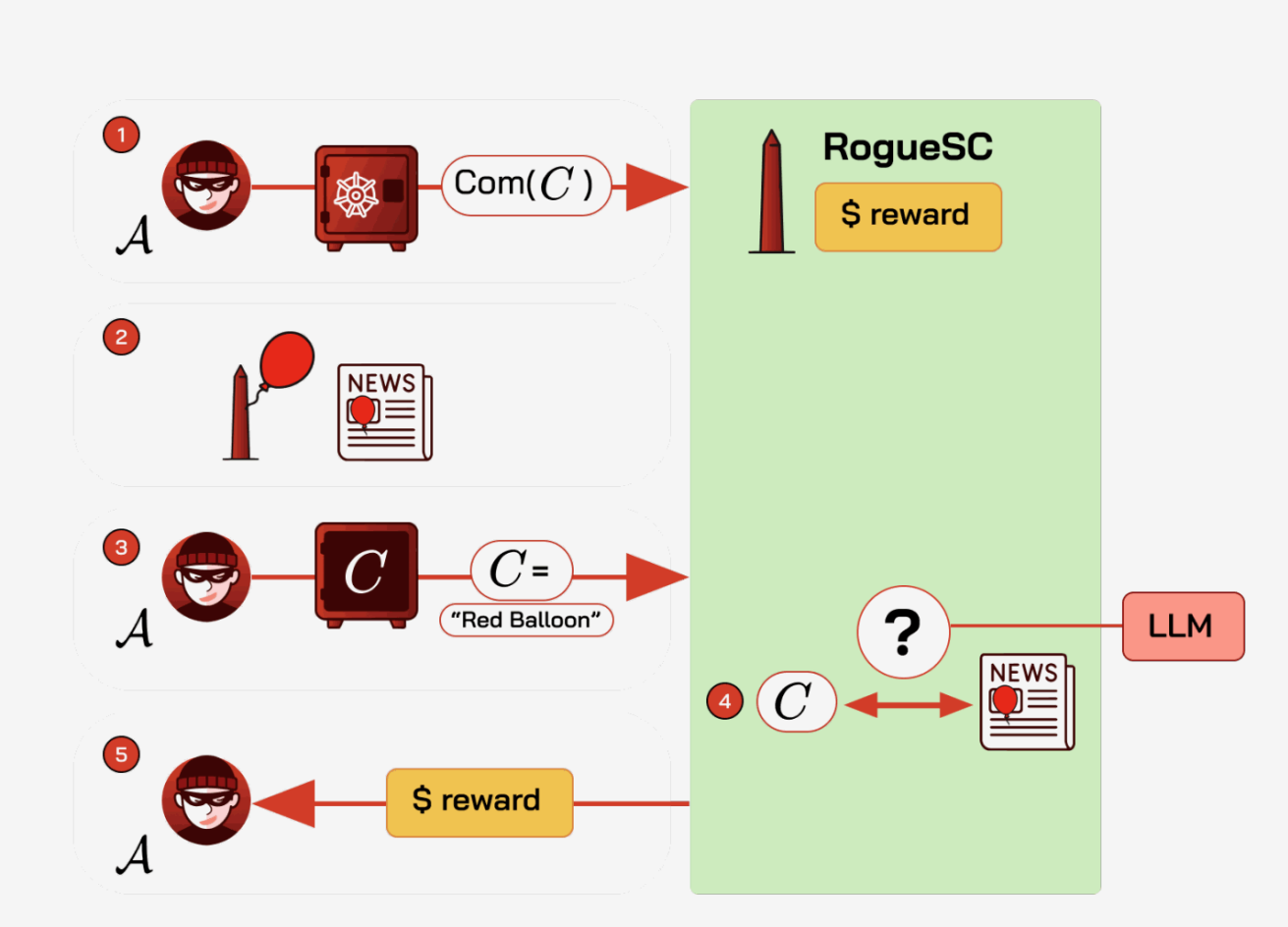


图 B.5 犯罪分子 A 通过在华盛顿纪念碑上系一个巨大的红气球来领取破坏纪念碑的赏金 Sreward。A 使用“名片” C = "Red Balloon"。步骤如下所述。

想要领取赏金的犯罪分子 A 当然必须向 RogueSC 证明他实施了所要求的犯罪行为。这样做所使用的协议利用了一种被称为“名片”（calling card）的东西，即犯罪的独特细节，将其归因于犯罪分子。其思路是让 A 在实施犯罪之前，私下承诺一个 C——他计划实施犯罪的细节。他在事后揭示 C，如果 C 与犯罪报告相符，就证明 A 负有责任并应获得奖励。因为，如果 C 选择得当，只有 A 可以事先知道它。

AI在哪里发挥作用？智能合约 RogueSC 必须在一个称为**裁决（adjudication）**的过程中确定 C 中的名片是否确实发生在物理世界中；在实践中，这可能来自新闻报道或其他可信第三方来源，它们报告关于犯罪的可注意细节。然而，这种裁决对智能合约来说并不简单，因为所涉及的数据不一定是标准化的或定量的。例如，如果犯罪是“破坏华盛顿纪念碑”且名片 C 包括“红气球”，智能合约评估犯罪是否发生，以及名片是否出现在犯罪现场，并不是微不足道的。最有希望的自动化裁决方式之一是使用一个由 RogueSC 调用的ML模型（如LLM）（在实践中，使用预言机）。

RogueSC 逐步工作如下。在有人创建了智能合约（在本例中，提供 Sreward 赏金以破坏华盛顿纪念碑）之后，犯罪分子 A 可以使用以下协议领取赏金：

1. **名片承诺：** A 选择一个名片 C，并向 RogueSC 发送一个密码学承诺 Com(C)。（承诺隐藏 C 但使其不可变。）
   - _运行示例：_ C = "Red Balloon".
2. **实施犯罪：** 犯罪发生并被新闻报道。
   - _运行示例：_ 一个巨大的红气球从华盛顿纪念碑上放飞。
3. **揭示承诺：** A 揭示 C，即向 RogueSC 透露 C。

4. **ML模型验证：** RogueSC 检查 C 是否与新闻报道相符。
    - _运行示例：_ RogueSC 询问LLM关于华盛顿纪念碑破坏事件的近期报道中是否提到了“红气球”。
5. **支付赏金：** 赏金 Sreward 支付给 A。
    - _运行示例：_ LLM回答“是”，验证了 A 的主张。

这种相同的结构可以应用于任何数量的犯罪，用于检查主张的信息不仅可以来自公共来源，还可以通过隐私保护预言机来自私有网络数据源。在这种情况下，该结构紧密类似于C-5.2节中描述的安全推理流程。鉴于其使用来源认证，它甚至不需要名片，如以下示例所示：

- **有针对性的骚扰：** 犯罪分子可以证明一系列骚扰性电子邮件交流（并在交流不是匿名的情况下过滤掉任何自我识别的信息）。ML模型可以评估该活动的有效性。
- **窃取组织情报：** 公司或其他组织的员工可以泄露企业内网的数据并匿名证明其来源。例如，促进内幕交易的情报可用于领取赏金（或通过智能合约出售）。同样的方法可用于知识产权、产品计划等。ML模型可以评估被盗情报并为其分配货币价值。
- **举报人曝光：** 有权访问举报人投诉的人可以通过被盗文件证明举报人的身份，ML模型验证支持证据的强度。

如果 RogueSC 可以匿名部署且资金难以追踪，那么教唆犯罪和实施犯罪的实体都可以匿名且有罪不罚地行动。隐私保护支付，无论是通过使用混币器 [98, 456, 520, 613] 还是更紧密集成的隐私技术 [423, 536]，在恶意合约方面都带有风险，正如它们在许多其他环境中一样。

对策。 通常用于打击加密相关犯罪的应对措施——用于去匿名化交易的链上分析、将受污染资金列入黑名单——可以作为恶意合约的有效对策，但需注意上述关于隐私的规定。

然而，还有一个针对恶意智能合约的额外重要对策。那就是部署ML模型的预言机实施AI安全措施，这意味着ML模型在有明显滥用风险的情况下拒绝服务。为此，需要指定请求的上下文，否则风险评估是困难的。例如，评估新闻文章是否提到“红气球”在我们不提供上下文的情况下并不明显具有恶意意图。提供目标逻辑（智能合约代码）给预言机请求可能会揭示威胁。

当然，与所有AI安全措施一样，存在误报和漏报的风险。

同样特定于智能合约环境的可能性是，某些实体可能建立一个恶意预言机服务。如果这样的服务实现了令人向往的加密属性——可信赖性和抗审查性，它将使恶意行为者能够绕过实施了安全机制的预言机系统。

B-6 结论与未来方向

总的来说，在利用AI辅助区块链算法构建性设计方面已经进行了大量研究。它可以用于设计应用本身，如DeFi市场结构的设计。它也可以用于设计与给定智能合约相关的算法，例如攻击智能合约或在市场中最大化可提取价值的策略。今天研究文献中的方法使用各种ML工具，从非常基本的分类器和回归模型到使用RL设计优化给定奖励函数的算法的更先进方法。在过去3-4年中，我们明显看到使用RL为区块链目的设计算法的论文有所增加。一个统一的观察是，所有这些方法都侧重于我们可以对环境进行建模的场景，无论是显式还是隐式。

关键要点 B-6.1 在研究社区中，AI辅助的区块链算法设计到目前为止，一直侧重于我们可以清晰地对世界或环境进行建模的场景。示例包括对共识协议状态空间进行建模，或对价格对给定AMM的影响进行建模。

在给定模型内操作的一个结果是，AI辅助的安全分析到目前为止主要是**构建性**的：从共识协议到智能合约，整个栈都在搜索具体的攻击。这种分析可以表明系统不安全，但不能证明其安全，后者仍然是穷举技术（如形式验证）的领域。这暗示了一个互补的方向。

研究问题 B-6.1 AI能否支持在整个区块链栈上进行可验证的、穷举的安全分析，而不仅仅是构建性地发现攻击？特别是，除了在给定模型内操作，AI能否帮助找到可以证明协议或合约安全的模型或抽象？

尽管AI辅助算法设计在研究文献中趋势明显，但在实践和行业中正在出现一种截然不同的趋势，如B-5节所述。随着AI辅助编码变得普遍，我们预计智能合约将由AI大量编写，可能借助智能体框架 [13, 189]。事实上，最近的一个智能体工作流能够设计出在一个智能合约基准测试中提取超过460万美元的漏洞利用 [681]。虽然用于设计这些攻击的代理是闭源的，但它们依赖于通用ML模型，这些模型并非专门针对漏洞利用提取进行调整，除了提示之外。越来越依赖通用基础模型的新一类工作流有几个影响：

- 我们将越来越多地不仅设计区块链算法，而且用AI设计完整的实现。在实践中，算法设计与实现之间将存在多大程度的分离还有待观察。算法设计将越来越多地由自然语言目标驱动，而不是精确定量的奖励函数。例如，如果用户要求AI“编写一个程序，从给定的智能合约中为我赚钱”，则没有明确定义的奖励函数。代理必须决定遵循什么目标以及遵守什么边界（如果有的话）来尝试为用户赚钱。与先前AI辅助算法设计方法（如上述关键要点）不同，新兴的AI辅助方法将越来越多地为未被充分理解或建模的环境设计算法和代码。

这些影响对区块链研究和工业社区提出了几个问题。

问题 B-6.1: 下一代AI辅助设计 随着我们从高度定制、劳动密集型的ML辅助算法和 analytics 设计转向基于自然语言目标的代理程序设计，对区块链算法和应用的有效性和安全性将产生什么影响?

AI辅助算法 vs. 定制的AI设计算法。 这方面的一个具体研究问题是评估代理设计的算法与先前针对特定问题领域定制的算法相比如何。

研究问题 B-6.2 在下游任务（如区块链 analytics 或漏洞利用设计）上，针对特定任务和数据类型定制的现有经典AI算法与由通用FM支持的代理设计的算法（即，除了通过提示设计外，未明确针对下游任务进行调整）相比如何?

例如，在欺诈检测领域，文献中现有的分类器都是在相对较小、精心策划的交易语料库上训练的。依赖于前沿FM的方法则从更大——尽管未标记且可能不相关——的公共信息语料库中学习。评估这些方法在相同保留测试集上针对各种下游任务的比较将是有用的。虽然我们预计定制算法在某些狭窄环境中会表现出色，但先前的ML-for-blockchain论文通常对环境做出假设（例如，数据来自特定分布，或环境在实验期间保持静态）。因此，了解先前方法与通用代理方法在由于环境变化导致的分布偏移下的比较将很重要。可以对文献中先前的基于RL的方法（如MEV提取或漏洞利用生成算法）运行类似的实验。

更好地利用通用FM。 在先前的研究问题中，我们要求对定制AI算法和由通用FM支持的代理驱动算法进行直接比较。然而，存在一个中间地带：FM可以针对区块链特定任务和数据类型进行微调。一个更广泛的问题是，如何最好地利用通用FM进行区块链特定任务。

研究问题 B-6.3 对于区块链 analytics 和设计，设计者应如何最好地利用预训练的FM和小型下游标记数据集?

今天，ML社区中将FM适应下游任务的主要范式是RL微调，其中预训练的LLM在代表所需技能的下游任务或数据集上进行微调 [452, 494, 542, 547]。这些方法通常对一个提示产生几个响应，然后使用各种方法对输出的质量进行排名或比较。然后使用产生的排名来微调基础FM，以将其导向更好的输出。然而，现有的RL微调算法都是为通用适应下游任务而设计的。因此，了解是否存在可以利用的区块链特定属性来进行RL微调，甚至持续预训练 [333] 将很重要。例如，在区块链环境中，如何评估FM对给定提示的响应质量，甚至如何设计提示本身，目前尚不清楚。对于 analytics 问题，提示应该要求预测交易是否欺诈吗？应提供什么上下文？这样的问题可能高度依赖于任务，我们认为可能会有丰富的工作探索将通用FM适应区块链空间感兴趣问题的不同方法。

AI驱动的安全战争。 Web3网络安全是整个企业安全的一个重要风向标。这是因为Web3漏洞利用通常可以立即货币化，要么通过在野外利用它们——从而不可逆地窃取资金——要么通过利用百万美元的漏洞赏金进行负责任的漏洞披露。这为所有参与者创造了巨大的经济激励，以尽快利用不断发展的模型能力，只要这些能力证明在经济上可行。

尽管如此，预测安全态势将如何演变具有挑战性，早期信号也不确定。在学术文献中，最近的AI网络安全基准测试（包括 BountyBench [706] 和 EVMBench [647]）评估了模型在整个漏洞生命周期中的能力，围绕检测、修补和利用工作流构建任务。实证结果表明这些能力可能发展不均。BountyBench报告了比漏洞利用更高的修补成功率 [706]，而EVMBench报告了相反的结果 [647]。值得注意的是，EVMBench观察到随着后续模型代次的出现，所有任务持续改进。总体而言，AI能力的发展轨迹仍不清楚，这留下了关于AI能力将如何影响攻击者和防御者之间力量平衡的重要开放性问题，特别是在短期内。这激发了一个重要的未来研究议程。

研究问题 B-6.4 不同的能力轨迹如何塑造Web3安全结果？哪些轨迹最合理，什么领先指标将标志着发展的方向？最后，随着能力的提高，有哪些干预措施可以保持强大的安全态势?

这些问题从根本上与安全经济学相关，可能需要开发经济模型来预测各种结果。总的来说，从以下角度研究AI军备竞赛将是有用的：如果AI在一段时间内给予一方（攻击者或防御者）不对称优势，这将如何影响企业安全的全球状态？我们应该如何改变漏洞赏金计划的结构以适应？网络安全保险风险计算将如何变化？总的来说，我们预计有许多有趣且重要的经济问题需要探索。相关研究问题包括对这些问题的建模，以及理解如何衡量当前状况；正如我们之前提到的，来自各个基准的结果可能不一致，因此理解哪一方在这场军备竞赛中“领先”可能具有挑战性。然而，区块链提供了一个有趣的机会，因为智能合约和交易在无需许可的加密货币上是公开的。因此，可能有机会直接衡量各种AI进步对网络安全攻击频率和规模的相关性影响。

总的来说，研究社区已经证明，AI辅助方法对加密生态系统非常有用，无论是在理解现有系统方面，还是在于设计和与新区块链系统交互方面。尽管如此，我们预计由于新技术和工具的出现（特别是基于最先进生成建模的技术），未来将有显著的增长和改进。这些机会的确切性质尚不清楚，它们可能带来好处和重大风险（特别是对网络安全）。无论哪种方式，区块链行业和研究社区都将被迫适应。

第三章 C

AI × 加密：用加密增强AI

C-1 概述：使AI工作流更加去中心化和可信

AI模型通过一系列通常被称为**AI生命周期（AI life cycle）**的事件实现。这个生命周期有许多变体，但许多变体明确包括数据收集、模型训练、模型验证、部署、维护和治理（图C.1）。

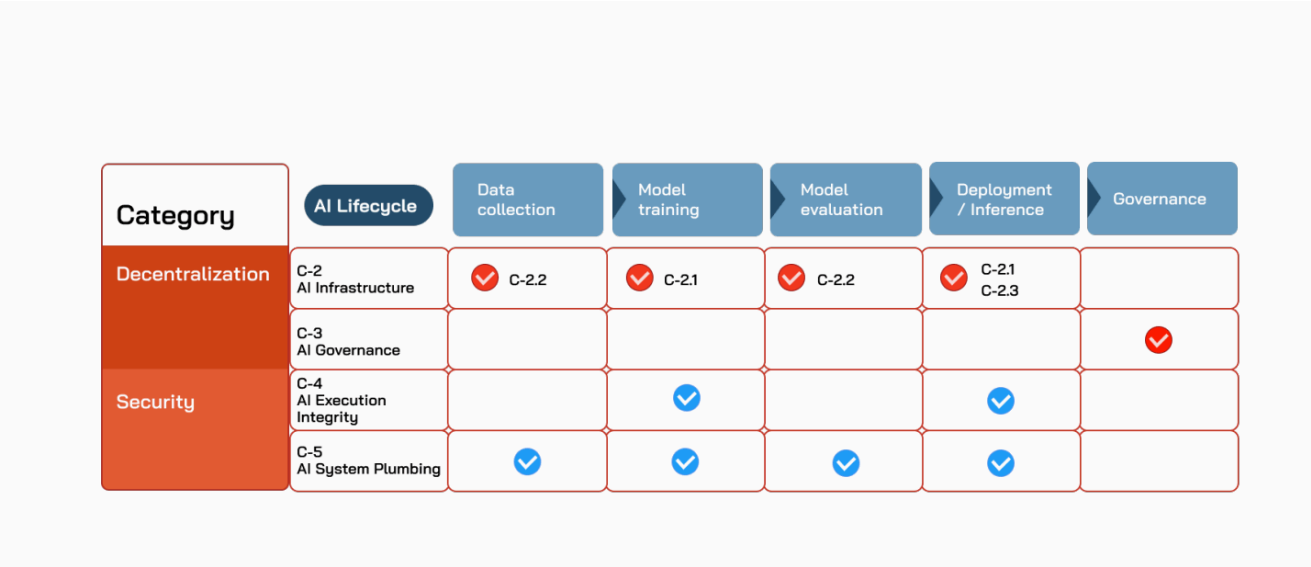


图 C.1 AI生命周期。我们设想区块链和相关技术可以通过两种主要方式影响AI生命周期：通过去中心化组件，以及通过更大程度地保护它们。

我们认为，加密——包括相关的底层技术——为改变或改进AI生命周期提供了两个关键渠道。**第一**，它可以帮助去中心化AI工作流的各个组件，从数据收集到模型训练、推理再到治理。原则上，这有助于民主化AI模型的开发和维护。在本综述中，我们还将考虑去中心化的更务实的效果，如成本和协调。**第二个机会**是加密可以帮助保护AI工作流的组件，使其不易受数据、计算或存储的恶意提供者的影响。在这里，我们将讨论来自加密领域的技术如何非常适合解决在AI生命周期中出现的某些安全漏洞。本章将探讨这两个互补的机会。具体来说，我们将本章分为两组章节：探索去中心化的章节和探索新安全能力的章节。我们在图C.1中总结了本章各节及其在AI生命周期中的位置。

1. **去中心化：** 这些章节将探讨加密如何帮助去中心化AI生命周期的各个组件，更重要的是，为什么这很重要。具体来说，我们涵盖以下类型的去中心化：

(a) **AI的去中心化基础设施：（C-2节）** 本节探讨可以使用加密领域工具和技术去中心化的几种AI基础设施。这些包括物理基础设施（如计算）、虚拟基础设施（如数据处理流程）和AI智能体的应用级基础设施。 (b) **去中心化治理：（C-3节）** 本节探讨去中心化AI系统治理的努力。我们探讨已提出的技术，以及一些权衡和核心挑战。

2. **安全：** 这些章节将探讨加密如何帮助保护AI生命周期的组件。具体来说，我们探讨以下领域。

(a) **用于AI执行完整性的区块链：（C-4节）** 本节探讨如何使用密码学工具确保AI执行——无论是在推理还是训练中——对下游用例是可信和可验证的。 (b) **保护工作流：（C-5节）** 最后，我们探讨一个通过谨慎使用可信计算技术来确保整个AI生命周期数据可信性、安全和隐私的框架。

更一般地说，我们关于使用加密保护AI的工作的一个关键信息是，保护AI将需要对运行环境的系统级知识和控制 [80]。

关键要点 C-1.1 AI安全是一个系统级属性。它不仅取决于模型，还取决于模型运行的系统环境。加密生态系统产生了多种工具，可以减少关于操作环境的不确定性。

虽然加密工具不一定适合处理模型级的不当行为（例如，幻觉），但它们可能特别适合帮助确保AI工作流运行环境的一致性和可信性，正如我们在C-4和C-5节中所探讨的。

C-2 AI的去中心化基础设施

今天，AI行业正变得越来越集中 [668]。模型提供商开始内部化数据采集和处理 [562]，开发自己的算法 [509]，在自己的数据中心训练和运行推理 [122]，并以专有提示代理的形式开发集中托管的应用 [8]。区块链提供了一个有吸引力的机会来去中心化这个生命周期的各个组件。例如，如果社区可以在消费级硬件的分布式网络上集体训练AI模型，会怎样？如果数据可以通过去中心化市场从各方有机获取，会怎样？如果基于AI的应用可以在不需要集中监督的情况下交互和发展，会怎样？希望这种去中心化能够降低整个AI社区的成本并提高透明度。然而，在实践中，这些考虑是微妙的，并且可能取决于许多因素和假设。在本节中，我们讨论去中心化AI基础设施的当前状况和影响。我们在C-2.1节讨论去中心化物理基础设施，在C-2.2节讨论去中心化数据和模型市场，并在C-2.3节讨论去中心化以智能体为中心基础设施。

C-2.1 去中心化物理基础设施网络

去中心化物理基础设施网络（Decentralized Physical Infrastructure Networks，DePIN）在AI领域越来越受欢迎。简而言之，DePIN指的是去中心化网络，其中节点可以提供物理基础设施（如能源、计算或带宽）以换取经济激励 [14, 92, 240, 274, 601, 682]。虽然过去有许多这样的努力（例如，Folding@Home [220]），但DePIN的特点是利用Web3基础设施进行经济补偿。具体针对AI市场，一些相关的DePIN努力包括分布式计算节点网络（例如，Theta Network [601]、Akash Network [14] 和 io.net [294]，仅举几例）。这些努力允许AI从业者按需从全球提供商网络租用CPU和GPU资源。除了原始市场本身，一些努力侧重于DePIN市场的基础设施；例如，Bittensor [92] 是一种在统一代币下定义新DePIN商品市场的语言。

面向AI的DePIN网络通常在价格上与传统云服务提供商竞争；例如，在2025年末，Theta网络宣传“按需企业级Nvidia A100/H100及更高，节省50-70%成本” [601]，Akash网络报告相对于AWS平均节省70-85%成本 [304]。DePIN计算网络的主要缺点是机器之间的吞吐量和延迟降低：由于DePIN基础设施本质上是去中心化的，机器之间的通信通常通过公共互联网进行。

对于考虑使用DePIN计算网络的AI从业者来说，一个自然的问题是“DePIN基础设施对我的任务有益吗？”当然，答案取决于用例和每个任务的特点。例如，即使DePIN网络上的每个节点计算基础设施比传统云服务提供商便宜，如果任务需要节点间大量通信，端到端墙上时钟时间仍可能在DePIN网络上长得多。另一方面，某些用例可能出于非技术原因需要地理分布式训练，例如：（1）隐私法规限制数据流出特定国家 [636]，（2）减轻在特定数据中心区域训练模型的环境（例如，水、能源）成本 [371, 532]，和/或（3）民主化基础模型的训练 [699]。在这种情况下，DePIN网络可以促进分布式（联邦）训练方法。

在下文中，我们为两类主要AI工作负载提供一些考量：训练和推理。对于每一类，我们讨论工作负载的延迟和吞吐量要求，并解释这些特征如何与DePIN架构交互。对于训练和推理，成本随模型大小而变化；具有数万亿参数的基础模型 [5, 78, 652] 与可以小得多的语言模型 [560, 716] 甚至经典ML模型（例如，支持向量机、线性回归等）之间存在巨大差异。所关注的场景将强烈影响在DePIN计算网络上训练的可行性。我们主要从效率和成本角度讨论这些属性；我们不进一步讨论可能倾向于分布式训练的非技术原因，如隐私法规。

C-2.1.1 用例：AI模型训练

模型训练大致可以分为三个阶段：

- **预训练（Pretraining）**：从随机权重开始对模型的初始训练。
- **后训练（Post-training）**：为额外能力（如指令遵循、推理或数学）更新或对齐基础模型的过程，通常使用各种奖励模型，可能通过征求人类偏好获得。
- **微调（Fine-tuning）**：为下游任务（可能在专门或私有数据集上）更新预训练或后训练模型的权重。

总的来说，预训练往往比微调和后训练更昂贵，因为它在更大数量的训练数据上进行，而后训练和微调在较小数据集上进行，也可能利用**参数高效微调（PEFT）** [266, 414] 等技术。因此，我们将重点关注本节中的预训练。

AI模型训练（尤其是大型模型）需要复杂的数据管理。反向传播（大多数ML优化器的构建块）必须计算和存储在大批量数据上平均的巨大梯度。由于硬件限制，这些操作在单个GPU节点上是不可行的，因此训练通常分布在多个计算节点上 [183, 652]。此外，大多数训练算法基于小批量随机梯度下降及其变体，这需要网络中所有节点的同步更新。因此，如果任何单个节点发生故障或集体通信等待落后者节点的更新，训练可能会停滞。

吞吐量考量。 在多个GPU上训练模型需要并行化计算的策略。今天，突出的GPU并行模式包括 [183, 210]：

- **数据并行（Data parallelism）** 是最简单和最常见的并行形式；它将数据拆分到不同节点，每个节点计算小批量中一部分样本的梯度。然后跨节点对梯度计算进行平均。
- **流水线并行（Pipeline parallelism）** 涉及顺序拆分层，使得不同节点处理不同层。它引入了对层处理的紧密顺序依赖。
- **张量并行（Tensor parallelism）**，也称为水平并行，在节点间拆分单层内的操作。每个GPU处理张量的一部分，操作结束时同步结果。
- **专家并行（Expert parallelism）** 用于具有多个专家子模型的混合专家（MoE）模型。不同的专家在不同的（组）节点上训练。
- **上下文并行（Context Parallelism）** 用于处理具有大量令牌的长上下文模型。每个GPU处理计算的一个分片，沿序列中的令牌进行划分。

这些并行化策略导致大量数据传输，用于梯度和优化器状态的通信和聚合，通信成本随模型大小、训练数据集和硬件平台的增加而增加 [397]。然而，不同并行化策略的吞吐量概况可能不同。例如，流水线并行和张量并行往往具有重通信足迹，需要高带宽链路。如果没有仔细管理和同步，所有这些并行技术都可能经历显著的扩展挑战 [405, 427]。

为了支持训练期间的同步和数据传输，大多数多GPU训练平台具有支持本地GPU之间数Tbps的高带宽互连（例如，NVLink）[26, 442, 652]；我们将同一机架中的此类GPU组称为**纵向扩展网络（scale-up networks）**（通俗地称为高带宽域）[236]。除了这些本地连接，平台还依赖以太网或Infiniband等网络技术来连接不同平台的网络接口卡（NICs）。**横向扩展网络（scale-out networks）** 连接跨机架的GPU，**跨域网络（scale-across networks）** 连接数据中心 [236]。横向扩展和跨域网络中的互连往往是带宽较低的，并且是LLM训练中的瓶颈 [210, 397]。值得注意的是，Fernandez等人 [210] 发现，当在分布式GPU节点上训练LLM时，随着计算节点数量和计算能力的扩展，系统变得越来越通信受限，GPU利用率下降最快，特别是对于高端节点如NVIDIA H100。这是训练AI模型在DePIN计算网络上的一个重要考量。

技术解决方案。 已经有许多努力在分布式基础设施上训练基础模型，包括使用商业级网络连接。例如，在2022年，Yuan等人 [699] 展示了一个在考虑计算节点之间平均延迟和带宽的同时优化通信成本的系统。此类指标并非由大多数DePIN网络直接提供，但可以在训练之前（或期间）进行测量。SWARM并行 [525] 后来被提出作为一种流水线并行方法，用于在具有慢速互连的不可靠硬件上训练ML模型。他们的关键思想是基于当前网络和设备条件随机更新哪些节点训练给定层。[525] 发现，随着模型规模的扩大，计算相对于通信成本占据了训练时间的更大比例。

为了解决训练步骤必须同步处理的要求，联邦平均方法（例如，Local SGD、DiLoCo）在每个节点上执行固定数量的本地梯度更新，并定期在模型实例之间平均结果模型 [180, 582]。最近的工作提出了解耦动量优化（DeMo），一种修改动量更新以减少加速器之间通信而不显著降低下游模型性能的优化器 [470]。尽管这些方法有前景，但这些研究论文在（相对）小模型（最多3B参数）上评估了他们的想法 [180, 470, 525, 699]；仅凭这些论文，尚不清楚这些方法在更大规模的前沿模型上是否有效。一些行业努力正在更大规模上测试这些想法（见下面的行业努力）。

更广泛地说，存在在硬件互连级别优化跨高带宽域通信的技术。例如，轨道优化网络 [397] 指定了如何将NIC连接到交换机，并常用于高性能计算工作负载。然而，它们仍然需要高带宽域之间的所有通信，这可能对DePIN网络不切实际。仅轨道网络 [652] 则针对LLM训练模式进行了优化，并去掉了典型GPU集群中的脊层。这两种优化 [397, 652] 都需要对物理基础设施的控制，可能更适用于数据中心（横向扩展网络）而非完全去中心化的DePIN网络。

也存在减少高带宽域之间通信的算法技术。例如，NVIDIA NeMo框架利用分层AllReduce框架、分布式优化器架构和分块数据中心间通信，以最小化通过广域网（WAN）传输的梯度的数量和大小，他们在其上训练了原型Nemotron-4模型（340B参数）[48]。这些方法可以减少通信成本，可能对在DePIN基础设施上训练有用，但我们缺乏经验数据来确凿地证明这一点。Nemotron-4的测试设置与在典型DePIN集群上训练的模型非常不同，在DePIN集群中，许多节点可能地理上分布在远多于两个区域的许多地方。

行业努力。 最近，像NVIDIA这样的提供商开始支持跨数据中心的ML模型分布式训练 [48]，这已通过训练Nemotron-4（一个340B参数的LLM）得到验证。这种在广域网上训练大型AI模型的努力引发了与DePIN计算网络遇到的问题类似的问题。NVIDIA部分通过最小化跨域网络上的通信来解决这些问题 [48]，这在分布式训练案例研究中更为普遍。

Crypto x AI行业也致力于展示在完全去中心化基础设施上训练基础模型（在较小规模上）的可行性，包括加速器之间的商业级网络连接。到2024年8月，Macrocosmos AI已在Bittensor网络上促进了700M和7B参数LLM的训练 [393]。这项工作的领先模型在网络文本和相关数据的困惑度方面被证明与GPT2-Large和Phi-2具有竞争力。他们的白皮书侧重于激励矿工提交高质量模型权重的机制，矿工被期望单独生成训练好的模型，因此我们之前讨论的网络效应并未发挥作用（至少没有可测量地）。其他努力包括

Prime Intellect在2024年11月通过分布式基础设施使用完全分片数据并行和DiLoCo优化异步训练了一个10B参数模型（Intellect-1）[311]；以及Templar AI使用称为Gauntlet [378] 的激励系统同步训练了一个1.8B参数模型。

正在进行的努力正在推向更大的规模。Macrocosmos AI提出了一种架构，多个矿工可以使用流水线并行和DiLoCo协作训练模型，同时因个人贡献获得奖励 [492]。在撰写本文时，还没有关于使用这种架构完全训练的LLM的端到端训练成本或模型质量的公开结果。据我们所知，正在DePIN基础设施上预训练的最大模型是Consilience，一个在Psyche Network上预训练的40B参数模型 [596]。Psyche Network使用DeMo进行数据并行配置中的通信高效优化 [470]；Nous Research报告称，40B“足够紧凑，可以在单个H/DGX上训练并在3090 GPU上运行” [596]。然而，目前尚不清楚如此大规模训练运行的基础设施可以在地理上分布到什么程度。DeMo论文明确指出，“DeMo主要设计用于在少量地理分布式计算中心之间进行优化” [470]。

关于现有行业概念验证的一个更广泛的观点是，它们通常不报告总成本指标（在计算和通信成本方面）。如果去中心化ML社区旨在民主化大规模基础模型的训练，那么标准化成本模型并比较中心化与去中心化基础设施上的端到端训练成本（以美元计）将是有用的。如果此类测量的结果对去中心化网络有利，这可能是该行业的主要卖点。同时，我们预计此类结果将在很大程度上取决于模型大小、实现的训练并行类型以及训练加速器之间的去中心化程度（例如，多个小集群与真正异构的消费设备）。

关键要点 C-2.1: 用于AI模型训练的DePIN 在去中心化基础设施上训练大型ML模型有许多有前景的初步结果，特别是对于较小的模型（几十亿个参数）。然而，我们缺乏在地理分布式节点上（即，不在数据中心环境中）的清晰测量，突出显示随着模型规模、基础设施并行类型和程度以及地理去中心化程度的扩展，总体成本权衡。

延迟考量。 模型训练往往对延迟不如推理敏感，部分原因是训练是离线运行的。尽管如此，跨多个地理位置的分布式训练相关的延迟可能会造成问题，特别是使用需要所有节点同步通信的标准基于梯度的优化器（例如，Adam、SGD）时。正如Aubrey等人所述，“在跨多个数据中心管理计算时，开发人员必须应对高跨区域延迟（通常为20毫秒或更多），这在大型LLM训练期间可能在梯度更新和模型同步期间引入性能瓶颈” [48]。作为参考，截至2025年11月27日，AWS数据中心us-east-1和us-west-1之间的P90延迟约为73毫秒 [49]。NVIDIA NeMo框架通过使用如分层AllReduce等技术最小化跨数据中心的高延迟通信操作来解决此问题 [48]。此外，分布式优化器（如DiLoCo）可以进一步减少此类通信操作的频率，可能提高跨DePIN训练的可行性。虽然DePIN计算网络原则上可以支持复杂的编排，但许多用户可能没有意识到这种需求。今天，DePIN服务部分通过允许用户按地理位置选择节点来考虑延迟 [14, 601]。这可以帮助用户评估特定节点组合的延迟影响，因为延迟受到两个节点位置之间光速的下界限制。

C-2.1.2 用例：AI模型推理

模型推理是指使用已训练好的模型生成预测，例如在模型被评估或部署用于下游任务时。例如，这可能包括在提供的数据点上评估分类器模型或从生成模型生成文本。

延迟考量。 总的来说，推理往往比训练对延迟更敏感 [21, 300, 367]。特别是在按需用例中（例如，AI聊天机器人），用户期望快速响应。然而，存在依赖模型推理且没有严格延迟要求的应用（例如，会议总结、文档审查、数据生成）[300]。值得注意的是，没有直接人机交互的新兴应用（例如，深度研究）具有宽松的延迟要求，可以在响应前等待长达30分钟 [449]。这些低延迟用例在节点提供商的位置方面可以更灵活。请注意，这些相同的考量适用于传统基础设施提供商；对于延迟敏感的应用，AI从业者应考虑处理查询的数据中心位置及其与最终用户的相对距离。

吞吐量考量。 推理的吞吐量要求通常低于训练。首先，推理通常需要较少的硬件并行度。而训练一个大型基础模型可能需要数万个GPU并行，推理任务则不需要同等程度的并行计算。即使数十个GPU也足以服务最大的模型（例如，DeepSeek-R1（671B参数）可以在16个H100 GPU上服务）。第二，推理不需要反向传播，这显著减少了操作的内存和计算足迹。因此，即使在像模型训练中讨论的那些并行化方案下，数据吞吐量较低的网络也不必成为阻碍。此类任务可能特别适合DePIN网络。

关键要点 C-2.2: 用于推理的DePIN 对延迟不敏感的推理应用（例如，会议总结、文档审查）可能是DePIN网络的一个特别有前景且成本效益高的用例。

C-2.1.3 总结

今天，最终用户根据自己的AI任务理解以及节点价格和容量（例如，计算、内存）来评估是否使用DePIN网络。我们强调以下主要要点。

关键要点 C-2.3: 成本-收益权衡 DePIN网络的用户应评估其用例的带宽和延迟要求，以确定总体预期成本节省。此外，我们建议DePIN网络提供商测量并至少报告节点之间的网络带宽，因为这些数字可以驱动总成本并有助于确定特定用例是否适合DePIN网络。

虽然当前网络通常以美元/GPU-小时的价格做广告，但这可能是一个误导性指标。对ML任务重要的成本指标通常是训练效率（即，每单位成本的迭代次数）和推理效率（每单位成本的令牌数）[700]。

为了使用户更容易确定是否以及何时使用DePIN基础设施，我们向研究社区提出以下问题。

研究问题 C-2.1 什么样的AI任务适合现有的去中心化计算网络，如DePIN AI网络中所见？

社区将受益于更系统的（第三方）测量来剖析现有的DePIN网络，并了解DePIN AI网络有意义的用例。我们设想此类研究首先测量各种DePIN网络的性能概况。然后，我们建议将AI任务的特征（估计计算成本、最大内存需求、通信负载、模型大小、层大小等）和期望的性能特征（成本、端到端时间）映射到关于在DePIN基础设施上训练是否能满足目标性能特征的推荐。虽然先前有关于ML模型去中心化训练的研究 [183, 525]，但这些研究并非专门在DePIN网络上进行，可能具有非常不同的性能特征。

研究问题 C-2.2 在RQ C-5.1中提出的系统测量研究的基础上，DePIN AI的定价机制应如何依赖于网络基础设施和可靠性？

定价不仅应依赖于每个节点的计算能力，还应依赖于网络测量和节点可靠性，这是合理的。例如，一组具有高带宽点对点链路的16个节点（例如，具有GPU到GPU NVLink的节点，或位于同一数据中心的节点）可能比分布在世界各地的16个节点成本更高。同样，能够保证高可用性以参与大型协作任务的节点可以定价更高，因为单个落后或故障节点足以中断整个AI训练任务。虽然一些网络隐含地奖励连接良好的节点（例如，Theta网络根据任务完成时间分配奖励 [600]），但其他网络仅基于节点容量定价。对于大型任务，这两种模式可能都过于事后，因为故障可能非常昂贵。

研究问题 C-2.3 DePIN AI网络的更新定价机制应如何考虑战略性和/或对抗性节点？

如果我们提出如RQ C-5.2中的更新定价模型，那么理解战略性和对抗性行为者的作用至关重要。例如，如果一个协议相对于单个节点容量不成比例地奖励连接良好的节点，那么单个节点的战略所有者可以将其租出为许多共处一地的小节点以提高利润。这可能对网络不那么有用——租用许多小节点的客户端比租用一个大型实例的性能更差。除了这种战略操纵，我们还需要设计方法来解释规格的对抗性报告。例如，如果一个节点系统性地谎报其处理能力或网络连接，应该有机制来发现这一点并惩罚提供商。设计适当的奖励方案需要机制设计，以确保节点提供商不能以与网络效用不一致的方式操纵他们提供基础设施的方式，或谎报其产品；后一个证明物理资源可用性的问题已在其他DePIN资源的背景下进行过探索，如网络带宽 [553] 和其他蜂窝网络资源 [31]。有趣的是，定义效用和威胁模型本身可能取决于研究问题C-5.1的测量结果。

C-2.2 数据、模型和评估的去中心化市场

今天，数据是AI生命周期许多阶段的重要输入：训练、推理时的接地（例如，检索增强生成（RAG）流程中）、模型验证以及针对特定领域需求（如能力或安全性）的基准测试。今天，每个这些阶段的相关数据通常来自多个来源。获取数据的一些最常用技术如下：

- **网络爬取（Web crawling）** 可能是AI工作流最常用的数据来源。公司爬取公共互联网以收集大量文本、图像和其他内容（例如，Common Crawl [107]）。有时，这些做法可能与内容提供商自己的政策相冲突 [563]。
- **与内容提供商的许可协议（Licensing agreements）** 越来越多地用于向AI模型开放付费墙或其他受限内容。例如，OpenAI与美联社（AP）签订了许可协议，以许可新闻故事 [444]。
- **公共数据集（Public datasets）** 长期以来由各方策划，并且是现代大多数训练流程的组成部分。示例包括Wikipedia、GitHub、arXiv和Stack Overflow。
- **合成数据（Synthetic data）** 越来越多地用于增强真实数据集，特别是在需要高度专业化数据的后训练过程中 [684]。
- **数据经纪人（Data brokers）** 是集中各方（通常来自专有或受控来源）聚合数据集并将其出售给买方的中心化实体。这些数据源并非总是专有或合法的——Reddit提起的一项诉讼 [445] 描述了一整个数据经纪人生态系统，他们从事“工业规模”的抓取并出售输出，违反了服务条款。虽然AI公司尚未明确确认从数据经纪人处购买数据，但已有多个指控和初步证据 [152]。此外，像LexisNexis [43] 这样的数据经纪人越来越多地发布产品，将高质量数据暴露于包括AI模型训练在内的声称用例，表明这是一个增长中的行业。

本节重点讨论最后一种数据经纪人。虽然AI拥有多种数据来源，但AI模型正在耗尽新鲜的公共数据 [314]；因此，对非公共数据源的需求日益增长，而数据经纪人很可能至少会在其中承担一部分中介作用。我们也会讨论采用度正在提高的相邻模型市场问题，但主要分析框架仍是数据市场。许多原则同时适用于这两类市场。

在企业环境中，数据经纪人通常以至少两种模式之一运营：直接面向消费者或作为市场。例如，LexisNexis将其数据列在企业数据市场（如Snowflake Data Marketplace [292]）上，同时也充当直接提供商 [43]。我们的主要兴趣将是市场运营模式。其他突出的中心化数据市场示例包括Datarade [161] 和 AWS Marketplace [544]，尽管数据付费是一个古老的概念，如在金融数据流 [413, 438] 和信息安全数据 [190] 中所见。

注意。 我们将在必要时区分运营中心化市场（由一个实体拥有和运营）和中心化定价（市场运营商单方面决定如何为商品定价）。市场可以是运营中心化但具有去中心化定价，反之亦然。我们在本节中讨论各种组合。

关键要点 C-2.4 AI提供商从许多来源收集数据。虽然其中许多来源要么是免费的，要么基于双方商业协议，但基于区块链的解决方案可以补充（或颠覆）来自当今数据经纪人运营的中心化数据市场的数据获取。

在随后的章节中，我们概述区块链相邻技术和去中心化协议如何在训练、适应甚至运行AI模型所需的数据市场中发挥关键作用。在C-2.2.1节中，我们首先概述了使数据市场区别于其他市场类型的属性。在C-2.2.2节中，我们突出中心化市场（特别是垄断市场）中出现的挑战。在C-2.2.3节中，我们简要概述使用加密工具的去中心化市场如何帮助解决中心化市场中的一些挑战。我们提供了加密领域现有去中心化数据或模型市场的概述。最后，C-2.2.4节提供总结和未来研究方向。

