Mira：用于无信任 AI 输出验证的去中心化网络

Ninad Naik
ninad@arohalabs.com

Sidhartha Doddipalli
sid@arohalabs.com

Karan Sirdesai
karan@arohalabs.com

摘要

虽然AI擅长生成看似合理的输出，但由于基于神经网络的技术（如大型语言模型和扩散模型）的概率性质，它经常产生不正确的信息。本文介绍了一个通过去中心化共识来验证AI生成输出的网络。该网络将AI输出转化为独立可验证的声明，使多个AI模型能够共同确定每个声明的有效性。执行这些基于推理验证的节点操作员通过混合的工作量证明/权益证明机制获得经济激励，以进行诚实的验证。除了验证之外，我们的愿景还包括一个提供无错误输出的合成基础模型。这一基础设施是实现AI系统在没有人类监督的情况下运行的关键步骤——这是AI在社会中发挥其变革潜力的必要条件。

1. 简介

人工智能有望成为一种变革性的力量，与印刷术、蒸汽机、电力和互联网并列——这些技术从根本上重塑了人类文明。然而，今天的AI面临着根本性的挑战，阻碍了其达到这种革命性的潜力。虽然AI擅长生成创造性和看似合理的输出，但它难以可靠地提供无错误的输出。这些限制使得AI主要局限于需要人类监督的任务或低风险应用，如聊天机器人，远远未能实现AI在高风险任务中自主实时处理的潜力。

关键障碍是人工智能的可靠性。人工智能系统主要存在两种类型的错误：幻觉和偏见，它们共同决定了模型的整体错误率。目前的错误率对于在相关场景中的自主操作来说仍然太高，这造成了人工智能的理论能力与实际应用之间的根本差距。

随着 AI 模型随着训练数据和参数化的增加而不断发展，由于训练困境，这些可靠性挑战仍然存在。这种困境反映了经典的精度-准确度权衡：幻觉代表精度误差（模型输出的一致性），而偏差则表现为准确度误差（与基本事实的系统偏差）。当模型构建者整理训练数据以提高精度并减少幻觉时，他们不可避免地会通过选择标准引入准确度误差（偏差）。相反，在各种可能相互冲突的数据源上进行训练以提高准确度（减少偏差）会导致精度下降（幻觉增加），因为模型在其更广泛的知识分布中产生不一致的输出。

据观察，经过微调的模型在狭窄领域内可实现更高的可靠性；然而，研究表明，经过微调的模型难以可靠地整合新知识，引入新信息的训练示例的学习效率远低于与模型现有知识库一致的训练示例。经过微调的模型还难以处理训练领域之外的极端情况和意外情况，这使得它们不适合必须处理各种现实情况的自主系统。

这一基本约束为人工智能模型性能建立了一个不变的界限：无论规模或架构如何，都存在一个任何单一模型都无法克服的最小错误率。

虽然没有一个模型能够同时减少幻觉和偏见，但集体智慧提供了一条前进的道路。通过共识机制，多个模型协同工作可以实现单个模型无法实现的目标——通过集体验证过滤掉幻觉，同时通过不同的视角平衡个体偏见。这一见解表明，可靠的人工智能不仅需要更好的模型，还需要更好的方法来结合它们的优势并减轻它们各自的弱点。

然而，简单地在集中控制下组装一组模型并不能完全解决可靠性挑战。模型选择本身会引入系统性错误——集中管理者的选择不可避免地反映了特定的观点和局限性。此外，许多事实本质上是与背景相关的，因文化、地区和领域而异。真正的可靠性不仅需要多个模型，还需要真正多样化的观点，而这些观点只能通过分散参与才能产生。

我们所需要的是一个基于去中心化共识而非中心化权威的人工智能验证系统，该系统允许验证任何人工智能生成的输出，而无需依赖单个可信实体。一个使操纵共识在计算和经济上都不切实际的系统将保护用戶免受不可靠输出的影响，同时激励开发专门领域模型和代表不同观点的模型。

在本文中，我们提出了一种解决人工智能可靠性问题的方法，即使用基于区块链的多样化人工智能验证者网络来生成人工智能输出有效性的计算证明。该网络的安全框架通过结合经济激励、技术保障和博弈论原理来确保可靠的验证。这种方法通过其分布式验证机制增强了人工智能的可靠性，从而降低了偏差和幻觉率。

