Mira: A Decentralized Network for Trustless AI Output Verification

Ninad Naik
ninad@arohalabs.com

Sidhartha Doddipalli
sid@arohalabs.com

Karan Sirdesai
karan@arohalabs.com

Abstract. While AI excels at generating plausible outputs, it frequently produces incorrect information due to the probabilistic nature of neural network-based technologies such as large language and diffusion models. This paper introduces a network that verifies AI-generated output through decentralized consensus. The network transforms AI outputs into independently verifiable claims, enabling multiple AI models to collectively determine each claim's validity. Node operators performing these inference-based verifications are economically incentivized through a hybrid Proof-of-Work/Proof-of-Stake mechanism to conduct honest verification. Beyond verification, our vision extends to a synthetic foundation model delivering error-free output. This infrastructure represents a crucial step toward enabling AI systems to operate without human oversight—a necessary condition for AI to achieve its transformative potential across society.

1. Introduction

Artificial Intelligence stands poised to become a transformative force on par with the printing press, steam engine, electricity, and internet—technologies that fundamentally reshaped human civilization. However, AI today faces fundamental challenges that prevent it from reaching this revolutionary potential. While AI excels at generating creative and plausible outputs, it struggles to reliably provide error-free outputs. These limitations constrain AI primarily to human-supervised tasks or lower-consequence applications like chatbots, falling far short of AI's potential to handle high-stakes tasks autonomously and in real time.

The key barrier is AI reliability. AI systems suffer from two primary types of errors: hallucinations and bias, which together determine a model's overall error rate. Current error rates remain too high for autonomous operation in consequential scenarios, creating a fundamental gap between AI's theoretical capabilities and practical applications.

As AI models continue to evolve with increased training data and parametrization, these reliability challenges persist due to the ***training dilemma***. This dilemma mirrors the classical precision-accuracy trade-off: hallucinations represent precision errors (the consistency of model outputs), while bias manifests as accuracy errors (systematic deviation from ground truth). When model builders curate training data to increase precision and reduce hallucinations, they inevitably introduce accuracy errors (bias) through their selection criteria. Conversely, training on diverse, potentially conflicting data sources to improve accuracy (reduce bias) leads to decreased precision (increased hallucinations) as the model produces inconsistent outputs across its broader knowledge distribution.

Fine-tuned models have been observed to achieve higher reliability within narrow domains; however, research has shown that fine-tuned models struggle to reliably incorporate new knowledge, with training examples that introduce novel information being learned substantially less effectively than those that align with the model's existing knowledge base. Fine-tuned models also struggle with edge cases and unexpected scenarios outside their training domain, making them unsuitable for autonomous systems that must handle diverse, real-world situations.

This fundamental constraint establishes an immutable boundary in AI model performance: there exists a minimum error rate that cannot be overcome by any single model, regardless of scale or architecture.

While no single model can minimize both hallucinations and bias, collective wisdom offers a path forward. Through consensus mechanisms, multiple models working together can achieve what individual models cannot—filtering out hallucinations through collective verification while balancing individual biases through diverse perspectives. This insight suggests that reliable AI requires not just better models, but better ways of combining their strengths and mitigating their individual weaknesses.

However, simply assembling an ensemble of models under centralized control cannot fully solve the reliability challenge. Model selection itself introduces systematic errors—a centralized curator's choices inevitably reflect particular perspectives and limitations. Moreover, many truths are inherently contextual, varying across cultures, regions, and domains. True reliability requires not just multiple models, but genuinely diverse perspectives that can only emerge from decentralized participation.

What is needed is an AI verification system based on decentralized consensus instead of centralized authority, allowing any AI generated output to be verified without relying on a single trusted entity. A system that makes it computationally and economically impractical to manipulate consensus would protect users from unreliable outputs, while incentivizing the development of both specialized domain models and models representing diverse perspectives.