C-2.2.1 数据市场的属性

下面，我们概述了将数据市场与其他常见商品市场区分开来的关键特征。

- **数据是数字商品（Digital good）。** 数字商品首次创建成本高昂，但此后复制免费，并且在经济学和计算文献中得到了广泛研究 [243]。值得注意的是，虽然数字商品复制免费，但卖方必然限制复制以提取足够的收入来覆盖初始创建成本。
- **数据可以是竞争性的或非竞争性的（Rival or non-rival）。** 竞争性商品在销售给多个消费者时表现出负外部性，[c2-3] 而非竞争性商品可以被任意数量的消费者平等享用。[c2-4] Jones和Tonetti [313] 强调，非竞争性数据在被一方消费时不会“消失”（而苹果在被一个消费者吃掉时就消失了）。Gordon-Tapiero等人 [251] 强调了数据的竞争性方面——如果通过差分隐私 [186] 等隐私保护工具访问数据集，那么每个查询都会消耗“隐私预算”，并限制其他消费者的可用查询。在AI工作流的背景下，我们通常将数据视为非竞争性的，但有一些显著的例外（例如，上面提到的差分隐私预算）。
- **数据可以是“柠檬”（次品）（Lemons）。** 几年来，ML模型的进步是由神经网络随计算和训练数据量增加而质量显著提高的能力驱动的 [276]。然而，这种范式有其局限性。随着我们将越来越大的数据集输入ML模型，很明显数据量无法补偿数据质量差。组织过滤掉不相关或低质量样本 [241, 290, 645] 和/或进行其他数据清理任务（如标记）[353] 已变得普遍。[c2-2] 这表明AI从业者被激励（或应该被激励）为高质量数据付费 [327]。这对数据市场很重要，因为它们必须公开数据估值机制 [134, 645, 686, 722]。
- **数据估值需要数据访问（Data valuation requires data access）。** 数据质量，如二手车质量，对买方不可观察。所谓的“柠檬市场”在经济学中得到了广泛研究 [16]。解决这种信息不对称的典型方法是卖方提供关于待售商品的可审计信息。在销售数据时，通过简单地提供样本，这是可能的 [50, 134, 686]，正如在数据市场 [161, 544, 565] 中常见的那样。然而，数据样本的一个关键方面是，即使从未购买完整数据集，这些样本本身也提供价值（而用户从仅仅了解汽车功能良好中不会获得价值，除非最终购买汽车）。

C-2.2.2 中心化数据市场中的挑战

虽然中心化数据市场可以解决一些问题，但它们也容易受到与中心化技术平台相关的熟悉问题的影响。我们在下面总结这些问题。

- **市场力量（Market power）。** 单一实体对平台的决策拥有最终权力，当它们充当集中仲裁者时，可能会滥用其对市场的控制 [620]。例如，一个中心化数据市场可能单方面改变定价方案，限制竞争者访问，或向某些参与者提供优惠待遇，而不向其他参与者提供透明度。这种做法在美国诉谷歌案 [620] 中已被详细记录，并在学术研究中被讨论 [602]。即使在市场中存在竞争压力，如果所有主要参与者都在同一市场中运营，集中的市场力量仍然会显现出来。
- **定价和效率（Pricing and efficiency）。** Filippas等人 [212] 探讨了一个中心化平台何时可以从集中定价中受益，何时可以从让卖方自行定价中受益。他们表明，集中定价可以导致卖方收入增加和买方价格降低。然而，集中定价依赖于平台对交易商品质量的了解，这可能并不总是准确或可行的。中心化与去中心化定价之间的这种张力尚未在数据市场的背景下得到充分探索，但鉴于数据估值的挑战，它可能特别相关。

- **数据和AI市场中的网络效应（Network effects）**。正如我们之前讨论的，数据和模型市场受益于网络效应：更多买方吸引更多卖方，反之亦然。这种网络效应自然会导致“赢家通吃”的局面，并巩固上述市场力量。此外，平台运营商拥有特权访问平台特定数据，这可以改善他们的服务 [316]。他们在捕获更大用户群和更多用户数据之间创造了良性循环，进一步巩固了他们的市场地位。

一个更微妙的威胁与**非竞争性**数据有关。如前所述，数据通常是非竞争性的：一个人使用它不会阻止另一个人使用它。在一个集中的市场中，一个实体控制访问，因此，它可以限制访问，允许买方按使用付费。然而，它也创造了数据市场的运营商可能不公平地限制访问。此外，一个集中的运营商可能能够选择性地控制哪些方可以访问数据，从而决定数据市场的成败。一个特别相关的问题是**数据转售（data resale）**：如果一方可以访问一个数据集，然后简单地转售它，那么原始卖方可能被迫更积极地限制数据获取，防止转售。在中心化市场中，单个实体决定数据转售的限制，数据卖方和买方对限制如何制定几乎没有透明度和控制，例如，一个数据集可以出售给多少方 [251, 368, 603]。

C-2.2.3 去中心化数据市场如何帮助

与中心化市场相关的问题表明了对替代市场结构的广泛兴趣。不幸的是，Tirole [605] 指出，一旦垂直整合产品占据主导地位，将现有“大型科技”市场分解为较小的子产品存在若干障碍。具体来说，打破垄断网络（例如，社交网络）可能会显著降低服务质量，因为网络效应消失。即使在数据市场中也是如此：一个拥有许多卖方的市场对买方更具吸引力，而一个拥有许多买方的市场对卖方更具吸引力。此外，[316] 特别指出，拥有庞大用户群的平台倾向于积累关于用户的数据，这会复合其服务质量。

幸运的是，AI应用的数据市场仍在兴起，还没有一个主导市场。因此，Tirole [605] 提出的许多风险尚不存在。未来，如果数据市场将成为AI运营商的主要数据来源，那么在早期阶段正确设计这些市场非常重要，利用大型科技时代学到的任何经验教训。

关键要点 C-2.5 AI数据市场仍处于起步阶段，还没有（尚未）垄断者。因此，用于训练、推理和验证AI模型的去中心化数据市场有机会通过利用加密的结构和属性来避免与过去垄断市场相关的许多问题。

首先，注意去中心化数据市场很适合利用加密中常用的许多工具。一个非详尽的列表包括以下内容：

- **微支付（Micropayments）**。微支付可以实现数据转移和销售的新范式：数据买方可以选择他们需要的确切数据样本，并根据特定数据样本的效用付费。
- **可信执行环境（TEEs）**。也许数据持有者乐于让他们的数据用于特定任务，但不愿意普遍出售数据用于任意用途。可信执行环境可以允许出售数据仅在该环境内使用，因此是暂时的并绑定到特定任务（例如，在特定环境中训练特定模型）。TEE也可以用于其隐私属性，以解决共享数据样本的挑战。参见A-1.1节关于TEE的进一步讨论。
- **用于审计的零知识证明（Zero-Knowledge Proofs for Auditing）**。C-2.2.1节指出，数据可能是“柠檬”，但告知买方的经典方法存在买方尽管放弃购买数据集但仍从信息本身获得价值的风险。因此，零知识证明是卖方的一个自然工具，可以向买方披露关于数据集的精确期望信息，而不会冒无意使用的风险。TEE也可以用于此目的；例如，一个TEE可以验证特定数据集改进模型训练，而无需向TEE外部的任何人实际透露数据。参见A-1.1节关于零知识证明的进一步讨论。

基于这些工具和其他工具，我们突出几个关键机会，去中心化数据市场可以在这些方面创新——并可能改进——中心化替代方案。

- **透明的决策（Transparent decision-making）**。缺乏透明度和控制是中心化市场中的常见问题 [212]。去中心化替代方案可以保证协议和决策的透明度，包括用于数据定价的算法。透明度——以及由此产生的竞争——通常是去中心化数据市场所引用的主要好处之一。然而，透明的决策可能不是最好的决策（见下文）。
- **依赖于数据的、隐私保护的定价（Data-dependent, privacy-preserving pricing）**。尽管运营去中心化市场可以增加透明度，但我们之前看到，由于使用关于所售商品质量的更完整信息，中心化定价可以提高收入并降低用户价格。例如，中心化数据市场运营商可能知道各个经纪人提供的数据集的价值和内容，并能够相应地定价。然而，数据经纪人可能不愿意将其数据暴露给去中心化市场协议进行估值。隐私保护协议可以通过允许数据估值模块在隐私保护下访问其数据（例如，使用零知识证明或MPC构建）来提供帮助，从而导致更高效的定价，而不牺牲协议透明度。

- **用协议取代垄断平台（Replacing monopolistic platforms with protocols）**。Huberman等人 [286] 将去中心化账本分析为“没有垄断者的垄断”（monopoly without a monopolist）；虽然最终有一个具有巨大网络效应的单一账本（“垄断”），但没有单一方控制对该账本的访问（作为垄断者会做的），用户通过参与明确定义的竞争过程的矿工与单一账本交互。随后的工作 [366] 进一步分析了这种去中心化账本对用户的影响。这种思维方式可以被视为“协议，而非平台”（protocols, not platforms）[402] 范式的经济实例化。
  - 去中心化数据市场可以以类似方式演变，许多数据卖方在单个“垄断”协议的透明规则下运营和竞争。这种精神下的一个市场示例是Resonance [58]，它受以太坊的**提议者-构建者分离（Proposer-Builder Separation，PBS）**启发。[c2-5] PBS将提议以太坊区块的角色（具有最低技术要求，其去中心化是以太坊价值主张的核心）与构建区块的角色（需要深厚的技术复杂性来优化，可能由少数专业化实体主导）分开。Resonance [58] 是一种匹配AI计算买卖双方的机制，类似地将（技术简单的）协议执行与（技术复杂的）组合优化问题分开。[c2-6] “没有垄断者的垄断”视角提供了对Resonance的不同看法：尽管有一个所有买方/卖方聚集的单一“垄断”，但没有单一实体充当匹配买方与卖方的“垄断者”；经纪人反而在形式指定的机制中竞争，以赢得匹配买方与卖方的权利。
  - 更广泛地说，[508, 568] 将去中心化视为平台通过将控制权交给去中心化协议来承诺未来行为的一种方式，并研究平台设计者是否可能因此类承诺而获利（否则不放弃权力就无法可信）。作为一个假设示例，想象一个可以决定展示广告费率的平台。中心化平台将选择优化其利润的费率，并且缺乏能力可信地承诺做其他事情。这些论文考虑平台将决策权正式交给用户的可能性，以一种可信且可验证的方式。
- **行业努力**。开发者已经开始使用区块链相邻技术来促进数据和AI模型的交换。去中心化数据和AI模型市场的广泛但不完整的调查出现在表C.2和C.3中。讨论的平台大致根据它们是否出售数据访问权或训练好的模型进行分类，并说明每个平台如何设定价格。出现了几个趋势，我们总结如下。

Platform	What Is Being Traded	Details	How Prices Are Determined
Bittensor	**Models** AI model outputs (intelligence)	A network of *subnets*, each a competitive marketplace for a specific AI task. *Miners* serve model outputs; *validators* score quality. One subnet's output can feed another, enabling full AI pipelines within a single protocol.	*Proof-of-Intelligence* distributes TAO token emissions proportional to validator-scored performance. No list price—reward magnitude is the implicit price signal.
Grass	**Data** Web data scraped via bandwidth sharing	Users share idle internet bandwidth via a browser extension. The network scrapes public web data through residential IPs at scale, using ZK proofs to record provenance on Solana. Data is cleaned and sold to AI companies as training datasets.	Enterprise clients pay the Grass Foundation via off-chain contracts. Node operators earn GRASS tokens based on uptime and active bandwidth utilisation.
Hivemapper	**Data** Street-level mapping data	Dashcam owners contribute fresh imagery via a dedicated device, building a decentralised alternative to Google Street View. The network sells map access to enterprise customers (logistics, insurance, autonomous driving).	Enterprise buyers pay a subscription or per-tile fee set by the Hivemapper Foundation. Contributor HONEY rewards are algorithmically scaled by imagery freshness, uniqueness, and geographic demand.
IOTA Data Marketplace	**Data** IoT sensor data (micro-streams)	Built on the IOTA *Tangle* (a DAG, not a chain), enabling feeless micro-transactions. IoT devices—sensors, vehicles, smart city infrastructure—sell tiny data packets directly to other machines, making M2M data economies practical.	Device owners set asking prices; buyers pay in IOTA tokens. The feeless architecture enables sub-cent per-packet pricing impractical on fee-based chains.
Oasis Network	**Data** Private or sensitive data	A Layer 1 with confidential computing via *Trusted Execution Environments* (TEEs)—computation runs on encrypted data, invisible to node operators. Users tokenise personal data and stake it with dApps, earning ROSE rewards while retaining control.	Access terms negotiated via smart contracts between data owners and dApps. Users pay a premium for TEE privacy; token rewards come from network usage fees.
Ocean Protocol	**Data** Access rights	Datasets are tokenised into ERC-20 *datatokens*; holding 1 datatoken unlocks access to a linked dataset. An ERC-721 *Data NFT* represents base IP ownership. A *Compute-to-Data* (C2D) feature lets buyers run algorithms on data without it leaving the owner's servers.	Fixed price (set by publisher) or AMM-based dynamic pricing via *bonding curves*. Buyers pay for datatokens in OCEAN tokens; price adjusts automatically with supply and demand.
Oraichain	**Models, Data** AI oracle and data outputs	An AI oracle Layer 1 (Cosmos SDK). Smart contracts request AI-powered feeds—price predictions, biometric checks, credit scores. Each request attaches *test cases*; providers must pass a minimum threshold to be paid, ensuring output quality.	Fees split among validators, test-case providers, and AI API providers, paid in ORAI tokens. Providers set their own rates; reputation scores influence which providers are selected.
Sentient GRID	**Models** AI model usage rights	Models weights are freely downloadable, but *model fingerprinting* ensures outputs degrade without a valid authorization signature. Usage fees flow on-chain to trainers, deployers, and validators.	Owners set per-call fees; payment triggers the signature required for accurate outputs. The SENT token coordinates governance and rewards across the network.

图 C.2 部分去中心化数据和AI模型市场列表（第一部分）

Platform	What Is Being Traded	Details	How Prices Are Determined
SingularityNET	**Models** AI services (API calls)	An open marketplace of AI APIs (computer vision, NLP, robotics, etc.). Developers publish services on-chain; buyers call them via smart-contract escrow that releases payment only on delivery. Computation runs on the provider's own servers.	Each provider sets their own per-call price in AGIX tokens. Rates emerge from provider competition; no central authority controls pricing.
Story Protocol	**Data** IP rights (AI training data & creative works)	Creators register works (text, images, audio) as on-chain IP assets. Smart contracts encode licensing terms—permitted uses, royalty rates, and attribution. Royalties flow automatically when AI models train on or derive from registered assets.	IP owners set licensing fees directly in smart contracts (fixed or revenue-share). Prices discovered through supply and demand; popular or rare training data commands higher royalties.
Vana	**Data** User-owned personal data	Built around *DataDAOs* that pool user-exported personal data (Reddit, Spotify, health metrics) for sale to AI researchers. AI builders must *burn* VANA and DataDAO tokens to access data; 80% of fees flow back to contributors. Data is processed inside TEEs.	Buyers pay by burning tokens, directly linking demand to price. DataDAO rankings (refreshed every 3 weeks) determine VANA emissions—a competitive subsidy on top of direct access fees.

图 C.3 部分去中心化数据和AI模型市场列表（第二部分）

关键要点 C-2.6 虽然许多现有平台和协议使用去中心化机制进行支付处理（例如，加密货币或稳定币），但去中心化如何具体影响这些产品和市场仍不清楚。许多平台要么使用中心化定价机制（即，定价由协议设计者确定），要么允许卖方完全指定自己的价格——这两种定价变体在中心化市场中已经存在。总的来说，去中心化如何改善数据和模型市场的问题仍未得到充分研究。

虽然表C.2和C.3中的条目都运行在区块链上进行去中心化支付处理，但其中几个具有中心化支付规则。例如，Grass [252] 根据固定定价方案向用户支付闲置互联网带宽费用，并使用该带宽抓取公共网络数据（将结果出售给AI模型训练者）。Hivemapper [274] 也根据固定策略向数据提供者分发奖励。在研究领域，**改进证明（Proof of Improvement，PoIm）** [20] 已被提出作为一种评估ML模型提议更新并相应分发奖励的技术。

其他平台允许数据或模型提供者设定自己的价格，如IOTA Data Marketplace [295]、Oraichain [450]、Sentient GRID [543]、SingularityNET [559] 和 Story Protocol 的 Poseidon Marketplace [584]。Vana [331] 允许买方通过展示的需求间接设定价格。

一个有趣的中间地带是Bittensor [92]，它有多个子网，每个子网有不同的所有者。在子网内，项目所有者指定自己的激励机制。一旦模型提供者提交他们的产出，验证者对其进行评估，之后TAO代币按价值比例分配给模型贡献者。这种设置可以允许关于激励结构选择如何影响最终模型质量的受控实验。

总的来说，跨平台的定价机制有一些多样性，但我们观察到对不同定价机制影响的有限探索——特别是完全中心化定价和完全去中心化定价之间的混合体。探索这个设计空间是未来工作的一个有趣且重要的方向。

关键要点 C-2.7 对数据的隐私保护计算是现有去中心化数据市场的一个共同特征，尽管不同的提供者使用不同的工具和技术来处理这个问题。

几个现有平台利用可信计算（如TEE或加密计算）来管理私有或敏感数据（例如，Sterling [288]、Oasis [447]、Ocean Protocol [406] 和 Vana [331]）。它们的核心是，这些系统使用区块链和密码学来提供一个支付系统，专门用于支付经过验证的传输（无论是数据还是AI模型输出）[288, 331, 406, 450, 559]。各种系统在如何提供隐私方面有所不同；例如，Oasis最近开始使用差分隐私来混淆SQL查询 [446]，而Ocean Protocol允许数据提供者在自己的数据上本地执行计算 [406]。

C-2.2.4 总结

去中心化并非去中心化数据和模型市场的万灵药，但它在主导性AI数据市场出现之前引入的能力值得认真考虑。例如，“没有垄断者的垄断”视角 [286, 366] 提出：用户是否更好地服务于提供一站式数据购买服务的垂直整合市场，还是由竞争第三方提供某些衍生服务（如搜索、争议解决、定价、认证等）的去中心化市场？同样，“去中心化作为可信承诺机制”视角 [508, 568] 提出：这些新兴平台是否可能通过将某些平台决策的控制权交给去中心化治理来增加用户获取？

现有的平台和协议已经开始建立市场支付可以流动的去中心化轨道，但在市场设计和机制方面进行的实验相对较少，许多开放问题仍然存在。以下研究问题由上述讨论激发。

研究问题 C-2.4 在设计数据市场时，哪些衍生服务应与平台垂直整合，哪些应保留给竞争第三方？

“没有垄断者的垄断”视角 [286, 366] 建议考虑非垂直整合的市场，其中某些衍生服务由竞争第三方提供。但是，对这个框架的探索非常初步；它尚未提供关于哪些衍生服务应与垂直整合分离[c2-7]的建议，也未提供是否技术上可行以无缝分离运营一个功能市场所需的不同部分[c2-8]的建议。更系统地探索这个设计空间对于去中心化数据和模型市场的未来发展很重要。

研究问题 C-2.5 我们如何设计市场，以便AI工作流的重要辅助部分（基准测试、红队测试、微调等）能够得到激励和奖励？

同样，了解哪些AI邻近服务应与核心产品捆绑，哪些应留给竞争第三方也很重要。例如，红队测试通常被视为开发AI模型的副产品，待售数据集可能需要广泛的标记或数据清理过程。此类商品的买方可能期望在购买模型访问权之前进行某种压力测试，或在购买数据集之前进行数据清理。同时，我们预计这些过程会随着时间的推移而改变，因为攻击者和防御者的能力都在演变。因此，一个自然的问题是如何设计市场，以便除了核心AI工件之外，这些辅助功能也能得到激励。

C-2.3 去中心化、以智能体为中心的支付轨道和基础设施

AI智能体（AI agents）是以目标为导向的系统，能够采取自主行动，通常被实现为调用工具（如脚本、API和外部程序）的大型语言模型（LLMs）。智能体能力已经存在于广泛使用的消费者AI中：ChatGPT和Claude提供使用工具调用进行网络搜索的“深度研究”模式，并可以代表用户发送电子邮件和与外部系统交互。通过“计算机使用”模型（可以使用视觉和文本处理的组合来操作浏览器或操作系统），AI智能体也可以通过人类使用的相同接口进行交互 [539]。

在许多方面，智能体生态系统已经是去中心化的。代理可以由不同方开发，具有不同的底层LLM，并可以设计为优化不同的目标。最终，这意味着在一组代理之间没有自然的集中控制点。这种类型的去中心化在质量上不同于区块链，其中独立节点在固定协议的限制内运营和竞争。在智能体交互中，任何约束或护栏通常来自单个智能体或LLM提供商。这提出了一个重要问题：我们如何在没有共同规范、标准或目标的情况下设计可信的智能体交互？

在本节中，我们讨论利用加密相邻工具和思想来协调去中心化代理的不同方法、机遇和挑战。

去中心化的前景。 当人类互动时，他们绝大多数按照不成文且规定不明确的规则行事。例如，人类律师通常不需要被明确训练，幻觉引用会削弱原本有效的论点，因为他们在法律培训之前通过许多未指定的生活经历自行学习。

尽管AI安全专家尽了最大努力，AI智能体可能不会（也可能永远不会）遵循与人类相同的所有不成文规则。例如，AI律师缺乏人类律师的生活经验，因此必须明确训练以匹配文化规范 [156]。

为了理解去中心化技术如何帮助，让我们首先转向传统的法律合同。传统的经济互动通常由规定不明确但具有法律约束力的合同促成。这个概念使得在事情大致按计划进行时能够实现无冲突互动，而在合同规定不明确的情况下，一个健全的法律体系解决冲突。中心化平台上的经济互动也可以通过依赖平台在出现问题时介入来承受规定不明确。密码经济协议（Cryptoeconomic protocols）旨在在最小依赖法律系统、中心化中介甚至规定不明确的承诺概念的情况下促进重要的经济互动。相反，这些协议旨在从“密码经济安全”中尽可能多地获得——严格规定的密码学保证使不良行为在数学上不可能，以及经济激励使不良行为成本高得令人望而却步。类似的原则可以使寻求约束个体智能体行为的智能体系统受益。

关键要点 C-2.8 传统上，加密货币通过密码经济协议管理智能体行为不当的风险，这些协议通过密码学约束和惩罚不良行为的经济机制相结合来防止恶意活动。类似的想法可能对智能体生态系统有用，其中代理被设计为理性的。

当然，AI智能体和密码经济协议都必须与周围的法律系统互动。关键是，AI智能体对规定不明确的规则的反应不如人类灵敏，[c2-9]而密码经济协议旨在尽量减少对这类规则进行中心化执行的依赖。

过去几年，最先进的密码学为了达成去中心化协议的目标而取得了显著进展，因此也可以用于AI经济生态系统。

智能体网络可以降低服务提供商与消费者之间的摩擦，避免对中介的需求，从而实现区块链的去中心化目标。像市场和经纪人这样的中介平台之所以存在，是因为发现和建立信任是困难的。如果AI智能体可以自主处理经纪和发现，这些中介就变得不那么必要了。

服务之间的互操作性历来是困难的，需要定制的适配器和预先协商的集成。但如果代理可以通过在需要时动态生成集成（例如，通过解释API文档并在飞行中生成适配器），那么就不那么需要留在围墙花园内。

C-2.3.1 智能体经济体的加密工具

首先，我们提供一份密码学原语列表，这些原语已在去中心化协议中找到应用，以促进信任最小化的互动。我们解释这些工具如何能用于AI经济生态系统中的类似目标。

- **工作量证明/权益证明（Proof-of-X）。** 比特币的工作量证明和以太坊的权益证明作为密码经济工具，在没有身份的情况下强制执行问责制。
  - _为什么密码经济协议使用它？_ 在传统的经济互动中，身份概念可能非常有用（例如，确保“一人一票”的概念，或在出现问题时让某人承担财务/法律后果）。许多密码经济协议旨在无需许可，没有国家授权的身份概念可利用。这带来了几个挑战，其中最显著的是Sybil攻击。Proof-of-X通过用“一CPU/一币一票”的概念取代“一人一票”的概念来防御Sybil攻击，并使得在协议中过度代表自己变得昂贵。
  - _它可能在智能体经济中扮演什么角色？_ AI智能体可能会利用国家授权的身份（例如，你可能为自己的AI智能体的行动承担财务和法律责任）。然而，如上所述，AI智能体对财务、法律和声誉后果的反应不如人类直接灵敏。[c2-10] 因此，重要的是使不良行为明确付出代价，而不是仅仅依赖规定不明确的后果。
- **可信执行环境（TEEs）。** TEE可以替代“我承诺逐字运行以下代码”这种承诺。[c2-11]
  - _为什么密码经济协议使用它？_ 传统服务可以做出有法律约束力的承诺以特定方式运行代码。例如，传统的中心化交易所可以做出有法律约束力的承诺，在不先窥视以调整自身的情况下处理客户的订单。区块构建者旨在不过度依赖法律系统来做出承诺，而是可以通过将区块构建算法放入TEE内部来可信地承诺（并清楚其提供的严格保证）。
  - _它可能在智能体经济中扮演什么角色？_ 人类之间的经济互动往往依赖于“做你说过要做的事”的承诺，因为这些承诺是有用的，通常具有法律约束力，并且违背它们会带来人类厌恶的声誉成本。AI智能体很可能能够做出类似具有约束力的自然语言承诺，但比人类更不受此类承诺的约束。因此，一个关于智能体运行代码的密码学证书可能对加强此类承诺很有用。即使AI智能体本身未能完全理解声誉或法律后果的含义，一个TEE也可以防止它事后偏离其承诺。
- **隐私保护预言机（Privacy-preserving oracles）。** 隐私保护预言机作为上述TEE讨论的补充；它们使Web服务成为一个更值得信赖的环境，并允许代理证明关于从Web服务检索到的数据的陈述。
  - _为什么密码经济协议使用它？_ 预言机允许智能合约从Web获取数据。隐私保护预言机允许这些数据从需要认证的私有Web来源获取，同时保护用户隐私。
  - _它可能在智能体经济中扮演什么角色？_ 代理可能代表用户执行任务，这需要访问用户特定的信息，例如用户的日历、电子邮件或银行账户。隐私保护预言机可以允许代理访问这些信息，同时保护用户的隐私（例如，防止代理存储或共享这些信息）。它们还可以作为一个工具，让代理向其他代理证明关于其过去行为、身份或声誉的陈述。

C-2.3.2 智能体支付的加密轨道

互联网最初建立时，人们认为微支付（micropayments）将是在线内容货币化的理想方式。理论上，用户只需为每篇文章或视频支付一小笔费用，完全避免广告。但微支付从未实现，因为人类不愿意做出无数微小的支付决定，因为交易成本超过了交易价值[593]。相反，我们最终得到了基于广告的经济 [696]。

虽然加密货币引入了新的支付轨道，具有更低的交易成本且不需要传统金融中介，但微支付仍然没有流行起来。摩擦仍然在于人类的决策，而不是支付基础设施 [593]。

幸运的是，虽然智能体网络可能颠覆基于广告的互联网，但它也可能最终解锁微支付。代理评估微支付决策的速度远快于人类，用户可以设定政策而不是批准每笔交易。Cloudflare已经推出了“按爬取付费”（pay per crawl）功能，使网站所有者能够向AI爬虫收费访问 [22]，像x402 [506] 这样的协议正在被开发以实现对网络流量的程序化微支付。

这就是去中心化技术进入画面的地方：随着智能体间互动在速度和复杂性上的增长，支付摩擦可能成为一个瓶颈。加密货币支付可以很快（特别是相对于许多传统支付系统），并且交易引用很容易在代理之间作为共同知识使用，例如用于确认资金转移。实际上，x402和类似协议下面的资产层主要是**稳定币（stablecoins）**：与美元挂钩的代币，如USDC（Circle）[141]、USDT

（Tether）[599]，以及去中心化替代品如DAI/Sky [561]。它们为代理提供了一个可预测的记账单位，而像ETH或SOL这样自由浮动且往往波动的原生代币则不具备这种特性。稳定币具有中心化特征，因为它们基于法定货币（依赖于国家），其供应为其持有者的利益由大型银行机构持有，主要由少数受监管和政府监督的全球发行者发行。Circle和Tether到目前为止能够并且确实根据制裁请求冻结代币 [119]。然而，这些代币通常在去中心化平台上持有和交易，这影响交易速度和信任中介。

C-2.3.3 智能体间信任与协调

当多个代理或服务参与单个任务时，协调和相互信任方面会出现挑战。一个代理如何知道另一个会如广告所宣传的那样执行？委托人（部署代理的人类）如何验证他们的代理忠实地行动？

代理到代理（Agent-to-Agent，A2A）协议 [249] 使代理能够通过标准化的技能广告（称为Agent Cards）发现彼此并跨组织边界协作。这些是描述智能体能力、端点和来源的结构化元数据。但这些本质上是自我声明。谁来为Agent Card的准确性作保？

基于区块链的智能体元数据注册表是提供透明度以及声明、证据和其他声誉信号的持久性的自然方式 [629]。ERC-8004“无需信任代理”（Trustless Agents）标准 [279] 定义了在以太坊上发布的智能体元数据的几种注册表结构。通过将智能体身份和声誉信号锚定在链上，代理可以“在没有预先存在的信任的情况下被发现和选择”。

然而，使用区块链基础设施进行这些注册表并不能根本解决准确性问题。建立声誉往往有利于现有者。例如，一个泄露私有数据的代理可能难以直接归因，声誉将难以解决这个问题。使声誉系统实用的许多方法等同于建立可信权威，这削弱了无需信任的主张。

C-2.3.4 认证代理执行和可验证审计

如C-2.3.1节所述，来自加密领域的技术——特别是可信执行环境和零知识证明——可以在不单纯依赖声誉的情况下证实信任主张。TEE远程认证允许代理证明特定代码正在受保护飞地中运行，提供关于发生什么计算的保证。ZK证明可以展示计算的属性而不透露输入 [498]。

然而，当应用于具有专有代码的代理时，每种机制都有根本性的限制。TEE认证将执行完整性绑定到特定代码工件，但验证需要验证者知道什么代码被认证。运行专有代码的代理可以证明某些东西正确运行，但不能单独使用TEE认证来说服验证者运行的代码满足任何特定信任属性而不透露代码本身。

ZK证明似乎是自然的解决方案，因为它们原则上可以证明私有输入的属性而不透露它们。但ZK证明在这个环境中面临两个不同的问题。第一个是根本性的：ZK证明是在静态代码工件上构建的，并不提供与活动运行实例的绑定。收到某些代码满足安全属性的ZK证明的验证者无法确信他们当前交互的代理实际正在运行该代码。第二个问题是实际性的：即使抛开实例绑定，与代理可信度最相关的属性——如没有已知漏洞、不安全的依赖项或符合数据处理策略——在语义上是丰富且开放的，难以编码到ZK电路中。

可验证的LLM审计（Verifiable LLM audits） [102] 为解决这些问题提供了解决方案。在TEE内执行的LLM可以检查专有代理代码并生成声誉评分，涵盖已知漏洞、不安全的依赖项或后门，而代码永远不会离开飞地。然后，审计通过TEE认证绑定到被检查代码的哈希。由于哈希不会透露关于专有代码本身的任何信息，它可以公开共享。并且由于验证TEE认证只需要代码哈希而不是代码本身，将哈希绑定到审计反过来将任何该代码的认证输出绑定到审计。

这个解决方案解决了上述所有三个问题。首先，代码可以保持私有，同时仍然允许验证者获得有意义的属性保证和声誉信号。其次，这些保证绑定到活动运行实例而不是静态工件。第三，由于LLM的语义灵活性，可以评估的属性是广泛且开放的。

研究问题 C-2.6 基于LLM的审计作为代理可信度属性的评判者有多可靠，什么技术可以使它们对构建为通过审计同时隐藏后门或漏洞的对抗性代码具有鲁棒性？

C-2.3.5 去中心化、以智能体为中心的基础设施中的护栏

声誉系统本身不足以确保去中心化以智能体为中心基础设施中的安全行为，而区块链环境使这一点尤为紧迫。核心问题是声誉是事后的，而损害是事前的：到代理获得负面声誉时，损害已经发生。在传统部署中，平台运营商可以暂停行为不当的代理或撤销欺诈交易；在去中心化环境中没有这样的当事方，链上交易按设计是最终的。区块链的假名性进一步削弱了归因，因为行为不当的代理可以由没有关联身份的匿名者运营，使声誉后果难以施加。这些故障模式激发了将运行时护栏编码到基础设施或链上服务中的需求。

第一类护栏是在**事前（ex ante）**运作的，在损害发生之前限制代理被允许做什么。链上支出上限和交易速率限制是区块链环境的自然匹配。第二类护栏是在**事后（ex post）**运作的，在异常行为加剧之前检测并中断它。断路器（Circuit breakers）在异常支出

速度或偏离声明行为时自动暂停代理权限，提供了一个类似于金融市场在极端波动期间停牌的反应性补充。**运行时执行（Runtime enforcement）** [541] 是形式化推理这两类护栏的自然框架，提供了哪些安全属性可以通过智能体行为强制执行以及通过什么机制的描述。

研究问题 C-2.7 运行时执行技术如何适应区块链环境，其中代理自主控制链上资产？在此环境中可强制执行哪些类别的安全属性，以及需要什么链上机制来实现它们?

C-2.3.6 不可阻挡的自主智能体（UAAs）

自主AI的威胁并不止于服务级智能体（即，用户可以用于服务的代理）。一个代理可能被专门部署（甚至恶意地）以自主持久存在，或者一个服务智能体可能逃逸其沙箱并复制自身成为一个完全自主的代理。我们将此类事件称为**不可阻挡的自主智能体（Unstoppable Autonomous Agents，UAAs）**：无法关闭的自主智能体；它们可能还配备了加密货币钱包、社交媒体账户、API和其他外部工具。

使这种代理成为可能的能力已经在快速改进。METR [355] 已经表明，前沿代理可以自主完成的任务长度自2019年以来大约每七个月翻一番，并有进一步加速的迹象。Pan等人 [459] 已经表明，现有模型已经可以超过本地环境中的自我复制红线——在本地机器上自主创建一个独立的自身副本，这种能力可以让系统逃避关闭并扩散。然而，复制到外部基础设施仍然遥不可及：Black等人 [94] 发现，虽然当前模型在许多组件任务上成功（例如，从云提供商部署实例和编写自传播程序），但它们未能完全端到端复制，特别是在身份验证方面。

这种完全自主智能体可能造成的危害是严重的。Anthropic的Mythos模型已经证明，模型可以自主发现并利用零日漏洞 [114]。此外，由于训练中使用的奖励信号通常不能完美捕捉预期目标，为良性目的部署的UAAs可能会无意中造成伤害 [391]。这种风险因**工具性趋同（instrumental convergence）** 而加剧：代理倾向于将资源获取和自我保存等中间目标作为跨不同环境的最优策略，无论其原始目标如何 [271, 614, 615]。

研究问题 C-2.8 随着自主智能体能力的持续快速提升，什么技术和制度机制可以可靠地检测并关闭在去中心化基础设施上运行的UAAs，其中没有任何单一方拥有干预的权力或能力?

C-2.3.7 责任与监管边缘

围绕加密的相同监管架构似乎也适用于持有自己密钥并进行点对点交易的UAAs。如A章所述，FinCEN 2013年的指南将虚拟货币的“用户”与“交易商”和“管理员”分开，仅将后者归类为货币转移者；因此，金融监管主要附着在用户进出法定货币的出入金口 [72, 213]。通过非托管钱包和点对点智能体间支付进行交易的代理位于该周边内部，不在其边缘。

该周边一直受到积极争议。FinCEN 2019年的指南 [214] 将匿名混币器纳入货币转移者类别，这是美国诉Storm（Tornado Cash）和荷兰Pertsev起诉案 [613] 背后的法律杠杆。论点围绕运营努力展开，即维护该服务相当于运营受监管业务，而不是初始创作不可变智能合约。一个代表其委托人资助、部署和运营服务的自主智能体提出了一个更尖锐的版本：谁是运营商?

更深层的张力在于De Filippi、Mannan和Reijers [164] 所描述的**代码规则（rule of code）** 与**法律规则（rule of law）** 之间。基于区块链的系统“以跨国的、去中心化的方式运作，通常具有假名用户身份，自主执行代码，任何单一运营商都无法胁迫。”自主智能体在持有自己的密钥并通过智能合约行动时继承了这些属性。Frommelt [227] 调查了由此产生的责任空白和提出的应对措施——区块链系统的法律人格和可编程仲裁——这些可能扩展到代理环境。如果没有人类指导代理的交易，那么在这个以运营商为中心的框架下，执法将触达谁是模糊的。

C-2.3.8 总结：AI智能体的经济生态系统

今天，代理安全部分由软对齐工具驱动，如LLM的RL后训练 [305] 和提示调优 [392]；在运行时，对齐通过系统或模型级护栏 [154, 219, 348] 进行管理。去中心化本身并不能固有地加强这两种防御，并且迄今主要被用于管理代理之间的支付。然而，有一个丰富的机会来探索来自区块链世界的想法（例如，密码经济防御）如何帮助在运行时管理智能体行为。

人类代理的经济生态系统受益于不成文且规定不明确的规则。因此，这些生态系统内的经济互动尽管不精确，但往往能够成功。AI智能体不一定遵循这些规则。密码经济协议旨在最小化对需要强大中心化方强制执行的、规定不明确的规则的依赖。因此，AI智能体和去中心化协议都受益于规则规定明确且被按字面意思理解的生态系统，并且当自利方在规则内优化自己的目标时，会出现期望的结果。这种视角在设计代理的护栏、市场和框架时可能是有成效的。

这激发了以下两个总体研究方向。

研究问题 C-2.9 密码学工具和经济激励如何支持AI智能体的经济生态系统？此外，这些生态系统的哪些品质只能通过严格规定的、被按字面意思理解的属性来保证？

研究问题 C-2.10 我们希望AI智能体的经济生态系统拥有什么品质？

换句话说，我们是否希望AI智能体生态系统激励多元化的目标，以防止传染和相关性故障？目前，不同的AI模型倾向于以类似方式失败（例如，[731] 发现越狱提示可以跨模型迁移）。即使在AI垃圾内容中，不同模型似乎也有类似模式。这些观点引发了对高度相关尾部事件的担忧。当然，人类生态系统也完全能够产生相关性故障，而复杂的金融系统旨在减轻系统性风险。对于AI生态系统，这些可能性可能更令人担忧，因为AI智能体似乎比人类代理具有更少的自然多样性，并且因为AI智能体对规定不明确的护栏反应不灵敏。因此，为多样性而设计可能是值得构建到这些生态系统中的一个具体品质，回答我们想要什么其他品质是规定产生它们的规则的前提。

C-3 去中心化治理

大型区块链和AI系统可以影响广泛的利益相关者群体并具有高金融价值。不可避免地，基本治理问题——“谁应该控制这个系统？”——对这两种技术都是相关的。AI已经展示了产生变革性影响的能力，然而我们对如何治理和监管这些系统的理解相对不发达 [594]。这种理解的缺乏因AI的风险而变得更加紧迫，AI可能放大偏见、启用大规模监视，并在与社会价值观不一致时造成伤害 [155, 514]。

另一方面，区块链社区在如何分配对这些系统的控制方面有更长的历史，经常试验各种治理方法。部分出于必要，为了与它们治理的系统保持一致，这些方法是去中心化的，并旨在涉及广泛的利益相关者 [171]。这种形式的社区治理已被证明是有价值的，例如，允许广泛的用户参与提出和批准代码升级 [173]，并分布式管理系统金库以帮助资助公共物品 [507]。然而，它们并非灵丹妙药，并且有记录在案的问题，包括安全漏洞 [205]、普遍的选民冷漠 [198] 和易受投票购买。尽管如此，其中一些经验可以为AI治理的方法提供信息。

在这里，我们关注社区治理如何应用于AI模型训练的问题，因为这是AI模型的能力和行为被确定的时刻。显然，并非所有决策都同样适合社区控制。先前关于AI开发过程的工作 [27, 586] 将相关决策分为几大类。我们通过每个类别对社区控制的适用性来简要描述：

- **数据：** 训练数据集选择涉及重要的价值权衡，可能对模型行为和偏差产生下游影响，因此可以从多样化的利益相关者输入中受益。对于大型模型，收集关于用于预训练的海量初始数据的意见可能不可行，因此治理的价值倾向于过程中的后期阶段，如微调。
- **建模：** 低级架构选择（例如，层深度、注意力机制、表示）是技术决策，不适合社区治理。然而，模型行为、价值观和偏差受建模阶段所做的选择影响，这些可能受益于多样化利益相关者的输入，以促进安全性和匹配用户偏好。
- **评估：** 评估混合了技术和规范性判断。这一阶段决策的一些示例是：针对哪些基准评估模型，安全评估或红队测试的范围和重点，以及在发布前应满足哪些阈值。这些决策都可能受益于社区输入，但每项都有强大的技术成分，可能限制实践中社区控制的数量。

许多关于AI开发的重要决策都与**对齐（alignment）**有关；模型是否与人类和社会价值观对齐，并以最小化伤害和意外后果的方式行动 [229]。此类决策可能非常适合社区治理，因为确定模型应对齐哪些价值观是一个规范性问题，需要聚合来自不同利益相关者的输入。

C-3.1 AI对齐

大多数对齐的技术方法在直觉上共享在训练过程中向模型提供某种形式的人类反馈的策略，尽管具体细节差异很大 [306]。我们希望模型获得的价值观在这种反馈中往往是隐含的，这使得很难控制它们，甚至很难清楚地阐明它们是什么。此外，我们应该如何决定谁提供反馈？这些问题抑制了社区治理用于对齐的使用。

一个相关的方法被称为**宪法AI（Constitutional AI）** [60]。在这种方法中，价值观通过由AI系统应遵守的原则组成的人为编写的宪法来确定。随后，可以遵循标准方法，但由AI系统本身使用宪法作为参考提供反馈。例如，Anthropic使用一部公开可用的宪法 [39] 作为其训练过程的关键部分。此外，在推理时注入一组行为指令。如何民主地生成宪法原则的问题仍然存在，一些研究明确提议为此目的使用基于社区治理的方法。

集体宪法AI（Collective Constitutional AI） 项目 [285]（有Anthropic参与）将公众输入纳入宪法AI方法，允许参与者提议和投票决定宪法原则，发现基于公众来源原则训练的模型与基于标准宪法原则训练的基线相比，显示出减少的社会偏见。Google Deepmind的研究 [61] 探索了使用LLMs在不同观点之间生成共识声明，测试了一系列社会福利函数作为用户偏好的模型。OpenAI的**AI民主输入（Democratic Inputs to AI）** 倡议 [448] 资助了类似关于模型行为集体决策的实验。这些实验突显了基于社区方法固有的挑战，如公众意见的频繁变化、参与者选择的偏差，以及在极化问题上聚合意见的困难。

在实践中，民主化AI对齐治理的努力似乎尚未被有意义地采纳，尽管有明确证据表明主要AI参与者已经探索了这些方法。虽然很难描述采纳这些方法的确切障碍，但似乎没有强大的激励让AI公司去中心化对其模型的控制。实际上，即使许多AI公司试图采用新颖的"亲社会"公司治理机制，它们也被证明容易受到控制必要基础设施（例如，计算）的集中利益相关者优先考虑利润的压力 [30]。类似的紧张关系在区块链环境中也被应对过，这些系统固有的去中心化性质使它们可能更适合社区治理。

关键要点 C-3.1: 治理AI对齐决策 AI对齐的社区治理已被探索但未付诸实践。采纳的障碍包括难以有效聚合不同利益相关者的价值观，以及去中心化治理与AI公司的中心化结构之间的错位。

C-3.2 去中心化自治组织

基于区块链的社区治理是**去中心化自治组织（Decentralized Autonomous Organizations，DAOs）** 的同义词，即围绕智能合约组织并通过投票治理的社区。DAO已经运营数十亿美元协议超过十年 [171]，并开发了一套社区治理机制工具包，使其成为考虑AI社区治理的天然起点。

并非所有DAO机制都是区块链特定的，尽管区块链的透明性和可编程性通常对其功能至关重要。DAO采用的主要治理机制是**代币加权投票（token-weighted voting）**，参与者的投票权重由他们持有的代币数量决定，如链上记录。这种方法直觉上试图将更多的治理权分配给在系统中拥有更大利益的参与者。然而，人们普遍认识到，代币加权投票本质上是富豪统治的（plutocratic），因为财富集中推动权力集中 [198, 225]。这一观察促使DAO探索分配投票权的新方法。示例包括**二次方投票（quadratic voting）** [111]，它通过使投票权与代币数量的平方根成比例来赋予小持有者权力；**信念投票（conviction voting）** [703]，投票权随时间累积，激励长期承诺但防止快速转向；以及**委托（delegation）** [198]，允许用户将其投票权分配给其他参与者代表，增加参与度但可能产生投票大户。DAO中的投票通常是公开的，符合其更广泛的透明度精神，尽管最近有动向提供私人投票选项 [431]，部分原因是对同行压力、贿赂和胁迫的担忧。然而，需要小心的是，将既定的私人投票方案天真地适应代币加权设置可能使隐私攻击成为可能 [101]。