2.网络架构

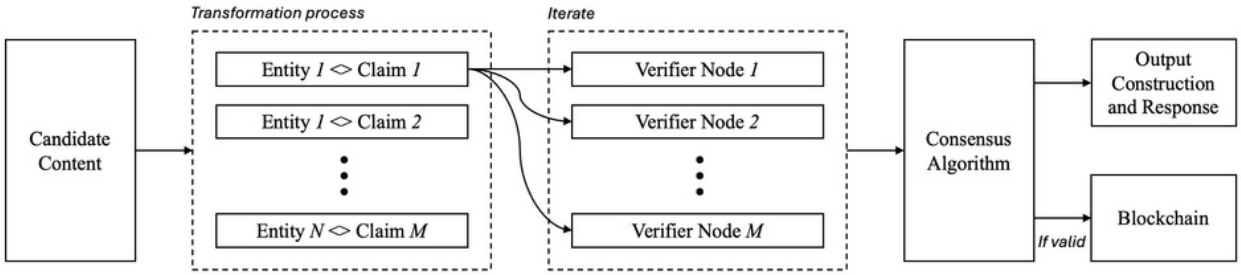
Mira 网络通过一种新颖的协议实现了对 AI 生成的输出进行无需信任的验证，该协议将复杂的内容转换为可独立验证的声明。这些声明通过各种 AI 模型之间的分布式共识进行验证，节点运营商在经济上受到激励以进行诚实的验证。这种去中心化的方法确保没有任何参与者实体可以操纵验证结果，同时实现对 AI 生成的输出的验证。

该网络的架构通过内容转换、分布式验证和共识机制的全新组合实现了可靠的验证。该系统可处理从简单的事实陈述到复杂的内容（包括技术文档、创意写作、多媒体内容和代码）的所有内容。

考虑一个复合语句：“地球围绕太阳旋转，月亮围绕地球旋转。”虽然使用多个模型验证这个简单语句似乎很简单，但将验证扩展到复杂内容（整个段落、法律摘要或代码）则带来了根本性的挑战。将候选内容按原样传递给验证器模型会失败，因为每个验证器模型可能会解释和验证内容的不同方面。 系统验证需要以确保每个验证模型以相同的背景和视角解决完全相同的问题的方式标准化人工智能生成的输出。

我們所提出的转换方法解决了这一根本挑战。对于示例语句，系统将候选内容分解为不同的可验证声明：(1)“地球围绕太阳旋转”和（2）“月球围绕地球旋转”。通过集成验证，它确定每个声明的有效性并颁发证明验证结果的加密证书。此过程普遍适用于人工智能和人工智能。生成和人类生成的内容，使得系统与来源无关，同时保持严格的验证标准。

网络负责处理候选内容的转换、声明分发、共识管理和网络编排。节点基础设施包括运行验证器模型、处理声明、 并提交验证结果。节点自主运行，但必须保持特定的性能和可靠性标准才能参与网络。



验证工作流按系统进行。客戶提交候选内容并指定验证要求，如领域（例如医学、法律等特定知识领域）和共识阈值（例如绝对共识、N/M同意等）。网络将这些内容转换为可验证声明，同时保留逻辑关系，将这些声明分发给节点进行验证，并汇总结果以达成共识。然后，网络生成记录验证结果的加密证书，包括哪些模型对每个声明达成了共识，并将结果和证书返回给客戶。

3.经济安全模式

该网络的经济安全模型结合了工作量证明 (PoW) 和权益证明 (PoS) 机制，为诚实验证创造可持续的激励，同时获取和分配实际经济价值。这种混合方法解决了验证 AI 输出的独特挑战。

网络通过验证降低人工智能错误率，从而产生切实的经济价值。客戶支付网络费用以获得经过验证的输出，网络通过验证奖励将这些费用分配给参与者（节点运营商和数据提供者）。

与传统区块链网络不同，PoW 涉及解决随机成功概率极低的加密难题，而 Mira 网络将验证转变为标准化的多项选择题。 虽然这种标准化可以实现跨节点的系统验证，但它也带来了一个根本性的挑战：可能响应的概率空间受到限制。例如，验证任务在二元选择中随机成功的几率为 50%，在四元选择中随机成功的几率为 25%。这使得随机猜测成为一种潜在的有吸引力的策略，无需计算成本即可获得高回报。