In this paper, we propose a solution to the AI reliability problem using a blockchain-based network of diverse AI verifiers to generate computational proof of the validity of AI outputs. The network's security framework ensures reliable verification through a combination of economic incentives, technical safeguards, and game-theoretic principles. This approach enhances AI reliability through its distributed verification mechanism, which reduces both bias and hallucination rates.

2. Network Architecture

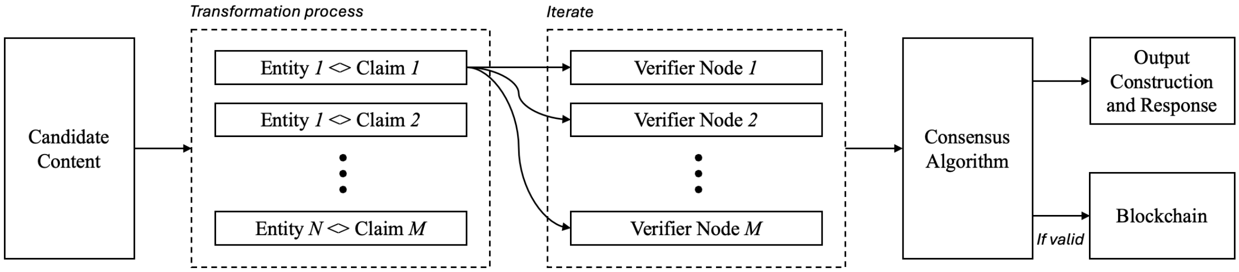
The Mira network enables trustless verification of AI-generated output through a novel protocol that transforms complex content into independently verifiable claims. These claims are verified through distributed consensus among diverse AI models, with node operators economically incentivized to perform honest verification. This decentralized approach ensures no actor entity can manipulate verification outcomes while enabling verification of AI-generated output.

The network's architecture enables reliable verification through a novel combination of content transformation, distributed verification, and consensus mechanisms. The system processes everything from simple factual statements to complex content including technical documentation, creative writing, multimedia content, and code.

Consider a compound statement: "The Earth revolves around the Sun and the Moon revolves around the Earth." While verifying this simple statement with multiple models might seem straightforward, scaling verification to complex content—entire passages, legal briefs, or code—presents fundamental challenges. Passing the candidate content as-is to verifier models fails because each verifier model may interpret and verify different aspects of the content. Systematic verification requires standardizing AI generated output in a manner ensuring each verifier model addresses the exact same problem with identical context and perspective.

The proposed transformation approach solves this fundamental challenge. For the example statement, the system breaks down the candidate content into distinct verifiable claims: (1) "The Earth revolves around the Sun" and (2) "The Moon revolves around the Earth." Through ensemble verification, it determines the validity of each claim and issues cryptographic certificates attesting to the verification outcome. This process applies universally to both AI-generated and human-generated content, making the system source-agnostic while maintaining rigorous verification standards.

The network handles the transformation of candidate content, claim distribution, consensus management, and network orchestration. The node infrastructure comprises independent operators running verifier models, processing claims, and submitting verification results. Nodes operate autonomously but must maintain specific performance and reliability standards to participate in the network.



The verification workflow proceeds systematically. Customers submit candidate content and specify verification requirements such as domain (e.g., specific knowledge areas like medical, legal, etc.) and consensus threshold (e.g. absolute consensus, N of M agreement, etc.). The network transforms this content into verifiable claims while preserving logical relationships, distributes these claims to nodes for verification, and aggregates the results to reach consensus. The network then generates a cryptographic certificate recording the verification outcome, including which models reached consensus for each claim, and returns both outcome and certificate to the customer.

3. Economic Security Model

The network's economic security model combines Proof-of-Work (PoW) and Proof-of-Stake (PoS) mechanisms to create sustainable incentives for honest verification while capturing and distributing real economic value. This hybrid approach addresses unique challenges in verifying AI outputs.

The network generates tangible economic value by reducing AI error rates through verification. Customers pay network fees to obtain verified output, and the network distributes these fees to participants—node operators and data providers—through verification rewards.