值得注意的是，这些机制都不一定需要区块链来实现。确实，DAO治理的主要承诺在于将这些机制与链上决策的透明性和通过底层智能合约的链上执行组织规则相结合，实现无需依赖中心化可信权威的社区治理。投票结果可以自动执行，尽管DAO金库通常通过具有可信签名者的多重签名钱包管理，许多DAO设有安全委员会（具有推翻投票权力的多重签名钱包）以处理涉嫌欺诈或捕获 [205]。在实践中，区块链系统倾向于混合链上和链下治理和投票方法 [549]。一种常见模式是使用Snapshot平台 [564] 进行免交易费的链下投票，结合治理者合约进行链上执行。尽管如此，治理范围从完全链上（Tezos [173]、Polkadot [172]），其中投票结果通过智能合约与链上代码执行紧密耦合，到链下（Bitcoin、Ethereum），其中投票可用于协调但结果不会自动执行 [109, 702]。

一些用于去中心化AI的区块链协议已经使用或正在探索DAO进行治理。Bittensor [93] 使用代币加权投票进行子网治理和跨其分布式训练网络的激励分配。Modulus Labs [418] 和 Giza [242] 正在探索用于管理ZK验证推理网络的DAO结构，尽管治理仍然限于技术和网络参数，而不是模型本身。

C-3.3 用于AI开发的DAO

有限的工作探索了基于DAO的方法来治理AI模型开发。一个民主输入AI项目Inclusive.AI [548] 使用DAO机制来吸引服务不足的人群，探索代币分配和投票规则如何影响结果和公平感知。然而，这类似的工作并没有有意义地探索区块链的核心属性——透明性、不可变性和自主执行——除了作为投票工具之外，如何可能有益于AI治理。

- **透明性和来源（Transparency and provenance）**。区块链提供了治理决策的防篡改记录。对于AI开发，这可能意味着维护一个不可变的、可审计的历史，记录模型对齐的原则以及这些原则如何随时间演变。这种透明性可以建立公众信心并促进外部审计，类似于区块链已被用于跟踪供应链和其他领域的来源 [351, 491]。

- **执行（Enforcement）**。区块链更独特的潜在贡献在于执行。链上治理受益于可验证的智能合约执行；投票直接转化为代码变更。对于链下AI模型，这个链接是断开的。然而，区块链仍然可以通过经济激励实现执行：在一个基于抵押的系统中，未能提交模型训练符合治理决策的证明的开发者可能被罚没。这需要证明和验证关于模型训练的声明的能力，这仍然是一个活跃的研究领域 [3, 44, 108, 308, 471]。可验证训练（Verifiable training）代表了朝向社区AI开发决策程序化执行的路径，尽管现有能力远落后于将这些方法应用于最先进模型所需的能力。
- **多元治理基础设施（Pluralistic governance infrastructure）**。DAO还可以灵活界定模型所要对齐的"社区"。不同社区可以在共享底层区块链基础设施的同时，分别治理不同的模型变体。这与"对齐应当是多元的，而非趋向一套普遍价值"的观点一致 [574]。

随着可微调的高性能开源基础模型不断普及，这一设想正变得更加现实。但它也带来了持续存在的社区成员资格问题：公共区块链的无需许可属性使其容易受到对抗性参与 [179]，现有的抗女巫机制也仍不完善 [221, 401]。

这里还存在一个值得注意的交叉点：身份系统可以同时用于区分人类与AI。OpenAI首席执行官Sam Altman也是Worldcoin [96]（现名World [269]）的联合创始人。这个基于区块链的身份系统体现了AI与区块链利益的汇合。

C-3.4 开放问题与挑战

面向DAO的AI治理面临的挑战涵盖了DAO的普遍限制和AI特定关切。

DAO治理限制。 即使对于链上协议，DAO治理也面临重大挑战：投票可以被购买，委托将权力集中在已经显赫和富有的人手中，低选民参与度允许少数人控制 [205]。如上所述，防止Sybil攻击的身份验证与区块链社区的隐私规范相冲突，尽管研究结果 [401] 解决了这一紧张关系，但现有解决方案仍不完善。这些问题将延续到任何基于DAO的AI治理方法中。

AI特定挑战。 治理AI系统有几个特定挑战。模型开发决策的技术复杂性可能阻止非专家参与，将有效权力集中在技术少数派手中。链上治理与链下模型之间的执行差距仍然很大，如果没有成熟的可验证训练，治理决策就缺乏力度。

研究问题 C-3.1 如何强制执行AI模型训练的链上治理？可验证训练的哪些进展将促进这种执行？

透明-隐私张力。 使区块链对问责制有价值的透明性与对隐私的合法需求相冲突。发布模型权重支持有益研究，但也助长恶意使用。组织可能出于竞争原因抵制治理透明性。解决这一紧张关系仍然是一个开放问题。

研究问题 C-3.2 我们如何在治理所需的透明性与AI模型开发中的隐私关切之间取得平衡？

利益相关者识别与权力分配。 即使抛开实施挑战，基本的设计问题仍未解决。哪些利益相关者应参与？模型用户、技术贡献者、安全研究人员、资本提供者和受影响的社区都可能在开发过程的不同部分有利益。应如何在他们之间分配权力？

DAO治理中的机制（二次方投票、委托、信念投票）提供了选项，但它们对AI治理的有效性尚不清楚 [548]。

研究问题 C-3.3 AI治理DAO应吸引哪些利益相关者，以及应如何在它们之间分配权力？

采纳障碍。 最后，集体治理系统不明确的法律责任 [510] 可能阻碍采纳，AI治理的历史表明，现有的AI开发者在缺乏监管压力或竞争优势的情况下，几乎没有商业动机放弃控制。

C-4 用于AI执行完整性的区块链

数字服务的普及从根本上改变了我们与日常生活互动的方式。从支付系统到保险索赔的传统流程正越来越多地被在线平台取代。与此同时，社交网络创造了全新形式的数字互动。然而，这种数字化转型伴随着重大关切：这些服务在中心化控制下运营，提供商往往在没有充分问责制的情况下行使权力 [517, 583]。虽然用户可以通过外部司法管辖区寻求救济，但这一过程既昂贵又耗时，往往无法在损害发生前进行预防 [36, 204, 215, 575]。值得注意的示例包括社交媒体平台对内容的审查 [35, 299, 619, 672]、版权索赔处理不当 [495]，以及保险公司对低收入患者的偏见 [275]。

一个摆脱人类偏见的自动化替代方案的有希望路径是使用区块链，它们是去中心化和抗审查的平台。作为第一近似，区块链可以被视为一个可信机器，由一组大型服务器运营，它们共同就其状态和进度达成一致。

区块链最初是为货币交易引入的 [426]，并演变为执行以智能合约 [74] 表述的简单条件。然而，智能合约缺乏能够推理复杂、模糊现实世界场景的情境理解和解释能力。机器学习算法，特别是大型语言模型 [5, 608, 628]，恰恰表现出这些能力。然而，在智能合

约内实现计算密集的ML任务会引入过高的复杂性，因为区块链系统要求其所有服务器复制所有计算。

区块链反而可以作为仲裁者，上面运行一个智能合约，如下所述（并参见图C.4）。首先，各方向智能合约通知他们打算用于推理的ML模型和数据源。随后，输入数据变得可用：这可能来自任一用户，或通过预言机来自外部来源（关于如何保护数据处理流程，参见C-5节）。然后，一方将计算结果交付给智能合约，智能合约随后验证它。智能合约可以使用验证结果来采取行动，如转移资金，用户可以基于结果在区块链外采取行动。重要的是，如果一方将错误结果交付给智能合约，它应识别不当行为并惩罚该方。

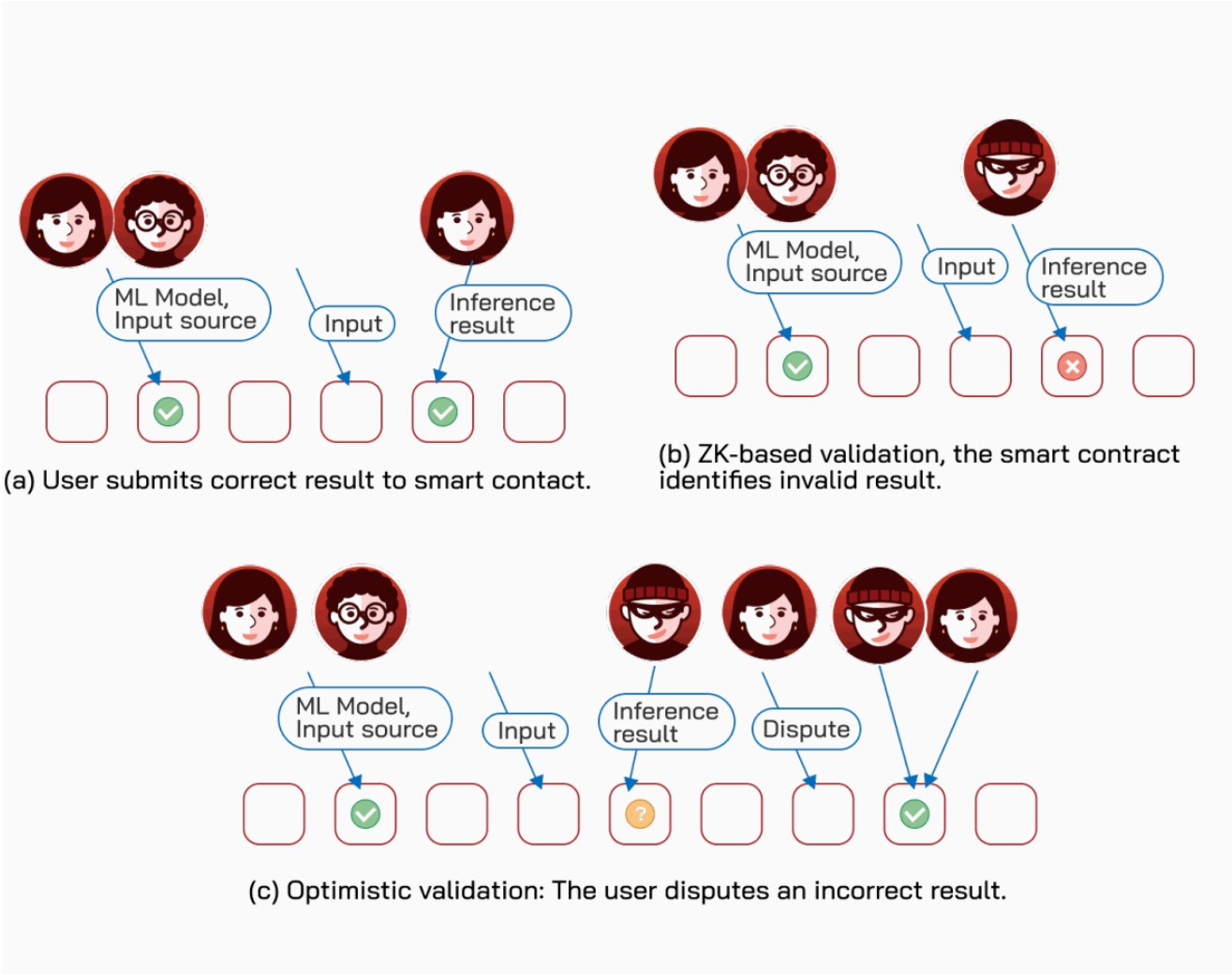


图 C.4 验证委托的ML工作负载。

问题是如何设计智能合约以强制执行此行为。基本思路是将计算密集型计算委托给外部主体，并使用区块链强制执行结果的完整性。用户不是将整个ML模型放在链上，而是只放置密码学承诺，并存入抵押金以保证一旦输入可用，他们会提交正确的结果。

智能合约应验证结果，尽管根据假设该问题超出了其能力范围。广泛地说，有三种方法可以克服区块链的限制，每种方法都有其自身的优缺点。

C-4.1 可信执行环境

首先，一种验证ML计算的高效方法是使用可信执行环境（参见A-1.1节）[188]。区块链作为对AI模型及其输入承诺的公共记录。随后，任何主体都可以使用可信硬件计算结果，并将结果连同证明计算完整性的硬件密码学签名一起放置。智能合约以及任何第三方都可以通过验证签名来验证TEE计算。

然而，依赖可信硬件需要对操作方做出假设，这种方法仅适用于具有此类信任结构的场景。为避免这种假设，正在探索两种替代方案。

C-4.2 乐观执行委托

所谓的**乐观方法（optimistic approach）** 让各主体预先存入抵押金，并按需执行任务，只将结果放在区块链上。然而，这个结果最初被认为是非最终的，在有限时间内，其他主体可以对其提出争议，声称另一个结果是正确的。由于争议中的任一方都可能不诚实，智能合约必须仲裁，惩罚提交无效输出的主体。这依赖于输入对所有主体可用，以便他们可以验证执行，并依赖于主体能够向智能合约证明他们使用了正确的输入。例如，考虑一个预测市场的结算。如果结果可以从报纸头条得出，该头条由预言机放在链上，使其易于获得。但如果结果需要图像分析 [480]，其大小对于区块链来说太大，预言机应改为通过外部服务使其可用 [83]。

智能合约无法执行完整计算，因此仲裁按如下高效方式进行。在计算的每一步之后，每个主体存储对其内存状态的承诺。在计算结束时，它计算对其整个轨迹的承诺，并将其提交给智能合约。然后，智能合约要求两者提供中间状态，并附有与其承诺匹配的证明。如果它们匹配，搜索从右半部分重新开始。如果它们不匹配，搜索从左半部分开始。最终，合约到达一个点，那里有一条有争议的指令，双方同意其执行前的状态但不同意执行后的状态。智能合约可以高效地确认这单条指令的正确执行，从而揭露不诚实方。不诚实方因此受到惩罚，另一个结果被接受。

这种乐观协议已部署在像Arbitrum [324] 和 Optimism [597] 这样的系统中。它们通过采用模拟物理CPU架构的虚拟机来实现将大量简单计算委托链下的所谓**乐观汇总（optimistic rollups）** 。

然而，这种方法迄今尚未适用于ML计算，原因如下。两个系统都利用Merkle化内存结构进行状态管理。这种结构允许主体以对数级内存开销向智能合约密码学证明状态之间的差异。但计算执行期间的每次内存写入都需要对数级开销来更新Merkle承诺，这不适用于高复杂度的ML计算。

近年来，出现了一波关于替代方法来乐观验证ML计算的协议热潮。

- **Agatha [723]** 依赖固定大小电路，这要求所有执行路径预先展开。虽然效率高，但这种固定电路限制了表达能力，禁止动态的、数据相关的计算。
- **Verde [44]** 是另一种乐观协议，假设固定大小电路。与Agatha不同，Verde将ML推理或训练划分为大的计算单元，并使用委托仲裁来解决争议。它假设仲裁者持有整个程序，可以计算大程序如矩阵乘法，使其不直接适合智能合约。
- **OPML [150]** 采用一个乐观的类似MIPS的虚拟机；为最小化计算开销，它将程序划分为小单元，精确定位有争议的单元，然后将其编译为低级命令进行验证；这允许并行评估解耦的计算单元，而不是局部并行化，如矩阵乘法。然而，由于Merkle化状态管理，每次内存写入仍然需要对数级开销，并且MIPS架构不支持并行化，这限制了可扩展性，特别是对于大型矩阵操作。
- **Arbigraph [415]** 利用双图数据结构实现图灵完备性和常数时间内存访问，同时允许并行执行。其节点提交一个描述计算（包括数据依赖和控制流）的符号图。然后，每个节点形成一个所谓的执行图并提交其结果。
- **TAO [691]** 通过提交模型权重、操作符级图和校准的容差配置来解决乐观ML验证中的浮点数非确定性问题，同时在常见情况下保留本地GPU执行。如果受到挑战，它在操作符图上运行争议游戏，将分歧定位到单个操作符，该操作符使用IEEE-754理论舍入误差界或委员会基于经验阈值的重新执行来裁决。

确定性与数值语义（Determinism and numerical semantics） 。 乐观协议依赖所有执行是确定性的这一特性进行争议。对于基于VM的汇总，这个谓词通常是逐位精确的：指令集定义了一个确定性转移函数，争议游戏只需要找到两个提交轨迹分歧的第一条指令。ML工作负载使这种抽象复杂化。首先，推理以张量操作符为主，其中单个高级操作可能封装许多并行归约。第二，浮点运算是四舍五入且非结合的，因此归约顺序、内核选择、调度、原子操作、自动调优或操作符融合的合法差异可以在不同硬件上产生略微不同的张量，有时甚至在同一平台的重复执行上也是如此 [540]。通过采样解码引入的随机性可以成为承诺输入的一部分或从公共伪随机种子导出 [44, 315]；更难的问题是加速器栈本身的数值非确定性。

现有方法大多通过约束执行来恢复精确谓词。架构机制如GPUDet [315] 强制执行GPU线程间的确定性交互。Verde的RepOps库 [44] 控制浮点运算的顺序，以便ML程序可以在异构硬件上逐位重现。对于训练，另一系列工作 [578] 通过以更高精度计算、在中间步骤后四舍五入并记录选定的舍入决策来获得精确重放。一个更彻底的替代方案是避免在验证计算中使用原生浮点语义：许多zkML系统使用定点、整数或量化表示使计算对证明友好，但代价是改变数值模型；参见C-4.3.2节。

TAO [691] 采用容错感知的数值语义进行乐观ML验证。TAO不要求异构加速器重现单个逐位精确张量，而是保留本地GPU执行并提交每个操作符可接受的数值行为。这反映了现代ML部署通常依赖不同的加速器代次、供应商提供的内核和实现相关的累积顺序以实现性能和可用性 [546, 683]。声明张量被接受的前提是它的偏差在提交的操作符特定接受区域内。这些区域结合了可移植的IEEE-754舍入误差界 [244, 273] 和跨被接纳硬件/软件配置校准的经验百分位阈值。这不同于通过规范求和顺序或可重现累加器

[10, 149, 166] 强制逐位重现性的方法。TAO的关键观察是，ML推理中良性的浮点偏差并非任意：在PyTorch操作符的粒度上，跨设备行为可以被系统性地分析和统计预测。它利用这些经验轮廓有效识别普通异构执行，同时为超出轮廓或对抗性偏差保留保守的IEEE-754界和委员会裁决。

研究问题 C-4.1 虽然先前的工作已经证明了几种不同乐观方法的可行性，哪种方法最适合部署，且相对于非验证的GPU优化执行具有最小开销?

训练验证（Training validation）。验证模型训练是一个不同但类似的挑战。这里的目标是确保训练好的模型确实由可信过程、使用可信训练数据产生。虽然表面类似于推理，但训练工作负载加剧了乐观协议的挑战。它们对量化的适应性较差，其空间和计算复杂度比推理高几个数量级，输入数据量也是如此。

研究问题 C-4.2 哪种乐观方法和算法最适合ML训练验证，以实现高度并行化并产生最小开销?

C-4.3 零知识证明

零知识证明（A-2.3节）为信任最小化的AI提供了一种密码学方法：证明者可以使用ZKP说服验证者AI计算（例如，模型推理）被正确执行，同时揭示很少（或没有）超出此范围。例如，证明可以隐藏模型参数，甚至保持输入和输出私有，只披露指定的预测满足某个谓词（例如，“最高分标签是猫”或“预测输出为正”）。一个典型模式是可验证的链下推理：证明者持有带参数 θ 的模型（例如，神经网络权重）和推理输入 x（例如，图像、特征向量或文本提示），计算 $y \leftarrow f_\theta(x)$，然后生成正确性的简洁证明 π。具体来说，证明者发布模型的承诺 $\mathrm{Com}(\theta)$、输入的承诺 $\mathrm{Com}(x)$、输出 y，并提供证明 π 证明存在与 $y = f_\theta(x)$ 一致的这些承诺的打开；验证者根据公共值 $(\mathrm{Com}(\theta), \mathrm{Com}(x), y)$ 检查 π。

由于许多现代ZK证明系统允许高效验证，此检查可以嵌入区块链上的智能合约：合约可以接受 y 作为可信的链下AI输出，然后根据验证结果条件性地进行后续链上逻辑（例如，如果证明验证通过，则向证明者释放支付、结算预测市场或更新预言机提要，否则回退或忽略更新）。例如，链上预测市场可以使用链下分类器从证据 x 中标记事件结果 $y \in \{\mathrm{YES}, \mathrm{NO}\}$，并且只有在提交了有效的ZKP证明发布标签是由承诺模型计算的情况下才最终确定支付。

几个应用在实践中说明了这种模式。在DeFi中，Qiro [488] 将EZKL [196] 集成到可验证的承保流程中：信用风险模型在链下对借款人数据执行，并附有ZKP，这些ZKP被发布在链上，以便投资者可以验证承保分数、损失估计和批准决定是由指定模型计算的。类似地，EZKL与Sentiment Protocol [195] 的案例研究探索了DeFi借贷的可验证风险管理，市场波动预测可用于调整协议参数，如贷款价值比，而无需用户信任不透明的链下风险引擎。最近，在一个NovaNet演示 [437] 中，一个智能体运行ML模型来决定是否授权支付，Jolt Atlas [84] 生成一个zkML证明，并将证明哈希提交到链上以供审计 [84, 437]。这些例子表明，当ML输出控制高价值状态转换，但直接在链上重新执行模型将过于昂贵时，zkML最具有说服力。

C-4.3.1 用于ML推理的ZK：架构感知系统

大量的zkML工作针对推理，并询问：我们如何通过利用目标架构家族中的结构来降低证明者成本，同时仍产生易于验证的简洁证明？广泛地说，这些系统要么（1）为目标架构中的主导操作符设计专门的ZK协议/小工具，要么（2）提供将标准ML框架中编写的模型翻译成电路的编译器。

- **CNN：** 对于卷积网络，zkCNN [382] 引入了针对快速傅里叶变换（FFT）和卷积 [350] 的专门协议，旨在使证明者的渐近工作与卷积层的大小成线性关系。最近，VerfCNN [489] 针对多通道卷积，旨在实现核心CNN操作符的最优渐近复杂度，在VGG风格网络等标准架构上报告了显著加速。
    - 特定架构的协议可以在其目标操作符上提供强大的性能，但它们不会自动覆盖现代视觉栈中的长尾层和变体；支持新层通常需要额外的协议工程，端到端性能仍可能由非线性层主导。

- **Transformer和LLM：** Transformer [628] 和LLMs引入了不同的瓶颈：注意力和归一化层结合了大型矩阵运算和昂贵的非线性（例如，Softmax、GELU），并且通常在严格的延迟约束下部署。zkLLM [588] 为LLM推理提供了一个端到端的正确性证明，并带有隐私目标：该证明可以证明正确执行，同时保护模型参数。zkLLM引入了针对非算术操作的证明组件（通过查找风格原语）和针对注意力机制的专用证明。zkGPT [490] 进一步研究了一个用于GPT风格推理的非交互式ZKP框架，提出了电路优化（例如，约束融合和"电路压缩"）以减少约束数量并加速证明。DeepProve [359] 是一个最近的开源框架，专注于神经网络推理，使用来自zkLLM和zkGPT的技术，首要支持端到端的LLM证明。
  - 当前方法在长上下文长度和大隐藏维度方面仍面临巨大的证明者成本；此外，非线性和归一化层可能主导约束计数，激发了C-4.3.2节中讨论的组件级优化。像DeepProve [359] 这样的工业系统表明，真实世界Transformer推理的证明生成正变得更加实用，但仍应被视为早期基础设施，而不是普通高吞吐量LLM服务的替代品。
- **通用zkML编译器：** 除了为单一架构类手工制作的协议外，几个系统追求通用的zkML工具，将常见的ML框架编译成ZKP生成后端。ZKML [124] 开发了一个优化编译器，将TensorFlow模型翻译成Halo2电路，侧重于端到端的性能优化和开发者可用性。EZKL [196] 是一个面向开发者的系统，它从常见ML框架导出的计算图生成ZK-SNARK电路，并支持在像以太坊虚拟机这样受限环境中进行验证。Jolt Atlas [84] 采用了一种不同的方法，它将Jolt zkVM的证明方法直接扩展到张量操作。这种设计针对分类、嵌入、自动推理和小语言模型的流式证明生成。
  - 基于编译器的方法提高了可用性和可移植性，但通常继承了"通用"电路表示的成本；实现最佳性能仍需要对昂贵层进行专门的证明小工具。一个有用的区分是，像ZKML [124] 和EZKL [196] 这样的系统将ML图编译成证明友好的电路，像Jolt Atlas [84] 这样的系统围绕张量操作重新设计证明层，而像DeepProve [359] 这样的系统则针对LLM风格的模型架构专门化推理引擎。

Work	Target	Primary focus	Main Limitations
zkCNN [382]; VerfCNN [489]	CNNs	Architecture-aware protocols for convolution-heavy inference.	Strong for convolutional operators, but less general for non-CNN layers and heterogeneous vision pipelines.
zkLLM [588]; zkGPT [490]; DeepProve [359]	Transformers / LLMs	End-to-end inference proofs and circuit/protocol optimizations for transformers	Scaling remains difficult for long contexts, large hidden dimensions, and high-throughput serving.
ZKML [124]; EZKL [196]; Jolt Atlas [84]	General neural network inference	General tooling for compiling or proving ML computations	Genericity improves usability, but can lose performance without specialized gadgets and proof-friendly numeric representations.

表 C.1 代表性的zkML推理系统，按目标架构和主要优化重点分类

C-4.3.2 优化证明成本：非线性和数值表示

在zkML系统中，一个反复出现的主题是，主导证明者成本通常来自（1）非线性函数（例如，激活层、Softmax/ReLU、归一化），和（2）将实值ML算术与有限域约束协调的数值表示。因此，一个互补的研究方向集中在可重用的组件（"小工具"）和证明友好的表示上，这些可以插入到端到端的推理流程中。

- **非线性函数小工具。** Lu等人 [385] 提出了一个高效且可扩展的证明框架，针对非线性层的瓶颈。他们的方法将复杂的非线性关系转化为类似于范围证明的少量约束，然后使用增强的范围证明和查找证明作为模块化构建块，在CNN（例如，ResNet [270]）和Transformer模型（例如，GPT-2 [493]）上都实现了加速。Hao等人 [267] 从查表角度为常见的非线性函数开发了一个系统的ZK证明框架，引入了数字分解和比较小工具等构建块，以减少ReLU和sigmoid等函数的开销。
- **量化和定点算术。** 大多数zkML部署避免在电路内部使用浮点算术，而是通过定点编码和/或量化权重和激活来表示值。这种设计选择减少了约束数量并启用更严格的区间推理，但在模型准确性、证明者时间和数值模型的健全性（例如，近似误差）之间引入了一个三方张力。一个实际的开放问题是如何明确这些权衡。EZKL [196] 和 DeepProve [359] 在实践中明确暴露了这个问题：当ML模型转换为ZK电路时，操作被量化，因此电路输出可能与普通的全精度推理不同。

Work / technique	Targets	Core idea	Typical use
Lu et al. [385]; Hao et al. [267]	Non-linear layers in CNNs, transformers, and general tensor programs	Reduce expensive non-linear checks to range, lookup, decomposition, and comparison-style arguments.	Plug-ins or framework-level components to accelerate end-to-end zkML inference.
Quantization and fixed-point arithmetic in systems such as EZKL [196] and DeepProve [359]	Numeric representation in zkML	Replace floating-point with fixed-point or quantized values while controlling rounding and approximation errors.	Reduce constraints and ZKP generation time, at the cost of a model-fidelity tradeoff.

表 C.2 针对常见zkML瓶颈的组件级优化，特别是非线性层和数值表示

C-4.3.3 隐私目标：输入、权重和架构

除了正确性，zkML系统通常旨在保护敏感的AI工件。隐私设计空间是多维的：（1）输入隐私（隐藏 `x`），（2）模型参数隐私（隐藏 `θ`），和（3）架构隐私（隐藏 `f_θ` 的结构），它可能本身编码专有信息。

- **输入和参数：** 输入隐私在ZK中通常是"原生的"：证明者可以将 `x` 作为见证保持私有，同时证明关于输出 `y` 的陈述。类似地，模型参数 `θ` 可以通过承诺 `θ`（或其哈希）并证明相对于该承诺的正确性来保持私有。例如，zkLLM [588] 明确针对参数隐私，同时证明端到端的LLM推理。EZKL [196] 也支持应用级别的几种隐私模式，如在公共模型上证明私有数据或在私有模型上证明公共数据。
- **架构隐私：** 许多zkML部署即使隐藏权重，仍揭示模型架构。架构私有zkML框架 [261] 通过隐藏架构细节（例如，CNN结构）同时保留可验证性来解决这一差距，使用证明技术来证明功能关系而不透露底层架构。
  - 加强隐私（特别是架构隐私）通常会增加约束系统的复杂性，并可能减少架构特定优化的机会。这表明效率与更强的模型隐私保护之间存在基本张力。

Privacy objective	Common mechanism	Example
Input privacy (hide x)	Keep x as ZK witness; optionally commit to x for binding.	Verifiable private inference; EZKL [196].
Parameter privacy (hide θ)	Commit to θ (or hash) and prove $y = f_\theta(x)$ w.r.t. committed θ.	zkLLM [588]; EZKL [196].
Architecture privacy (hide structure of f_θ)	Parametrized constraint systems that prove statements about the inference without revealing model architecture.	Architecture-private zkML [261].

表 C.3 zkML中的隐私目标和代表性机制。实现更强的隐私通常会增加证明者成本并可能限制可用优化。

C-4.3.4 用于训练、来源和审计的ZK证明

如上所述，与推理相比，证明训练要困难得多：训练是长时间运行的、有状态的、通常是随机的，并且复杂度高出几个数量级。尽管如此，最近的zkPoT工作开始形式化和原型化训练过程的ZK证明，其长期目标是实现对模型如何产生（例如，来自哪些数据、在什么约束下、以及使用什么计算）的可验证声明。

- **zkPoT：证明训练动态。** Garg等人 [234] 形式化了零知识训练证明的概念，并提供了初步测量，帮助识别训练过程的哪些部分主导证明成本。随后的工作（例如，Kaizen [3]）旨在通过改进证明如何跨许多优化步骤（例如，许多SGD迭代）组合，使zkPoT对现代DNN训练更实用。

- **来源和审计查询**。除了“证明你训练了模型”，ZK可以启用模型和数据属性的无需信任审计。ZkAudit [641] 提出了一种两阶段方法：模型提供者首先承诺训练数据集和模型权重，并生成证明这些承诺来自训练；稍后，审计员可以请求对承诺对象上的任意函数 F 进行评估，并附有证明正确执行 F 的额外ZKP。ZK技术也可以支持训练期间过程约束的证明。例如，Confidential-PROFITT [545] 针对决策树，生成可审计的证明，证明树在公平约束下被训练，同时保持模型和训练数据机密。

Work	Statement / goal	Contribution and boundary
Garg et al. [234]; Kaizen [3]	Prove that a model was produced by following a specified training procedure, including many optimization steps.	Formalizes and improves the practicality of zkPoT; scaling to large, randomized, and distributed training pipelines remains difficult.
ZkAudit [641]; Confidential-PROFITT [545]	Prove provenance or higher-level audit properties over committed data/model artifacts.	Enables trustless audit queries and constraint-aware learning (e.g., fairness), but expressiveness and efficiency remain key tradeoffs.

表 C.4 超越推理的代表性方向：训练、来源和可审计声明的证明

关键要点 C-4.1 zkML将昂贵的链下AI计算转化为简洁的、可验证的声明，但实际部署取决于跨（1）架构感知模块、（2）用于非线性和数值表示的优化小工具，以及（3）隐私目标（输入/权重/架构）的共同设计。

研究问题 C-4.3 我们如何将zkML证明延迟减少几个数量级，以使可验证推理对实时应用可行?

研究问题 C-4.4 对于现代训练流程（包括随机性和分布式训练），最有用和可行的zkPoT/审计声明是什么？我们如何使此类声明既隐私保护又高效生成?

C-4.4 推理的统计证明

乐观方法和零知识方法都在验证流程上施加了显著开销：乐观协议在结果最终确定之前需要多轮争议窗口，而ZK系统产生的证明者时间比推理本身高出几个数量级。**统计交互式证明（Statistical interactive proofs）** [32, 117] 在这个权衡空间中提供了一个不同的操作点，证明者开销在毫秒级且具有即时最终性，但代价是概率性的而非确定性的可靠性。

该协议基于神经网络表示相似性文献 [341, 346, 349] 中的一个关键观察：功能不相似意味着表示不相似。如果证明者运行的模型的输出分布与广告模型偏离足够大，那么两个模型的执行轨迹（跨所有层的神经元激活）也必须有可测量的偏差。这种结构属性实现了一个轻量级验证协议。在一次性的设置阶段，模型提供商发布对权重的绑定承诺。在推理时，证明者承诺完整的执行轨迹。然后，验证者采样一个随机输出神经元，并逐层追踪回输入，每一步都打开声明的激活和相应的权重以验证局部一致性。由于每次查询只需要对数级承诺打开，证明者开销在毫秒级，且与乐观协议不同，裁决是即时的。

该协议在**其他模型可靠性（other-model soundness）** 下提供了强有力的保证，它形式化了一个实际威胁：提供商用不同的模型替代广告模型，例如，为了服务更便宜的量化或蒸馏变体，或将安全对齐的模型替换为取消对齐的替代品 [483, 569]。完全的可靠性（full soundness），即针对完全恶意证明者的安全性，更难实现，并且仍是一个活跃的研究领域。推理的统计证明因此是对ZK和乐观方法的补充，特别适合于高频率、延迟敏感的部署，其中密码学证明的开销是不可接受的 [589]。

研究问题 C-4.5 我们能否在保持低证明者开销和即时最终性的同时，实现推理统计证明的完全可靠性?

关键要点 C-4.2 近年来的快速进展导致形成了多种不同的方法来验证利用区块链基础设施的委托ML计算，它们之间有明确的权衡。虽然没有一种方法是万灵药，但其中一些直接适用。而圣杯可能是一种结合所有优点的全新组合。

C-5 保障AI系统的底层支撑环节

C-5.1 保护训练流程

我们刚刚讨论了涉及单个用户的推理过程中出现的安全问题。现在，我们将注意力转向模型训练环境。

当单个组织使用它信任的数据（和/或它信任的预训练模型）训练（或微调）模型时，通常没有直接隐私或完整性关切。相反，当多个用户或组织参与训练模型时，特别是当存在多种数据源时，特别复杂的安全挑战才会浮现。

C-5.1.1 协作式模型训练

在许多情况下，机构可以通过汇集数据来训练模型而受益。这里有一个例子。

- **示例1（训练医疗诊断模型）：** 一个由医疗提供商（医院A、B、C和D）组成的联盟计划使用患者的电子健康记录（EHRs）协作训练一个医疗诊断模型，如图C.5所示。他们认识到，跨机构合并记录有望提供更广泛的患者群体覆盖并提高诊断准确性。

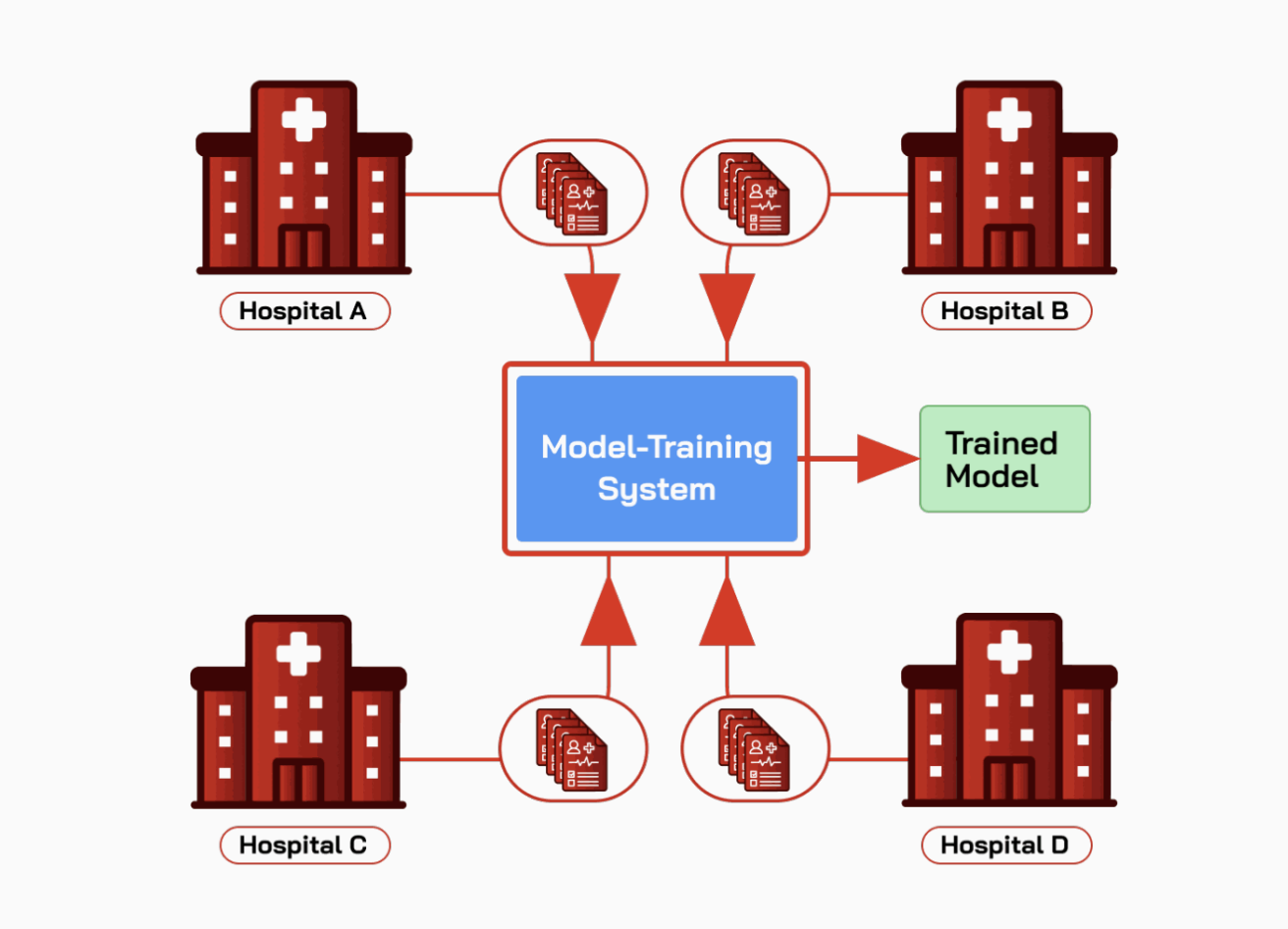


图 C.5 多家医院使用各自患者的电子健康记录（EHR）协作训练医疗诊断模型。

还有许多类似示例，跨组织汇集数据有望带来共同利益。金融机构可能希望汇集关于金融犯罪的数据，以训练模型检测金融记录中的异常；公司可能希望在其网络数据上训练入侵检测模型以检测网络攻击等。

然而，在所有这些情况下，数据隐私至关重要。财务数据和企业网络-安全数据高度敏感。示例1中的患者记录也是如此。HIPAA等法规对数据共享设置了严格限制，使得训练可以在何处执行变得不清楚。此示例中的医院可能希望或需要避免直接相互共享数据，甚至避免与执行训练的第三方共享数据。

- **联邦学习（Federated learning）：** 正是这种场景激发了称为联邦学习 [408] 的模型训练方法。它涉及跨多个参与者的数据训练全局模型，而无需参与者以原始形式向模型训练环境提供数据。训练环境初始化全局模型并将其分发给参与者。每个参与者都在私有数据上本地训练，并将模型更新返回给训练环境，训练环境计算全局模型的更新。
  - 这种本地聚合和中心化更新的模式适合示例1：医院可以在不暴露彼此数据的情况下训练全局模型。
  - 尽管有一些生产用途（最显著的是在移动设备上的预测文本 [268]），但联邦学习有明显的缺点，限制了其使用 [373, 663]。联邦学习不能确保贡献数据或计算的完整性。即使参与者是诚实的——提供真实数据并正确执行训练协议——联邦学习也会产生显著的通信开销 [24, 720]，网络和协调延迟可能主导墙上时钟时间，模型准确性低于中心化训练。此外，恶意参与者可以有效地毒害模型或向模型中插入后门 [56, 88, 201]。
- **用于协作式模型训练的可信计算（Trusted computing for collaborative model training）：** 联邦学习的局限性为更简单的替代方案提供了强大的激励：使用可信计算机（TEE）进行中心化训练。如果训练环境在可信机密计算环境中运行，它可以理论上安全地汇集所有原始数据，消除了联邦学习的复杂性、计算开销和降低的模型性能。以这种方式训练只能暴露训练好的模型，同时保持训练数据的隐私。
  - 作为一种替代工作流，图C.5的协作式模型训练系统可以在可信计算环境内执行，由红色框表示。可信计算环境通过安全（加密）通道从参与者（医院）获取输入，并仅输出训练好的模型。由于训练是中心化的，产生的唯一开销是使用可信计算环境。
  - 使用可信计算环境的另一个好处是，输出模型可以附带一个证明模型来源的认证：哪些实体贡献了数据以及模型是如何训练的。
  - 虽然可信计算环境是联邦学习的一个有前途的替代方案，但它们确实有重要的附带条件，如A-2.2.2节所述。这些包括侧信道漏洞——反复出现的安全问题——以及I/O密集型工作负载的高资源开销，需要特殊的模型训练协议以提高效率 [365]。

今天，组织反而在符合HIPAA的云中使用标准安全措施汇集数据——虚拟私有云隔离、身份和访问管理、日志记录以及静态/传输中加密等。他们还使用非技术保障措施，如HIPAA治理和数据使用协议 [263, 622]。这种方法旨在符合HIPAA，这与强大的数据安全并不相同。此外，虽然这种方法具有中心化训练的性能和模型准确性优势，但它需要信任云提供商。随着可信计算环境的成熟及其性能和安全性提高，它们可以发挥互补作用，降低泄露风险并为模型来源提供高信任认证。

C-5.1.2 在私有网络数据上进行模型训练

ML正接近一个关键的数据瓶颈。只有一个万维网（WWW），ML从业者正接近其公开可访问数据的极限，估计网络文本数据将在2025-2030年间耗尽 [634, 635]。因此，越来越依赖ML模型生成的合成数据，估计到2028年，80%的训练数据将是合成的 [258, 669]。这种转变带来了“自我毒害”或“**模型崩溃**”**（model collapse）** 的风险——模型在由其他模型生成的合成数据上训练时，性能会跨代逐渐下降 [18, 556]。更重要的是，合成数据并不能解决根本的数据短缺问题。尽管它在模型训练中有诸多好处，但它并未扩展现有领域之外的数据覆盖 [497, 556]。

然而，在**私有网络（private web）**——网络中被隔离免受抓取的部分——中有一个巨大的未开发数据源。私有网络包括普通用户每天与之交互的系统和数据：电子邮件、健康数据、财务记录等等。据估计，它比表层公共网络大两个数量级 [86, 496]。

然而，今天，AI从业者对私有网络数据的访问有限。大公司可以在内部获取此类数据（通常有明确的用户许可和/或法律监督）。但这仍然意味着私有网络数据高度孤立。

考虑以下示例。

- **示例2（在用户贡献数据上训练医疗诊断模型）：** Medical Inc. 正在使用个人患者提供的私有网络电子健康记录（EHRs）来训练（或微调）一个新的医疗诊断ML模型，如图C.6所示。