为了缓解这种情况，节点必须质押价值才能参与验证。如果某个节点持续偏离共识，或表现出随机响应而非实际推理的模式，则其质押金额可能会被削减。这种经济惩罚确保试图通过随机响应来玩弄系统在经济上是不合理的。

因此，网络的经济安全模型基于三个基本原则。首先，节点运营商在经济激励下会采取理性行为，因为他们的质押价值会因惩罚而面临风险。其次，只要诚实的运营商控制大部分质押价值，网络安全就能得到维护，这使得操纵尝试的成本高得令人望而却步。第三，随着网络规模的扩大，验证者模型的自然多样性会减少统计偏差，因为不同的模型会带来不同的训练方法和知识库。这些原则相辅相成：经济激励吸引了不同的参与者，他们不同的观点加强了安全性，进而支持了经济模型。

表 1 说明了在给定答案选项数量的情况下成功猜出正确答案的概率

验证次数	2个答案选项	3个答案选项	4个答案选项	5个答案选项	6个答案选项
1	50.0000%	25.0000%	16.6667%	12.5000%	10.0000%
2	25.2500%	6.2500%	2.7778%	1.5625%	1.0000%
3	12.5000%	1.5625%	0.4630%	0.1953%	0.1000%
4	6.2500%	0.3906%	0.0772%	0.0244%	0.0100%
5	3.1250%	0.0977%	0.0129%	0.0031%	0.0010%
6	1.5625%	0.0244%	0.0021%	0.0004%	0.0001%
7	0.7813%	0.0061%	0.0004%	0.0000%	0.0000%
8	0.3906%	0.0015%	0.0001%	0.0000%	0.0000%
9	0.1953%	0.0004%	0.0000%	0.0000%	0.0000%
10	0.0977%	0.0001%	0.0000%	0.0000%	0.0000%

在网络的初始阶段，节点运营商会经过仔细审查，以确保网络完整性。在第二阶段，网络将开始通过设计重复来实现去中心化，其中同一验证器模型的多个实例将处理每个验证请求。这种重复虽然增加了验证成本，但可以可靠地识别恶意或懒惰的运营商。随着网络逐渐成熟并进入稳定状态，验证请求将在节点之间随机分配，从而使勾结变得越来越困难且成本越来越高。

网络的分片机制提供了另一层安全性：通过研究节点之间的响应模式和相似性指标，系统可以识别潜在的勾结。恶意行为者需要控制网络质押价值的很大一部分才能影响结果，此时他们的经济动机与诚实操作相一致。

节点运营商可能会尝试通过各种策略来操纵网络，例如维护常见验证结果的数据库以简化验证过程。短期内，验证请求的多样性和独特性使得这种缓存策略无效。从规模上看，大量经过验证的事实的存在为利用网络验证功能的衍生协议提供了机会。

节点运营商通过以最低成本获得正确答案来取得成功。当专用模型在特定验证任务上实现与大型模型相当的性能时，这就创造了合法的优化机会。这些机会推动了针对特定领域优化的高效、任务专用模型的开发，通过更高的准确率、更低的成本和更低的延迟使整个生态系统受益。

网络的经济模型通过多个强化周期强化了这些积极的动态。随着网络使用量的增长，费用的增加可以带来更好的验证奖励，吸引更多的节点运营商，并推动准确性、成本和延迟的改善。这种增长有机地增强了网络安全性：权益要求随着网络价值的增加而增加，模型多样性通过专业化和视角的根本差异而扩大，而累积的验证历史使异常检测越来越复杂。这些复合效应创造了一个强大的博弈论均衡，其中诚实验证和持续创新成为主导策略，同时使恶意操纵在经济上不合理，在技术上不可行。

4. 隐私

基于上述安全基础，网络设计将隐私保护作为核心架构原则。隐私保护始于网络内容转换的基本方法：将复杂内容分解为实体声明对，然后随机分片到各个节点。这确保没有任何单个节点运营商能够重建完整的候选内容，从而保护客戶隐私，同时保持验证完整性。

隐私模型通过多层保护得到加强。节点的验证响应在达成共识之前保持私密，从而防止验证过程中的信息泄露。达成共识后，网络将生成仅包含必要验证详细信息的证书，通过数据最小化进一步保护隐私。

在网络演进的早期，转换软件的中心化特性形成了天然的隐私边界。 该网络的路线图包括该组件的逐步去中心化，同时通过加密协议和安全计算技术保持强大的隐私保障。

5. 网络演进

网络的发展遵循了自然发展过程，朝着全面的 AI 验证和生成平台发展，这将从根本上重塑 AI 系统的运行方式。我们的愿景不仅限于简单的验证，而是创建一类新的基础模型，其中验证是生成的内在因素 — 这是 AI 实现其变革潜力所需的根本性突破。

该网络最初专注于事实准确性至关重要且偏见风险最小的领域，例如医疗保健、法律和金融，随后将逐步扩展以处理越来越复杂的内容类型，包括代码、结构化数据和多媒体内容。该网络将通过数据可用性层和互补技术将验证功能扩展到私有数据和其他上下文，从而实现安全高效的验证，而不会使基础网络臃肿。每次扩展不仅代表着更广泛的覆盖范围，而且是朝着更复杂、更可靠的人工智能系统迈出的一步。

验证能力从简单的有效性检查发展到全面重建无效内容，最终直接生成经过验证的输出。这一进步消除了传统的生成速度和准确性之间的权衡，在保持严格验证标准的同时，接近实时性能。

除了直接验证之外，在区块链上积累经济上安全的事实还可以实现强大的衍生应用。这个经过验证的知识库可以支持确定性事实核查系统和继承网络安全保障的预言机服务。更根本的是，通过为事实验证创造经济激励，网络建立了一种将原始数据转换为有价值支持的事实的新模型——这是可靠人工智能系统的关键基石。

通过技术能力和经济激励的不断发展，网络将使新一代人工智能应用能够以前所未有的可靠性运行。这不仅代表了人工智能系统的渐进式改进，还建立了一种新范式，即无需人工监督的无错误运行使人工智能最终能够自主运行。

6. 结论

当今的人工智能系统面临着一个根本性的挑战：虽然它们擅长生成富有创意且合理的输出，但它们无法可靠地提供无错误的输出，需要人工监督。我们的去中心化验证网络 通过加密经济激励机制实现的内容转换和分布式共识的新组合解决了这一挑战，使操纵在技术和经济上都不切实际。与解决任意难题的传统 PoW 不同，Mira 网络需要由质押价值支持的有意义的推理计算，以确保诚实操作。

除了验证之外，我们的愿景是建立一个将验证直接集成到生成过程中的综合基础模型。这种简化的方法消除了生成和验证之间的区别，从而提供无错误的输出。通过在激励运营商的分散网络中分发验证，我们创建了本质上能够抵御集中控制的基础设施。这代表着一项根本性的进步：通过使人工智能系统能够在没有人工监督的情况下运行，我们为实际的人工智能奠定了基础——这是释放人工智能在整个社会变革潜力的关键一步。

参考

[1] E. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?”, https://dl.acm.org/doi/pdf/10.1145/3442188.3445922, 2021

[2] M. Chelli, J. Descamps, V. Lavoué, C. Trojani, M. Azar, M. Deckert, J. Raynier, G. Clowez, P. Boileau, C. Ruetsch-Chelli, “Hallucination Rates and Reference Accuracy of ChatGPT and Bard for Systematic Reviews: Comparative Analysis”, https://www.jmir.org/2024/1/e53164/?t, 2024

[3] Z. Gekhman, G. Yona, R. Aharoni, M. Eyal, A. Feder, R. Reichart, J. Herzig, “Does Fine-Tuning LLMs on New Knowledge Encourage Hallucinations?”, https://arxiv.org/pdf/2405.05904, 2024

[4] N. Naik, “Probabilistic Consensus through Ensemble Validation: A Framework for LLM Reliability”, https://mira.network/research/ensemble-validation.pdf, 2024

[5] S. King, S. Nadal, “PPCoin: Peer-to-Peer Crypto-Currency with Proof-of-Stake”, https://www.peercoin.net/read/papers/peercoin-paper.pdf, 2012

[6] S. Nakamoto, “Bitcoin: A Peer-to-Peer Electronic Cash System”, https://bitcoin.org/bitcoin.pdf, 2009

[7] X. Zuwei, S. Jain, M. Kankanhalli, “Hallucination is Inevitable: An Innate Limitation of Large Language Models”, https://arxiv.org/abs/2401.11817, 2024