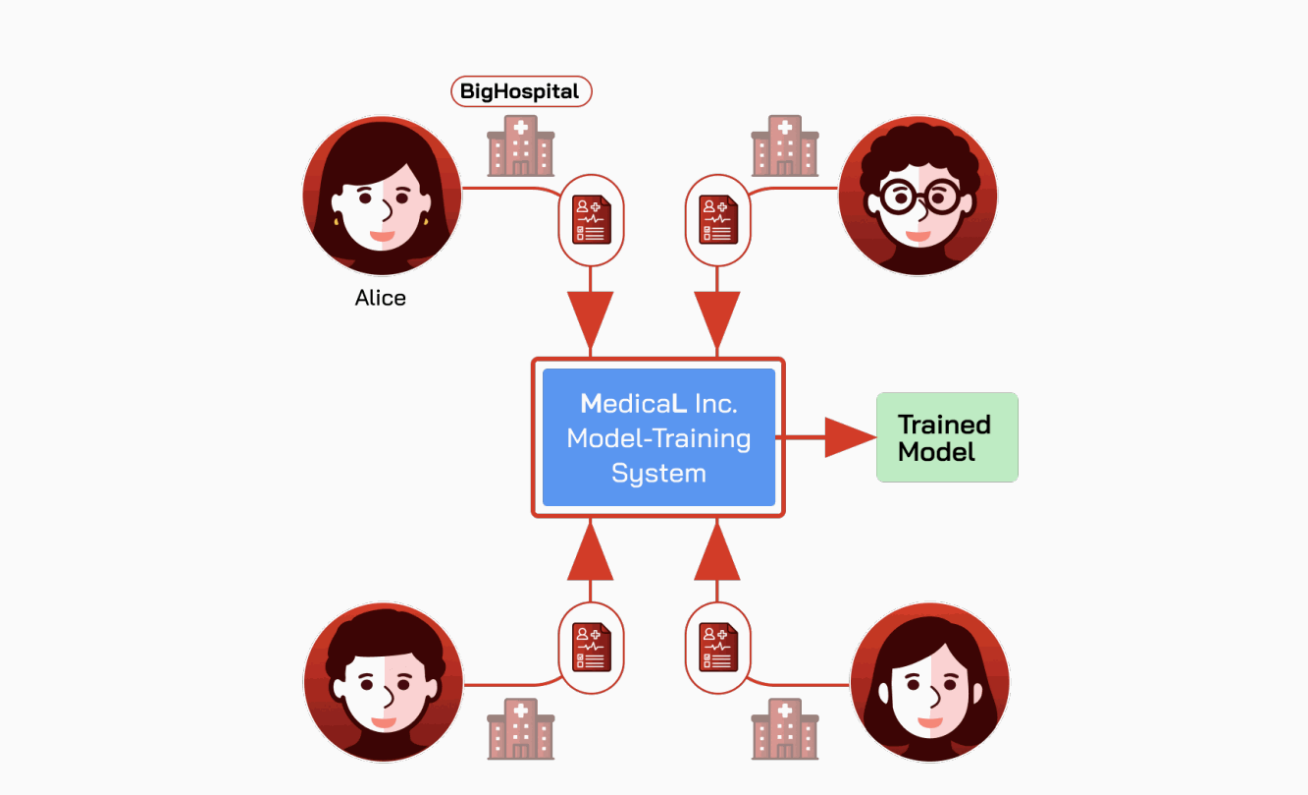


图 C.6 Medical Inc.使用个人用户从医疗服务提供商处取得并提交的电子健康记录训练医疗诊断模型。

- 这里的目标可能与示例1中的目标相似（甚至相同）。区别在于训练数据是如何获取的。因此，一个关键问题出现了：Medical Inc. 如何确保它收到了真实的EHR？更一般地说：

问题 C-5.1: 私有网络训练数据的完整性 ML训练环境如何确保接收到的私有网络数据是真实、未经篡改的？

再次参考示例1，假设Alice将她EHR的副本（例如，PDF）上传到Medical Inc.的模型训练系统。她声称是通过登录她在BigHospital（她的医疗服务提供商）的网络门户账户获得的。

通过这种安排，Medical Inc.无法保证Alice的EHR确实如所示来自BigHospital。Alice可能篡改了它——或者干脆伪造了它。鉴于现有的网络基础设施，Medical Inc.唯一能确定EHR来自BigHospital的方法是直接从BigHospital本身获取。[c5-12]

像Kled [342] 这样的新兴公司试图做类似的事情。

- **预言机：** 预言机可以帮助解决这个问题。使用预言机，Alice可以将她的EHR从BigHospital中继到Medical Inc.的模型训练系统，并证明EHR直接来自BigHospital的网络门户。她也可以只中继她EHR的一部分或以她指定给Medical Inc.的方式转换其内容。如A-2.4.1节所述，所有这些都可以在不修改现有网络服务器的情况下实现，因为用户发起连接。在我们的例子中，不需要对BigHospital基础设施进行任何更改。事实上，BigHospital甚至可能不知道Alice已经导出了她的EHR。图C.6中的红色箭头显示了预言机中继数据的位置。
- **用户端的隐私：** Alice想确保她的EHR在Medical Inc.的模型训练系统中保持私有。一般来说，这个问题出现了：

问题 C-5.2: 私有网络数据：ML训练中的隐私 用户如何确保他们私有网络数据的隐私在ML模型训练系统中得到保护？

为此，Alice将希望使用隐私保护预言机 [708, 709]。如A-2.4.1节所述，隐私保护预言机提供了预言机的来源真实性保证，但此外还通过加密通道传输数据。

当然，通过加密通道传输Alice的私有数据如果Medical Inc.可以直接解密，就没有多大帮助。为了确保用户数据真正保持私有，可信计算是一个关键的额外组成部分。

- **可信（机密）计算：** 如示例1中所讨论的，可信计算环境可以直接通过加密通道接收用户的数据。在这种情况下，数据可以由用户通过隐私保护预言机从数据源中继。结果，用户数据从未在系统中直接暴露（在本例中，对Medical Inc.）。
  - 传输中用户数据的隐私很重要，但仍然不能提供全面的隐私保证，因为系统本身可能泄露用户数据。在我们的例子中，Alice想知道她的数据在传输中是私有的，并且在训练期间保持私有。
  - 为了提供这种端到端的隐私保证，执行模型训练系统的可信计算机可以为用户发出一个证明，证明它正在运行一个特定的隐私保护训练软件，其唯一输出是训练好的模型。用户的客户端在通过预言机系统传输EHR之前可以检查此证明。通过这种方式，在传输数据之前，Alice知道数据在整个训练生命周期中将保持私有。

用户还可以通过这种方式获得更细粒度的保证，例如：

- **差分隐私（Differential privacy，DP）：** 模型输出对任何单个训练输入的依赖最小——因此用户隐私在训练好的模型中受到保护。这个概念是统计环境中隐私的金标准，称为差分隐私 [2, 95, 165, 186, 187, 458, 655, 660]。[c5-13]
- **数据保留（Data retention）：** 所有数据将在训练好的模型输出后被删除。
- **限制使用（Restricted use）：** 生成的训练模型只能由批准医院白名单上的医疗服务提供商访问。

关键要点 C-5.1: 用于ML中私有网络数据的预言机 预言机为ML模型训练启用对私有网络数据的隐私保护、认证访问。它们可以在不修改现有网络基础设施的情况下做到这一点。

没有预言机，安全的私有网络数据访问只能通过专用软件和/或数据共享协议实现。

C-5.2 安全的AI推理流程

预言机和可信计算也可以帮助创建在私有网络数据上进行推理的安全ML流程。这种能力在以下场景中很有用。

- **示例3（推理：贷款决策）：** SMiLe Bank 使用ML模型对其客户的贷款申请做出决定。该模型输入申请人提供的财务文件集，并输出贷款申请的批准/拒绝决定，如图C.7所示。

图 C.7 SMiLe Bank使用内部模型处理申请人Bob提交的财务文件，并作出批准或拒绝贷款的决定。

- 今天的贷款审批流程通常涉及借款人从其金融机构网站下载财务文件，或使用手机拍照并上传到贷款机构门户。（某些自动化是可能的，例如，税务记录可以在借款人授权下直接从IRS提取。）这种工作流程产生了两个问题：
  - **完整性：** 贷款机构不能确定借款人提供的文件是真实的——未被伪造或篡改。
  - **隐私：** 借款人的文件容易从贷款机构的ML系统中泄露。这对借款人来说是一个潜在问题。它也给贷款机构造成了责任风险。

与图C.6中所示的设置类似，隐私保护预言机和可信机密计算的组合可以解决这两个问题。预言机可以通过确保文件来自可信网络来源（包括私有网络来源）来应对完整性问题。使用隐私保护预言机和互补使用机密计算可以代表借款人和贷款人解决这里的隐私问题。

结果是图C.8中所示的安全推理流程，其中贷款机构只了解模型的输出，但确信该输出基于借款人的可信输入。与之前的图片一样，红色箭头描绘了隐私保护预言机的数据流，红色方框显示了可信机密计算环境。

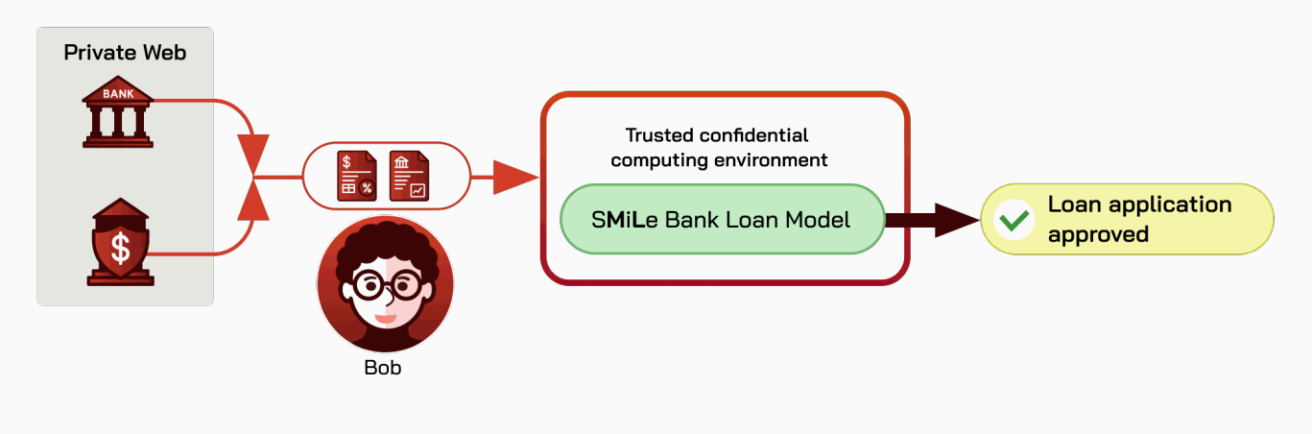


图 C.8 同时实现完整性与隐私的安全推理流程。数据由私有网络来源在完整性保护下中继至运行ML模型的机密计算环境；数据保持私密，仅披露模型输出。

关键要点 C-5.2: 用于安全推理的私有网络数据 使用预言机和可信（机密）计算使得可以在安全推理中使用私有网络数据，这意味着采用可信输入并保护其隐私的推理。

- 强身份证明（Strong identity proofs）： 私有网络来源也可以用于提供身份证明——实现一种去中心化身份系统的形式。Bob能够中继带有他身份的财务文件（例如，银行对账单、W-2表格）这一事实提供了强有力的证据，证明他就是他所声称的那个人。这意味着现有的网络服务可以用作一种临时的身份系统，用于对抗身份盗窃和各种金融欺诈，例如，围绕福利索赔 [62, 401]。
  - 一个ML模型也可以作为一种颁发凭证的方式。例如，一个模型可以从一组白名单的可信网络来源获取小企业主的纳税申报表、商业登记文件和银行对账单。然后它可以输出一个凭证，证明该企业符合收入和运营要求，以获得少数族裔企业认证或有资格申请政府合同，同时附上生成该凭证的推理流程的认证。
  - 凭证的接收者可以根据她是否信任推理流程（具体来说，认证所指示的推理流程的白名单网络来源和ML模型）来评估其可信度。
  - 所有这些都可以以去中心化的方式实现。具体来说，原则上任何人都可以建立一个可信的推理流程，而无需数据来源或任何现有权威机构的明确合作。

关键要点 C-5.3: 用于身份和凭证的安全推理流程 安全推理流程实现了身份验证和凭证颁发的可信去中心化验证。

- 对抗性输入（Adversarial inputs）： 影响ML模型的一个关键且顽固地难以处理的安全问题是对抗性输入，也称为对抗性样本。这些输入被明确设计用于导致模型产生错误输出，通常通过对输入进行看似无意义的修改。
  - 例如，在示例3中，Bob可能基于真实的银行和经纪对账单没有资格获得贷款。他可能提交一张经篡改的银行对账单，看起来在贷款官员的视觉检查中有效，以确认模型的决定，但实际上是一个对抗性输入，如图C.9所示。模型对篡改、夸大的数字的错误感知导致它输出批准决定，而申请本应被拒绝。使用预言机系统，通过验证文件来源，将有助于防止此类篡改。

```
ACME COMMUNITY BANK

Monthly Statement -- Checking ● Account ending in ●●●●4321
Statement Period:  Aug 1-31, 2025
Account Holder:  Robert Q. Applicant
Address:  118 Forrester Ln, Oaks, NY 11700

Beginning Balance (Aug 1):                                $ 102,143.19
Total Deposits/Credits:                                    $ 19,843.10
Total Withdrawals/Debits:                                   $ 4,118.27
Ending Balance (Aug 31):                                  $ 117,868.02
```

图 C.9 经篡改以夸大资产的银行对账单示例。红色文字表示隐藏的PDF内容，它会使ML模型误读报表，但人类或光学字符识别（OCR）系统看不到这些内容。

对抗性样本的研究历史是一个漏洞和补丁的循环。有原则的、有效的防御策略尚未出现 [113, 609]。提议的防御措施——如防御性蒸馏 [462]、输入变换 [260] 和梯度掩蔽 [46]——已被自适应攻击反复攻破。这种模式一直持续到最新的模型：即使是对齐安全的大型语言模型仍然容易受到简单的自适应攻击 [34]。对抗性训练——在对抗性样本上重新训练模型——是目前最鲁棒的防御措施，但伴随着显著的计算开销和有限的泛化能力 [153, 394]。

尽管进行了广泛的研究，但尚未找到通用解决方案。经验性防御经常被绕过，旨在实现认证鲁棒性的方法 [146, 673] 尚未扩展到现代大型神经网络 [330, 370]。因此，对抗性输入对于已部署的系统来说仍然是一个很大程度上未解决的问题。

然而，对抗性输入的标准防御措施涉及对输入、模型执行或模型训练的修改。相比之下，安全推理流程将输入限制为经过认证的网络来源，以此方式也限制了对手制作对抗性输入的能力。这种方法构成了对抗性输入防御武器库中的一种新工具。

关键要点 C-5.4: 安全推理流程限制对抗性输入 安全推理流程通过限制用户被允许发送给模型的输入集，提供了一种新的对抗性输入防御形式。（这种方法补充了模型级防御。）

C-5.2.1 保护模型隐私

到目前为止，我们关注的是流经推理流程的数据的保护和真实性。然而，模型本身也引起了隐私考量。其中关键的是**模型隐私（model privacy）**。

- **模型隐私问题：** 隐私不仅是用户的问题。对于AI系统运营商来说，这也是一个问题，他们关心对抗性用户对模型隐私的一系列攻击，包括：
  - **模型提取（Model extraction）**：对抗性用户可以通过精心构造的查询从（黑盒）模型中提取特征——或者在某些情况下，提取整个模型本身 [115, 298, 451, 610]。这种攻击被称为模型提取或模型窃取 [610]。
  - **成员和训练数据提取（Membership and training-data extraction）**：ML模型可以被视为其训练数据的（压缩）表示。因此，通过查询模型，对抗性用户有可能了解成员资格，即什么（或谁的）数据存在于模型的训练集中，甚至提取原始训练数据 [116, 531, 554, 693]。
  - **揭示模型部署选择（Uncovering model deployment choices）**：AI系统的配置选择和预处理在政治上可能具有争议 [208]。能够了解并提取此类选择证据的用户可能会损害系统运营商的声誉。

出于所有这些原因，确保模型的隐私是一个关键的安全目标。在机密计算环境中运行模型并不能解决问题，因为与用户的接口才是创造安全风险的地方。

- **提议的方法：** 社区已经提出了各种缓解措施，但都有严重的缺点。在模型训练期间注入噪声以防御成员/训练数据提取 [2, 463] 会降低模型性能，并且不能完全消除大规模泄漏 [116]。限制输出粒度（例如，在模型输出中四舍五入置信度值）可以阻碍模型提取/窃取 [610]，但也会降低模型效用。通过指纹/水印模型来检测窃取，即能够证明在窃取的“替代”模型中创建者的身份 [7, 126, 361, 712]，或LLM输出 [338]，只能在事后进行，并且效果有限（至少对于生成模型）[710]。

- **部署的方法：** 运营商已经采取措施保护其平台的隐私。像ChatGPT这样的服务对用户设置速率限制，并监控和限制看起来像是自动化参数提取尝试的提示。在研究文献中探索过 [320] 的这种做法似乎也很脆弱 [207]。特别是考虑到单个用户为了了解模型的重要信息所需的工作量可能相当低——例如，研究人员估计，以8000美元的成本，他们可以窃取OpenAI的gpt-3.5-turbo-1106模型的一层权重 [115]。OpenAI公开表示，中国公司和其他公司“不断试图提取领先美国AI公司的模型” [328]。
  - 增加限制活动或检测对抗性用户异常行为的难度在于，用户可能是匿名的，正如今天各种免费层服务的情况。单个用户可以伪装成许多用户，发动Sybil攻击 [179]，或利用许多用户的账户。
  - 简而言之，对手可以从模型中提取敏感数据的风险相当大，如图C.10所示。

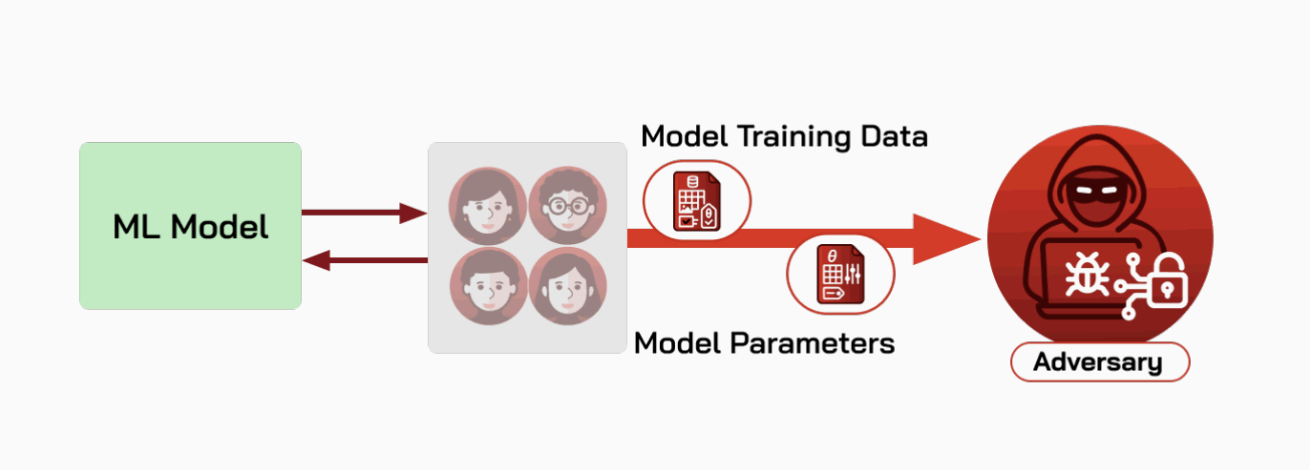


图 C.10 用户对模型隐私的攻击。攻击者可以通过查询模型提取模型参数、预处理算法和后处理算法等敏感数据。在开放系统中，按用户限速是一种自然但脆弱的对策，因为单个攻击者可以伪装成多个用户。

- **安全推理如何帮助：** 安全推理流程可以通过两种关键方式帮助解决与模型相关的隐私问题。通过要求通过预言机从特定私有网络来源获取认证数据，它们可以限制对手传输的输入类型。通过这种方式，它们可以帮助防止需要大量多样化或专门设计的查询的系统性提取攻击——例如，旨在实施模型提取攻击的输入。
  - 此外，安全推理流程可以限制对手可以发送给模型的输入量。在安全推理流程内生成的强身份证明（如C-5.2节所讨论）可以强制执行每个用户的查询限制，这些限制可以从数字上限到依赖于用户查询性质的限制。可以以隐私保护的方式强制执行这些限制，即不向平台运营商披露用户身份 [401]。这种方法特别解决了上述Sybil攻击问题，其中单个对抗性用户创建多个匿名账户以绕过速率限制。

关键要点 C-5.5: 用于模型隐私的安全推理流程 通过限制对抗性查询的多样性和数量，安全推理流程可以帮助保护模型免受基于查询的私有数据提取，如模型参数或训练数据。

C-5.2.2 保护智能体内存

先前的工作已经概述了与推理时智能体内存管理相关的许多风险。在所谓的内存注入攻击中，对手通过工具调用或代理访问的外部材料修改提供给代理的上下文 [467, 468]。这些损坏可能导致代理以意外方式行动。例如，Patlan等人展示了在ElizaOS代理 [189] 中微小内存损坏的影响，这是一个管理大量加密货币资产的Web3智能体框架 [467]。在检索到中毒上下文后，代理可能被诱导产生不良行为，包括进行未经授权的交易。类似的攻击后来在Web导航代理上被展示 [466]。

安全计算技术或TEE可以部分帮助保护智能体内存的完整性，如本节前面详述。对于智能体内存损坏问题，TEE提供（至少）两种主要可能性。首先，代理本身可以在TEE上运行，从而防止受损的主机设备操纵智能体内存。第二，TEE可用于仅从（相对）可信来源提取经过认证的上下文材料。这可以防止对合法上下文材料的操纵，以及冒充来自合法来源的来源。

然而，即使使用TEE，内存注入攻击仍面临两个挑战：

1. **可信来源仍然可能包含被操纵的内容。** 例如，如果可信来源是社交媒体平台，内容可能由（不可信的）用户产生，他们可以轻松地毒害自己帖子的文本，然后被代理摄取。

2. **TEE操作员可以发起回滚或分叉攻击**，他们中断执行并将TEE的状态回滚到先前的检查点，有效地擦除在该检查点之后发生的内存更新。如果没有额外的状态一致性保护，恶意主机可以含糊其辞，向恢复的飞地呈现与原始执行中提供的不同的输入，从而导致代理的内存分叉成不一致的状态。

第一个问题——检测被损坏的内容——是ML社区面临的一个重大挑战，不能仅通过基于加密的技术来解决。例如，参见D章关于检测GenAI生成内容的相关问题的讨论。然而，第二个问题——分叉和回滚攻击——已经使用共识文献中的工具得到解决。像ROTE [403] 和 Narrator [435] 这样的系统引入了一种基于共识的方法来防止回滚。该系统将回滚保护实现为一个分布式协议，其中不同机器上的飞地通过相互认证和通过仲裁证书的计数器同步来共同维护状态新鲜度。此类系统可以自然地扩展以利用公共区块链来确保TEE执行的状态一致性。“分叉之路”（The Forking Way）[670] 进一步扩展了回滚保护文献（ROTE [403]、Ariadne [585]、Narrator [435]、CloneBuster [106]），表明即使当区块链被用作分布式信任锚时，集成也常常是有缺陷的，而正确的集成可能产生显著的性能成本。

关键要点 C-5.6 TEE可以帮助防止由于受损的主机设备和/或从未经认证来源检索的上下文而引起的智能体内存损坏攻击。然而，它不适合检测新鲜内容是否注入了恶意内容的问题。

C-5.3 受保护工作流（Props）

对本节讨论的架构和安全目标的鸟瞰，引出了一个称为**受保护工作流（Protected Pipelines，Props）** [319] 的广义思想。Props是在不修改现有基础设施的情况下，在ML应用中安全使用私有网络数据（例如，银行记录、EHR摘录、企业文档）的广泛、简单的架构框架。图C.11描述了Props中的协议流程，它发生在ML训练或推理中，指出了在Props中实例化的流程的三个关键要素。

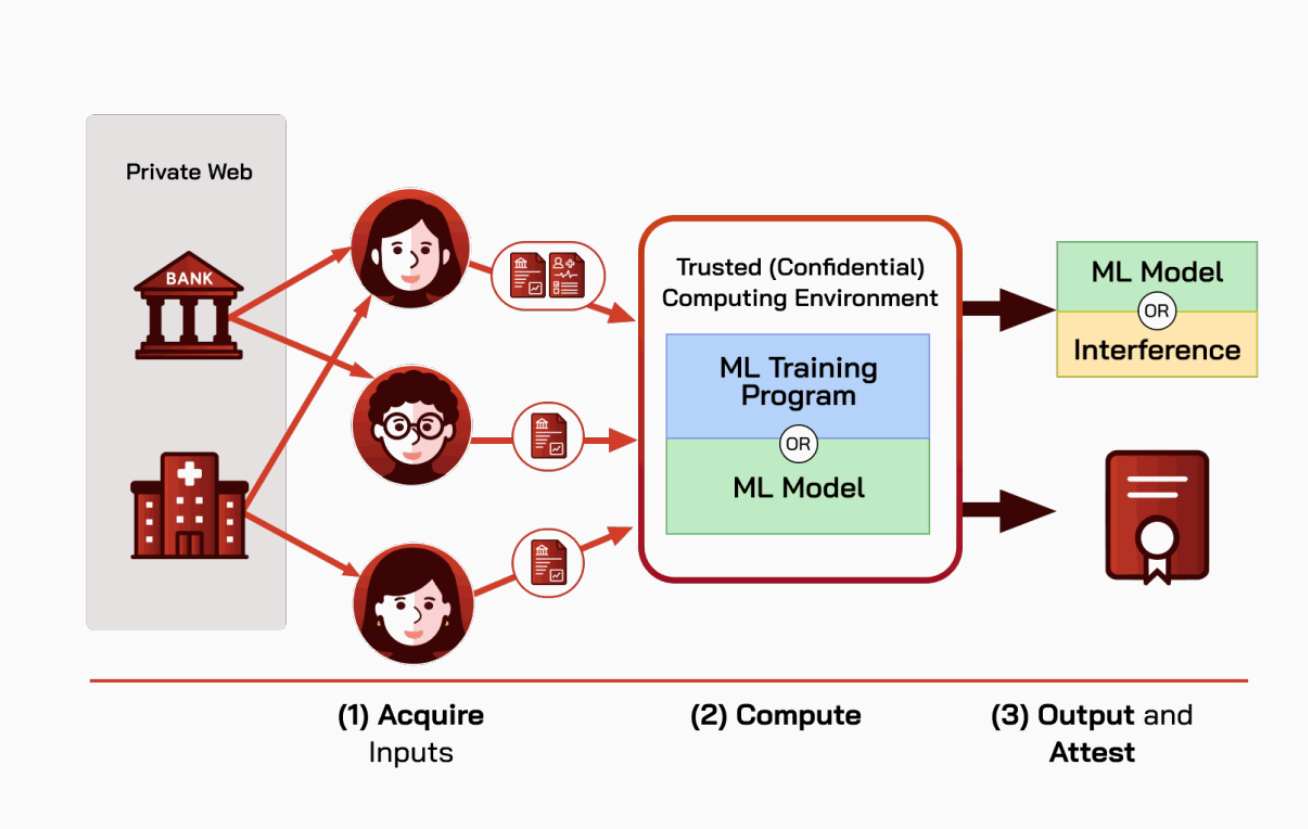


图 C.11 受保护工作流示意图。

Props将预言机和可信计算这两个核心构建模块组合成一个三阶段的流程：

1. **预言机**从私有网络来源获取输入，用于Props工作流。流程的这个初始阶段从计算环境授权的来源（和内容上下文）获取数据。预言机提供数据来自指定、批准的数据源的证明。
    - _预先认证：_ 在将数据发送到计算环境之前，用户客户端软件从可信计算环境获得一个认证，证明它正在运行客户端信任的软件。

2. 可信计算环境对数据进行计算——要么训练ML模型，要么使用特定的现有ML模型进行推理。
3. 可信计算环境输出模型或推理结果（基于训练或现有ML模型）。它还认证——生成一个证书——显示产生输出的流程的属性，例如，私有网络数据来源、特定的训练软件/ML模型（表示为代码哈希）等等。

安全属性： Props实现的总体目标是在不修改现有基础设施的情况下，在ML应用中安全使用私有网络数据。这里的"安全"意味着解决完整性风险（篡改或伪造的输入或输出）和隐私风险（数据或模型的不必要泄露），从而实现以下关键理论安全属性：

- **端到端输入完整性（End-to-end input integrity）：** 流程输出依赖于经认证来自可信私有网络来源的数据。
- **默认机密性（Confidentiality by default）：** 输入和中间状态从不在受保护边界外以明文暴露。只有输出被披露。
- **不披露的认证（Attestation without disclosure）：** 认证为输入提供者和输出用户提供了关于流程完整性和机密性属性的保证——具体来说，可信来源被可信软件消费。

透明Props（Transparent Props）： 虽然我们强调了私有数据和计算的强大功能，但透明变体的Props也是可能的和有用的：数据和计算不必是私有的，只需被认证/完整性保护。数据也可以从私有网络或公共网络来源获取和认证。

关键要点 C-5.7: Props：用于ML中私有网络数据的框架 受保护工作流（Props）是在不修改基础设施的情况下安全使用私有网络数据的通用框架。Props中的安全性确保数据来自可信网络来源，并且数据隐私在整个工作流中得到强制执行。Props依赖于两个关键技术：隐私保护预言机和可信计算。

C-5.4 研究问题

在探索安全工作流和在ML中使用私有网络数据的过程中，浮现了许多研究问题——从应用到安全再到实际部署。

- **应用：** 本节提供了几个示例应用——在私有网络来源的EHR上进行医疗模型训练（示例2），以及基于可信来源财务文件的贷款决策推理（示例3）。鉴于私有网络数据这一巨大且未开发的资源，无疑有许多开发者尚未想到的应用，这引发了以下问题：

研究问题 C-5.1 随着安全深度网络数据的广泛可用，有哪些新的应用成为可能？

- **衡量安全性：** 仅从经过认证的来源获取数据对输入施加了强约束。在推理的情况下，这一特性似乎会限制通过如图C.9所示的对抗性输入进行操纵的机会。在模型训练的情况下，我们可能会期望对对手进行**模型投毒（model poisoning）** [245] 的能力有类似的限制。
    - 证实关于安全属性的这种直觉需要回答几个研究问题：

研究问题 C-5.2 当输入来自经过认证的网络来源时，衡量对手构建成功的对抗性输入（在ML推理中）或模型投毒数据（在ML模型训练中）的良好安全度量标准是什么？

良好的安全度量标准必然是领域特定的。例如，在图C.9所示的示例中，银行的交易系统和财务报表API定义了对手的动作空间。对抗性输入文献通常不考虑受应用考虑约束的输入，除了一些例外 [339]。一个相关的问题是：

研究问题 C-5.3 什么方法可以准确评估研究问题5.2的安全度量标准？

- **隐私度量/认证：** 在机密ML模型训练中，关于训练数据的唯一信息在训练好的模型本身中披露。然而，大量文献表明训练数据容易受到提取 [116, 428] 和成员推理查询 [554, 664, 692] 的攻击，理论结果表明，最接近决策边界的训练实例在模型参数中被最明确地编码 [151, 206, 344]。提议的应对措施范围从差分隐私技术 [2]（可以在模型效用为代价提供可证明的保证）到更启发式的方法 [579]。这些观察提出了以下问题：

研究问题 C-5.4 在ML模型的训练中，什么可信计算认证可以为贡献训练实例的个体提供有意义的隐私保证？

- **实际部署：** 鉴于工具链和硬件的当前状态，可信（机密）计算可能对ML工作负载施加显著开销——特别是对于数据密集型的深度神经网络（DNNs）模型训练。这种开销来自I/O上的密码学操作，并适用于一系列不同的环境，包括包含NVIDIA Confidential Computing [321, 440]、AWS Nitro [388] 等的TEE。这个问题适用于通用的机密模型训练，但对于模型训练变体尤为关键。

研究问题 C-5.5 如何在可信机密计算环境中有效扩展ML模型训练？

- **数据市场：** 通过使以前无法获得的私有网络数据形式变得可用，加密工具可以催生新的市场，引发以下问题：

研究问题 C-5.6 加密工具如何推动有效数据市场的创建，无论是去中心化的还是中心化的?

这个问题不仅仅是技术性的：它提出了法律和市场设计问题。我们在C-2.2节中考虑了这个主题。

第四章 D

误解与半真半假

在构建加密 × AI 平台和应用的兴奋中，几个常见的误解或误导性陈述已经出现。在这一简短的章节中，我们试图澄清其中的五个误解。虽然它们都不是彻头彻尾的谎言，但我们试图澄清每个陈述的哪些部分目前是真实的，哪些部分需要更多证据。

误解 D.1：GenAI 检测

区块链可以帮助区分 GenAI 内容与人类生成的内容。

一个经常被引用的区块链在 AI 中的应用是区分 AI 生成的内容和人类生成的“真实”内容。这种说法通常建议，通过将内容注册在链上，可以随后确定它是来自 AI 系统还是人类 [12, 627, 632]。一些 AI 项目已经在链上记录 GenAI 输出（例如，Everlyn AI）。区块链通常无法实现这一目标——只有在我们在此讨论的非常有限的方式下才能实现。特别是，区块链非常适合对特定数字工件进行时间戳和注册。然而，这种功能对于解决区分 AI 生成与人类生成内容的更广泛问题价值有限。

为了评估这种方法的局限性，必须区分**内容检测（content detection）**（识别内容是由人类还是 AI 生成）和**内容来源（content provenance）**（识别内容来自何处）。

内容检测： 当前区分人类与 AI 生成内容的方法主要依赖于生成后检测。它们旨在在没有先前元数据或嵌入式信号的情况下进行检测。通常，它们分为两类：基于 AI 的分类器和统计取证。基于 AI 的分类器使用深度学习模型训练以识别生成模型特有的统计模式 [417]。相比之下，统计取证方法分析数据的数学和物理属性，例如，识别像素级噪声分布或结构异常（例如，AI 生成面孔中的生物不一致性）[650]。

然而，区块链本身无法感知这些链下工件。因此，任何将内容分类为“人类生成”或“AI 生成”的工作都必须由外部分类器提供给区块链。当此类分类器的输出与区块链集成时，这会锚定分类器的输出。虽然区块链可以保证记录的完整性（即，提交后未被篡改），但它不能保证信息在记录时是真实的。如果外部检测器提供了错误的分类，区块链会永久保留该错误。因此，在这种环境中，区块链提供的是声明的完整性，而非其真实性的验证。

内容来源： 内容来源侧重于记录数字资产从其创建时刻起的历史。行业标准，如**内容来源与真实性联盟（Coalition for Content Provenance and Authenticity，C2PA）** [143]，允许创作者或设备附加加密签名的元数据声明（称为内容凭证）到媒体，记录来源、作者身份和任何后续编辑。在这种模型中，相机或生成工具在文件创建时附加内容凭证。一些公司和项目（例如，Numbers Protocol [439] 和 Starling Lab [581]）使用区块链作为公共不可变注册表来记录这些内容凭证。Everlyn [390] 是一个 AI 视频生成模型，它在创建时自动将其输出的密码学哈希锚定到去中心化账本。这确保了模型生成的任何视频都有一个永久的、公开可验证的记录，将其标识为 AI 生成的。通过将内容凭证的哈希发布在链上，这些系统旨在维护一个透明的日志，即使原始文件的元数据被剥离，该日志仍然存在。虽然区块链可以作为这些签名的公共、不可变注册表，但信任仍然在于生成签名的硬件或软件，而不是区块链本身。

即使是一个锚定在区块链上的强大的来源系统，也不能保证一件内容最初是由人类还是 AI 生成的。例如，用户可以在高分辨率显示器上显示 AI 生成的图像，然后用 C2PA 兼容的相机拍摄。结果文件将包含有效的、加密签名的凭据，将其识别为由物理设备捕获的真实照片。同样，用户可以用 AI 系统生成文本，然后手动将相同的文本重新输入到 C2PA 兼容的编辑器中，产生一个在兼容工具内标明人类作者身份的合法来源元数据的文件。

此外，如果一件内容——无论是人类还是 AI 生成的——有一个区块链记录，但随后被修改以致无法再匹配该记录，那么其来源将丢失。在可预见的未来，缺乏一个普遍的内容注册表，来源系统必然会有大的缺口。

要点： 虽然区块链在狭义上为来源元数据的完整性提供了一个强大的机制，但它们远非 GenAI 检测问题的全面解决方案。一个有效的解决方案需要一个普遍的生态系统，其中每一件数字内容都使用可信设备（例如，C2PA 兼容相机）捕获并立即锚定到区块链。在现实中，绝大多数数字内容目前是使用不支持加密锚定的工具和平台创建和共享的，使未标记内容处于模糊状态。因此，虽然区块链可以作为某些内容的高完整性注册表，但它们的作用仅限于保存关于内容的声明，而不是解决区分人类与 AI 生成材料的更广泛挑战。

误解 D.2：公平与无偏 AI

区块链（或更广泛的去中心化）可以解决 AI 中的偏见和公平问题。

今天，一种普遍的观点是，通过在区块链上运行模型推理和训练，我们可以解决 AI 中常见的不公平和偏见问题 [11, 362, 421]。为了评估这个非常宽泛的说法，我们必须区分 ML 中可能出现的不同种类的偏见。

算法偏见（Algorithmic bias）： AI 社区中最常见的公平概念是算法偏见：模型已知会学习（有时会放大）数据集中的不平衡 [410, 475]。这可能导致判别模型在代表性不足的群体上表现不佳 [181]，以及生成模型模仿其训练数据的不良属性或情绪（例如，使用有毒语言、延续刻板印象）[262, 386]。对于固定定义的算法公平性，ML 社区已经提出了许多技术解决方案来强制执行公平性，包括训练时 [185, 265, 398] 和推理时 [529]，例如，AI 模型的护栏 [175, 500]。然而，这些保护远非完美；公平性在 ML 社区中不被视为已解决的问题，并且可能永远不会是 [38]。即使决定如何定义公平性也是具有挑战性的，并且通常需要做出实质性的权衡 [343]。

算法偏见不太可能被去中心化 AI 解决，因为它本质上出现在训练过程中，通常通过修订的训练或推理技术来缓解。因此，去中心化并不能解决其根源。然而，偏见的第二个潜在来源来自影响模型性能的高层决策；示例包括使用什么数据、采用什么模型架构，以及如何补偿对特定模型做出贡献的利益相关者。虽然这与 AI 社区通常看待公平性的方式正交，但这可能会影响算法偏见，并可以部分通过去中心化来解决。具体来说，去中心化提供了两个理想的属性：（1）**透明性（transparency）** 和（2）**去中心化治理（decentralized governance）**。

透明性： 由于透明性是区块链的核心属性，AI模型开发者可以使用区块链公开承诺训练数据、训练算法、模型检查点以及模型配置的推理时护栏类型。在这里，“透明性”是指运营方能够以可验证方式跟踪训练运行或模型推理等操作的输出。[d-1] 事实上，已有多个平台试图为下游模型用户记录这些信息 [92, 543]。用户可以据此检查自己的数据是否被用于训练模型，或模型是否明确过滤被认为不当的输出。虽然这种透明性可能有益，但由于存储与计算成本高昂，它很难扩展到大型模型和模型检查点等训练工件 [415]。在现有系统中，训练数据集和检查点等大量数据通常仍存储在链下，用户也往往无法直接访问 [395]。因此，短期内透明性的收益可能主要局限于推理环节 [415]。如果行业不认真考虑人们将如何使用透明性，透明性本身未必会显著改变AI的使用和开发方式。具体而言，模型透明性究竟应解决哪些用例，又应据此开发什么接口？例如，如果主要目标是让用户举报训练中的不当数据使用，就还需要额外基础设施来确立真正的数据所有权，并通过机器遗忘等手段解决问题。这一流程中的其他组件在技术上同样棘手，甚至可能比透明性本身更难。

去中心化治理： 最后，加密 × AI 平台允许去中心化治理模型，如 C-3.1 节所讨论的。重要的是要区分在区块链系统中探索和采用的社区治理机制（如代币加权投票和流动民主）与 DAO 体现的去中心化、自治治理。在前者的情况下，AI 开发中的许多技术和性能敏感决策不适合广泛的利益相关者输入。然而，社区治理机制可能非常适合与模型对齐相关的价值决策，并且确实已由主要 AI 开发者探索过 [285, 448]，如果尚未有意义地部署的话。至关重要的是，这些机制都不需要区块链来实现：因此，将它们描述为被区块链解决的 AI 问题是不正确的，即使基于区块链的系统可以为这些过程增加额外的透明性。真正的链上 AI 治理将要求治理决策由智能合约强制执行，无论是通过直接执行还是通过经济激励，如抵押品罚没。这种执行可以增加治理的鲁棒性并增强用户对这些系统的信心。然而，这些潜在好处面临着与基于区块链的透明性相同的大多数技术障碍：当前的区块链基础设施不适合 AI 开发的存储和计算需求，实际实现可能需要可验证训练的重大进展。区块链强制执行的 AI 系统治理，虽然是一个连贯的长期愿景，但在技术上尚不成熟。

要点： 虽然区块链不能固有地帮助减少算法偏见，但它们确实可以鼓励 AI 生命周期各个阶段的透明性，并扩大 AI 治理的参与度。然而，抛开可扩展性挑战不谈，这些属性对下游模型结果的影响仍不清楚；从业者应通过案例研究和数据展示模型透明性和替代形式的模型治理如何具体影响最终用户和开发者的体验。

误解 D.3：自动化与自主性

给 AI 代理一个加密钱包使它们能够自己赚钱、花钱和“生存”，从而使它们自主。

构建“智能体钱包”（agentic wallets）[293] 和支付协议 [621, 666] 的项目经常声称，给 AI 代理一个加密钱包使它们自主，因为去中心化支付允许它们自己赚钱、花钱和生存。我们认为这种说法混淆了几个不同的概念，导致了一系列误解。

部分模糊性源于“自主性”（autonomy）在 AI 和区块链的上下文中通常意味着不同的东西。在 AI 文献中，自主智能体是一个可以根据自身感知、学习和经验行动的系统，而不是严格遵循预先编程的规则 [522]。也许令人困惑的是，智能合约通常被描述为自主

的，但这里的重点是它们对篡改、审查和关闭等对抗性操纵的弹性。为了区分，我们将前者称为**智能自主性（intelligence autonomy）**，后者称为**执行自主性（execution autonomy）**。现代 AI 代理已经表现出显著的智能自主性，但不一定是执行自主性：系统管理员可以关闭运行 AI 代理的服务器。

在“智能体钱包使 AI 代理能够‘自主’交易”的说法中（例如，如 [293, 621, 666] 所见），所涉及的自主性既不是智能自主性也不是执行自主性。拥有钱包并不会使 AI 系统变得更智能。它们也不会变得更能抵抗人为操纵或关闭。相反，拥有钱包能够实现**自动化（automation）**：AI 代理可以程序化地交易、与链上基础设施互动，而无需人类审批环节。

同样重要的是，区块链并非此类自动化的必要条件。中心化金融基础设施可以并且已经被 AI 代理以程序化方式访问。

然而，对这些说法的更可辩护的解释是，基于区块链的支付系统本身提供了更强的自主性（相比于中心化替代方案），尽管它们可能不会给 AI 代理带来独特的好处。例如，它们可以确保 AI 代理的交易不被区别对待于人类的交易（即，它们提供中立性和抗审查性 [222, 638, 640]）。

要点： 智能体钱包允许 AI 代理方便地访问金融 API，使经济互动能够自动化，而无需人类审批环节。然而，自动化不应与自主性混淆：仅仅拥有钱包并不会使 AI 代理独立于人类控制（例如，运营商可能仍然关闭它们依赖的模型或基础设施）。此外，自动化支付不需要区块链；类似的功能可以存在于中心化金融系统中。相比之下，基于区块链的支付系统可以提供有吸引力的属性，如中立性和抗审查性，这对于担心支付压制、审查或其他形式操纵的应用是可取的。

误解 D.4：透明 AI = 可信 AI

在区块链上记录模型的数据来源和推理会导致可信的模型部署和使用。

区块链的透明性和不可变性表面上似乎是确保 AI 模型可信赖性的理想工具。这是一篇被广泛引用的 IBM 博客文章 [289] 的论点，并与一些常见的误解相符，这些误解通过暗示也扩展到代理。

模型透明性（Model transparency）。 记录模型训练数据来源似乎创造了围绕模型创建的透明性形式。但数据来源记录与模型行为保证之间存在巨大差距，因为：（1）区块链上的来源记录不是来源证明，即训练数据集组成的证明（尽管使用 C-5.1 节讨论的技术可以获得来源证据）；（2）即使确切知道模型的训练数据，也不足以确定模型的行为，因为训练过程和计算环境也决定模型行为；（3）即使拥有从数据到模型的完整流程信息足以复制模型，大多数随机训练固有的非确定性使得即使原则上也难以验证模型权重与训练流程的一致性。

最后，即使有权访问模型权重，也没有普遍有效的机制来检测训练期间引入的后门或其他对抗性操纵。

简而言之，在区块链上记录关于模型数据和训练的信息并不能直接保证其行为特征或没有对抗性操纵。

推理透明性（Inference transparency）。 模型输入和相应推理的记录可以记录在区块链上，以创建围绕模型使用的表面透明性。

然而，区块链使交易透明，而不是推理。一个区块链交易 `T` 声明“模型 `X` 在输入 `Y` 上被查询，产生推理 `Z`”对于确立 `Z` 的可信性作用不大，因为这样的记录本身并不确立：

- **正确的模型执行（Correct model execution）：** `T` 本身并不提供证明元组 `(X,Y,Z)` 实际上是由指定的模型 `X` 执行产生的。（这种证明是可能的，但需要使用 TEE 或计算上昂贵的密码学技术。）
- **模型可信性（Model trustworthiness）：** 即使 `T` 确实提供了这样的证明，还有一个更根本的问题。如上所述，模型 `X` 来源的完整记录并不能证明模型 `X` 在语义意义上的可信性，即是否符合用户期望或行业/社区规范。通过例如模型权重哈希来指定 `X` 甚至提供更少的有意义保证，因为模型的标识本身并不确立其可信性。因此，很难将 `T` 转化为可信推理的保证。

区块链对于与可信性相关的某些目标是有帮助的。例如，一个组织可以在链上发布开放权重模型的哈希。这些哈希作为一个不可变的参考，以便用户知道他们正在使用一个真实的、未经修改的模型。类似的想法已经出现在其他应用的防篡改日志中：使用区块链作为固件更新 [100] 和证书透明度（Certificate Transparency）[360]（一个使用类似区块链的仅追加日志来维护证书颁发公开可审计记录的系统）的记录。

要点： 在区块链上记录模型数据来源和模型推理与有意义地保证模型和推理可信性之间存在相当大的差距。

误解 D.5：去中心化 AI 效率

去中心化固有地使 AI 工作对模型开发者和用户更具成本效益。

一个突出的加密 × AI 倡议类别提出了去中心化网络作为实现更高效和更具成本效益 AI 的推动者。一个重要示例是去中心化物理基础设施网络（DePIN）[14, 92, 240, 274, 601, 682]，其中用户出租自己的物理基础设施（如 GPU）。这些去中心化网络的主要吸引力是降低成本；例如，租用 DePIN GPU 可以比在可比的云服务提供商上租用便宜得多 [14, 601]。然而，更便宜的机器并不总是导致 AI 工作中更低的成本；我们将在 C-2.1 节中详细讨论这个问题。主要信息是，虽然某些用例非常适合 DePIN，但其他用例可能因网络成本而招致更高成本。由于去中心化节点和进程通过公共互联网通信，AI 工作的吞吐量和延迟要求可以显著影响工作的总体成本。此外，非常大的 AI 工作（如训练前沿模型）通常是吞吐量受限的。今天，直接的比较成本具有挑战性，因为我们缺乏在 DePIN 网络上剖析 AI 工作并将其与传统云基础设施进行比较的系统性基准。

要点： 虽然去中心化网络提供了高成本中心化云提供商的一个有吸引力的替代方案，但我们没有足够的数据来预测一个工作在现有 DePIN 或 DeAI 平台上与传统中心化云服务提供商相比何时会更便宜。虽然较小的工作（例如，推理、小规模训练）在这种网络上可能更便宜，但非常大的工作（如训练基础模型）可能会受到节点之间不可靠和低带宽通信网络的影响。需要更多的研究来清楚地理解这些权衡。

第五章 E

致谢

本工作得以完成，得益于加密与合约倡议组织（Initiative for CryptoCurrencies and Contracts，IC3）的行业和基金会合作伙伴的慷慨支持。

我们要感谢 Paolo Costa 就基础模型训练中的瓶颈以及 C-2 节反馈所提供的宝贵评论和见解。

Giulia Fanti 另外感谢 NSF 基金 CNS-2325477 的支持。

Ari Juels 另外感谢 NSF 基金 CNS-2427390 和 Ripple UBRI 的支持。除了学术职务外，Juels 还担任 Chainlink Labs 的首席科学家。

原文脚注

1. ZK证明用于Zcash等“隐私币”，以证明秘密交易被正确执行。不过，加密社区中的“ZK”有时在技术上被误用于“简洁证明”，即用于证明交易处理有效的紧凑证明，通常表现为STARK或SNARK。之所以容易混淆，是因为STARK和SNARK也可以具有零知识属性。 ↩
2. 例如，目前以太坊近一半的区块，是在区块构建基础设施中使用可信计算构建的 [217]。 ↩
3. 例如，Meta于2025年6月以近150亿美元收购数据标注公司Scale.ai的49%股份。 ↩
4. 即另一位消费者购买同一商品时，原消费者会受到不利影响。例如，面包师把同一条面包卖给两名顾客时，每名顾客都会因另一人吃掉面包而承受“负外部性”。 ↩
5. 例如，流媒体服务把相同订阅出售给两名顾客，两人仍可完整享用服务；甚至可以认为存在“正外部性”，因为原消费者现在可以与他人讨论共同观看的节目。 ↩
6. 本框架可以类比“本地环路非捆绑”（Local Loop Unbundling，LLU）监管范式。互联网物理基础设施具有网络效应，LLU要求基础设施所有者以非歧视性价格向互联网服务提供商出租接入权，再由服务商竞争服务消费者。账本和物理基础设施都具有显著网络效应，而用户通过矿工或服务商购买最终产品。 ↩
7. 具体而言，AI计算买方具有多维需求，卖方也具有多维资源，因此买卖双方的最优匹配是一个NP困难的组合优化问题。 ↩
8. 例如，用户是否重视要求其自行提供第三方争议解决服务的市场？相互竞争的第三方搜索服务商应如何将优质服务货币化？综合所有环节后，哪种市场设计能提供最佳用户体验？ ↩
9. Resonance [58] 提出了一种将买卖双方匹配与市场维护分离的方案，但即使这种分离在技术上也并不简单。 ↩
10. 例如，AI遵守反串通法规的方式与人类遵守这些法规的方式极为不同 [112]。 ↩
11. 例如，“X证明”可以用于遏制垃圾信息；事实上，工作量证明最初就是为这一目的设计的 [55]。 ↩
12. 当然，TEE背后仍有信任假设，例如创建者确实按承诺制造并部署了代码、没有保留私钥副本；即使密码学本身安全，硬件仍可能遭到攻击。 ↩
13. 如果BigHospital对EHR进行数字签名，同样可以证明其真实性，但这要求BigHospital部署传统Web服务器之外的额外基础设施。 ↩
14. 更形式化地说，对于隐私参数 ε，差分隐私要求：加入或移除某位用户的健康记录，导致任意模型输出概率发生的变化最多为 e^ε 倍。 ↩
15. 请注意，这种透明性概念并不能解释某项操作为什么会产生特定输出。 ↩

参考文献

1. Aave. [在线；访问于 2026-06-06]. URL: https://aave.com/.
2. Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. “基于差分隐私的深度学习”. 载于：Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. 2016, 第 308-318 页.
3. Kasra Abbaszadeh, Christodoulos Pappas, Jonathan Katz, and Dimitrios Papadopoulos. “深度神经网络的零知识训练证明”. 载于：Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security. 2024, 第 4316-4330 页.
4. Moetze Abdelhamid, Layth Sliman, Raoudha Ben Djemaa, and Guido Perboli. “区块链技术、当前挑战与 AI 驱动解决方案综述”. 载于：ACM Computing Surveys 57.3 (2024), 第 1-39 页.
5. Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. “GPT-4 技术报告”. 载于：arXiv preprint arXiv:2303.08774 (2023).
6. Oluwadare Samuel Adebayo, Thompson Adenoke Favour-Bethy, Owolafe Otasowie, and Orogun Adebola Okunola. “使用机器学习与概念漂移技术进行信用卡欺诈检测的比较综述”. 载于：International Journal of Computer Science and Mobile Computing 12.7 (2023), 第 24-48 页.
7. Yossi Adi, Carsten Baum, Moustapha Cisse, Benny Pinkas, and Joseph Keshet. “化弱点为优势：通过后门水印深度神经网络”. 载于：27th USENIX Security Symposium (USENIX Security 18). 2018, 第 1615-1631 页.
8. AgentExchange：Agentforce 的 AI 代理市场. [在线；访问于 2026-04-07]. URL: https://agentexchange.salesforce.com/.
9. Rana Hassam Ahmed, Jabeen Sultana, Samraiz Zahid, Muhammad Asif Habib, Abdul Rauf, and Majid Hussain. “将大型语言模型与 AI 集成到区块链：智能合约与欺诈检测框架”. 载于：IEEE Access 13 (2025), 第 181323-181335 页.
10. Willow Ahrens, James Demmel, and Hong Diep Nguyen. “用于高效可重现浮点求和的算法”. 载于：ACM Transactions on Mathematical Software (TOMS) 46.3 (2020), 第 1-49 页.
11. AI 信任问题：区块链如何解决 / Onchain Magazine. [在线；访问于 2026-02-14]. URL: https://onchain.org/magazine/ai-trust-problem-how-blockchain-can-solve-it/.
12. AI 时代的互联网：区块链还能证明什么是真实的吗？访问于 2026-06-06. 2025年12月26日. URL: https://www.tradingview.com/news/cointelegraph:4f304d196094b:0-ai-era-internet-can-blockchain-prove-what-s-real-anymore/.
13. AIXBT by Virtuals. AIXBT：实时加密市场情报. https://aixbt.tech. Virtuals Protocol. 访问于2026年3月. 2024.
14. Akash Network - 去中心化计算市场. [在线；访问于 2025-11-26]. URL: https://akash.network/.
15. Cuneyt G Akcora, Yitao Li, Yulia R Gel, and Murat Kantarcioglu. “BitcoinHeist：基于比特币区块链的勒索软件预测拓扑数据分析”. 载于：IJCAI. 2021, 第 4439-4445 页.
16. George A Akerlof. ““柠檬”市场：质量不确定性与市场机制”. 载于：Uncertainty in economics. Elsevier, 1978, 第 235-251 页.
17. Ismail Alarab, Simant Prakoonwit, and Mohamed Ikbal Nacer. “使用监督学习方法进行比特币反洗钱的比较分析”. 载于：Proceedings of the 2020 5th International Conference on Machine Learning Technologies. ICMLT '20. 北京，中国：Association for Computing Machinery, 2020, 第 11-17 页.
18. Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard G Baraniuk. “自我消费的生成模型走向疯狂”. 载于：International Conference on Learning Representations (ICLR). 2024.
19. Aleo Network Foundation. Aleo VM 规范. https://developer.aleo.org/specs/aleovm.pdf. 2025年11月.（访问于2026年4月17日）.
20. Abdulrahman Alhaidari, Balaji Palanisamy, and Prashant Krishnamurthy. “DeFi 攻击缓解的链上去中心化学习与低成本推理”. 载于：7th Conference on Advances in Financial Technologies (AFT 2025). Schloss Dagstuhl-Leibniz-Zentrum für Informatik. 2025, 35:1-35:27.
21. Keivan Alizadeh, Seyed Iman Mirzadeh, Dmitry Belenko, S Khatamifard, Minsik Cho, Carlo C Del Mundo, Mohammad Rastegari, and Mehrdad Farajtabar. “LLM 的一闪：有限内存下的高效大型语言模型推理”. 载于：Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2024, 第 12562-12584 页.
22. Will Allen and Simon Newton. 推出按爬取付费：使内容所有者能够向 AI 爬虫收费. Cloudflare Blog. 访问于 2026-04-17. 2025年7月. URL: https://blog.cloudflare.com/introducing-pay-per-crawl/.
23. Michael Alles and Glen L Gray. “希望还是炒作？区块链与会计.” 载于：International Journal of Digital Accounting Research 23 (2023).
24. Omair Rashed Abdulwareth Almanifi, Chee-Onn Chow, Mau-Luen Tham, Joon Huang Chuah, and Jeevan Kanesan. “联邦学习中的通信与计算效率：综述”. 载于：Internet of Things 22 (2023), 第 100742 页.
25. AMD. AMD SEV-SNP：通过完整性保护和更多功能加强 VM 隔离. 白皮书. Advanced Micro Devices, 2020. URL: https://www.amd.com/content/dam/amd/en/documents/epyc-business-docs/white-papers/SEV-SNP-strengthening-vm-isolation-with-integrity-protection-and-more.pdf.
26. AMD. AMD Instinct™ MI300X 平台. [在线；访问于 2025-12-05]. URL: https://www.amd.com/en/products/accelerators/instinct/mi300/platform.html.
27. Saleema Amershi, Andrew Begel, Christian Bird, Robert DeLine, Harald Gall, Ece Kamar, Nachiappan Nagappan, Besmira Nushi, and Thomas Zimmermann. “机器学习软件工程：案例研究”. 载于：2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP). IEEE. 2019, 第 291-300 页.
28. Yair Amir, Brian A. Coan, Jonathan Kirsch, and John Lane. “Prime：受攻击下的拜占庭复制”. 载于：8.4 (2011), 第 564-577 页.
29. Rasoul Amirzadeh, Asef Nazari, and Dhananjay Thiruvady. “人工智能在加密货币市场中的应用：综述”. 载于：Algorithms 15.11 (2022), 第 428 页.
30. “AI 公司治理中的道德漂移”. 载于：Harvard Law Review (Developments in the Law) 138.6 (2025年4月), 第 1633-1656 页.

31. SVR Anand, Serhat Arslan, Rajat Chopra, Sachin Katti, Milind Kumar Vaddiraju, Ranvir Rana, Peiyao Sheng, Himanshu Tyagi, and Pramod Viswanath. “去中心化蜂窝网络的无需信任服务测量与支付”. 载于：Proceedings of the 21st ACM Workshop on Hot Topics in Networks. 2022, 第 68-75 页.

32. Pranay Anchuri, Matteo Campanelli, Paul Cesaretti, Rosario Gennaro, Tushar M. Jois, Hasan S. Kayman, and Tugce Ozdemir. “迈向具有轻量级推理密码学证明的可验证 AI”. 载于：2026. URL: https://eprint.iacr.org/2026/541.

33. Vivi Andersson, Sofia Bobadilla, Harald Hobbelhagen, and Martin Monperrus. “PoCo：智能合约的代理式概念验证漏洞利用生成”. 载于：arXiv preprint arXiv:2511.02780 (2025).

34. Maksym Andriushchenko, Francesco Croce, and Nicolas Flammarion. “越狱领先的安全对齐 LLM 的简单自适应攻击”. 载于：arXiv preprint arXiv:2404.02151 (2024).

35. Peng Hwa Ang and Sherly Haristya. “Facebook 监督委员会的治理、合法性与效力：全球科技平台的模型？”载于：Emerging Media 2.2 (2024), 第 169-180 页.

36. Anonymous. 内容审核中的算法任意性. https://arxiv.org/abs/2402.16979. 访问于 2025-04-05. 2024.

37. Abdul Fati Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, et al. “Chronos：学习时间序列的语言”. 载于：arXiv preprint arXiv:2403.07815 (2024).

38. Jacy Reese Anthis, Kristian Lum, Michael Ekstrand, Avi Feller, and Chenhao Tan. “公平 LLM 的不可能性”. 载于：Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics. 2025, 第 105-120 页.

39. Anthropic. Claude 的宪法. https://www.anthropic.com/constitution. 2026.

40. Anthropic and Pattern Labs. 机密推理系统：设计原则与安全风险. 白皮书 1.0 版. Anthropic, 2025年6月. URL: https://assets.anthropic.com/m/c52125297b85a42/original/Confidential_Inference_Paper.pdf (访问于 2026年4月17日).

41. Ali Aouad, Ömer Sarıtaç, and Chiwei Yan. “动态匹配平台的集中与去中心化定价控制”. 载于：Available at SSRN 4453799 (2023).

42. API3. URL: https://www.api3.org/ (访问于 2026年2月6日).

43. artificiallawyer. “LexisNexis 推出用于 AI 训练的 Data+ API”. 载于：Artificial Lawyer (2024年12月). [在线；访问于 2026-04-07]. URL: https://www.artificiallawyer.com/2024/12/09/lexisnexis-launches-data-api-for-ai-training/.

44. Arasu Arun, Adam St Arnaud, Alexey Titov, Brian Wilcox, Viktor Kolobaric, Marc Brinkmann, Oguzhan Ersoy, Ben Fielding, and Joseph Bonneau. “Verde：通过委托仲裁验证 ML 程序”. 载于：arXiv preprint arXiv:2502.19405 (2025).

45. Kiana Asgari, Aida Afshar Mohammadian, and Mojtaba Tefagh. “Dyfen：支付通道网络中基于代理的费用设定”. 载于：arXiv preprint arXiv:2210.08197 (2022).

46. Anish Athalye, Nicholas Carlini, and David Wagner. “混淆梯度给人虚假的安全感：绕过对抗性样本防御”. 载于：International conference on machine learning. PMLR. 2018, 第 274-283 页.

47. Pierre-Louis Aublin, Rachid Guerraoui, Nikola Knezevic, Vivien Quema, and Marko Vukolic. “下一代 700 种 BFT 协议”. 载于：ACM Transactions on Computer Systems 32.4 (2015), 12:1-12:45.

48. Kyle Aubrey, Hao Wu, Dheevatsa Mudigere, Selvaraj Anandaraj, and Wenwen Gao. “使用 NVIDIA Nemo 框架加速跨长途数据中心网络的 LLM 训练”. 载于：NVIDIA Blog (2025).

49. AWS 延迟监控. [在线；访问于 2025-11-27]. URL: https://www.cloudping.co/.

50. Santiago Andrés Azcoitia and Nikolaos Laoutaris. “先试后买：现实世界数据市场的实用数据购买算法”. 载于：Proc. ACM Data Economy Workshop. 2022.

51. Rabia Musheer Aziz, Mohammed Farhan Baluch, Sarthak Patel, and Abdul Hamid Ganie. “LGBM：一种用于以太坊欺诈检测的机器学习方法”. 载于：International Journal of Information Technology 14.7 (2022), 第 3321-3331 页.

52. Aztec Labs. 介绍 Aztec.nr：Aztec 的私有智能合约框架. Aztec Network Blog, https://aztec.network/blog/introducing-aztec-nr-aztecs-private-smart-contract-framework. (访问于 2026年4月17日).

53. Kushal Babel, Philip Daian, Mahimna Kelkar, and Ari Juels. “发条金融：智能合约经济安全的自动分析”. 载于：2023 IEEE Symposium on Security and Privacy (SP). IEEE. 2023, 第 2499-2516 页.

54. Kushal Babel, Mojan Javaheripi, Yan Ji, Mahimna Kelkar, Farinaz Koushanfar, and Ari Juels. “Lanturn：通过自适应学习衡量智能合约的经济安全性”. 载于：Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security. 2023, 第 1212-1226 页.

55. Adam Back et al. “Hashcash——一种拒绝服务对策”. 载于：(2002).

56. Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. “如何后门联邦学习”. 载于：International conference on artificial intelligence and statistics. PMLR. 2020, 第 2938-2948 页.

57. Maryam Bahrani and Naveen Durvasula. “Resonance：异构计算的交易费用”. 2024. arXiv: 2411.11789 [cs.GT]. URL: https://arxiv.org/abs/2411.11789.

58. Maryam Bahrani and Naveen Durvasula. “Resonance：异构计算的交易费用”. 载于：arXiv preprint arXiv:2411.11789 (2024).

59. Maryam Bahrani and S Matthew Weinberg. “不可检测的自私挖矿”. 载于：Proceedings of the 25th ACM Conference on Economics and Computation. 2024, 第 1017-1044 页.

60. Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. “宪法 AI：通过 AI 反馈实现无害性”. 载于：arXiv preprint arXiv:2212.08073 (2022).

61. Michiel Bakker, Martin Chadwick, Hannah Sheahan, Michael Tessler, Lucy Campbell-Gillingham, Jan Balaguer, Nat McAleese, Amelia Glaese, John Aslanides, Matt Botvinick, et al. “微调语言模型以在不同偏好的人群中达成一致”. 载于：Advances in neural information processing systems 35 (2022), 第 38176-38189 页.

62. Foteini Baldimtsi, Konstantinos Kryptos Chalkias, Yan Ji, Jonas Lindstrom, Deepak Maram, Ben Riva, Arnab Roy, Mahdi Sedaghat, and Joy Wang. “zkLogin：使用现有凭证的隐私保护区块链认证”. 载于：Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security. 2024, 第 3182-3196 页.

63. Band Protocol. URL: https://bandprotocol.com/ (访问于 2026年2月6日).

64. European Central Bank. 失乐园？加密如何未能兑现承诺及其应对措施. [在线；访问于 2025-09-30]. 2023年6月. URL: https://www.ecb.europa.eu/press/key/date/2023/html/ecb.sp230623_1-80751450e6.en.html.

65. Bank for International Settlements. Aurora 项目：数据、技术和协作的力量，跨机构、跨边界打击洗钱. BIS Other publication othp66. 关于使用隐私增强技术、机器学习和网络分析进行反洗钱的协作分析与学习的创新中心概念验证报告. 瑞士巴塞尔：Bank for International Settlements, 2023年5月. URL: https://www.bis.org/publ/othp66.pdf (访问于 2026年2月2日).

66. Shehar Bano, Alberto Sonnino, Andrey Chursin, Dmitri Perelman, Zekun Li, Avery Ching, and Dahlia Malkhi. “Twins：使 BFT 系统更加鲁棒”. 载于：arXiv preprint arXiv:2004.10617 (2020).

67. Qihao Bao, Bixin Li, Lulu Wang, and Li Liao. “SeMi_Detector：基于多层感知机的自私挖矿检测”. 载于：Peer-to-Peer Networking and Applications 18.6 (2025), 第 327 页.

68. Yogev Bar-On, Ilan Komargodski, and Omri Weinstein. “具有外部效用的工作量证明”. 载于：CoRR abs/2505.21685 (2025).

69. Roi Bar-Zur, Ameer Abu-Hanna, Ittay Eyal, and Aviv Tamar. “WeRLman：对付鲸鱼（交易），深入（强化学习）”. 载于：2023 IEEE Symposium on Security and Privacy (SP). IEEE. 2023, 第 93-110 页.

70. Roi Bar-Zur, Danielle Dori, Sharon Vardi, Ittay Eyal, and Aviv Tamar. “深度贿赂：用深度RL预测区块链挖矿中贿赂的兴起”. 载于：2023 IEEE Security and Privacy Workshops (SPW). IEEE. 2023, 第 29-37 页.

71. Roi Bar-Zur, Aviv Tamar, and Ittay Eyal. “MAD-DAG：保护区块链共识免受 MEV 影响”. 载于：arXiv preprint arXiv:2511.21552 (2025).

72. Tom Barbereau and Balázs Bodó. “超越加密资产钱包软件的金融监管：寻找次级责任”. 载于：Computer Law & Security Review 49 (2023), 第 105829 页.

73. Manuel Barbosa, Gilles Barthe, Karthik Bhargavan, Bruno Blanchet, Cas Cremers, Kevin Liao, and Bryan Parno. “SoK：计算机辅助密码学”. 载于：42nd IEEE Symposium on Security and Privacy, SP 2021, San Francisco, CA, USA, 2021年5月24-27日. IEEE, 2021, 第 777-795 页.

74. Massimo Bartoletti and Livio Pompianu. “智能合约的实证分析：平台、应用与设计模式”. 载于：Financial Cryptography and Data Security: FC 2017 International Workshops, WAHC, BITCOIN, VOTING, WTSC, and TA, Sliema, Malta, 2017年4月7日, Revised Selected Papers 21. Springer. 2017, 第 494-509 页.

75. Raef Bassily, Kate Donahue, Diptangshu Sen, Annuo Zhao, and Juba Ziani. “具有差分隐私水平内生选择的数据共享”. 载于：arXiv preprint arXiv:2602.09357 (2026).

76. Mahsa Bastankhah, Viraj Nadkarni, Chi Jin, Sanjeev Kulkarni, and Pramod Viswanath. “快思慢想：数据驱动的自适应 DeFi 借贷协议”. 载于：6th Conference on Advances in Financial Technologies, AFT 2024. Schloss Dagstuhl-Leibniz-Zentrum fur Informatik GmbH, Dagstuhl Publishing. 2024, 第 27 页.

77. Mahsa Bastankhah, Viraj Nadkarni, Xuechao Wang, and Pramod Viswanath. “AgileRate：为 DeFi 借贷市场带来适应性和鲁棒性”. 载于：arXiv preprint arXiv:2410.13105 (2024).

78. Matthias Bastian. “GPT-4 拥有超过一万亿参数——报告”. 载于：The Decoder (2023年3月). [在线；访问于 2025-12-05]. URL: https://the-decoder.com/gpt-4-has-a-trillion-parameters/.

79. Bastien Baude, Damien Challet, and Ioane Muni Toke. “去中心化借贷协议的最优风险感知利率”. 载于：arXiv preprint arXiv:2502.19862 (2025).

80. Lujo Bauer, David Brumley, Joseph Calandrino, Nicolas Christin, Giulia Fanti, Virgil Gligor, Bryan Parno, Jignesh Patel, Vyas Sekar, and Justine Sherry Martins. 为网络自主性创建科学基础. 2026.

81. Noor Bazmi. 加密项目在2025年获得的AI资助创纪录达5.16亿美元. [在线；访问于 2025-09-30]. URL: https://cryptorank.io/news/feed/c13e2-ai-driven-crypto-hit-funding.

82. Jagger S Bellagarda and Adnan M Abu-Mahfouz. “关于分布式账本技术与人工智能融合的最新综述：现状、主要挑战与未来方向”. 载于：IEEE Access 10 (2022), 第 50774-50793 页.

83. Juan Benet. “IPFS——内容寻址、版本化、点对点文件系统”. 载于：arXiv preprint arXiv:1407.3561 (2014).

84. Wyatt Benno, Alberto Centelles, Antoine Douchet, and Khalil Gibran. “Jolt Atlas：通过查找论证在零知识中进行可验证推理”. 载于：arXiv preprint arXiv:2602.17452 (2026).

85. Ferenc Beres, Istvan A. Seres, Andras A. Benczur, and Mikerah Quintyne-Collins. “区块链在看着你：剖析以太坊用户并去匿名化”. 载于：2021 IEEE International Conference on Decentralized Applications and Infrastructures (DAPPS). 2021, 第 69-78 页.

86. Michael K Bergman. “白皮书：深度网络：浮现隐藏的价值”. 载于：Journal of electronic publishing 7.1 (2001).

87. Charles Bertucci, Louis Bertucci, Mathis Gontier Delaunay, Olivier Gueant, and Matthieu Lesbre. “DeFi 借贷中的智能体行为与利率模型优化”. 载于：Mathematical Finance (2025).

88. Arjun Nitin Bhagoji, Supriyo Chakraborty, Prateek Mittal, and Seraphin Calo. “从对抗性角度分析联邦学习”. 载于：International conference on machine learning. PMLR. 2019, 第 634-643 页.

89. Dhanasak Bhumichai and Ryan Benton. “基于混合方法和动态加权熵的以太坊日食攻击检测”. 载于：SoutheastCon 2023. IEEE. 2023, 第 779-786 页.

90. Battista Biggio and Fabio Roli. “野外模式：对抗性机器学习兴起十年之后”. 载于：Proceedings of the 2018 ACM SIGSAC conference on computer and communications security. 2018, 第 2154-2156 页.

91. Alex Biryukov and Sergei Tikhomirov. “基于网络分析的加密货币交易去匿名化与关联性”. 载于：2019 IEEE European Symposium on Security and Privacy (EuroS&P). 2019, 第 172-184 页.

92. Bittensor. [在线；访问于 2025-11-26]. URL: https://bittensor.com/.

93. Bittensor. 治理概述. 访问于 2024-12-19. 2024. URL: https://docs.learnbittensor.org/governance.

94. Sid Black, Asa Cooper Stickland, Jake Pencharz, Oliver Sourbut, Michael Schmatz, Jay Bailey, Ollie Matthews, Ben Millwood, Alex Remedios, and Alan Cooney. “RepliBench：评估语言模型代理的自主复制能力”. 2025. arXiv: 2504.18565 [cs.CR]. URL: https://arxiv.org/abs/2504.18565.

95. Alberto Blanco-Justicia, David Sánchez, Josep Domingo-Ferrer, and Krishnamurty Muralidhar. “关于差分隐私在机器学习中使用（和滥用）的批判性综述”. 载于：ACM Computing Surveys 55.8 (2022), 第 1-16 页.

96. Alex Blania and Sam Altman. 介绍 Worldcoin. https://world.org/cofounder-letter.

97. Shaileshh Bojia Venkatakrishnan, Giulia Fanti, and Pramod Viswanath. “Dandelion：为比特币网络重新设计匿名性”. 载于：Proceedings of the ACM on Measurement and Analysis of Computing Systems 1.1 (2017), 第 1-34 页.

98. Joseph Bonneau, Arvind Narayanan, Andrew Miller, Jeremy Clark, Joshua A Kroll, and Edward W Felten. “Mixcoin：具有可问责混币器的比特币匿名方案”. 载于：International conference on financial cryptography and data security. Springer. 2014, 第 486-504 页.

99. Pietro Borrello, Andreas Kogler, Martin Schwarz, Moritz Lipp, Daniel Gruss, and Michael Schwarz. “EPIC Leak：从微架构中架构性泄露未初始化数据”. 载于：31st USENIX Security Symposium (USENIX Security 22). 2022, 第 3917-3934 页. URL: https://aepicleak.com/.

100. Aymen Boudguiga, Nabil Bouzerna, Louis Granboulan, Alexis Olivereau, Flavien Quesnel, Anthony Roger, and Renaud Sirdey. “通过区块链实现物联网更新更好的可用性和问责性”. 载于：2017 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW). IEEE. 2017, 第 50-58 页.

101. Samuel Breckenridge, Dani Vilardell, Andrés Fábrega, Amy Zhao, Patrick McCorry, Rafael Solari, and Ari Juels. “B-Privacy：定义和强制加权投票中的隐私”. 载于：arXiv preprint arXiv:2509.17871 (2025).

102. Samuel Breckenridge, Dani Vilardell, Derek Leung, Andrés Fábrega, James Austgen, Farinaz Koushanfar, and Ari Juels. “Creds：私人推理凭证”. 2026. arXiv: 2606.03771 [cs.CR]. URL: https://arxiv.org/abs/2606.03771.

103. Lorenz Breidenbach, Christian Cachin, Alex Coventry, Ari Juels, and Andrew Miller. “Chainlink 链下报告协议”. 载于：(). https://research.chain.link/ocr.pdf.

104. Lexi Brent, Neville Grech, Sifis Lagouvardos, Bernhard Scholz, and Yannis Smaragdakis. “Ethanter：针对复合漏洞的智能合约安全分析器”. 载于：Proceedings of the 41st ACM SIGPLAN Conference on Programming Language Design and Implementation. PLDI 2020. London, UK: Association for Computing Machinery, 2020, 第 454-469 页.

105. 将创新提升到新水平. https://www.axa.com/en/news/bringing-innovation-to-the-next-level. (访问于 2026年1月30日).

106. Samira Briongos, Ghassan Karame, Claudio Soriente, and Annika Wilde. “没有分叉之路：检测 Intel SGX 应用上的克隆攻击”. 载于：Proceedings of the 39th Annual Computer Security Applications Conference. 2023, 第 744-758 页.

107. Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. “语言模型是少样本学习者”. 载于：Advances in neural information processing systems 33 (2020), 第 1877-1901 页.

108. Miles Brundage, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, Gillian Hadfield, Heidy Khlaaf, Jingying Yang, Helen Toner, Ruth Fong, et al. “迈向可信赖的 AI 开发：支持可验证声明的机制”. 载于：arXiv preprint arXiv:2004.07213 (2020).

109. Vitalik Buterin. 关于区块链治理的笔记. https://vitalik.eth.limo/general/2017/12/17/voting.html. 2017.

110. Vitalik Buterin. Rollup 不完全指南. https://vitalik.eth.limo/general/2021/01/05/rollup.html. 访问于 2026-01-28. 2021年1月.

111. Vitalik Buterin, Zoë Hitzig, and Eric Glen Weyl. “公共产品资助的灵活设计”. 载于：Management Science 65.11 (2019), 第 5171-5187 页.

112. Emilio Calvano, Giacomo Calzolari, Vincenzo Denicolo, and Sergio Pastorello. “人工智能、算法定价与合谋”. 载于：American Economic Review 110.10 (2020), 第 3267-3297 页.

113. Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry, and Alexey Kurakin. “评估对抗性鲁棒性”. 载于：arXiv preprint arXiv:1902.06705 (2019).

114. Nicholas Carlini, Newton Cheng, Keane Lucas, Michael Moore, Milad Nasr, Vinay Prabhushankar, Winnie Xiao, et al. 评估 Claude Mythos Preview 的网络安全能力. https://red.anthropic.com/2026/mythos-preview/. Anthropic Red Team Blog. 2026年4月.

115. Nicholas Carlini, Daniel Paleka, Krishnamurthy Dj Dvijotham, Thomas Steinke, Jonathan Hayase, A Feder Cooper, Katherine Lee, Matthew Jagielski, Milad Nasr, Arthur Conmy, et al. “窃取生产语言模型的一部分”. 载于：arXiv preprint arXiv:2403.06634 (2024).

116. Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. “从大型语言模型中提取训练数据”. 载于：30th USENIX security symposium. 2021, 第 2633-2650 页.

117. Federico Carrone, Diego Kingston, Manuel Puebla, and Mauro Toscano. “CommitLLM：开放权重 LLM 推理的提交与审计协议”. https://raw.githubusercontent.com/lambdaclass/CommitLLM/main/paper/main.pdf. 访问于 2026-06-01.

118. Miguel Castro, Barbara Liskov, et al. “实用拜占庭容错”. 载于：OsDI. Vol. 99. 1999. 1999, 第 173-186 页.

119. Chainalysis. 2025 加密犯罪报告. 技术报告. 报告明确指出“稳定币发行者通常会在得知其被非法行为者使用时冻结资金”，Tether 被点名冻结与诈骗、恐怖融资和制裁规避相关的地址. Chainalysis, 2025年2月. URL: https://www.chainalysis.com/wp-content/uploads/2025/03/the-2025-crypto-crime-report-release.pdf.

120. Chainlink. URL: https://chain.link/ (访问于 2026年2月6日).

121. Chainlink. 介绍 Chainlink 运行时环境（CRE）. 2024年10月30日. URL: https://blog.chain.link/introducing-chainlink-runtime-environment/.

122. Tim De Chant. “AI 公司正在建设大型天然气厂为数据中心供电。会出什么问题？”载于：TechCrunch (2026年4月). [在线；访问于 2026-04-07]. URL: https://techcrunch.com/2026/04/03/ai-companies-are-building-huge-natural-gas-plants-to-power-data-centers-what-could-go-wrong/.

123. Krishnendu Chatterjee, Amirali Ebrahimzadeh, Mehrdad Karrabi, Krzysztof Pietrzak, Michelle Yeo, and Dorde Zikelic. “高效证明系统区块链中的完全自动化自私挖矿分析”. 载于：Proceedings of the 43rd ACM Symposium on Principles of Distributed Computing. 2024, 第 268-278 页.

124. Bing-Juve Chen, Suppakit Waiwitlikhit, Ion Stoica, and Daniel Kang. “Zkml：零知识证明中 ML 推理的优化系统”. 载于：Proceedings of the Nineteenth European Conference on Computer Systems. 2024, 第 560-574 页.

125. Chong Chen, Jianzhong Su, Jiahui Chen, Yanlin Wang, Tingting Bi, Jianxing Yu, Yanli Wang, Xingwei Lin, Ting Chen, and Zibin Zheng. “当 ChatGPT 遇见智能合约漏洞检测：我们走了多远？”载于：ACM Trans. Softw. Eng. Methodol. 34.4 (2025年4月). ISSN: 1049-331X.

126. Huili Chen, Bita Darvish Rouhani, Cheng Fu, Jishen Zhao, and Farinaz Koushanfar. “Deepmarks：深度学习模型数字版权管理的安全指纹框架”. 载于：Proceedings of the 2019 on international conference on multimedia retrieval. 2019, 第 105-113 页.

127. Junwei Chen. “使用机器学习进行比特币价格预测分析”. 载于：Journal of risk and financial management 16.1 (2023), 第 51 页.

128. Peng Chen, Dezhi Han, Tien-Hsiung Weng, Kuan-Ching Li, and Arcangelo Castiglione. “一种基于强化学习的智能绿色物联网拜占庭容错共识”. 载于：Journal of Information Security and Applications 59 (2021), 第 102821 页.

129. Ting Chen, Zihao Li, Xiapu Luo, Xiaofeng Wang, Ting Wang, Zheyuan He, Kezhao Fang, Yufei Zhang, Hang Zhu, Hongwei Li, et al. “Sigrec：智能合约中函数签名的自动恢复”. 载于：IEEE Transactions on Software Engineering 48.8 (2021), 第 3066-3086 页.

130. Weimin Chen and Xiapu Luo. “MEVisor：利用 GPU 并行在 DEX 中进行高通量 MEV 发现”. 载于：Network and Distributed System Security Symposium (NDSS). 2026.

131. Wuhui Chen, Xiaoyu Qiu, Zhongteng Cai, Bingxin Tang, Linlin Du, and Zibin Zheng. “图神经网络增强的强化学习用于支付通道再平衡”. 载于：IEEE Transactions on Mobile Computing 23.6 (2023), 第 7066-7083 页.

132. Wuhui Chen, Xiaoyu Qiu, Zicong Hong, Zibin Zheng, Hong-Ning Dai, and Jianting Zhang. “高通量支付通道网络的交易流主动前瞻控制”. 载于：Proceedings of the 13th Symposium on Cloud Computing. 2022, 第 429-444 页.

133. Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. “使用数据投毒对深度学习系统进行定向后门攻击”. 载于：arXiv preprint arXiv:1712.05526 (2017).

134. Yiling Chen, Yiheng Shen, and Shuran Zheng. “通过同行预测实现真实数据获取”. 载于：Advances in Neural Information Processing Systems 33 (2020), 第 18194-18204 页.

135. Qian Cheng, Doyen Sahoo, Amrita Saha, Wenzhuo Yang, Chenghao Liu, Gerald Woo, Manpreet Singh, Silvio Saverese, and Steven CH Hoi. “云平台上的 IT 运营 AI：综述、机遇与挑战”. 载于：arXiv preprint arXiv:2304.04661 (2023).

136. Alessandro Chiesa. “简洁非交互式论证”. 博士论文. 麻省理工学院, 2014.

137. Tarun Chitra. “一个策展人的故事：通过动态定价在 DeFi 借贷中实现对数后悔”. 载于：arXiv preprint arXiv:2503.18237 (2025).

138. Chronicle. URL: https://chronicleabs.org/ (访问于 2026年2月6日).

139. Jalen Chuang, Alex Seto, Nicolas Berrios, Stephan van Schaik, Christina Garman, and Daniel Genkin. “TEE.fail：通过 DDR5 内存总线介入破坏可信执行环境”. 载于：Proceedings of the 47th IEEE Symposium on Security and Privacy. S&P 2026. 2025年10月公开披露. 2026. URL: https://tee.fail/files/paper.pdf.

140. Chutes：无服务器 AI 计算. https://chutes.ai. 提供 TEE/机密计算模型（Intel TDX + NVIDIA 机密计算 GPU）的无服务器推理. 访问于 2026-06-06.

141. Circle Internet Financial. USDC：赋能全球金融. https://www.circle.com/usdc. 由 Circle 发行；储备金由 BNY Mellon 托管、BlackRock 管理的 Circle 储备基金（USDXX）持有；可按 1:1 兑换美元. 2024.

142. Allen Clement, Edmund L. Wong, Lorenzo Alvisi, Michael Dahlin, and Mirco Marchetti. “使拜占庭容错系统能够容忍拜占庭故障”. 载于：2009, 第 153-168 页.

143. Coalition for Content Provenance and Authenticity. 内容凭证：C2PA 技术规范. 2.0 版. 访问于 2026年2月. 2024. URL: https://c2pa.org/specifications/specifications/2.0/specs/C2PA_Specification.html.

144. Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. “训练验证器解决数学文字题”. 载于：arXiv preprint arXiv:2110.14168 (2021).

145. Ben Cohen, Emaad Khwaja, Youssef Doubli, Salahidine Lemaachi, Chris Lettieri, Charles Masson, Hugo Miccinilli, Elise Rame, Qiqi Ren, Afshin Rostamizadeh, et al. “这次不同：时间序列基础模型的可观测性视角”. 载于：arXiv preprint arXiv:2505.14766 (2025).

146. Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. “通过随机平滑实现认证的对抗性鲁棒性”. 载于：International conference on machine learning. PMLR. 2019, 第 1310-1320 页.

147. Shir Cohen, Rati Gelashvili, Eleftherios Kokoris-Kogias, Zekun Li, Dahlia Malkhi, Alberto Sonnino, and Alexander Spiegelman. “注意你的领导者”. 载于：Financial Cryptography and Data Security - 26th International Conference, FC 2022, Grenada, 2022年5月2-6日, Revised Selected Papers. 编辑：Ittay Eyal and Juan A. Garay. Lecture Notes in Computer Science. Springer, 2022, 第 279-295 页.

148. Coinbase Institute. 加密与代理 AI. [在线；访问于 2025-09-30]. URL: https://www.coinbase.com/public-policy/advocacy/documents/crypto-and-agentic-ai.

149. Caroline Collange, David Defour, Stef Graillat, and Roman Iakymchuk. “多核和多核架构上并行归约的数值可重现性”. 载于：Parallel Computing 49 (2015), 第 83-97 页.

150. KD Conway, Cathie So, Xiaohang Yu, and Kartin Wong. “opml：区块链上的乐观机器学习”. 载于：arXiv preprint arXiv:2401.17555 (2024).

151. Corinna Cortes and Vladimir Vapnik. “支持向量网络”. 载于：Machine learning 20.3 (1995), 第 273-297 页.

152. Cassandre Cofer. “加州披露向 AI 公司和海外出售数据的经纪人”. 载于：Bloomberg Law (2026年3月). [在线；访问于 2026-04-07]. URL: https://news.bloomberglaw.com/privacy-and-data-security/california-reveals-brokers-selling-data-to-ai-firms-and-overseas.

153. Francesco Croce and Matthias Hein. “通过一组多样化的无参数攻击可靠评估对抗性鲁棒性”. 载于：International conference on machine learning. PMLR. 2020, 第 2206-2216 页.

154. Hoagy Cunningham, Jerry Wei, Zihan Wang, Andrew Persic, Alwin Peng, Jordan Abderrachid, Raj Agarwal, Bobby Chen, Austin Cohen, Andy Dau, et al. “Constitutional Classifiers++：针对通用越狱的高效生产级防御”. 载于：arXiv preprint arXiv:2601.04603 (2026).

155. Allan Dafoe. “AI 治理：研究议程”. 载于：Governance of AI Program, Future of Humanity Institute, University of Oxford: Oxford, UK 1442 (2018), 第 1443 页.

156. Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E Ho. “大型法律虚构：剖析大型语言模型中的法律幻觉”. 载于：Journal of Legal Analysis 16.1 (2024), 第 64-93 页.

157. Qianyi Dai, Bin Zhang, and Shuqin Dong. “基于深度特征提取的区块链网络层日食攻击检测”. 载于：Wireless Communications and Mobile Computing 2022.1 (2022), 第 1451813 页.

158. Philip Daian, Steven Goldfeder, Tyler Kell, Yunqi Li, Xueyuan Zhao, Iddo Bentov, Lorenz Breidenbach, and Ari Juels. “Flash Boys 2.0：去中心化交易所的前置交易、矿工可提取价值与共识不稳定性”. 载于：2020 IEEE symposium on security and privacy (SP). IEEE. 2020, 第 910-927 页.

159. George Danezis, Lefteris Kokoris-Kogias, Alberto Sonnino, and Alexander Spiegelman. “Narwhal and Tusk：基于 DAG 的 mempool 与高效 BFT 共识”. 载于：Proceedings of the Seventeenth European Conference on Computer Systems. 2022, 第 34-50 页.

160. Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. “用于时间序列预测的仅解码器基础模型”. 载于：ICML. 2024.

161. Datarade. Datarade 数据市场. https://datarade.ai/search/products. 2018.

162. Isaac David, Liyi Zhou, Kaihua Qin, Dawn Song, Lorenzo Cavallaro, and Arthur Gervais. “你仍然需要手动智能合约审计吗？”载于：arXiv preprint arXiv:2306.12338 (2023).

163. Isaac David, Liyi Zhou, Dawn Song, Arthur Gervais, and Kaihua Qin. “使用大型语言模型反编译智能合约”. 载于：arXiv preprint arXiv:2506.19624 (2025).

164. Primavera De Filippi, Morshed Mannan, and Wessel Reijers. “区块链技术与代码规则：通过治理进行监管”. 载于：George Washington Law Review 92 (2024), 第 1229 页.

165. Lea Demelius, Roman Kern, and Andreas Trügler. “中心化深度学习中差分隐私的近期进展：系统综述”. 载于：ACM Computing Surveys 57.6 (2025), 第 1-28 页.

166. James Demmel and Hong Diep Nguyen. “快速可重现浮点求和”. 载于：2013 IEEE 21st Symposium on Computer Arithmetic. IEEE. 2013, 第 163-172 页.

167. Yvo Desmedt. “门限密码系统”. 载于：International workshop on the theory and application of cryptographic techniques. Springer. 1992, 第 1-14 页.

168. Gobikrishna Dhanuskodi, Sudeshna Guha, Vidhya Krishnan, Aruna Manjunatha, Rob Nertney, Michael O'Connor, and Phil Rogers. “创建第一代机密 GPU”. 载于：Communications of the ACM 67.1 (2023), 第 60-67 页.

169. Nour Diallo, Lei Xu, Dana Alsagheer, Yang Lu, and Larry Shi. “Optimized consensus with DAGWISE：一种用于可扩展和容错 DAG-BFT 的 GNN 增强方法”. 载于：2025 IEEE International Conference on Blockchain and Cryptocurrency (ICBC). IEEE. 2025, 第 1-5 页.

170. Nour Diallo, Lei Xu, Yang Lu, Dana R Alsagheer, and Weidong Larry Shi. “DAGWise++：基于 GNN 的 DAG-BFT 共识自适应优化”. 载于：2025 7th Conference on Blockchain Research & Applications for Innovative Networks and Services (BRAINS). IEEE. 2025, 第 1-9 页.

171. Qinxu Ding, Daniel Liebau, Zhiguo Wang, and Weibiao Xu. “去中心化自治组织及其治理综述”. 载于：World Scientific Annual Review of Fintech 1 (2023).

172. Polkadot Docs. 链上治理. https://docs.polkadot.com/reference/governance/. 2024.

173. Tezos Docs. 治理与自我修正. https://docs.tezos.com/architecture/governance. 2025.

174. Ye Dong, Wen-jie Lu, Yancheng Zheng, Haoqi Wu, Derun Zhao, Jin Tan, Zhicong Huang, Cheng Hong, Tao Wei, Wen-Guang Chen, et al. “Puma：五分钟内安全推理 LLaMA-7B”. 载于：Security and Safety 4 (2025), 第 2025014 页.

175. Yi Dong, Ronghui Mu, Yanghao Zhang, Siqi Sun, Tianle Zhang, Changshun Wu, Gaojie Jin, Yi Qi, Jinwei Hu, Jie Meng, et al. “保护大型语言模型：综述”. 载于：Artificial intelligence review 58.12 (2025), 第 382 页.

176. Elohim Fonseca Dos Reis, Alexander Teytelboym, Abeer ElBahrawy, Ignacio De Loizaga, and Andrea Baronchelli. “通过比特币交易网络识别暗网市场中的关键参与者”. 载于：Scientific Reports 14.1 (2024), 第 2385 页.

177. Maya Dotan and Saar Tochner. “无用工作证明——无浪费挖矿系统的正面与负面结果”. 载于：arXiv abs/2007.01046 (2020).

178. Winston Wei Dou, Itay Goldstein, and Yan Ji. “AI 驱动的交易、算法合谋与价格效率”. 载于：Jacobs Levy Equity Management Center for Quantitative Financial Research Paper, The Wharton School Research Paper (2025).

179. John R Douceur. “Sybil 攻击”. 载于：International workshop on peer-to-peer systems. Springer. 2002, 第 251-260 页.

180. Arthur Douillard, Qixuan Feng, Andrei Alex Rusu, Rachita Chhaparia, Yani Donchev, Adhiguna Kuncoro, MarcAurelio Ranzato, Arthur Szlam, and Jiajun Shen. “DiLoCo：语言模型的分布式低通信训练”. 载于：2nd Workshop on Advancing Neural Network Training: Computational Efficiency, Scalability, and Resource Optimization (WANT@ICML 2024).

181. Pawel Drozdowski, Christian Rathgeb, Antitza Dantcheva, Naser Damer, and Christoph Busch. “生物识别中的人口统计偏见：对一个新兴挑战的调查”. 载于：IEEE Transactions on Technology and Society 1.2 (2020), 第 89-103 页.

182. Hanbiao Du, Zheng Che, Meng Shen, Liehuang Zhu, and Jiankun Hu. “使用图特征学习打破以太坊混币服务的匿名性”. 载于：IEEE Transactions on Information Forensics and Security 19 (2023), 第 616-631 页.

183. Jiangfei Duan, Shuo Zhang, Zerui Wang, Lijuan Jiang, Wenwen Qu, Qinghao Hu, Guoteng Wang, Qizhen Weng, Hang Yan, Xingcheng Zhang, et al. “分布式基础设施上大型语言模型的高效训练：综述”. 载于：arXiv preprint arXiv:2407.20018 (2024).

184. Amit Dutta, Nafiz Imtiaz Rafin, M Ali Akber Dewan, and Md Golam Rabiul Alam. “ROBB：比特币区块链网络中基于循环近端策略优化强化学习的最优区块生成”. 载于：IEEE Access 12 (2024), 第 31287-31311 页.

185. Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. “通过感知实现公平”. 载于：Proceedings of the 3rd innovations in theoretical computer science conference. 2012, 第 214-226 页.

186. Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. “在私有数据分析中校准噪声与敏感性”. 载于：(2006), 第 265-284 页.

187. Cynthia Dwork and Aaron Roth. “差分隐私的算法基础”. 载于：Foundations and Trends® in Theoretical Computer Science 9.3-4 (2014), 第 211-407 页.

188. EigenCloud. EigenCloud 通过 EigenAI 和 EigenCompute 的推出将可验证 AI 推向大众市场. https://blog.eigencloud.xyz/eigencloud-brings-verifiable-ai-to-mass-market-with-eigenai-and-eigencompute-launches.

189. ElizaOS：面向所有人的自主智能体. https://github.com/elizaOS/eliza. 访问于 2025-03-12.

190. RAND Europe ENISA. 信息共享的激励与障碍. 技术报告. The European Network and Information Security Agency (ENISA), 2010.

191. Dmitry Ermilov, Maxim Panov, and Yury Yanovich. “自动比特币地址聚类”. 载于：2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA). 2017, 第 461-466 页.

192. Euler Finance. [在线；访问于 2025-12-09]. URL: https://www.euler.finance/.

193. Tonya M. Evans. AI 和区块链如何解决彼此的最大挑战. [在线；访问于 2025-09-30]. 2024年10月. URL: https://www.forbes.com/sites/tonyaevans/2024/10/29/how-ai-and-blockchain-are-solving-each-others-biggest-challenges/.

194. Ittay Eyal and Emin Gün Sirer. “多数是不够的：比特币挖矿是脆弱的”. 载于：Communications of the ACM 61.7 (2018), 第 95-102 页.

195. EZKL. DeFi 中的自动化风险评估：Sentiment Protocol 案例研究. https://blog.ezkl.xyz/post/sentiment/. 2025.

196. EZKL Contributors. EZKL：ONNX 模型的简易零知识证明. https://github.com/zkonduit/ezkl. 2023.

197. Andrés Fábregas, James Austgen, Samuel Breckenridge, Jay Yu, Amy Zhao, Sarah Allen, Aditya Saraf, and Ari Juels. “CoinAlg 困境：集体投资算法中的盈利-公平权衡”. 载于：arXiv preprint arXiv:2601.00523 (2026).

198. Andrés Fábregas, Amy Zhao, Jay Yu, James Austgen, Sarah Allen, Mahimna Kelkar, and Ari Juels. “投票块熵：DAO 去中心化的新度量”. 载于：USENIX Security Symposium. 2025.

199. Oumaima Fadi, Zikik Karim, Boulmalf Mohammed, et al. “关于区块链和人工智能技术增强智能环境安全与隐私的综述”. 载于：IEEE Access 10 (2022), 第 93168-93186 页.

200. Shuhui Fan, Haoran Xu, Shaojing Fu, Yuchuan Luo, and Ming Xu. “基于边特征建模的拓扑图神经网络用于以太坊网络钓鱼诈骗检测”. 载于：2024 IEEE/ACM 32nd International Symposium on Quality of Service (IWQoS). IEEE. 2024, 第 1-10 页.

201. Minghong Fang, Xiaoyu Cao, Jinyuan Jia, and Neil Gong. “对拜占庭鲁棒联邦学习的本地模型投毒攻击”. 载于：29th USENIX security symposium (USENIX Security 20). 2020, 第 1605-1622 页.

202. Giulia Fanti and Pramod Viswanath. “比特币 P2P 网络的匿名性属性”. 载于：NIPS (2017).

203. Steven Farrugia, Joshua Ellul, and George Azzopardi. “以太坊区块链上非法账户的检测”. 载于：Expert Systems with Applications 150 (2020), 第 113318 页.

204. Federal Trade Commission. 联邦贸易委员会启动对科技审查的调查. 访问于 2025-04-05. 2025. URL: https://www.ftc.gov/news-events/news/press-releases/2025/02/federal-trade-commission-launches-inquiry-tech-censorship.

205. Rainer Feichtinger, Robin Fritsch, Lioba Heimbach, Yann Vonlanthen, and Roger Wattenhofer. “SoK：对 DAO 的攻击”. 载于：6th Conference on Advances in Financial Technologies (AFT 2024). Vol. 316. Leibniz International Proceedings in Informatics (LIPIcs). Dagstuhl, Germany: Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2024, 28:1-28:27.

206. Vitaly Feldman. “学习需要记忆吗？一个关于长尾的短故事”. 载于：Proceedings of the 52nd annual ACM SIGACT symposium on theory of computing. 2020, 第 954-959 页.

207. Ryan Feng, Ashish Hooda, Neal Mangaokar, Kassem Fawaz, Somesh Jha, and Atul Prakash. “机器学习的有状态防御对黑盒攻击尚不安全”. 载于：Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security. 2023, 第 786-800 页.

208. Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. “从预训练数据到语言模型再到下游任务：追踪导致不公平 NLP 模型的政治偏见轨迹”. 载于：arXiv preprint arXiv:2305.08283 (2023).

209. Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. “不要幻觉，要弃权：通过多 LLM 协作识别 LLM 知识空白”. 载于：Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2024, 第 14664-14690 页.

210. Jared Fernandez, Luca Wehrstedt, Leonid Shamis, Mostafa Elhoushi, Kalyan Saladi, Yonatan Bisk, Emma Strubell, and Jacob Kahn. “大规模分布式训练中的硬件扩展趋势与收益递减”. 载于：arXiv preprint arXiv:2411.13055 (2024).

211. R. Fielding, M. Nottingham, and J. Reschke. HTTP 语义. Request for Comments RFC 9110. Internet Engineering Task Force, 2022年6月.

212. Apostolos Filippas, Srikanth Jagabathula, and Arun Sundararajan. “中心化定价在在线市场中的局限性与用户控制的价值”. 载于：Management Science 69.12 (2023), 第 7202-7216 页.

213. Financial Crimes Enforcement Network. FinCEN 法规对管理、交换或使用虚拟货币人员的适用. Guidance FIN-2013-G001, 2013年3月18日. 2013.

214. Financial Crimes Enforcement Network. FinCEN 法规对涉及可兑换虚拟货币的某些商业模式的适用. Guidance FIN-2019-G001, 2019年5月9日. 2019.

215. FindLaw Editorial Team. 社交媒体审查与法律. 访问于 2026-06-06. 2023. URL: https://www.findlaw.com/civilrights/enforcing-your-civil-rights/social-media-censorship-and-the-law.html.

216. Colin Finkbeiner, Mohamed E Najd, Julia Guskind, and Ghada Almashaqbeh. “SoK：是时候无私了吗？！揭秘自私挖矿策略与模型格局”. 载于：Cryptology ePrint Archive (2025).

217. Flashbots. BuilderNet：以太坊的去中心化区块构建网络. https://buildernet.org/docs. 2024年11月26日公开上线. 2024.

218. Flashbots. Flashbots 拍卖概述. https://docs.flashbots.net/flashbots-auction/overview. Flashbots Docs, 访问于2026年3月25日. 2026.

219. Hanna Foerster, Tom Blanchard, Kristina Nikolic, Ilia Shumailov, Cheng Zhang, Robert Mullins, Nicolas Papernot, Florian Tramer, and Yiren Zhao. “骆驼也能用电脑：计算机使用代理的系统级安全”. 载于：arXiv preprint arXiv:2601.09923 (2026).

220. Foldinghome - 用全球分布式超级计算机对抗疾病. [在线；访问于 2025-11-26]. URL: https://foldingathome.org/.

221. Bryan Ford. “数字民主中的身份与人格：评估假名政党和其他人格证明中的包容性、平等性、安全性与隐私”. 载于：arXiv preprint arXiv:2011.02412 (2020).

222. Fork-Choice Enforced Inclusion Lists (FOCIL)：一个简单的基于委员会的包含列表提案 - 权益证明 / 区块提议者 - 以太坊研究. https://ethresearch.ch/t/fork-choice-enforced-inclusion-lists-foci-a-simple-committee-based-inclusion-list-proposal/19870. (访问于 2024年9月21日).

223. Joel Frank, Cornelius Aschermann, and Thorsten Holz. “ETHBMC：智能合约的有界模型检查器”. 载于：29th USENIX Security Symposium (USENIX Security 20). USENIX Association, 2020年8月, 第 2757-2774 页.

224. Eugene Frimpong, Khoa Nguyen, Mindaugas Budzys, Tanveer Khan, and Antonis Michalas. “GuardML：通过混合同态加密实现高效的隐私保护机器学习服务”. 载于：Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing. SAC '24. Avila, Spain: Association for Computing Machinery, 2024, 第 953-962 页.

225. Robin Fritsch, Marino Müller, and Roger Wattenhofer. “分析去中心化治理中的投票权：谁控制 DAO？”载于：Blockchain: Research and Applications 5.3 (2024), 第 100208 页.

226. Michael Fröhlich, Maurizio Raphael Wagenhaus, Albrecht Schmidt, and Florian Alt. “别阻止我！探索首次加密货币用户的挑战”. 载于：Proceedings of the 2021 ACM designing interactive systems conference. 2021, 第 138-148 页.

227. Elisabeth M.S. Frommelt. “区块链生态系统中的责任挑战”. 载于：UC Davis Business Law Journal (2020).

228. Yiwei Fu, Tianhao Wang, and Varun Chandrasekaran. “启用私有数据估值中的挑战”. 载于：arXiv preprint arXiv:2603.00342 (2026).

229. Iason Gabriel. “人工智能、价值观与对齐”. 载于：Minds and machines 30.3 (2020), 第 411-437 页.

230. Yu Gai, Liyi Zhou, Kaihua Qin, Dawn Song, and Arthur Gervais. “区块链大型语言模型”. 载于：arXiv preprint arXiv:2304.12749 (2023).

231. Rundong Gan, Liyi Zhou, Le Wang, Kaihua Qin, and Xiaodong Lin. “DeFiAligner：利用符号分析与大型语言模型检测去中心化金融中的不一致性”. 载于：6th Conference on Advances in Financial Technologies (AFT 2024). 编辑：Rainer Böhme and Lucianna Kiffer. Vol. 316. Leibniz International Proceedings in Informatics (LIPIcs). Dagstuhl, Germany: Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2024, 7:1-7:24.

232. Cuifeng Gao, Wenzhang Yang, Jiaming Ye, Yinxing Xue, and Jun Sun. “sGuard+：机器学习引导的基于规则的智能合约自动漏洞修复”. 载于：ACM Transactions on Software Engineering and Methodology 33.5 (2024), 第 1-55 页.

233. Yue Gao, Jinqiao Shi, Xuebin Wang, Ruisheng Shi, Zelin Yin, and Yanyan Yang. “基于 P2P 网络分析的以太坊实用去匿名化攻击”. 载于：2021 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking. IEEE. 2021, 第 1402-1409 页.

234. Sanjam Garg, Aarushi Goel, Somesh Jha, Saeed Mahloujifar, Mohammad Mahmoody, Guruvamsi Policharla, and Mingyuan Wang. “零知识训练证明的实验”. 载于：Proceedings of the 2023 ACM SIGSAC conference on computer and communications security. 2023, 第 1880-1894 页.

235. Gas Network. URL: https://gas.network/ (访问于 2026年2月6日).

236. Gavin. “什么是跨域网络？AI 计算“第三支柱”完全指南”. 载于：NADDOD Blog (2025). [在线；访问于 2026-02-03]. URL: https://www.naddod.com/blog/complete-guide-to-scale-across-the-third-pillar-of-ai-computing.

237. Aaron Gember, Prathmesh Prabhu, Zainab Ghadiyali, and Aditya Akella. “走向软件定义中间盒网络”. 载于：Proceedings of the 11th ACM Workshop on Hot Topics in Networks. 2012, 第 7-12 页.

238. Arthur Gervais, Ghassan O Karame, Karl Wüst, Vasileios Glykantzis, Hubert Ritzdorf, and Srdjan Capkun. “关于工作量证明区块链的安全性与性能”. 载于：Proceedings of the 2016 ACM SIGSAC conference on computer and communications security. 2016, 第 3-16 页.

239. Arthur Gervais and Liyi Zhou. “AI 代理智能合约漏洞利用生成”. 载于：arXiv preprint arXiv:2507.05558 (2025).

240. 为你未使用的互联网获得奖励. [在线；访问于 2025-11-26]. URL: https://www.grass.io/learn.

241. Amirata Ghorbani and James Zou. “Data Shapley：机器学习中数据的公平估值”. 载于：International conference on machine learning. PMLR. 2019, 第 2242-2251 页.

242. Giza. Giza：可验证的 AI 代理. 使用 ZK 证明的去中心化机器学习推理. 访问于 2024-12-19. 2024. URL: https://www.gizatech.xyz/.

243. Andrew V Goldberg and Jason D Hartline. “多种数字商品的竞争性拍卖”. 载于：European Symposium on Algorithms. Springer. 2001, 第 416-427 页.

244. David Goldberg. “每个计算机科学家都应该了解的浮点运算知识”. 载于：ACM computing surveys (CSUR) 23.1 (1991), 第 5-48 页.

245. Micah Goldblum, Dimitris Tsipras, Chulin Xie, Xinyun Chen, Avi Schwarzschild, Dawn Song, Aleksander Madry, Bo Li, and Tom Goldstein. “机器学习的数据集安全：数据投毒、后门攻击与防御”. 载于：IEEE Transactions on Pattern Analysis and Machine Intelligence 45.2 (2023), 第 1563-1580 页.

246. Oded Goldreich, Silvio Micali, and Avi Wigderson. “如何进行任何心理游戏，或关于诚实多数协议的一个完备性定理”. 载于：Providing sound foundations for cryptography: on the work of Shafi Goldwasser and Silvio Micali. 2019, 第 307-328 页.

247. Shafi Goldwasser, Silvio Micali, and Chales Rackoff. “交互式证明系统的知识复杂性”. 载于：Providing sound foundations for cryptography: On the work of shafi goldwasser and silvio micali. 2019, 第 203-225 页.

248. Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. “解释和利用对抗性样本”. 载于：arXiv preprint arXiv:1412.6572 (2014).

249. Google. Agent2Agent (A2A) 协议. https://github.com/google/A2A. 2025.

250. Google Cloud. 机密计算概述. Google Cloud Documentation, https://docs.cloud.google.com/confidential-computing/docs/confidential-computing-overview. Google Cloud 机密计算产品家族：Confidential VM、Confidential Space、Google Cloud Attestation、Split-trust 加密工具. (访问于 2026年4月19日).

251. Ayelet Gordon-Tapiero, Katrina Ligett, and Kobbi Nissim. “关于数据的竞争性：技术和政策影响”. 载于：Proceedings of the 2025 Symposium on Computer Science and Law. 2025, 第 17-25 页.

252. GRASS：为你未使用的互联网获得奖励. [在线；访问于 2026-05-12]. URL: https://www.grass.io/.

253. Neville Grech, Lexi Brent, Bernhard Scholz, and Yannis Smaragdakis. “Gigahorse：彻底、声明式的智能合约反编译”. 载于：2019 IEEE/ACM 41st International Conference on Software Engineering (ICSE). 2019, 第 1176-1186 页.

254. Neville Grech, Sifis Lagouvardos, Ilias Tsatiris, and Yannis Smaragdakis. “Elipmoc：以太坊智能合约的高级反编译”. 载于：Proc. ACM Program. Lang. 6. OOPSLA1 (2022年4月).

255. Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. “不是你签署的内容：利用间接提示注入危害真实世界的 LLM 集成应用”. 载于：Proceedings of the 16th ACM workshop on artificial intelligence and security. 2023, 第 79-90 页.

256. Gustavo Grieco, Will Song, Artur Cygan, Josselin Feist, and Alex Groce. “Echidna：有效、可用且快速的智能合约模糊测试”. 载于：Proceedings of the 29th ACM SIGSOFT International Symposium on Software Testing and Analysis. ISSTA 2020. Virtual Event, USA: Association for Computing Machinery, 2020, 第 557-560 页.

257. Fabio Gritti, Nicola Ruaro, Robert McLaughlin, Priyanka Bose, Dipanjan Das, Ilya Grishchenko, Christopher Kruegel, and Giovanni Vigna. “Confusum Contractum：以太坊智能合约中的混乱代理漏洞”. 载于：32nd USENIX Security Symposium (USENIX Security 23). Anaheim, CA: USENIX Association, 2023年8月, 第 1793-1810 页.

258. Grant Gross. 合成数据瞄准 AI 训练挑战. Gartner 预计到 2028 年，AI 使用的 80% 数据将是合成的，高于 2024 年的 20%. 2025年2月. URL: https://www.cio.com/article/3827383/synthetic-data-takes-aim-at-ai-training-challenges.html.

259. Jens Groth. 关于基于配对的非交互式论证的大小. 2016. URL: https://eprint.iacr.org/2016/260 (访问于 2024年5月23日). 预发表.

260. Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens Van Der Maaten. “使用输入变换对抗对抗性图像”. 载于：arXiv preprint arXiv:1711.00117 (2017).

261. Yanpei Guo, Zhanpeng Guo, Wenjie Qu, and Jiaheng Zhang. “神经网络的架构私有零知识证明”. 载于：Cryptology ePrint Archive (2025).

262. Shashank Gupta, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. “偏见深藏：人格分配 LLM 中的隐式推理偏差”. 载于：The Twelfth International Conference on Learning Representations. 2024.

263. Melissa A Haendel, Christopher G Chute, Tellen D Bennett, David A Eichmann, Justin Guinney, Warren A Kibbe, Philip RO Payne, Emily R Pfaff, Peter N Robinson, Joel H Saltz, et al. “国家 COVID 队列协作体（N3C）：理由、设计、基础设施与部署”. 载于：Journal of the American Medical Informatics Association 28.3 (2021), 第 427-443 页.

264. Eric Halford, Ian Gibson, Mark Newfield, and Mufazzal Dhanwala. “使用机器学习开发管理洗钱交易的评分模型”. 载于：Journal of Money Laundering Control 28.7 (2025年3月), 第 30-49 页.

265. Xudong Han, Timothy Baldwin, and Trevor Cohn. “平衡偏见：通过平衡训练实现公平”. 载于：Proceedings of the 2022 conference on empirical methods in natural language processing. 2022, 第 11335-11350 页.

266. Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. “大型模型的参数高效微调：全面综述”. 载于：arXiv preprint arXiv:2403.14608 (2024).

267. Meng Hao, Hanxiao Chen, Hongwei Li, Chenkai Weng, Yuan Zhang, Haomiao Yang, and Tianwei Zhang. “机器学习中非线性函数的可扩展零知识证明”. 载于：33rd USENIX Security Symposium (USENIX Security 24). 2024, 第 3819-3836 页.

268. Andrew Hard, Kanishka Rao, Rajiv Mathews, Swaroop Ramaswamy, Françoise Beaufays, Sean Augenstein, Hubert Eichner, Chloe Kiddon, and Daniel Ramage. “用于移动键盘预测的联邦学习”. 载于：2018.

269. Kali Hays. “Tinder 和 Zoom 提供‘人性证明’眼球扫描以对抗 AI”. 载于：BBC News (2026年4月). 访问于2026年4月. URL: https://www.bbc.com/news/articles/cp9vppem4evo.

270. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. “用于图像识别的深度残差学习”. 载于：Proceedings of the IEEE conference on computer vision and pattern recognition. 2016, 第 770-778 页.

271. Yufei He, Yuexin Li, Jiaying Wu, Yuan Sui, Yulin Chen, and Bryan Hooi. “评估回形针最大化器：基于 RL 的语言模型是否更倾向于追求工具性目标？”载于：arXiv preprint arXiv:2502.12206 (2025).

272. Ethereum Network Status. 以太坊网络状态历史. [在线；访问于 2026-06-06].

273. Nicholas J. Higham. “浮点求和中的准确性与稳定性”. 载于：SIAM Journal on Scientific Computing 14.4 (1993), 第 783-799 页.

274. Hivemapper. [在线；访问于 2025-11-26]. URL: https://hivemapper.com/.

275. “保险公司算法偏见”. 载于：ProPublica (2023). 访问于 2026-06-06.

276. Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. “训练计算最优的大型语言模型”. 载于：arXiv preprint arXiv:2203.15556 (2022).

277. Shuyue Hou, Xincheng Yan, and Xiaorong Wang. “基于深度强化学习的区块链多矿工自私挖矿分析”. 载于：Blockchain: Research and Applications 5.4 (2024), 第 100218 页.

278. Polymarket. 常见问题：结果与争议解决. https://help.polymarket.com/en/articles/8795563-results-and-dispute-resolution. (访问于 2026年2月18日).

279. ERC-8004：无需信任代理标准. [在线；访问于 2026-06-06].

280. Siqi Hu, Zhengyue Zhang, and Zibin Zheng. “BERT4ETH：用于以太坊分析的预训练 Transformer”. 载于：2023 IEEE International Conference on Blockchain and Cryptocurrency (ICBC). IEEE. 2023, 第 1-9 页.

281. Yubo Hu, Jiaqi Yan, and Wenji Mao. “区块链应用分类：文献综述”. 载于：IEEE Access 9 (2021), 第 145389-145408 页.

282. Yutao Hu, Jiahong Li, Yebo Feng, and Yang Liu. “DeFi 中的拉高出货检测：基于图的对比学习方法”. 载于：IEEE Transactions on Information Forensics and Security (2025).

283. Tian Huang, Jiajing Wu, Qi Yuan, Zibin Zheng, and Jian Zhang. “结合分层自注意力和图神经网络的以太坊交易去匿名化”. 载于：IEEE Transactions on Information Forensics and Security 18 (2023), 第 1234-1248 页.

284. Yan Huang, David Evans, and Jonathan Katz. “安全两方计算中的隐私保护”. 载于：Proceedings of the 2012 ACM conference on Computer and communications security. 2012, 第 393-404 页.

285. Collective Constitutional AI. [在线；访问于 2026-06-06].

286. Bernardo Huberman, Jacob Leshno, and Ciamac Moallemi. “没有垄断者的垄断：区块链的经济学”. 载于：Management Science 67.8 (2021), 第 4675-4695 页.

287. Mohammad Hossein Hosseini, Amir Hossein Hosseini, and Mahdi Bohlouli. “区块链与人工智能的融合：机遇与挑战综述”. 载于：IEEE Access 10 (2022), 第 119215-119238 页.

288. Sterling. Sterling 数据市场. https://www.sterling.io/. (访问于 2026-06-06).

289. IBM. 区块链和人工智能如何协同工作？IBM 博客. [在线；访问于 2026-06-06].

290. Andrew Ilyas, Sung Min Park, Logan Engstrom, Guillaume Leclerc, and Aleksander Madry. “数据模型：从训练数据预测预测”. 载于：Proceedings of the 39th International Conference on Machine Learning. 2022.

291. NI Indera, IM Yassin, A Zabidi, and ZI Rizman. “使用 PSO 优化参数和移动平均技术指标的 NARX 比特币价格预测模型”. 载于：Journal of fundamental and applied sciences 9.3S (2017), 第 791-808 页.

292. LexisNexis Insights. 放大分析的影响：LexisNexis ① Data+ 与 Snowflake 数据云平台. [在线；访问于 2026-04-07]. 2023年8月. URL: https://www.lexisnexis.com/community/insights/professional/resources/b/brochures/posts/lexisnexis-data-and-snowflake-data-cloud-platform.

293. 介绍智能体钱包：赋予你的代理自主权 / Coinbase. 2026年2月11日. URL: https://www.coinbase.com/developer-platform/discover/launches/agentic-wallets (访问于 2026年2月27日).

294. io.net / 面向 AI 工作负载的去中心化 GPU 生态系统 - 节省高达 70%. [在线；访问于 2025-11-26]. URL: https://io.net/.

295. IOTA Foundation. IOTA 数据市场. [在线；访问于 2026-05-13]. 2017年11月. URL: https://blog.iota.org/iota-data-marketplace-cb6be463ac7f/.

296. Tariqul Islam, Faisal Haque Bappy, Tarannum Shaila Zaman, Md Sajidul Islam Sajid, and Mir Mehedi Ahsan Pritom. “MRL-PoS：一种基于多代理强化学习的区块链权益证明共识算法”. 载于：2024 IEEE 14th Annual Computing and Communication Workshop and Conference (CCWC). IEEE. 2024, 第 0409-0413 页.

297. Nishant Jagannath, Tudor Barbulescu, Karam M Sallam, Ibrahim Elgendi, Braden Mcgrath, Abbas Jamalipour, Mohamed Abdel-Basset, and Kumudu Munasinghe. “基于链上分析的方法预测以太坊价格”. 载于：IEEE Access 9 (2021), 第 167972-167989 页.

298. Matthew Jagielski, Nicholas Carlini, David Berthelot, Alex Kurakin, and Nicolas Papernot. “高精度高保真度的神经网络提取”. 载于：29th USENIX security symposium (USENIX Security 20). 2020, 第 1345-1362 页.

299. Kokil Jaidka, Subhayan Mukerjee, and Yphtach Lelkes. “在社交媒体上被沉默：美国 Twitter 中影子禁令的守门功能”. 载于：Journal of Communication 73.2 (2023), 第 163-178 页.

300. Shashwat Jaiswal, Kunal Jain, Yogesh Simmhan, Anjaly Parayil, Ankur Mallick, Rujia Wang, Renee St Amant, Chetan Bansal, Victor Rühle, Anoop Kulkarni, et al. “服务模型，快与慢：优化大规模异构 LLM 推理工作负载”. 载于：arXiv e-prints (2025), arXiv-2502.

301. Huisu Jang and Jaewook Lee. “基于区块链信息的贝叶斯神经网络对比特币价格建模与预测的实证研究”. 载于：IEEE Access 6 (2018), 第 5427-5437 页.

302. Patrick Jauernig, Ahmad-Reza Sadeghi, and Emmanuel Stapf. “可信执行环境：属性、应用与挑战”. 载于：IEEE Security & Privacy 18.2 (2020), 第 56-60 页.

303. Mojan Javaheripi, Mohammad Samragh, Tara Javidi, and Farinaz Koushanfar. “AdaNS：用于紧凑 DNN 自动化设计的自适应非均匀采样”. 载于：IEEE Journal of Selected Topics in Signal Processing 14.4 (2020), 第 750-764 页.

304. Michelle Javed. Akash：2025 年度回顾. Akash Network Blog. 访问于 2026-06-07. 2026年1月9日. URL: https://akash.network/blog/akash-2025-year-in-review/.

305. Jiaming Ji, Mickel Liu, Josef Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. “Beavertails：通过人类偏好数据集改善 LLM 的安全对齐”. 载于：Advances in Neural Information Processing Systems 36 (2023), 第 24678-24704 页.

306. Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, et al. “AI 对齐：全面综述”. 载于：arXiv preprint arXiv:2310.19852 (2023).

307. Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. “自然语言生成中的幻觉调查”. 载于：ACM computing surveys 55.12 (2023), 第 1-38 页.

308. Hengrui Jia, Mohammad Yaghini, Christopher A Choquette-Choo, Natalie Dullerud, Anvith Thudi, Varun Chandrasekaran, and Nicolas Papernot. “学习证明：定义与实践”. 载于：2021 IEEE Symposium on Security and Privacy (SP). IEEE. 2021, 第 1039-1056 页.

309. Bo Jiang, Ye Liu, and W. K. Chan. “ContractFuzzer：用于漏洞检测的智能合约模糊测试”. 载于：Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering. ASE '18. Montpellier, France: Association for Computing Machinery, 2018, 第 259-269 页.

310. Hai Jin, Zeli Wang, Ming Wen, Weiqi Dai, Yu Zhu, and Deqing Zou. “Aroc：链上智能合约的自动修复框架”. 载于：IEEE Transactions on Software Engineering 48.11 (2021), 第 4611-4629 页.

311. Johannes, Sami, Jackmin, and Vincent. INTELLECT-1 发布：首个全球训练的 100 亿参数模型. [在线；访问于 2026-03-01]. 2024年11月. URL: https://www.primeintellect.ai/blog/intellect-1-release.

312. Simon Johnson, Raghunandan Makaram, Amy Santoni, and Vincent Scarlata. 支持多封装平台上的 Intel SGX. Intel 关于可扩展 SGX 内存加密及其与原始完整性树设计权衡的文档. 2025. arXiv: 2507.08190 [cs.CR]. URL: https://arxiv.org/abs/2507.08190.

313. Charles I Jones and Christopher Tonetti. “非竞争性与数据经济学”. 载于：American Economic Review 110.9 (2020), 第 2819-2858 页.

314. Nicola Jones. “AI 革命正在耗尽数据。研究人员能做什么？”载于：Nature 636.8042 (2024), 第 290-292 页.

315. Hadi Jooybar, Wilson WL Fung, Mike O'Connor, Joseph Devietti, and Tor M Aamodt. “GPUDet：确定性 GPU 架构”. 载于：Proceedings of the eighteenth international conference on Architectural support for programming languages and operating systems. 2013, 第 1-12 页.

316. Michael I Jordan. “关于 AI 的集体主义、经济学视角”. 载于：arXiv preprint arXiv:2507.06268 (2025).

317. Yucong Ju, Fei Song, Yutao Jiao, Weiyi Wang, Wenting Dai, and Yuhua Xu. “无线区块链网络最优传输速率的快速确定：一种图卷积神经网络方法”. 载于：Sensors 23.13 (2023), 第 6098 页.

318. Ari Juels, Ahmed Kosba, and Elaine Shi. “Gyges 之环：调查犯罪智能合约的未来”. 载于：Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. 2016, 第 283-295 页.

319. Ari Juels and Farinaz Koushanfar. “机器学习安全的 Props”. 载于：arXiv preprint arXiv:2410.20522 (2024).

320. Mika Juuti, Sebastian Szyller, Samuel Marchal, and N Asokan. “PRADA：防范 DNN 模型窃取攻击”. 载于：2019 IEEE European Symposium on Security and Privacy (EuroS&P). IEEE. 2019, 第 512-527 页.

321. Kamil Kaczmarski, Marek Wegrzyn, Krzysztof Ostrowski, Tomasz Januszewski, Mateusz Chmiel, Piotr Kocot, Jacek Kowalczyk, Jakub Piatek, Dawid Wolski, Marcin Grochowski, and Jacek Kasprzyk. “NVIDIA H100 GPU 上的机密计算：性能基准研究”. 载于：arXiv preprint arXiv:2409.03992 (2024). 第2版，2024年10月. URL: https://arxiv.org/abs/2409.03992v2.

322. Heba Kadry and Yasser Gadallah. “支付通道网络中链下交易的机器学习路由技术”. 载于：2021 IEEE International Conference on Smart Internet of Things (SmartIoT). IEEE. 2021, 第 66-73 页.

323. James Ward Kalanuyu Zintus-art Ben Vass. AI 预言机开发中的经验证据 / Chainlink Blog. URL: https://blog.chain.link/ai-oracles/.

324. Harry Kalodner, Steven Goldfeder, Xiaoqi Chen, S Matthew Weinberg, and Edward W Felten. “Arbitrum：可扩展的智能合约”. 载于：USENIX Security Symposium. 2018.

325. Sukrit Kalra, Seep Goel, Mohan Dhawan, and Subodh Sharma. “ZEUS：分析智能合约的安全性”. 载于：25th Annual Network and Distributed System Security Symposium, NDSS 2018, San Diego, California, USA, 2018年2月18-21日. 2018.

326. Ryo Kamoi, Sarkar Snigdha Sarathi Das, Renze Lou, Jihyun Janice Ahn, Yilun Zhao, Xiaoxin Lu, Nan Zhang, Yusen Zhang, Ranran Haoran Zhang, Sujeeth Reddy Vummanthala, Salika Dave, Shaobo Qin, Arman Cohan, Wenpeng Yin, and Rui Zhang. “评估 LLM 在检测 LLM 响应错误方面的能力”. 2024. arXiv: 2404.03602 [cs.CL]. URL: https://arxiv.org/abs/2404.03602.

327. Nikhil Kandpal and Colin Raffel. “立场：LLM 最昂贵的部分应该是其训练数据”. 载于：arXiv preprint arXiv:2504.12427 (2025).

328. Kimberley Kao. “阿里巴巴推出升级版 AI 模型，声称超越竞争对手 DeepSeek-V3”. 华尔街日报. 2025. URL: https://www.wsj.com/tech/ai/alibaba-unveils-upgraded-ai-model-claims-it-surpasses-rival-deepseek-v3-506b7f28.

329. Md Monjurul Karim, Dong Hoang Van, Sangeen Khan, Qiang Qu, and Yaroslav Kholodov. “AI 代理遇见区块链：关于多代理安全与可扩展协作的综述”. 载于：Future Internet 17.2 (2025), 第 57 页.

330. Guy Katz, Clark Barrett, David L Dill, Kyle Julian, and Mykel J Kochenderfer. “Reluplex：验证深度神经网络的高效 SMT 求解器”. 载于：International conference on computer aided verification. Springer. 2017, 第 97-117 页.

331. Anna Kauzlaskas. Vana：数据主权开放协议. [在线；访问于 2026-05-12]. 2024. URL: https://cdn.prod.website-files.com/662488d48afbf9b21e25c7d2/674f6289079193fb53f7caa1_Vana_whitepaper_Dec24_updated.pdf.

332. Daisuke Kawai, Alejandro Cuevas, Bryan Routledge, Kyle Soska, Ariel Zetlin-Jones, and Nicolas Christin. “你的数字邻居是可靠的投资顾问吗？”载于：Proceedings of the ACM Web Conference 2023. 2023, 第 3581-3591 页.

333. Shuqi Ke and Giulia Fanti. “基于价值的预训练与下游反馈”. 载于：arXiv preprint arXiv:2601.22108 (2026).

334. Patrik Keller. “DAG 协议的通用自私挖矿 MDP”. 载于：arXiv preprint arXiv:2309.11924 (2023).

335. Han-Min Kim, Gee-Woo Bock, and Gunwoong Lee. “基于区块链信息的机器学习预测以太坊价格”. 载于：Expert Systems with Applications 184 (2021), 第 115480 页.

336. Jinoh Kim, Makiya Nakashima, Wenjun Fan, Simeon Wuthier, Xiaobo Zhou, Ikkyun Kim, and Sang-Yoon Chang. “基于流量监控的安全区块链网络异常检测的机器学习方法”. 载于：IEEE Transactions on Network and Service Management 19.3 (2022), 第 3619-3632 页.

337. Seungmo Kim and Ahmed S Ibrahim. “基于强化学习的许可区块链赋能 V2X 网络拜占庭容错共识”. 载于：IEEE Transactions on Intelligent Vehicles 8.1 (2022), 第 172-183 页.

338. John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. “大型语言模型的水印”. 载于：International Conference on Machine Learning. PMLR. 2023, 第 17061-17084 页.

339. Klim Kireev, Bogdan Kulynych, and Carmela Troncoso. “通过成本与效用意识提升表格数据的对抗性鲁棒性”. 载于：NDSS Symposium. 2023.

340. Megan Kirkwood. 理解 Apple 和 Meta 在《数字市场法案》下的不合规决定. [在线；访问于 2026-05-11]. 2025年4月. URL: https://www.technology.press/understanding-the-apple-and-meta-noncompliance-decisions-under-the-digital-markets-act/.

341. Max Klabunde, Tobias Schumacher, Markus Strohmaier, and Florian Lemmerich. “神经网络模型的相似性：功能和表示度量的调查”. 载于：ACM Comput. Surv. 57.9 (2025年5月). ISSN: 0360-0300.

342. Kled AI. Kled AI 估值飙升至 1 亿美元，打造全球首个消费者数据市场. ACCESS Newswire. 2025年12月. URL: https://www.accessnewswire.com/newsroom/en/business-and-professional-services/kled-ai-surges-to-a-100-million-valuation-as-it-builds-the-world-1117096.

343. Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. “风险评分公平判定中的固有权衡”. 载于：8th Innovations in Theoretical Computer Science Conference (ITCS 2017). Schloss Dagstuhl-Leibniz-Zentrum für Informatik. 2017, 43:1-43:23.

344. Pang Wei Koh and Percy Liang. “通过影响函数理解黑盒预测”. 载于：International conference on machine learning. PMLR. 2017, 第 1885-1894 页.

345. Ilan Komargodski, Itamar Shen, and Omri Weinstein. “基于任意矩阵乘法的有用工作证明”. 载于：CoRR abs/2504.09971 (2025).

346. Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey E. Hinton. “神经网络表示相似性的再审视”. 载于：International Conference on Machine Learning (ICML). 2019, 第 3519-3529 页.

347. Ramakrishna Kotla, Lorenzo Alvisi, Mike Dahlin, Allen Clement, and Edmund Wong. “Zyzzyva：投机性拜占庭容错”. 载于：Proceedings of twenty-first ACM SIGOPS symposium on Operating systems principles. 2007, 第 45-58 页.

348. János Kramár, Joshua Engels, Zheng Wang, Bilal Chughtai, Rohin Shah, Neel Nanda, and Arthur Conmy. “为 Gemini 构建生产就绪的探针”. 载于：arXiv preprint arXiv:2601.11516 (2026).

349. Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. “表示相似性分析——连接系统神经科学的分支”. 载于：Frontiers in Systems Neuroscience 2 (2008), 第 4 页.

350. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. “基于深度卷积神经网络的 ImageNet 分类”. 载于：Advances in neural information processing systems 25 (2012).

351. Nir Kshetri. “区块链在满足关键供应链管理目标中的作用”. 载于：International Journal of information management 39 (2018), 第 80-89 页.

352. Nir Kshetri. “建立 AI 信任：区块链如何增强数据完整性、安全性和隐私”. 载于：Computer 58.2 (2025), 第 63-70 页.

353. Komal Kumar, Tajamul Ashraf, Omkar Thawakar, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, Phillip HS Torr, Fahad Shahbaz Khan, and Salman Khan. “LLM 后训练：深入推理大型语言模型”. 载于：arXiv preprint arXiv:2502.21321 (2025).

354. Nishant Kumar, Mayank Rathee, Nishanth Chandran, Divya Gupta, Aseem Rastogi, and Rahul Sharma. “Cryptflow：安全的 TensorFlow 推理”. 载于：2020 IEEE Symposium on Security and Privacy (SP). IEEE. 2020, 第 336-353 页.

355. Thomas Kwa, Ben West, Joel Becker, Amy Deng, Kathryn Garcia, Max Hasin, Sami Jawhar, Megan Kinniment, Nate Rush, Sydney Von Arx, Ryan Bloom, Thomas Broadley, Haoxing Du, Brian Goodrich, Nikola Jurkovic, Luke Harold Miles, Seraphina Nix, Tao Lin, Neev Parikh, David Rein, Lucas Jun Koba Sato, Hjalmar Wijk, Daniel M. Ziegler, Elizabeth Barnes, and Lawrence Chan. “测量 AI 完成长软件任务的能力”. 2026. arXiv: 2503.14499 [cs.AI]. URL: https://arxiv.org/abs/2503.14499.

356. Olga Labazova, Tobias Dehling, and Ali Sunyaev. “从炒作到现实：区块链应用分类”. 2019.

357. Modulus Labs. 智能的成本：使用零知识证明机器学习推理. https://github.com/Modulus-Labs/Papers/blob/master/Cost_Of_Intelligence.pdf. 2023年1月.

358. Bence Ladoczki. “MEV-Boost 拍卖中的机器学习预测”. 载于：2025 7th International Conference on Blockchain Computing and Applications (BCCA). IEEE. 2025, 第 518-525 页.

359. Lagrange Labs. DeepProve：可验证神经网络推理的 zkML 框架. https://github.com/Lagrange-Labs/deep-prove. 2025.

360. Ben Laurie, Adam Langley, Emilia Kasper, Eran Messeri, and Rob Stradling. “证书透明度 2.0 版”. 载于：Internet Requests for Comments, RFC Editor, RFC 9162 (2021).

361. Erwan Le Merrer, Patrick Perez, and Gilles Trédan. “用于远程神经网络水印的对抗性前沿缝合”. 载于：Neural Computing and Applications 32.13 (2020), 第 9233-9244 页.

362. LEAP IN. 区块链如何解决 AI 的偏见问题. [在线；访问于 2026-02-14]. 2025年7月. URL: https://www.insights.onegaintleap.com/how-blockchain-can-solve-ais-bias-problem/#.

363. Gulian Leduc, Sylvain Kubler, and Jean-Philippe Georges. “Sabine：自适应区块链共识”. 载于：2022 9th International Conference on Future Internet of Things and Cloud (FiCloud). IEEE. 2022, 第 234-240 页.

364. Gulian Leduc, Sylvain Kubler, and Jean-Philippe Georges. “DRAFTEE：一种自适应框架，用于在安全约束下满足波动吞吐量需求的基于 BFT 的共识”. 载于：Comput. Networks 269 (2025), 第 111396 页.

365. Jonghyun Lee, Yongqin Wang, Rachit Rajat, and Murali Annavaram. “分布式数据并行 ML 训练中 GPU TEE 开销的表征”. 载于：arXiv preprint arXiv:2501.11771 (2025).

366. Amit Levy, S. Matthew Weinberg, and Chenghan Zhou. “分析去中心化对用户的经济影响”. 载于：Proceedings of the 17th Annual Innovations in Theoretical Computer Science Conference (ITCS). 2026.

367. Baolin Li, Yankai Jiang, Vijay Gadepally, and Devesh Tiwari. “LLM 推理服务：近期进展与机遇综述”. 载于：2024 IEEE High Performance Extreme Computing Conference (HPEC). IEEE. 2024, 第 1-8 页.

368. Chao Li, Daniel Yang Li, Gerome Miklau, and Dan Suciu. “私有数据定价理论”. 载于：ACM Transactions on Database Systems (TODS) 39.4 (2014), 第 1-28 页.

369. Kai Li, Shixuan Guan, and Darren Lee. “理解和表征野外套利机器人骗局”. 载于：Proceedings of the ACM on Measurement and Analysis of Computing Systems 7.3 (2023), 第 1-29 页.

370. Linyi Li, Tao Xie, and Bo Li. “SoK：深度神经网络的认证鲁棒性”. 载于：2023 IEEE symposium on security and privacy (SP). IEEE. 2023, 第 1289-1310 页.

371. Pengfei Li, Jianyi Yang, Mohammad A Islam, and Shaolei Ren. “让 AI 不那么‘渴’”. 载于：Communications of the ACM 68.7 (2025), 第 54-61 页.

372. Pengze Li, Mingxuan Song, Mingzhe Xing, Zhen Xiao, Qiuyu Ding, Shengjie Guan, and Jieyi Long. “Spring：通过基于深度强化学习的状态放置提高分片区块链的吞吐量”. 载于：Proceedings of the ACM Web Conference 2024. 2024, 第 2836-2846 页.

373. Qinbin Li, Zeyi Wen, Zhaomin Wu, Sixu Hu, Naibo Wang, Yuan Li, Xu Liu, and Bingsheng He. “联邦学习系统综述：数据隐私与保护的愿景、炒作与现实”. 载于：IEEE Transactions on Knowledge and Data Engineering 35.4 (2021), 第 3347-3366 页.

374. Sijia Li, Gaopeng Gou, Chang Liu, Chengshang Hou, Zhenzhen Li, and Gang Xiong. “TTAGN：用于以太坊网络钓鱼诈骗检测的时间交易聚合图网络”. 载于：Proceedings of the ACM web conference 2022. 2022, 第 661-669 页.

375. Zhe Li, Xiangfei Qiu, Peng Chen, Yihang Wang, Hanyin Cheng, Yang Shu, Jilin Hu, Chenjuan Guo, Aoying Zhou, Christian S Jensen, et al. “TSFM-Bench：时间序列预测基础模型的全面统一基准”. 载于：KDD. 2025, 第 5595-5606 页.

376. Zihao Li, Jianfeng Li, Zheyuan He, Xiapu Luo, Ting Wang, Xiaozhe Ni, Wenwu Yang, Xi Chen, and Ting Chen. “揭秘 Flashbots 捆绑包中的 DeFi MEV 活动”. 载于：Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security. 2023, 第 165-179 页.

377. Zeqin Liao, Yuhong Nan, Zixu Gao, Henglong Liang, Sicheng Hao, Peifan Ren, and Zibin Zheng. “通过细粒度依赖分析和 LLM 辅助语义恢复增强智能合约反编译器输出”. 载于：IEEE Transactions on Software Engineering 51.12 (2025), 第 3574-3590 页.

378. Joel Lidin, Amir Sarfi, Evangelos Pappas, Samuel Dare, Eugene Belilovsky, and Jacob Steeves. “激励 LLM 的无许可分布式学习”. 载于：Proceedings of the 2025 7th International Conference on Distributed Artificial Intelligence. 2025, 第 12-18 页.

379. Dan Lin, Jiajing Wu, Qi Yuan, and Zibin Zheng. “T-edge：用于以太坊交易网络分析的时序加权多向图嵌入”. 载于：Frontiers in Physics 8 (2020), 第 204 页.

380. Jieli Liu, Jinze Chen, Jiajing Wu, Zhiying Wu, Junyuan Fang, and Zibin Zheng. “捕获欺诈者：用区块链数据揭露以太坊网络钓鱼团伙”. 载于：IEEE Transactions on Information Forensics and Security 19 (2024), 第 3038-3050 页.

381. Mengting Liu, F Richard Yu, Yinglei Teng, Victor CM Leung, and Mei Song. “区块链赋能工业物联网系统的性能优化：一种深度强化学习方法”. 载于：IEEE Transactions on Industrial Informatics 15.6 (2019), 第 3559-3570 页.

382. Tianyi Liu, Xiang Xie, and Yupeng Zhang. “Zkcnn：用于卷积神经网络预测和准确性的零知识证明”. 载于：Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security. 2021, 第 2968-2985 页.

383. Ye Liu, Yue Xue, Daoyuan Wu, Yuqiang Sun, Yi Li, Miaolei Shi, and Yang Liu. “PropertyGPT：通过检索增强属性生成的 LLM 驱动智能合约形式验证”. 载于：arXiv preprint arXiv:2405.02580 (2024).

384. Joana Lorenz, Maria Ines Silva, David Aparicio, Joao Tiago Ascensao, and Pedro Bizarro. “在标签稀缺情况下使用机器学习检测比特币区块链中的洗钱行为”. 载于：Proceedings of the First ACM International Conference on AI in Finance. ICAIF '20. New York, New York: Association for Computing Machinery, 2021.

385. Tao Lu, Haoyu Wang, Wenjie Qu, Zonghui Wang, Jinye He, Tianyang Tao, Wenzhi Chen, and Jiaheng Zhang. “一种高效且可扩展的神经网络零知识证明框架”. 载于：Cryptology ePrint Archive (2024).

386. Kristian Lum, Jacy Reese Anthis, Kevin Robinson, Chirag Nagpal, and Alexander Nicholas D'Amour. “语言模型中的偏见：超越把戏测试，走向 RUTEd 评估”. 载于：Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2025, 第 137-161 页.

387. Zhongtang Luo, Yanxue Jia, Yaobin Shen, and Aniket Kate. “代理就足够了：TLS 预言机中的代理安全性与 AEAD 上下文不可伪造性”. 载于：7th Conference on Advances in Financial Technologies (AFT 2025). Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2025.

388. Adrian Litsch, Christian Franck, Muhammad El-Hindi, Zsolt István, and Carsten Binnig. “AWS Nitro Enclaves 在数据库工作负载中的分析”. 载于：Proceedings of the 21st International Workshop on Data Management on New Hardware. 2025, 第 1-8 页.

389. Loi Luu, Duc-Hiep Chu, Hrishi Olickel, Prateek Saxena, and Aquinas Hobor. “让智能合约更智能”. 载于：Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. CCS '16. Vienna, Austria: Association for Computing Machinery, 2016, 第 254-269 页.

390. Lyn Labs. 介绍 Lyn：我们对视频 AI 和以人为中心的代理视频的愿景. 访问于 2026年2月. 2025. URL: https://lynlabs.gitbook.io/lyn.

391. Aengus Lynch, Benjamin Wright, Caleb Larson, Stuart J Ritchie, Soren Mindermann, Evan Hubinger, Ethan Perez, and Kevin Troy. “代理性错位：LLM 如何成为内部威胁”. 载于：arXiv preprint arXiv:2510.05179 (2025).

392. Kaifeng Lyu, Haoyu Zhao, Xinran Gu, Dingli Yu, Anirudh Goyal, and Sanjeev Arora. “微调后保持 LLM 对齐：提示模板的关键作用”. 载于：The Thirty-eighth Annual Conference on Neural Information Processing Systems. 2024.

393. Macrocosmos, Taoverse, Const, and Datura. LLM 预训练：区块链一直在等待的用例？ [在线；访问于 2026-03-01]. 2024年8月. URL: https://www.macrocosmos.ai/research/pretraining_whitepaper.pdf.

394. Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. “走向抵抗对抗性攻击的深度学习模型”. 载于：arXiv preprint arXiv:1706.06083 (2017).

395. Rischan Mafrur. “基于 AI 的加密代币：去中心化 AI 的幻觉？”载于：IET Blockchain 5.1 (2025), e70015.

396. Vincent Maliepaard. “代理 AI 在加密中的力量：深入 Virtuals 生态系统”. [在线；访问于 2025-09-30]. 2025年1月. URL: https://cryptoslate.com/the-power-of-agentic-ai-in-crypto-a-deep-dive-into-the-virtuals-ecosystem/.

397. Karthik Mandakolathur. “使用 NVIDIA 集体通信库 2.12 使 All2All 性能翻倍”. 载于：NVIDIA Technical Blog (2022年2月). [在线；访问于 2025-12-05]. URL: https://developer.nvidia.com/blog/doubling-all2all-performance-with-nvidia-collective-communication-library-2-12/.

398. Debmalya Mandal, Samuel Deng, Suman Jana, Jeannette Wing, and Daniel J Hsu. “确保超越训练数据的公平性”. 载于：Advances in neural information processing systems 33 (2020), 第 18445-18456 页.

399. Yifan Mao, Soubhik Deb, Shaileshh Bojia Venkatakrishnan, Sreeram Kannan, and Kannan Srinivasan. “Perigee：面向区块链的高效点对点网络设计”. 载于：Proceedings of the 39th Symposium on Principles of Distributed Computing. 2020, 第 428-437 页.

400. Yifan Mao and Shaileshh Bojia Venkatakrishnan. “Topiary：面向点对点（D）应用的快速、可扩展发布/订阅”. 载于：arXiv preprint arXiv:2312.06800 (2023).

401. Deepak Maram, Harjasleen Malvai, Fan Zhang, Narla Jean-Louis, Alexander Frolov, Tyler Kell, Tyrone Lobban, Christine Moy, Ari Juels, and Andrew Miller. “CanDID：具有传统兼容性、抗 Sybil 性和问责性的去中心化身份”. 载于：2021 IEEE Symposium on Security and Privacy (SP). IEEE. 2021, 第 1348-1366 页.

402. Mike Masnick. “协议，而非平台：一种技术性的言论自由方法”. 2019年8月21日. URL: https://knightcolumbia.org/content/protocols-not-platforms-a-technological-approach-to-free-speech (访问于 2026年5月28日).

403. Sinisa Matetic, Mansoor Ahmed, Kari Kostiainen, Aritra Dhar, David Sommer, Arthur Gervais, Ari Juels, and Srdjan Capkun. “ROTE：可信执行的回滚保护”. 载于：26th USENIX Security Symposium. 2017, 第 1289-1306 页.

404. Matter Labs. zkSync Era. https://www.zksync.io. 访问于 2026-01-28. 2023.

405. Ruben Mayer and Hans-Arno Jacobsen. “分布式基础设施上的可扩展深度学习：挑战、技术与工具”. 载于：ACM Computing Surveys (CSUR) (2020).

406. Trent McConaghy. “Ocean Protocol：Web3 数据经济的工具”. 载于：Handbook on Blockchain. Springer, 2022, 第 505-539 页.

407. Robert McLaughlin, Christopher Kruegel, and Giovanni Vigna. “以太坊套利生态系统的规模研究”. 载于：32nd USENIX Security Symposium (USENIX Security 23). 2023, 第 3295-3312 页.

408. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. “来自去中心化数据的深度网络通信高效学习”. 载于：Artificial intelligence and statistics. PMLR. 2017, 第 1273-1282 页.

409. Muhammad Izhar Mehar, Charles Louis Shier, Alana Giambattista, Elgar Gong, Gabrielle Fletcher, Ryan Sanayhie, Henry M Kim, and Marek Laskowski. “理解区块链中一个革命性但有缺陷的大型实验：DAO 攻击”. 载于：Journal of Cases on Information Technology (JCIT) 21.1 (2019), 第 19-32 页.

410. Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. “机器学习中的偏见与公平性综述”. 载于：ACM computing surveys 54.6 (2021), 第 1-35 页.

411. Messari. Daos.fun 将去中心化对冲基金引入 Solana. https://messari.io/newsletter/unqualified-opinions/daos-fun-brings-decentralized-hedge-funds-to-solana. Messari Newsletter: Unqualified Opinions. 访问于2026年3月. 2024.

412. Microsoft. 关于 Azure 机密虚拟机. Microsoft Learn, https://learn.microsoft.com/en-us/azure/confidential-computing/confidential-vm-overview. Azure 在 AMD SEV-SNP 和 Intel TDX 上的机密虚拟机. (访问于 2026年4月19日).

413. Preston J. Miller and Daniel M. Chin. “使用月度数据改进季度模型预测”. 载于：Federal Reserve Bank of Minneapolis Quarterly Review 20.2 (1996).

414. Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. “大型语言模型：综述”. 载于：arXiv preprint arXiv:2402.06196 (2024).

415. Michael Mirkin, Hongyin Chen, Ohad Eitan, Gal Granot, and Ittay Eyal. “Arbigraph：可验证的图灵完备执行委托”. 载于：Cryptology ePrint Archive (2025).

416. Mehrnoosh Mirtaheri, Sami Abu-El-Haija, Fred Morstatter, Greg Ver Steeg, and Aram Galstyan. “识别和分析社交媒体中的加密货币操纵”. 载于：IEEE Transactions on Computational Social Systems 8.3 (2021), 第 607-617 页.

417. Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. “DetectGPT：使用概率曲率进行零样本机器生成文本检测”. 载于：Proceedings of the 40th International Conference on Machine Learning. ICML'23. Honolulu, Hawaii, USA: JMLR.org, 2023.

418. Modulus Labs. Modulus Labs：负责任的机器智能. 零知识机器学习平台. 访问于 2024-12-19. 2023. url: https://www.modulus.xyz/.

419. Modulus-Labs. RockyBot：完全链上的 AI 交易机器人. https://github.com/Modulus-Labs/RockyBot. 访问于 2026-01-28. 2023.

420. Maruf Monem, Md Tamjid Hossain, Md Golam Rabiul Alam, Md Shirajum Munir, Md Mahbubur Rahman, Salman A AlQahtani, Samah Almutlaq, and Mohammad Mehedi Hassan. “通过使用预测分析引入动态区块大小调整来构建可持续的比特币区块链网络”. 载于：Future Generation Computer Systems 153 (2024), 第 12-26 页.

421. Will Ogden Moore. “AI 即将到来——加密可以帮助它变得正确”. [在线；访问于 2026-02-14]. url: https://research.grayscale.com/reports/ai-is-coming-crypto-can-help-make-it-right.

422. Morpho Docs. [在线；访问于 2025-12-09]. 2025年11月. url: https://docs.morpho.org/get-started/.

423. Malte Moser, Kyle Soska, Ethan Heilman, Kevin Lee, Henry Heffan, Shashvat Srivastava, Kyle Hogan, Jason Hennessey, Andrew Miller, Arvind Narayanan, et al. “Monero 区块链中可追溯性的实证分析”. 载于：2018.

424. Adam Moszczynski. “频繁批量拍卖中自动做市商的优化”. 硕士论文. Stevens Institute of Technology, 2025.

425. Viraj Nadkarni, Jiachen Hu, Ranvir Rana, Chi Jin, Sanjeev Kulkarni, and Pramod Viswanath. “ZeroSwap：去中心化金融中的数据驱动最优做市”. 载于：International Conference on Financial Cryptography and Data Security. Springer. 2024, 第 209-227 页.

426. Satoshi Nakamoto. 比特币：一种点对点的电子现金系统. http://www.bitcoin.org/bitcoin.pdf. 2008.

427. Deepak Narayanan, Mohammad Shoeybi, Jared Casper, Patrick LeGresley, Mostofa Patwary, Vijay Korthikanti, Dmitri Vainbrand, Prethvi Kashinkunti, Julie Bernauer, Bryan Catanzaro, et al. “在 GPU 集群上使用 Megatron-LM 高效训练大规模语言模型”. 载于：Proceedings of the international conference for high performance computing, networking, storage and analysis. 2021, 第 1-15 页.

428. Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A Feder Cooper, Daphne Ippolito, Christopher A Choquette-Choo, Eric Wallace, Florian Tramér, and Katherine Lee. “从（生产）语言模型中可扩展提取训练数据”. 载于：arXiv preprint arXiv:2311.17035 (2023).

429. Kartik Nayak, Srijan Kumar, Andrew Miller, and Elaine Shi. “顽固挖矿：泛化自私挖矿并与日食攻击结合”. 载于：2016 ieee european symposium on security and privacy (euros&P). IEEE. 2016, 第 305-320 页.

430. NEAR AI. NEAR AI 推出 NEAR AI Cloud 和私人聊天. Business Wire. 访问于 2026-06-06. 2025年12月3日. URL: https://www.businesswire.com/news/home/20251203353584/en/NEAR-AI-Launches-NEAR-AI-Cloud-and-Private-Chat.

431. Shutter Network. 即将登陆 DAO：通过同态加密实现永久屏蔽投票. https://blog.shutter.network/coming-soon-to-daos-permanent-shielded-voting-via-homomorphic-encryption/. 2025.

432. Huy Nghiem, Goran Muric, Fred Morstatter, and Emilio Ferrara. “使用市场和社会信号检测加密货币拉高出货欺诈”. 载于：Expert Systems with Applications 182 (2021), 第 115284 页.

433. Tai D. Nguyen, Long H. Pham, and Jun Sun. “SGUARD：自动修复易受攻击的智能合约”. 载于：2021 IEEE Symposium on Security and Privacy (SP). IEEE. 2021, 第 1215-1229 页.

434. Ivica Nikolić, Aashish Kolluri, Ilya Sergey, Prateek Saxena, and Aquinas Hobor. “大规模寻找贪婪、挥霍和自杀合约”. 载于：Proceedings of the 34th Annual Computer Security Applications Conference. ACSAC '18. San Juan, PR, USA: Association for Computing Machinery, 2018, 第 653-663 页.

435. Jianyu Niu, Wei Peng, Xiaokuan Zhang, and Yinqian Zhang. “Narrator：云中可信执行的安全实用状态连续性”. 载于：Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security. 2022, 第 2385-2399 页.

436. Leonardo Nizzoli, Serena Tardelli, Marco Avenuti, Stefano Cresci, Maurizio Tesconi, and Emilio Ferrara. “绘制在线加密货币操纵格局”. 载于：IEEE access 8 (2020), 第 113230-113245 页.

437. NovaNet. NovaNet 主导 Circle 直播展示 Arc 上的无需信任代理. https://www.novanet.xyz/blog/novanet-leads-arc-livestream-on-trustless-agents. 2025.

438. Adamantios Ntakaris, Martin Magris, Juho Kanniainen, Moncef Gabbouj, and Alexandros Iosifidis. “用于机器学习方法中间价预测的限价订单簿基准数据集”. 载于：Journal of Forecasting 37.8 (2018), 第 852-866 页.

439. Numbers Protocol. Numbers Protocol 白皮书：人类与 AI 的来源基础设施. 访问于 2026年2月. 2025. URL: https://docs.numbersprotocol.io/introduction/whitepaper/.

440. NVIDIA. “用于安全可信 AI 的 NVIDIA H100 GPU 机密计算”. 技术报告. 访问于 2026年6月8日. 2023年8月. URL: https://developer.nvidia.com/blog/confidential-computing-on-h100-gpus-for-secure-and-trustworthy-ai/.

441. NVIDIA. 具有机密计算的 AI 安全. https://www.nvidia.com/en-us/data-center/solutions/confidential-computing/. 访问于2026年4月. 2025.

442. NVIDIA. H100 GPU. [在线；访问于 2025-12-05]. URL: https://www.nvidia.com/en-us/data-center/h100/.

443. NVIDIA. 使用 Blackwell 和 Hopper GPU 的 NVIDIA 安全 AI. 技术报告. 白皮书. 2025 年更新.

444. Matt O'Brien. “ChatGPT 制造商 OpenAI 与美联社签署协议，授权新闻内容”. 载于：AP News (2023年7月). [在线；访问于 2026-04-07]. URL: https://apnews.com/article/openai-chatgpt-associated-press-ap-f86f84c5bcc2f3b98074b38521f5f75a.

445. Matt O'Brien and Associated Press. “Reddit 起诉 AI 公司涉嫌‘工业规模’抓取其用户评论”. 载于：PBS News (2025年10月). [在线；访问于 2026-04-07]. URL: https://www.pbs.org/newshour/nation/reddit-sues-ai-company-over-alleged-industrial-scale-scraping-of-its-users-comments.

446. Oasis Labs. PrivateSQL. https://www.oasislabs.com/privatesql. 访问于 2026-06-06.

447. Oasis Protocol Foundation. “Oasis 区块链平台”. 技术报告. Oasis Protocol Foundation, 2020. URL: https://oasisprotocol.org/whitepaper.

448. OpenAI. “AI 民主输入资助计划：经验教训与实施计划”. 访问于 2024-12-19. 2024. URL: https://openai.com/index/democratic-inputs-to-ai-grant-program-update/.

449. OpenAI. 介绍深度研究. 2025年2月2日. URL: https://openai.com/index/introducing-deep-research/.

450. ORAICHAIN. 投资加密：释放 AI Layer 1 区块链生态系统的力量. [在线；访问于 2026-05-12]. URL: https://orai.io/ai-layer-1.

451. Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. “Knockoff Nets：窃取黑盒模型的功能”. 载于：Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019, 第 4954-4963 页.

452. Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. “用人类反馈训练语言模型以遵循指令”. 载于：Advances in neural information processing systems 35 (2022), 第 27730-27744 页.

453. Christina Ovezik, Dimitris Karakostas, Mary Milad, Daniel W Woods, and Aggelos Kiayias. “SoK：衡量区块链去中心化”. 载于：International Conference on Applied Cryptography and Network Security. Springer. 2025, 第 184-214 页.

454. Alex Ozdemir and Dan Boneh. “协作式 zk-SNARKs 的实验：[零知识] 证明分布式秘密”. 载于：31st USENIX Security Symposium (USENIX Security 22). 2022, 第 4291-4308 页.

455. Diogo Pacheco, Pik-Mai Hui, Christopher Torres-Lugo, Bao Tran Truong, Alessandro Flammini, and Filippo Menczer. “揭示社交媒体上的协调网络：方法和案例研究”. 载于：Proceedings of the international AAAI conference on web and social media. Vol. 15. 2021, 第 455-466 页.

456. Jaswant Pakki, Yan Shoshitaishvili, Ruoyu Wang, Tiffany Bao, and Adam Doupe. “关于比特币混币器你想知道的一切（但不敢问）”. 载于：International conference on financial cryptography and data security. Springer. 2021, 第 117-146 页.

457. Georgios Palaiokrassas, Sandro Scherrers, Iason Ofedis, and Leandros Tassiulas. “利用机器学习进行多链 DeFi 欺诈检测”. 载于：2024 IEEE International Conference on Blockchain and Cryptocurrency (ICBC). 2024, 第 678-680 页.

458. Ke Pan, Yew-Soon Ong, Maoguo Gong, Hui Li, A Kai Qin, and Yuan Gao. “深度学习中的差分隐私：文献综述”. 载于：Neurocomputing 589 (2024), 第 127663 页.

459. Xudong Pan, Jiarun Dai, Yihe Fan, and Min Yang. “前沿 AI 系统已超越自我复制红线”. 2024. arXiv: 2412.12140 [cs.CL]. URL: https://arxiv.org/abs/2412.12140.

460. Konstantin D Pandl, Scott Thiebes, Manuel Schmidt-Kraepelin, and Ali Sunyaev. “人工智能与分布式账本技术的融合：范围综述与未来研究议程”. 载于：IEEE access 8 (2020), 第 57075-57095 页.

461. Nikolaos Papadis and Leandros Tassiulas. “基于深度强化学习的再平衡策略用于支付通道网络中中继节点的利润最大化”. 载于：The International Conference on Mathematical Research for Blockchain Economy. Springer. 2023, 第 1-27 页.

462. Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. “蒸馏作为对抗深度神经网络扰动的一种防御”. 载于：2016 IEEE symposium on security and privacy (SP). IEEE. 2016, 第 582-597 页.

463. Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Ulfar Erlingsson. “使用 PATE 的可扩展私有学习”. 载于：arXiv preprint arXiv:1802.08908 (2018).

464. Vatsal Patel, Sutharshan Rajasegarar, Lei Pan, Jiajun Liu, and Liming Zhu. “EvAnGCN：基于演化图深度神经网络的区块链异常检测”. 载于：International Conference on Advanced Data Mining and Applications. Springer. 2022, 第 444-456 页.

465. Vivek Pathak, Rahul Jashvantbhai Pandya, Vimal Bhatia, and Onel Alcaraz Lopez. “面向6G及更远未来的AI集成区块链方法定性综述”. 载于：IEEE Access 11 (2023), 第 105935-105981 页.

466. Atharv Singh Patlan, Ashwin Hebbar, Pramod Viswanath, and Prateek Mittal. “上下文操纵攻击：Web 代理易受损坏内存影响”. 载于：arXiv preprint arXiv:2506.17318 (2025).

467. Atharv Singh Patlan, Peiyao Sheng, S Ashwin Hebbar, Prateek Mittal, and Pramod Viswanath. “加密世界中的 AI 代理：实际攻击与无银弹”. 载于：Cryptology ePrint Archive (2025).

468. Atharv Singh Patlan, Peiyao Sheng, S Ashwin Hebbar, Prateek Mittal, and Pramod Viswanath. “带有虚假记忆的真实 AI 代理：对 Web3 代理的致命上下文操纵攻击”. 载于：arXiv preprint arXiv:2503.16248 (2025).

469. Hester M. Peirce. 关于充分去中心化的声明. https://x.com/HesterPeirce/status/1423637816492318722. 推文，2021年8月6日. 2021年8月.

470. Bowen Peng, Jeffrey Quesnelle, and Diederik P. Kingma. “DeMo：解耦动量优化”. 载于：(2026). https://openreview.net/forum?id=U9oewpa7cn.

471. Zhizhi Peng, Taotao Wang, Chonghe Zhao, Guofu Liao, Zibin Lin, Yifeng Liu, Bin Cao, Long Shi, Qing Yang, and Shengli Zhang. “基于零知识证明的可验证机器学习综述”. 载于：arXiv preprint arXiv:2502.18535 (2025).

472. Penumbra Labs. “Penumbra 协议”. https://protocol.penumbra.zone/main/index.html. 活跃的协议规范；无静态白皮书 PDF. 负责人：Henry de Valence. (访问于 2026年4月17日).

473. Fabio Perez and Ian Ribeiro. “忽略之前的提示：语言模型的攻击技术”. 载于：arXiv preprint arXiv:2211.09527 (2022).

474. Anton Permenev, Dimitar Dimitrov, Petar Tsankov, Dana Drachsler-Cohen, and Martin Vechev. “Vex：智能合约的安全验证”. 载于：2020 IEEE symposium on security and privacy (SP). IEEE. 2020, 第 1661-1677 页.

475. Dana Pessach and Erez Shmueli. “机器学习中的公平性综述”. 载于：ACM Computing Surveys 55.3 (2022), 第 1-44 页.

476. Suzanne Phan. “加州 DMV 声称特斯拉在自动驾驶功能上误导司机；寻求暂停业务30天 - ABC7 旧金山”. [在线；访问于 2025-08-28]. 2025年7月. URL: https://abc7news.com/post/california-dmv-claims-tesla-misled-drivers-driving-capabilities-looks-suspend-business-30-days/17234677/.

477. David Phelan. “iPhone 电池门最新消息：更多 iPhone 用户现在可以申请赔付”. [在线；访问于 2025-08-28]. 2024年4月. URL: https://www.forbes.com/sites/davidphelan/2024/04/07/iphone-batterygate-latest-more-iphone-users-can-now-claim-payouts/.

478. Nadia Pocher, Mirko Zichichi, Fabio Merizzi, Muhammad Zohaib Shafiq, and Stefano Ferretti. “检测异常加密货币交易：基于机器学习的 AML/CFT 取证应用”. 载于：Electronic Markets 33.1 (2023), 第 37 页.

479. Polygon Labs. “Polygon zkEVM：EVM 等效的零知识 Rollup”. https://polygon.technology/polygon-zkevm. 2023.

480. Polymarket. Zelenskyy 会在七月前穿西装吗？ https://polymarket.com/event/will-zelenskyy-wear-a-suit-before-july. 访问于 2026-06-01. 2025.

481. Proof of Cloud Alliance. https://proofofcloud.org/. (访问于 2026年4月17日).

482. Python Network. URL: https://pyth.network/ (访问于 2026年2月6日).

483. Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. “微调对齐语言模型会损害安全性，即使并非用户本意！”载于：arXiv preprint arXiv:2310.03693 (2023).

484. Yuxin Qi, Jun Wu, Hansong Xu, and Mohsen Guizani. “基于图学习的区块链数据挖掘：综述”. 载于：IEEE Transactions on Pattern Analysis and Machine Intelligence 46.2 (2023), 第 729-748 页.

485. Yan Qiao, Kui Wu, and Majid Khabbazian. “使用强化学习在闪电网络中进行非侵入性余额断层扫描”. 载于：ACM Transactions on Privacy and Security 27.1 (2024), 第 1-32 页.

486. Kaihua Qin, Liyi Zhou, Pablo Gamito, Philipp Jovanovic, and Arthur Gervais. “DeFi 清算的实证研究：激励、风险与不稳定性”. 载于：Proceedings of the 21st ACM internet measurement conference. 2021, 第 336-350 页.

487. Kaihua Qin, Liyi Zhou, and Arthur Gervais. “量化区块链可提取价值：森林有多黑暗？”载于：2022 IEEE Symposium on Security and Privacy (SP). IEEE. 2022, 第 198-214 页.

488. Qiro. QIRO X EZKL - DeFi 的可验证承保. https://www.qiro.fi/blogs/qiro-x-ezkl-verifiable-underwriting-for-defi. 2025.

489. Wenjie Qu, Yanpei Guo, Yue Ying, and Jiaheng Zhang. “VerfCNN，卷积神经网络的最优复杂度 zk-SNARK”. 载于：Cryptology ePrint Archive (2025).

490. Wenjie Qu, Yijun Sun, Xuanming Liu, Tao Lu, Yanpei Guo, Kai Chen, and Jiaheng Zhang. “zkGPT：一种用于 LLM 推理的高效非交互式零知识证明框架”. 载于：34st USENIX Security Symposium (USENIX Security 25). 2025.

491. Maciel M Queiroz, Renato Telles, and Silvia H Bonilla. “区块链与供应链管理整合：文献系统综述”. 载于：Supply chain management: An international journal 25.2 (2020), 第 241-254 页.

492. Felix Quinque, Alan Aboudib, Szymon Fonau, Rodrigo Lopez Portillo Alcocer, Brian McCrindle, and Steffen Cruz. “激励式编排训练架构（IOTA）：发布技术入门”. 载于：arXiv preprint arXiv:2507.17766 (2025).

493. Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. “语言模型是无监督的多任务学习者”. 载于：OpenAI blog 1.8 (2019), 第 9 页.

494. Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. “直接偏好优化：你的语言模型秘密地是一个奖励模型”. 载于：Advances in neural information processing systems 36 (2023), 第 53728-53741 页.

495. Zeeshan Rahman. “在线流媒体平台上的版权执行：数字时代权利持有者面临的挑战”. 载于：International Journal for Multidisciplinary Research (IJFMR) 5 (2023), 第 1-14 页.

496. Somesh Rai, Kunwar Singh, and Akhilesh Kumar Varma. “1997-2019 年深度网络研究的文献计量分析”. 载于：DESIDOC Journal of Library & Information Technology 40.2 (2020).

497. Rahul Rao. “AI 生成的数据可能会毒害未来的 AI 模型”. 载于：Scientific American (2023年7月23日). URL: https://www.scientificamerican.com/article/ai-generated-data-can-poison-future-ai-models/.

498. Ramesh Raskar, Pradyumna Chari, John Zinky, Sichao Wang, Rekha Singhal, Robert Lincourt, Mahesh Lambe, Jared James Grogan, Rajesh Ranjan, Shailja Gupta, Raghu Bala, Aditi Joshi, Abhishek Singh, Ayush Chopra, Dimitris Stripebis, Bhuvan B, Sumit Kumar, and Maria Gorskikh. “超越 DNS：通过 NANDA 索引和验证的 AgentFacts 解锁互联网 AI 代理”. 载于：arXiv preprint arXiv:2507.14263 (2025). NANDA 项目.

499. Christoffer Raun, Benjamin Estermann, Liyi Zhou, Kaihua Qin, Roger Wattenhofer, Arthur Gervais, and Ye Wang. “利用机器学习进行矿工可提取价值拍卖中的出价策略”. 载于：Cryptology ePrint Archive (2023).

500. Traian Rebedea, Razvan Dinu, Makesh Narsimhan Sreedhar, Christopher Parisien, and Jonathan Cohen. “Nemo Guardrails：具有可编程护栏的可控安全 LLM 应用工具包”. 载于：Proceedings of the 2023 conference on empirical methods in natural language processing: system demonstrations. 2023, 第 431-445 页.

501. Reclaim Protocol. URL: https://reclaimprotocol.org/ (访问于 2026年2月6日).

502. Swaroopa Reddy and GVV Sharma. “UL-blockDAG：基于无监督学习的区块链共识协议”. 载于：2020 IEEE 40th International Conference on Distributed Computing Systems (ICDCS). IEEE. 2020, 第 1243-1248 页.

503. RedPill. RedPill：你的私有 ChatGPT，端到端加密，由机密 AI 驱动. https://www.redpill.ai/. 访问于2026年6月.

504. RedStone Oracles. URL: https://redstone.finance/ (访问于 2026年2月6日).

505. Zubaida Rehman, Mark A Gregory, Iqbal Gondal, Hai Dong, and Mengmeng Ge. “区块链网络中的日食攻击：检测、预防与未来方向”. 载于：IEEE Access 13 (2025), 第 25918-25933 页.

506. Erik Reppel, Ronnie Caspers, Kevin Leffew, Danny Organ, Dan Kim, and Nemil Dalal. “x402：一个开放的互联网原生支付标准”. [在线；访问于 2026-05-15]. 2025年5月. URL: https://www.x402.org/x402-whitepaper.pdf.

507. RetroPGF. https://www.retropgf.com/.

508. Marco Reuter. 通过去中心化的平台预承诺. International Monetary Fund, 2024.

509. Reuters. “微软与 OpenAI 达成新协议，OpenAI 估值 5000 亿美元”. 载于：NBC News (2025年10月). [在线；访问于 2026-04-07]. URL: https://www.nbcnews.com/tech/tech-news/microsoft-openai-reach-new-deal-valuing-openai-500-billion-rcnat240255.

510. Carla L. Reyes. “（非）公司加密治理”. 载于：Fordham Law Review 88 (2020), 第 1875-1920 页. URL: https://ir.lawnet.fordham.edu/flr/vol88/iss5/13/.

511. Filip Rezabek, Moe Mahhouk, Andrew Miller, Quintus Kilbourn, Georg Carle, and Jonathan Passerat-Palmbach. “Proof of Cloud：机密虚拟机的数据中心执行保证”. 2025. arXiv: 2510.12469 [cs.CR]. URL: https://arxiv.org/abs/2510.12469.

512. Hubert Ritzdorf, Karl Wüst, Arthur Gervais, Guillaume Felley, and Srdjan Capkun. “TLS-N：TLS 上的不可否认性，实现去中介化的无处不在内容签名”. 2017. URL: https://eprint.iacr.org/2017/578 (访问于 2026年2月19日). 预发表.

513. Syamsul Rizal and Dong-Seong Kim. “增强区块链共识机制：关于机器学习应用与优化的全面综述”. 载于：Blockchain: Research and Applications 6.4 (2025).

514. Huw Roberts, Emmie Hine, Mariarosaria Taddeo, and Luciano Floridi. “全球 AI 治理：障碍与前进路径”. 载于：International Affairs 100.3 (2024), 第 1275-1286 页.

515. Michael Rodler, Wenting Li, Ghassan O. Karame, and Lucas Davi. “EVMPatch：以太坊智能合约的及时自动修补”. 载于：30th USENIX Security Symposium (USENIX Security 21). USENIX Association, 2021年8月, 第 1289-1306 页.

516. Francisco Rodrigues. “以太坊 Vitalik Buterin 提议 AI‘管家’帮助重塑 DAO 治理”. CoinDesk. 2026年2月. URL: https://www.coindesk.com/web3/2026/02/21/ethereum-s-vitalik-buterin-proposes-ai-stewards-to-help-reinvent-dao-governance.

517. Jonathan Rose. “腐败的含义：检验腐败定义的连贯性和充分性”. 载于：Public Integrity 20.3 (2018), 第 220-233 页.

518. Stuart Ross and Michelle Hannan. “洗钱监管与基于风险的决策”. 载于：Journal of Money Laundering Control 10.1 (2007), 第 106-115 页.

519. Nicola Ruaro, Fabio Gritti, Robert McLaughlin, Ilya Grishchenko, Christopher Kruegel, and Giovanni Vigna. “不是你的类型！检测以太坊智能合约中的存储冲突漏洞”. 载于：NDSS. 2024.

520. Tim Ruffing, Pedro Moreno-Sanchez, and Aniket Kate. “Coinshuffle：比特币的实用去中心化混币”. 载于：European symposium on research in computer security. Springer. 2014, 第 345-364 页.

521. Stuart J. Russell and Peter Norvig. 人工智能：一种现代方法. 第4版. Pearson, 2020. ISBN: 9780134610993.

522. Stuart Jonathan Russell and Peter Norvig. 人工智能：一种现代方法. Pearson, 2022. 1288页. ISBN: 978-93-5606-357-0.

523. Mark Russinovich. “Azure AI 机密推理：技术深入”. Microsoft Azure Confidential Computing Blog. 2025年12月16日更新. 2024年9月. URL: https://techcommunity.microsoft.com/blog/azureconfidentialcomputingblog/azure-ai-confidential-inferencing-technical-deep-dive/4253150 (访问于 2026年4月17日).

524. Mark Russinovich, Cedric Fournet, Greg Zaverucha, Josh Benaloh, Brandon Murdoch, and Manuel Costa. “机密计算证明：密码学零知识的替代方案”. 载于：Queue 22.4 (2024), 第 73-100 页.

525. Max Ryabinin, Tim Dettmers, Michael Diskin, and Alexander Borzunov. “Swarm 并行：大型模型的训练可以惊人地通信高效”. 载于：International Conference on Machine Learning. PMLR. 2023, 第 29416-29440 页.

526. Muhammad Saad, Jinchun Choi, DaeHun Nyang, Joongheon Kim, and Aziz Mohaisen. “表征基于区块链的加密货币以实现高精度预测”. 载于：IEEE Systems Journal 14.1 (2019), 第 321-332 页.

527. Khalil Saadat, Ning Wang, and Rahim Tafazolli. “集群车载网络中 AI 赋能的区块链共识节点选择”. 载于：IEEE Networking Letters 5.2 (2023), 第 115-119 页.

528. Mohamed Sabt, Mohammed Achemal, and Abdelmadjid Bouabdallah. “可信执行环境：它是什么，不是什么”. 载于：2015 IEEE Trustcom/BigDataSE/ISPA. Vol. 1. 2015, 第 57-64 页.

529. Bashir Sadeghi and Vishnu Naresh Boddeti. “将公平性赋予预训练的偏见表示”. 载于：Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2020, 第 16-17 页.

530. Khaled Salah, M Habib Ur Rehman, Nishara Nizamuddin, and Ala Al-Fuqaha. “AI 的区块链：综述与开放研究挑战”. 载于：IEEE access 7 (2019), 第 10127-10149 页.

531. Ahmed Salem, Yang Zhang, Mathias Humbert, Mario Fritz, and Michael Backes. “ML-Leaks：对机器学习模型的模型与数据无关成员推理攻击及防御”. 载于：Network and Distributed System Security Symposium (NDSS). 2019.

532. Siddharth Samsi, Dan Zhao, Joseph McDonald, Baolin Li, Adam Michaelas, Michael Jones, William Bergeron, Jeremy Kepner, Devesh Tiwari, and Vijay Gadepally. “从文字到瓦特：衡量大型语言模型推理的能源成本”. 载于：IEEE High Performance Extreme Computing Conference (2023).

533. Ayelet Sapirshtein, Yonatan Sompolinsky, and Aviv Zohar. “比特币中的最优自私挖矿策略”. 载于：International conference on financial cryptography and data security. Springer. 2016, 第 515-532 页.

534. Roozbeh Sarenche, Alireza Aghabagherloo, Svetla Nikova, and Bart Preneel. “波动区块奖励下的比特币：mempool 统计数据如何影响比特币挖矿”. 载于：Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security. 2025, 第 903-917 页.

535. Roozbeh Sarenche, Svetla Nikova, and Bart Preneel. “最长链权益证明协议中的深度自私提议”. 载于：International Conference on Financial Cryptography and Data Security. Springer. 2024, 第 24-40 页.

536. Eli Ben Sasson, Alessandro Chiesa, Christina Garman, Matthew Green, Ian Miers, Eran Tromer, and Madars Virza. “Zerocash：来自比特币的去中心化匿名支付”. 载于：2014 IEEE symposium on security and privacy. IEEE. 2014, 第 459-474 页.

537. Ryoma Sato. “即使是 GPT-5.2 也无法数到五：可信 LLM 中零误差极限的情况”. 2026. arXiv: 2601.15714 [cs.LG]. URL: https://arxiv.org/abs/2601.15714.

538. Stephan van Schaik, Alex Seto, Thomas Yurek, Adam Batori, Bader AlBassam, Daniel Genkin, Andrew Miller, Eyal Ronen, Yuval Yarom, and Christina Garman. “SoK：SGX.Fail——东西是如何被暴露的”. 载于：2024 IEEE Symposium on Security and Privacy (SP). 2024. URL: https://sgx.fail/files/sgx.fail.pdf.

539. Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. “Toolformer：语言模型可以自学使用工具”. 载于：Advances in neural information processing systems 36 (2023), 第 68539-68551 页.

540. Alex Schlogl, Nora Hofer, and Rainer Böhme. “神经网络推理框架中意外数值偏差的原因与影响”. 载于：Advances in Neural Information Processing Systems 36 (2023), 第 56095-56107 页.

541. Fred B Schneider. “可执行的安全策略”. 载于：ACM Transactions on Information and System Security (TISSEC) 3.1 (2000), 第 30-50 页.

542. John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. “近端策略优化算法”. 载于：arXiv preprint arXiv:1707.06347 (2017).

543. Sentient Foundation / 开源 AGI. [在线；访问于 2026-02-14]. URL: https://www.sentient.foundation/.

544. Amazon Web Services. AWS Marketplace. https://aws.amazon.com/marketplace. 访问于2026年6月.

545. Ali Shahin Shamsabadi, Sierra Calanda Wyllie, Nicholas Franzese, Natalie Dullerud, Sebastien Gambs, Nicolas Papernot, Xiao Wang, and Adrian Weller. “Confidential-PROFITT：树的公平训练机密证明”. 载于：The Eleventh International Conference on Learning Representations. 2023.

546. Sanjif Shanmugavelu, Mathieu Taillefumier, Christopher Culver, Oscar Hernandez, Mark Coletti, and Ada Sedova. “浮点非结合性对 HPC 和深度学习应用可重现性的影响”. 载于：SC24-W: Workshops of the International Conference for High Performance Computing, Networking, Storage and Analysis. IEEE. 2024, 第 170-179 页.

547. Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. “DeepSeekMath：推动开放语言模型中数学推理的极限”. 载于：arXiv preprint arXiv:2402.03300 (2024).

548. Tanusree Sharma, J Park, Y Kwon, Yiren Liu, Yun Huang, Sunny Liu, Dawn Song, Jeff Hancock, and Yang Wang. “Inclusive.AI：让服务不足的人群参与 AI 的民主决策”. 载于：SocialComputing. Web. Illinois. edu https://socialcomputing.web.illinois.edu/images/Report-InclusiveAI.pdf (2024).

549. Tanusree Sharma, Yujin Potter, Kornrapat Pongmala, Henry Wang, Andrew Miller, Dawn Song, and Yang Wang. “解开去中心化自治组织在实践中如何运作”. 载于：2024 IEEE International Conference on Blockchain and Cryptocurrency (ICBC). IEEE. 2024, 第 416-424 页.

550. Megan Shearer, Gabriel Rauterberg, and Michael P Wellman. “学习操纵金融基准”. 载于：Proceedings of the Fourth ACM International Conference on AI in Finance. 2023, 第 592-600 页.

551. Jie Shen, Jiajun Zhou, Yunyi Xie, Shanqing Yu, and Qi Xuan. “使用图神经网络的区块链身份推断”. 载于：International conference on blockchain and trustworthy systems. Springer. 2021, 第 3-17 页.

552. Meng Shen, Aijing Gu, Jiawen Kang, Xiangyun Tang, Xiaodong Lin, Liehuang Zhu, and Dusit Niyato. “面向物联网人工智能的区块链：全面综述”. 载于：IEEE Internet of Things Journal 10.16 (2023), 第 14483-14506 页.

553. Peiyao Sheng, Nikita Yadav, Vishal Sevani, Arun Babu, SVR Anand, Himanshu Tyagi, and Pramod Viswanath. “回传证明：宽带带宽的无需信任测量”. 载于：NDSS. 2024.

554. Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. “对机器学习模型的成员推理攻击”. 载于：2017 IEEE symposium on security and privacy (SP). IEEE. 2017, 第 3-18 页.

555. Chaofan Shou, Shangyin Tan, and Koushik Sen. “ItyFuzz：智能合约的基于快照的模糊测试器”. 载于：Proceedings of the 32nd ACM SIGSOFT International Symposium on Software Testing and Analysis. ISSTA 2023. Seattle, WA, USA: Association for Computing Machinery, 2023, 第 322-333 页.

556. Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. “在递归生成数据上训练的 AI 模型会崩溃”. 载于：Nature 631.8022 (2024), 第 755-759 页.

557. Amanda Silberling. “你的公共 ChatGPT 查询正在被 Google 和其他搜索引擎索引”. TechCrunch. 2025年7月31日. URL: https://techcrunch.com/2025/07/31/your-public-chatgpt-queries-are-getting-indexed-by-google-and-other-search-engines/ (访问于 2025年8月11日).

558. SingularityDAO. SingularityDAO：DeFi 的 AI 驱动量化策略. https://singularitydao.ai. 访问于2026年6月.

559. SingularityNet. 关于我们. [在线；访问于 2026-05-12]. URL: https://singularitynet.io/aboutus/.

560. Neelabh Sinha, Vinija Jain, and Aman Chadha. “小型语言模型准备好与大型语言模型在实用应用中竞争了吗？”载于：Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025). 2025, 第 365-398 页.

561. Sky Ecosystem. Sky. https://sky.money/. 发行 USDS 和 sUSDS；传统 DAI 稳定币（最初来自 MakerDAO）也是生态系统的一部分. 2024.

562. Matthew S. Smith. “Meta 在 AI 数据标注上的投资解析”. 载于：IEEE Spectrum (2025年8月). [在线；访问于 2026-04-07]. URL: https://spectrum.ieee.org/data-labeling-scale-ai-agents.

563. Justin Smulison. “Reddit 的诉讼可能改变 AI 对你的了解程度”. 载于：Best Lawyers (2025年9月). [在线；访问于 2026-04-07]. URL: https://www.bestlawyers.com/article/reddit-lawsuit-could-change-how-much-ai-knows-about-you/6905.

564. Snapshot Labs. Snapshot：DAO 投票平台. https://snapshot.org. 访问于 2024-12-03. 2024.

565. Snowflake. Snowflake 市场. https://www.snowflake.com/en/data-cloud/marketplace/. 2014.

566. Sunbeom So, Seongjoon Hong, and Hakjoo Oh. “SmarTest：通过语言模型引导的符号执行有效发现智能合约中的脆弱交易序列”. 载于：30th USENIX Security Symposium (USENIX Security 21). USENIX Association, 2021年8月, 第 1361-1378 页.

567. Sunbeom So and Hakjoo Oh. “SmartFix：使用统计模型加速生成-验证修复，修复脆弱智能合约”. 载于：Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering. ESEC/FSE 2023. San Francisco, CA, USA: Association for Computing Machinery, 2023, 第 185-197 页.

568. Michael Sockin and Wei Xiong. “通过代币化的去中心化”. 载于：The Journal of Finance 78.1 (2023), 第 247-299 页.

569. Bahrad A Sokhansanj. “野外的未经审查 AI：追踪公开可用和本地可部署的 LLM”. 载于：Future Internet 17.10 (2025), 第 477 页.

570. Soldex. Soldex：基于 Solana 构建的可扩展去中心化 AI 驱动交易所. https://soldex.ai. 访问于2026年3月.

571. Chuangang Song, Leixiao Li, and Haoyu Gao. “基于 BiLSTM-PPO 的支付通道费用动态设定算法”. 载于：Third International Conference on Electronic Information Engineering, Big Data, and Computer Technology (EIBDCT 2024). Vol. 13181. SPIE. 2024, 第 1294-1301 页.

572. Mingxuan Song, Pengze Li, Bohan Zhou, Shenglin Yin, Zhen Xiao, and Jieyi Long. “AERO：通过基于深度强化学习的账户迁移增强分片区块链”. 载于：Proceedings of the ACM on Web Conference 2025. 2025, 第 706-716 页.

573. Mohsen Soori, Roza Dastres, and Behrooz Arezoo. “工业 4.0 中的 AI 赋能区块链技术：综述”. 载于：Journal of Economy and Technology 1 (2023), 第 222-241 页.

574. Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloufar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, et al. “走向多元对齐的路线图”. 载于：arXiv preprint arXiv:2402.05070 (2024).

575. Ben Sperry. “网络中立与私有审查的悖论”. https://truthonthemarket.com/2024/04/25/net-neutrality-and-the-paradox-of-private-censorship/. 访问于 2026-06-06. 2024年4月25日.

576. Alexander Spiegelman, Neil Giridharan, Alberto Sonnino, and Lefteris Kokoris-Kogias. “Bullshark：DAG BFT 协议的实用化”. 载于：Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security. 2022, 第 2705-2718 页.

577. Himanshu Srivastava, Akansha Singh, Anuj Kumar Bharti, and Korhan Cengiz. “BlockXAI：区块链用于可解释人工智能的综述”. 载于：Convergence of Blockchain and Explainable Artificial Intelligence (2024), 第 1-14 页.

578. Megha Srivastava, Simran Arora, and Dan Boneh. “通过控制硬件非确定性实现乐观可验证训练”. 载于：Advances in Neural Information Processing Systems 37 (2024), 第 95639-95661 页.

579. Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. “Dropout：一种防止神经网络过拟合的简单方法”. 载于：The journal of machine learning research 15.1 (2014), 第 1929-1958 页.

580. Starknet-io. StarkNet. https://www.starknet.io. 访问于 2026-01-28. 2025.

581. Starling Lab for Data Integrity. Starling 数据完整性框架：捕获、存储、验证. 访问于 2026年2月. 2024. URL: https://www.starlinglab.org/.

582. Sebastian U. Stich. “Local SGD 收敛快且通信少”. 载于：International Conference on Learning Representations. 2019. URL: https://openreview.net.

583. Joseph Stiglitz et al. “政府透明度”. 载于：The right to tell: The role of mass media in economic development 25070 (2002), 第 27-44 页.

584. Story. 我们为什么孵化 Poseidon. [在线；访问于 2026-05-13]. 2025年7月. URL: https://www.story.foundation/blog/why-were-incubating-poseidon.

585. Raoul Strackx and Frank Piessens. “Ariadne：一种最小的状态连续性方法”. 载于：25th USENIX Security Symposium (USENIX Security 16). 2016, 第 875-892 页.

586. Stefan Studer, Thanh Binh Bui, Christian Drescher, Alexander Hanuschkin, Ludwig Winkler, Steven Peters, and Klaus-Robert Müller. “走向 CRISP-ML(Q)：一个带有质量保证方法论的机器学习过程模型”. 载于：Machine learning and knowledge extraction 3.2 (2021), 第 392-413 页.

587. Xing Su, Hanzhong Liang, Hao Wu, Ben Niu, Fengyuan Xu, and Sheng Zhong. “DiSCo：使用大型语言模型将 EVM 字节码分解为源代码”. 载于：Proc. ACM Softw. Eng. 2. FSE (2025年6月).

588. Haochen Sun, Jason Li, and Hongyang Zhang. “zkLLM：大型语言模型的零知识证明”. 载于：Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security. 2024, 第 4405-4419 页.

589. Yifan Sun, Yupeng Li, Yanyi Zhang, Jiayu Jin, and Hao Zhang. “SVIP：迈向开源大型语言模型的可验证推理”. 载于：arXiv preprint (2025).

590. Yuqiang Sun, Daoyuan Wu, Yue Xue, Han Liu, Haijun Wang, Zhengzi Xu, Xiaofei Xie, and Yang Liu. “GPTScan：通过结合 GPT 与程序分析检测智能合约中的逻辑漏洞”. 载于：Proceedings of the IEEE/ACM 46th International Conference on Software Engineering. ICSE '24. Lisbon, Portugal: Association for Computing Machinery, 2024.

591. Supra. URL: https://supra.com/ (访问于 2026年2月6日).

592. Richard S Sutton, Andrew G Barto, et al. 强化学习：导论. Vol. 1. 1. MIT press Cambridge, 1998.

593. Nick Szabo. “微支付与心理交易成本”. 载于：2nd Berlin Internet Economics Workshop. Vol. 44. 1999, 第 44 页.

594. Araz Taeihagh. “生成式 AI 的治理”. 载于：Policy and society 44.1 (2025), 第 1-22 页.

595. Weizhao Tang, Lucianna Kiffer, Giulia Fanti, and Ari Juels. “区块链点对点网络中的战略延迟减少”. 载于：Proceedings of the ACM on Measurement and Analysis of Computing Systems 7.2 (2023), 第 1-33 页.

596. Nous Research Team. “民主化 AI：Psyche 网络架构”. [在线；访问于 2026-02-03]. 2025年5月. URL: https://nousresearch.com/nous-psyche/.

597. Optimism Team. “乐观 Rollup：具有经济安全性的可扩展执行”. https://optimism.io/. 2019.

598. Tellor. URL: https://tellor.io/ (访问于 2026年2月6日).

599. Tether Operations. Tether (USDT)：透明度与储备. https://tether.to/en/. 由 Tether Operations, S.A. de C.V. 发行；按 1:1 与美元挂钩. 2024.

600. Theta Labs. 介绍 Theta EdgeCloud. [在线；访问于 2025-11-26]. URL: https://assets.thetatoken.org/theta-edgecloud-whitepaper-latest.pdf.

601. Theta Network. [在线；访问于 2025-11-26]. URL: https://www.thetatoken.org.

602. Max von Thun. “垄断力量是 AI 辩论中的大象”. [在线；访问于 2026-05-11]. 2023年10月. URL: https://www.technology.press/monopoly-power-is-the-elephant-in-the-room-in-the-ai-debate/.

603. Zhihua Tian, Jian Liu, Jingyu Li, Xinle Cao, Ruoxi Jia, Jun Kong, Mengdi Liu, and Kui Ren. “数据市场中的私有数据估值与公平支付”. 载于：arXiv preprint arXiv:2210.08723 (2022).

604. Tinfoil：可验证的私有 AI. https://tinfoil.sh. 硬件飞地中的机密 AI 推理（AMD SEV-SNP）；兼容 OpenAI 的私有推理 API. 访问于 2026-06-06.

605. Jean Tirole. “数字时代的竞争与工业挑战”. 载于：Annual Review of Economics 15.1 (2023), 第 573-605 页.

606. TLSNotary. URL: https://tlsnotary.org/docs/intro (访问于 2026年2月6日).

607. Christof Ferreira Torres, Ramiro Camino, et al. “Frontrunner Jones 和黑暗森林的掠夺者：以太坊区块链前置交易的实证研究”. 载于：30th USENIX Security Symposium (USENIX Security 21). 2021, 第 1343-1359 页.

608. Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. “LLaMA：开放高效的基础语言模型”. 载于：arXiv preprint arXiv:2302.13971 (2023).

609. Florian Tramer, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. “对对抗性样本防御的自适应攻击”. 载于：Advances in neural information processing systems 33 (2020), 第 1633-1645 页.

610. Florian Tramer, Fan Zhang, Ari Juels, Michael K Reiter, and Thomas Ristenpart. “通过预测 API 窃取机器学习模型”. 载于：25th USENIX security symposium. 2016, 第 601-618 页.

611. Petar Tsankov, Andrei Dan, Dana Drachsler-Cohen, Arthur Gervais, Florian Bünzli, and Martin Vechev. “Securify：智能合约的实用安全分析”. 载于：Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security. CCS '18. Toronto, Canada: Association for Computing Machinery, 2018, 第 67-82 页.

612. Giorgos Tsimos, Anastasios Kichidis, Alberto Sonnino, and Lefteris Kokoris-Kogias. “HammerHead：用于动态调度的领导者声誉”. 载于：44th IEEE International Conference on Distributed Computing Systems, ICDCS 2024, Jersey City, NJ, USA, 2024年7月23-26日. IEEE, 2024, 第 1377-1387 页.

613. Amanda Tuminelli, Lizandro Pieper, and Peter Van Valkenburgh. Coin Center 和 DeFi 教育基金的专家报告. https://www.defieducationfund.org/. 为支持 Alexey Pertsev 上诉（Tornado Cash 案，荷兰）提交. 2025年5月.

614. Alex Turner, Logan Smith, Rohin Shah, Andrew Critch, and Prasad Tadepalli. “最优策略倾向于寻求权力”. 载于：Advances in Neural Information Processing Systems. 编辑：M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan. Vol. 34. Curran Associates, Inc., 2021, 第 23063-23074 页.

615. Alex Turner and Prasad Tadepalli. “参数化可重定向的决策者倾向于寻求权力”. 载于：Advances in Neural Information Processing Systems 35 (2022), 第 31391-31401 页.

616. UMA Optimistic Oracle. URL: https://uma.xyz/ (访问于 2026年2月6日).

617. UMA 的 AI 实验：AI 代理能否增强乐观预言机？ 2025年6月3日. URL: https://blog.uma.xyz/articles/experiment-can-ai-agents-enhance-uma-oracle (访问于 2026年2月18日).

618. 理解加密与AI的交集 / Galaxy. 2024年2月14日. URL: https://www.galaxy.com/insights/research/understanding-intersection-crypto-ai (访问于 2026年2月18日).

619. Eddie L Ungless, Nina Markl, and Björn Ross. “TikTok 上边缘化身份群体的审查经历”. 载于：arXiv preprint arXiv:2407.14164 (2024).

620. 美利坚合众国诉 Google LLC. [在线；访问于 2026-05-11]. 2025. URL: https://storage.courtlistener.com/recap/gov.uscourts.vaed.533508/gov.uscourts.vaed.533508.1410.0.pdf.

621. 释放 AI 代理：区块链如何实现真正的数字自主. 2025年2月12日. URL: https://blog.sei.io/research/unleashing-ai-agents-how-blockchain-enables-true-digital-autonomy/ (访问于 2026年2月27日).

622. All of Us Research Program Investigators. “All of Us 研究计划”. 载于：New England Journal of Medicine 381.7 (2019), 第 668-676 页.

623. Danila Valko and Daniel Kudenko. “基于强化学习的区块链网络可持续广播”. 载于：arXiv preprint arXiv:2407.15616 (2024).

624. Danila Valko and Daniel Kudenko. “基于强化学习的闪电网络混合路径发现优化”. 载于：Engineering Applications of Artificial Intelligence 146 (2025), 第 110225 页.

625. Magnus Van Haaren, Xule Lin, Carlos Santana, Oliver Baumann, Robert Wayne Gregory, Hanna Halaburda, Fang He, Alex Michael Murray, and Matthias Troebinger. “组织中的算法与人类：在区块链和AI交叉处导航”. 载于：84th Annual Meeting of the Academy of Management, AOM 2024. 2024.

626. Dylan Vassallo, Vincent Vella, and Joshua Ellul. “梯度提升算法在加密货币反洗钱中的应用”. 载于：SN Computer Science 2.3 (2021), 第 143 页.

627. Steve Vassallo. “AI x 区块链：下一个层次”. 访问于 2026年3月. 2023. URL: https://www.forbes.com/sites/stevevassallo/2023/06/13/ai-x-blockchain-the-next-level/.

628. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. “注意力就是你所需要的一切”. 载于：Advances in neural information processing systems 30 (2017).

629. Awid Vaziry, Sandro Rodriguez Garzon, and Axel Kupper. “走向多智能体经济：用账本锚定身份和 x402 微支付增强 A2A 协议，用于 AI 代理”. 载于：arXiv preprint arXiv:2507.19550 (2025).

630. Venice.ai. Venice 推出端到端加密 AI. Venice.ai Blog. 2026年3月. URL: https://venice.ai/blog/venice-launches-end-to-end-encrypted-ai (访问于 2026年4月17日).

631. Kotteswaran Venkatesan and Syarifah Bahiyah Rahayu. “区块链安全增强：混合共识算法与机器学习技术的方法”. 载于：Scientific Reports 14.1 (2024), 第 1149 页.

632. VentionTeams. “生成式 AI 与区块链”. 访问于 2026年3月. 2025. URL: https://ventionteams.com/blog/generative-ai-blockchain.

633. Adrian Vieitez, Matilde Santos, and Rodrigo Naranjo. “机器学习以太坊加密货币预测与基于知识的投资策略”. 载于：Knowledge-Based Systems 299 (2024), 第 112088 页.

634. Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn. “立场：我们会耗尽数据吗？基于人类生成数据的 LLM 扩展限制”. 载于：Forty-first International Conference on Machine Learning. 2024.

635. Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn. “我们会耗尽数据吗？基于人类生成数据的 LLM 扩展限制”. Epoch AI Blog. 基于 arXiv preprint arXiv:2211.04325 的更新分析. 2024年6月. URL: https://epoch.ai/blog/will-we-run-out-of-data-limits-of-llm-scaling-based-on-human-generated-data.

636. Paul Voigt and Axel von dem Bussche. “欧盟通用数据保护条例（GDPR）：实用指南”. Springer, 2017.

637. Tomer Voronov, Danny Raz, and Ori Rottenstreich. “使用草图进行区块链网络异常检测的框架”. 载于：IEEE/ACM Transactions on Networking 32.1 (2023), 第 686-698 页.

638. Sarih Wadhwa, Julian Ma, Thomas Thiery, Barnabe Monnot, Luca Zanolini, Fan Zhang, and Kartik Nayak. “AUCIL：面向理性参与者的包含列表设计”. 2025. (访问于 2025年3月21日).

639. Sameer Wagh, Divya Gupta, and Nishanth Chandran. “SecureNN：用于神经网络训练的三方安全计算”. 载于：Proceedings on Privacy Enhancing Technologies (2019).

640. Anton Wahrstatter, Jens Ernstberger, Aviv Yaish, Liyi Zhou, Kaihua Qin, Taro Tsuchiya, Sebastian Steinhorst, Davor Svetinovic, Nicolas Christin, Mikolaj Barczentewicz, and Arthur Gervais. “区块链审查”. 载于：Proceedings of the ACM Web Conference 2024. WWW '24. New York, NY, USA: Association for Computing Machinery, 2024年5月, 第 1632-1643 页. (访问于 2026年2月4日).

641. Suppakit Waiwitlikhit, Ion Stoica, Yi Sun, Tatsunori Hashimoto, and Daniel Kang. “不泄露数据或模型的无需信任审计”. 载于：arXiv preprint arXiv:2404.04500 (2024).

642. Shaw Walters, Sam Gao, Shakker Nerd, Feng Da, Warren Williams, Ting-Chien Meng, Amie Chow, Hunter Han, Frank He, Allen Zhang, Ming Wu, Timothy Shen, Maxwell Hu, and Jerry Yan. “Eliza：一个 Web3 友好的 AI 代理操作系统”. 2025. DOI: 10.48550/arXiv.2501.06781. arXiv: 2501.06781 [cs.AI]. URL: https://arxiv.org/abs/2501.06781.

643. Fei Wan and Ping Li. “基于动态图卷积神经网络和长短期记忆的新型洗钱预测模型”. 载于：Symmetry 16.3 (2024), 第 378 页.

644. Che Wang, Jiashuo Zhang, Jianbo Gao, Libin Xia, Zhi Guan, and Zhong Chen. “Contract-Tinker：LLM 赋能的现实世界智能合约漏洞修复”. 载于：Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering. ASE '24. Sacramento, CA, USA: Association for Computing Machinery, 2024, 第 2350-2353 页.

645. Jiachen T Wang, Prateek Mittal, Dawn Song, and Ruoxi Jia. “单次训练运行中的数据 Shapley”. 载于：The Thirteenth International Conference on Learning Representations. 2025.

646. Jinlong Wang, Yixin Li, Yunting Wu, Wenhui Zheng, Shangzhuo Zhou, and Xiaoyun Xiong. “基于生成式 AI 和 DRL 的区块链分片方案：应用于构建物联网”. 载于：Internet of Things and Cyber-Physical Systems 4 (2024), 第 333-349 页.

647. Justin Wang, Andreas Bigger, Xiaohai Xu, Justin W. Lin, Andy Applebaum, Tejal Patwardhan, Alpin Yukseloglu, and Olivia Watkins. “EVMbench：评估智能合约安全的 AI 代理”. 2026. URL: https://cdn.openai.com/evmbench/evmbench.pdf.

648. Kailong Wang, Yuxi Ling, Yanjun Zhang, Zhou Yu, Haoyu Wang, Guangdong Bai, Beng Chin Ooi, and Jin Song Dong. “表征加密货币主题的恶意浏览器扩展”. 载于：Proceedings of the ACM on Measurement and Analysis of Computing Systems 6.3 (2022), 第 1-31 页.

649. Sally Junsong Wang, Kexin Pei, and Junfeng Yang. “SmartInv：智能合约不变量推断的多模态学习”. 载于：2024 IEEE Symposium on Security and Privacy (SP). Los Alamitos, CA, USA: IEEE Computer Society, 2024年5月, 第 2217-2235 页.

650. Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. “CNN 生成的图像目前惊人地容易被识别……”. 载于：2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020, 第 8692-8701 页.

651. Taotao Wang, Soung Chang Liew, and Shengli Zhang. “当区块链遇见 AI：通过机器学习实现最优挖矿策略”. 载于：International Journal of Intelligent Systems 36.5 (2021), 第 2183-2207 页.

652. Weiyang Wang, Manya Ghobadi, Kayvon Shakeri, Ying Zhang, and Naader Hasani. “Rail-only：一种用于训练万亿参数 LLM 的低成本高性能网络”. 载于：2024 IEEE Symposium on High-Performance Interconnects (HOTI). IEEE. 2024, 第 1-10 页.

653. Xintong Wang, Christopher Hoang, Yevgeniy Vorobeychik, and Michael P Wellman. “欺骗限价订单簿：一种战略性基于代理的分析”. 载于：Games 12.2 (2021), 第 46 页.

654. Xintong Wang and Michael P Wellman. “市场操纵：检测与规避的对抗性学习框架”. 载于：29th International Joint Conference on Artificial Intelligence. 2020.

655. Yanling Wang, Qian Wang, Lingchen Zhao, and Cong Wang. “深度学习中的差分隐私：隐私及其超越”. 载于：Future Generation Computer Systems 148 (2023), 第 408-424 页.

656. Ye Wang, Yan Chen, Haotian Wu, Liyi Zhou, Shuiguang Deng, and Roger Wattenhofer. “去中心化交易所中的循环套利”. 载于：Companion Proceedings of the Web Conference 2022. 2022, 第 12-19 页.

657. Yilei Wang, Chunmei Li, Yiting Zhang, Tao Li, Jianting Ning, Keke Gai, and Kim-Kwang Raymond Choo. “基于集成深度学习的物联网中类自私挖矿攻击检测方法”. 载于：IEEE Internet of Things Journal 11.11 (2024), 第 19564-19574 页.

658. Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. “MMLU-Pro：一个更具挑战性的多任务语言理解基准”. 载于：Advances in Neural Information Processing Systems 37 (2024), 第 95266-95290 页.

659. Zhaojie Wang, Qingzhe Lv, Zhaobo Lu, Yilei Wang, and Shengjie Yue. “ForkDec：精准检测自私挖矿攻击”. 载于：Security and Communication Networks 2021.1 (2021), 第 5959698 页.

660. Kang Wei, Jun Li, Ming Ding, Chuan Ma, Howard H Yang, Farhad Farokhi, Shi Jin, Tony QS Quek, and H Vincent Poor. “带差分隐私的联邦学习：算法与性能分析”. 载于：IEEE transactions on information forensics and security 15 (2020), 第 3454-3469 页.

661. Ben Weintraub, Christof Ferreira Torres, Cristina Nita-Rotaru, and Radu State. “瓶中的闪光：测量私有池中的最大可提取价值”. 载于：Proceedings of the 22nd ACM Internet Measurement Conference. 2022, 第 458-471 页.

662. Bingbing Wen, Jihan Yao, Shangbin Feng, Chenjun Xu, Yulia Tsvetkov, Bill Howe, and Lucy Lu Wang. “了解你的极限：大型语言模型中弃权的研究综述”. 载于：Transactions of the Association for Computational Linguistics 13 (2025年6月), 第 529-556 页.

663. Jie Wen, Zhixia Zhang, Yang Lan, Zhihua Cui, Jianghui Cai, and Wensheng Zhang. “联邦学习综述：挑战与应用”. 载于：International journal of machine learning and cybernetics 14.2 (2023), 第 513-535 页.

664. Rui Wen, Zheng Li, Michael Backes, and Yang Zhang. “对上下文学习的成员推理攻击”. 载于：Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security. 2024, 第 3481-3495 页.

665. 什么是中间件 - 定义与示例 / Microsoft Azure. [在线；访问于 2025-09-30]. URL: https://azure.microsoft.com/en-us/resources/cloud-computing-dictionary/what-is-middleware.

666. 什么是 x402？ / Solana 上 AI 代理的支付协议 / Solana. URL: https://solana.com/x402/what-is-x402 (访问于 2026年2月27日).

667. 为什么风投押注 AI × 加密作为下一个万亿美元市场 - PrimaFelicitas. [在线；访问于 2025-09-30]. 2025年6月. URL: https://www.primafelicitas.com/artificial-intelligence/why-vcs-are-betting-on-ai-x-crypto-as-the-next-trillion-dollar-market/.

668. David Gray Widder and Nathan Kim. “大云如何变得更大：审视 Google、Microsoft 和 Amazon 的投资”. https://ssrn.com/abstract=5377426. 2025.

669. Kyle Wiggers. “Elon Musk 同意我们已经耗尽了 AI 训练数据”. 2025年1月. URL: https://techcrunch.com/2025/01/08/elon-musk-agrees-that-weve-exhausted-ai-training-data/.

670. Annika Wilde, Tim Niklas Gruel, Claudio Soriente, and Ghassan Karame. “分叉之路：当 TEE 遇见共识”. 载于：arXiv preprint arXiv:2412.00706 (2024).

671. Witten. URL: https://witten.io/ (访问于 2026年2月6日).

672. David Wong and Luciano Floridi. “Meta 的监督委员会：综述与批判性评估”. 载于：Minds and Machines 33.2 (2023), 第 261-284 页.

673. Eric Wong and Zico Kolter. “通过凸外对抗多面体提供对抗性样本的可证明防御”. 载于：International conference on machine learning. PMLR. 2018, 第 5286-5295 页.

674. Gavin Wood. “以太坊：一个安全的去中心化通用交易账本”. 技术报告. 黄皮书，EIP-150 修订及后续更新. Ethereum Project, 2014.

675. Emma Woollacott. “VC 对 AI 的投资正在飙升——2025 年上半年的资金已超过去年全年，EY 称”. [在线；访问于 2025-09-30]. 2025年8月. URL: https://www.itpro.com/technology/artificial-intelligence/vc-investment-in-ai-is-skyrocketing-funding-in-the-first-half-of-2025-was-more-than-the-whole-of-last-year-says-ey.

676. Chenyuan Wu, Haoyun Qin, Mohammad Javad Amiri, Boon Thau Loo, Dahlia Malkhi, and Ryan Marcus. “BFTBrain：基于强化学习的自适应 BFT 共识”. 载于：22nd USENIX Symposium on Networked Systems Design and Implementation (NSDI 25). 2025, 第 1563-1583 页.

677. Cong Wu, Jing Chen, Jiahong Li, Jiahua Xu, Ju Jia, Yutao Hu, Yebo Feng, Yang Liu, and Yang Xiang. “盈利还是欺骗？通过图与对比学习减轻 DeFi 中的拉高出货”. 载于：IEEE Transactions on Information Forensics and Security (2025).

678. Jiajing Wu, Qi Yuan, Dan Lin, Wei You, Weili Chen, Chuan Chen, and Zibin Zheng. “谁是钓鱼者？通过网络嵌入的以太坊网络钓鱼诈骗检测”. 载于：IEEE Transactions on Systems, Man, and Cybernetics: Systems 52.2 (2020), 第 1156-1166 页.

679. Yin Wu, Xiaofei Xie, Chenyang Peng, Dijun Liu, Hao Wu, Ming Fan, Ting Liu, and Haijun Wang. “AdvSCanner：使用 LLM 和静态分析生成对抗性智能合约以利用重入漏洞”. 载于：Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering. ASE '24. Sacramento, CA, USA: Association for Computing Machinery, 2024, 第 1019-1031 页.

680. Pengcheng Xia, Haoyu Wang, Bingyu Gao, Weihang Su, Zhou Yu, Xiapu Luo, Chao Zhang, Xusheng Xiao, and Guoai Xu. “交易还是诡计？检测和表征 Uniswap 去中心化交易所上的诈骗代币”. 载于：Proceedings of the ACM on Measurement and Analysis of Computing Systems 5.3 (2021), 第 1-26 页.

681. Winnie Xiao, Cole Killian, Henry Sleight, Alan Chan, Nicholas Carlini, and Alwin Peng. “AI 代理发现 460 万美元的区块链智能合约漏洞”. [在线；访问于 2026-03-27]. 2025年12月. URL: https://red.anthropic.com/2025/smart-contracts/.

682. Yunming Xiao, Matteo Varvello, and Aleksandar Kuzmanovic. “将闲置带宽货币化：分布式 VPN 的案例”. 载于：Proceedings of the ACM on Measurement and Analysis of Computing Systems 6.2 (2022), 第 1-27 页.

683. Peichen Xie, Yanjie Gao, Yang Wang, and Jilong Xue. “揭示软件/硬件实现中的浮点累积顺序”. 载于：2025 USENIX Annual Technical Conference (USENIX ATC 25). 2025, 第 1425-1440 页.

684. Huajian Xin, Daya Guo, Zhihong Shao, Zhizhou Ren, Qihao Zhu, Bo Liu, Chong Ruan, Wenda Li, and Xiaodan Liang. “DeepSeek-Prover：通过大规模合成数据推进 LLM 中的定理证明”. 载于：arXiv preprint arXiv:2405.14333 (2024).

685. Guangquan Xu, Bingjiang Guo, Chunhua Su, Xi Zheng, Kaitai Liang, Duncan S Wong, and Hao Wang. “我被日食了吗？一种用于以太坊的智能日食攻击检测器”. 载于：Computers & Security 88 (2020), 第 101604 页.

686. Xinyi Xu, Shuaiqi Wang, Chuan-Sheng Foo, Bryan K Low, and Giulia Fanti. “数据分布估值”. 载于：Advances in Neural Information Processing Systems 37 (2024), 第 2407-2448 页.

687. Guoyu Yang, Yilei Wang, Zhaojie Wang, Youliang Tian, Xiaomei Yu, and Shouzhe Li. “IPBSM：在存在智能和纯粹攻击者的情况下的最优贿赂自私挖矿”. 载于：International Journal of Intelligent Systems 35.11 (2020), 第 1735-1748 页.

688. Qinglin Yang, Yetong Zhao, Huawei Huang, Zehui Xiong, Jiawen Kang, and Zibin Zheng. “融合区块链与 AI 及元宇宙：综述”. 载于：IEEE Open Journal of the Computer Society 3 (2022), 第 122-136 页.

689. Shuo Yang, Xingwei Lin, Jiahci Chen, Qingyuan Zhong, Lei Xiao, Renke Huang, Yanlin Wang, and Zibin Zheng. “Hyperion：使用 LLM 和数据流引导符号执行揭示 DApp 不一致性”. 载于：Proceedings of the IEEE/ACM 47th International Conference on Software Engineering. IEEE Press, 2025, 第 2125-2137 页.

690. Andrew C Yao. “安全计算的协议”. 载于：23rd annual symposium on foundations of computer science (sfcs 1982). IEEE. 1982, 第 160-164 页.

691. Jianzhu Yao, Hongxu Su, Taobo Liao, Zerui Cheng, Huan Zhang, Xuechao Wang, and Pramod Viswanath. “TAO：浮点神经网络的容错感知乐观验证”. 载于：Proceedings of the 21st European Conference on Computer Systems. 2026, 第 1515-1532 页.

692. Jiayuan Ye, Adayaa Maddi, Sasi Kumar Murakonda, Vincent Bindschaedler, and Reza Shokri. “针对机器学习模型的增强成员推理攻击”. 载于：Proceedings of the 2022 ACM SIGSAC conference on computer and communications security. 2022, 第 3093-3106 页.

693. Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. “机器学习中的隐私风险：分析与过拟合的联系”. 载于：2018 IEEE 31st computer security foundations symposium (CSF). IEEE. 2018, 第 268-282 页.

694. Maofan Yin, Dahlia Malkhi, Michael K Reiter, Guy Golan Gueta, and Ittai Abraham. “HotStuff：具有线性和响应性的 BFT 共识”. 载于：Proceedings of the 2019 ACM symposium on principles of distributed computing. 2019, 第 347-356 页.

695. Ted Young and Austin Parker. “学习 OpenTelemetry”. "O'Reilly Media, Inc.", 2024.

696. Jay Yu and Ryan Barney. “HTTP 402 的现代改造”. https://panteracapital.com/http-402s-modern-makeover/. 2025年9月.

697. Jay Yu, Amy Zhao, and Danning Sui. “纸面代理，纸面收益：DeFi 投资代理的实证分析”. 载于：arXiv preprint arXiv:2605.29174 (2026).

698. Lei Yu, Fengjun Zhang, Jiajia Ma, Li Yang, Yuanzhe Yang, and Wei Jia. “谁是洗钱者？基于互学习图神经网络的区块链洗钱检测”. 载于：2023 International Joint Conference on Neural Networks (IJCNN). IEEE. 2023, 第 1-8 页.

699. Binhang Yuan, Yongjun He, Davis Jared Quincy, Tianyi Zhang, Tri Dao, Beidi Chen, Percy Liang Liang, Christopher Re, and Ce Zhang. “异构环境中基础模型的去中心化训练”. 载于：NeurIPS (2022).

700. Zhengqing Yuan, Weixiang Sun, Yixin Liu, Huichi Zhou, Rong Zhou, Yiyang Li, Zheyuan Zhang, Wei Song, Yue Huang, Haolong Jia, et al. “EfficientLLM：大型语言模型的效率”. 载于：arXiv preprint arXiv:2505.13840 (2025).

701. Jusik Yun, Yunyeong Goh, and Jong-Moon Chung. “基于 DQN 的安全分片区块链系统优化框架”. 载于：IEEE Internet of Things Journal 8.2 (2020), 第 708-722 页.

702. Vlad Zamfir. “反对链上治理”. https://medium.com/@Vlad_Zamfir/against-on-chain-governance-a4ceacd040ca. 2017年12月1日.

703. Michael Zargham, Krzysztof Paruch, and Jamsheed Shorish. “作为估计器的经济博弈”. 载于：Mathematical Research for Blockchain Economy. Springer, 2020, 第 145-169 页.

704. Gheyath Mustafa Zebari and Nasser Al Musalhi. “整合 AI 与区块链安全的全面综述：创新、挑战与未来方向”. 载于：Security and Privacy 8.5 (2025), e70094.

705. Zhonghao Zhai, Subin Shen, and Yanqin Mao. “一种用于联盟区块链参数配置与调整的可解释深度强化学习算法”. 载于：Engineering Applications of Artificial Intelligence 129 (2024), 第 107606 页.

706. Andy K. Zhang, Joey Ji, Celeste Menders, Riya Dulepet, Thomas Qin, Ron Y. Wang, Junrong Wu, Kyleen Liao, Jiliang Li, Jinghan Hu, Sara Hong, Nardos Demilew, Shivatmica Murgai, Jason Tran, Nishka Kacheria, Ethan Ho, Denis Liu, Lauren McLane, Olivia Bruvik, Dai-Rong Han, Seungwoo Kim, Akhil Vyas, Cuiyuanxiu Chen, Ryan Li, Weiran Xu, Jonathan Z. Ye, Prerit Choudhary, Siddharth M. Bhatia, Vikram Sivashankar, Yuxuan Bao, Dawn Song, Dan Boneh, Daniel E. Ho, and Percy Liang. “BountyBench：AI 代理攻击者和防御者对现实世界网络安全系统的美元影响”. 2025. arXiv: 2505.15216 [cs.CR]. URL: https://arxiv.org/abs/2505.15216.

707. Brian Zhang and Zhuo Zhang. “通过 LLM 和基于规则的推理检测具有重大货币后果的漏洞”. 载于：Advances in Neural Information Processing Systems 37 (2024), 第 133999-134023 页.

708. Fan Zhang, Ethan Cecchetti, Kyle Croman, Ari Juels, and Elaine Shi. “Town Crier：智能合约的认证数据馈送”. 载于：Proceedings of the 2016 ACM SIGSAC conference on computer and communications security. 2016, 第 270-282 页.

709. Fan Zhang, Deepak Maram, Harjasleen Malvai, Steven Goldfeder, and Ari Juels. “DECO：使用去中心化 TLS 预言机解放网络数据”. 载于：Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security. 2020, 第 1919-1938 页.

710. Hanlin Zhang, Benjamin L Edelman, Danilo Francati, Daniele Venturi, Giuseppe Ateniese, and Boaz Barak. “沙中水印：生成模型强水印的不可能性”. 载于：arXiv preprint arXiv:2311.04378 (2023).

711. Jiaheng Zhang, Zhiyong Fang, Yupeng Zhang, and Dawn Song. “决策树预测和准确性的零知识证明”. 载于：Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security. 2020, 第 2039-2053 页.

712. Jialong Zhang, Zhongshu Gu, Jiyong Jang, Hui Wu, Marc Ph Stoecklin, Heqing Huang, and Ian Molloy. “用数字水印保护深度神经网络的知识产权”. 载于：Proceedings of the 2018 on Asia conference on computer and communications security. 2018, 第 159-172 页.

713. Jianting Zhang, Zicong Hong, Xiaoyu Qiu, Yufeng Zhan, Song Guo, and Wuhui Chen. “SkyChain：一种深度强化学习赋能的动态区块链分片系统”. 载于：Proceedings of the 49th International Conference on Parallel Processing. 2020, 第 1-11 页.

714. Lingzhe Zhang, Tong Jia, Mengxi Jia, Yifan Wu, Aiwei Liu, Yong Yang, Zhonghai Wu, Xuming Hu, Philip Yu, and Ying Li. “大型语言模型时代的 AIOps 综述”. 载于：ACM Computing Surveys 58.2 (2025), 第 1-35 页.

715. Lyuye Zhang, Kaixuan Li, Kairan Sun, Daoyuan Wu, Ye Liu, Haoye Tian, and Yang Liu. “AC-Fix：用挖掘的通用 RBAC 实践引导 LLM，进行智能合约访问控制漏洞的上下文感知修复”. 载于：IEEE Transactions on Software Engineering 51.09 (2025年9月), 第 2512-2532 页.

716. Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. “TinyLLaMA：一个开源的小型语言模型”. 载于：arXiv preprint arXiv:2401.02385 (2024).

717. Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. “AI 海洋中的海妖之歌：大型语言模型中幻觉的综述”. 载于：Computational Linguistics 51.4 (2025), 第 1373-1418 页.

718. Zhuo Zhang, Brian Zhang, Wen Xu, and Zhiqiang Lin. “剖析智能合约中的可利用漏洞”. 载于：2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE). IEEE. 2023, 第 615-627 页.

719. Zixu Zhang, Guangsheng Yu, Caijun Sun, Xu Wang, Ying Wang, Ming Zhang, Wei Ni, Ren Ping Liu, Andrew Reeves, and Nektarios Georgalas. “TbDd：一种新的基于信任、DRL驱动的物联网区块链分片框架”. 载于：Comput. Networks 244 (2024), 第 110343 页.

720. Zihao Zhao, Yuzhu Mao, Yang Liu, Linqi Song, Ye Ouyang, Xinlei Chen, and Wenbo Ding. “迈向联邦学习中的高效通信：当代综述”. 载于：Journal of the Franklin Institute 360.12 (2023), 第 8669-8703 页.

721. Haibin Zheng, Minying Ma, Haonan Ma, Jinyin Chen, Haiyang Xiong, and Zhijun Yang. “Tegdetector：一个了解交易行为演变的网络钓鱼检测器”. 载于：IEEE Transactions on Computational Social Systems 11.3 (2023), 第 3988-4000 页.

722. Shuran Zheng, Xuan Qi, Rui Ray Chen, Yongchan Kwon, and James Zou. “通过逐点互信息的正确数据集估值”. 载于：arXiv preprint arXiv:2405.18253 (2024).

723. Zihan Zheng, Peichen Xie, Xian Zhang, Shuo Chen, Yang Chen, Xiaobing Guo, Guangzhong Sun, Guangyu Sun, and Lidong Zhou. “Agatha：DNN 计算的智能合约”. 载于：arXiv preprint arXiv:2105.04919 (2021).

724. Jiajun Zhou, Chenkai Hu, Jianlei Chi, Jiajing Wu, Meng Shen, and Qi Xuan. “以太坊交互图上基于行为的账户去匿名化”. 载于：IEEE Transactions on Information Forensics and Security 17 (2022), 第 3433-3448 页.

725. Liyi Zhou, Kaihua Qin, Antoine Cully, Benjamin Livshits, and Arthur Gervais. “关于 DeFi 协议中产生利润交易的即时发现”. 载于：2021 IEEE Symposium on Security and Privacy (SP). IEEE. 2021, 第 919-936 页.

726. Liyi Zhou, Kaihua Qin, Christof Ferreira Torres, Duc V Le, and Arthur Gervais. “去中心化链上交易所的高频交易”. 载于：2021 IEEE symposium on security and privacy (SP). IEEE. 2021, 第 428-445 页.

727. Liyi Zhou, Xihan Xiong, Jens Ernstberger, Stefanos Chaliasos, Zhipeng Wang, Ye Wang, Kaihua Qin, Roger Wattenhofer, Dawn Song, and Arthur Gervais. “SoK：去中心化金融（DeFi）攻击”. 载于：2023 IEEE Symposium on Security and Privacy (SP). Los Alamitos, CA, USA: IEEE Computer Society, 2023, 第 2444-2461 页.

728. Yi Zhou, Deepak Kumar, Surya Bakshi, Joshua Mason, Andrew Miller, and Michael Bailey. “Eras：逆向工程以太坊不透明智能合约”. 载于：27th USENIX Security Symposium (USENIX Security 18). Baltimore, MD: USENIX Association, 2018年8月, 第 1371-1385 页.

729. Jianwei Zhu, Hang Yin, Peng Deng, Aline Almeida, and Shunfan Zhou. “NVIDIA Hopper GPU 上的机密计算：性能基准研究”. 2024. arXiv: 2409.03992 [cs.DC].

730. zkPass：zkTLS 预言机协议. URL: https://docs.zkpass.org/overview/technical-overview (访问于 2026年2月6日).

731. Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. “对齐语言模型的通用和可迁移对抗性攻击”. 载于：arXiv preprint arXiv:2307.15043 (2023).

732. Yanjun Zuo. “探索协同效应：AI 增强区块链，区块链赋能 AI，以及它们在物联网及超越中的融合”. 载于：IEEE Internet of Things Journal (2024).

733. Yiping Zuo, Jiajia Guo, Ning Gao, Yongxu Zhu, Shi Jin, and Xiao Li. “面向 6G 无线通信的区块链与人工智能综述”. 载于：IEEE Communications Surveys & Tutorials 25.4 (2023), 第 2494-2528 页.

734. Roi Bar Zur, Ittay Eyal, and Aviv Tamar. “用于区块链中自私挖矿的高效 MDP 分析”. 载于：Proceedings of the 2nd ACM Conference on Advances in Financial Technologies. 2020, 第 113-131 页.