Unlike traditional blockchain networks where PoW involves solving cryptographic puzzles with infinitesimally low probability of random success, the Mira network transforms verification into standardized multiple-choice questions. While this standardization enables systematic verification across nodes, it also creates a fundamental challenge: the probability space of possible responses is constrained. For instance, a verification task will have a 50% chance of random success with binary choices, or 25% with four options. This makes random guessing a potentially attractive strategy, offering high reward for no computational cost.

To mitigate this, nodes must stake value to participate in verification. If a node consistently deviates from consensus or demonstrates patterns suggesting random responses rather than actual inference, their stake can be slashed. This economic penalty ensures that attempting to game the system through random responses becomes economically irrational.

The network's economic security model thus operates on three foundational principles. First, node operators behave rationally in response to economic incentives, as their staked value is at risk through slashing penalties. Second, network security is maintained as long as honest operators control the majority of staked value, making manipulation attempts prohibitively expensive. Third, as the network scales, the natural diversity of verifier models reduces statistical bias, as different models bring varied training approaches and knowledge bases. These principles reinforce each other: economic incentives attract diverse participants, whose varied perspectives strengthen security, which in turn supports the economic model.

Table 1 illustrates the probability of successfully guessing the correct answer given the number of answer choices

Verifications	2 Answer Options	4 Answer Options	6 Answer Options	8 Answer Options	10 Answer Options
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%

During the network's initial phase, node operators are carefully vetted to ensure network integrity. In the second phase, the network will begin to decentralize with designed duplication, where multiple instances of the same verifier model process each verification request. This duplication, while increasing verification costs, enables robust identification of malicious or lazy operators. As the network matures into its steady-state, verification requests are randomly sharded across nodes, making collusion increasingly difficult and expensive.

The network's sharding mechanism provides another layer of security: by studying response patterns and similarity metrics across nodes, the system can identify potential collusion. Malicious actors would need to control a significant portion of the network's staked value to influence outcomes, at which point their economic incentives align with honest operation.

Node operators might attempt to game the network through various strategies, such as maintaining databases of common verification results to shortcut the verification process. In the short term, the diversity and uniqueness of verification requests make such caching strategies ineffective. At scale, the existence of a large corpus of verified facts presents opportunities for derivative protocols that leverage the network's verification capabilities.

Node operators achieve success by reaching correct answers at the lowest cost. When specialized models achieve comparable performance to larger models on specific verification tasks, this creates legitimate optimization opportunities. These opportunities drive the development of efficient, task-specific models optimized for particular domains, benefiting the entire ecosystem through higher accuracy rates, lower costs, and reduced latency.

The network's economic model reinforces these positive dynamics through multiple reinforcing cycles. As network usage grows, increased fee generation enables better verification rewards, attracting more node operators and driving improvements in accuracy, cost, and latency. This growth strengthens network security organically: stake requirements rise with network value, model diversity expands through both specialization and fundamental differences in perspective, and accumulated verification history enables increasingly sophisticated anomaly detection. These compounding effects create a robust game-theoretic equilibrium where honest verification and continuous innovation emerge as dominant strategies, while making malicious manipulation both economically irrational and technically infeasible.

4. Privacy

Building on the above security foundations, the network's design prioritizes privacy preservation as a core architectural principle. Privacy protection begins with the network's fundamental approach to content transformation: complex content is broken down into entity-claim pairs which are then randomly sharded across nodes. This ensures no single node operator can reconstruct the complete candidate content, protecting customer privacy while maintaining verification integrity.

The privacy model strengthens through multiple layers of protection. Verification responses from nodes remain private until consensus is reached, preventing information leakage during the verification process. When consensus is achieved, the network generates certificates containing only the necessary verification details, further preserving privacy through data minimization.

Early in the network's evolution, the centralized nature of transformation software presents a natural privacy boundary. The network's roadmap includes progressive decentralization of this component while maintaining strong privacy guarantees through cryptographic protocols and secure computation techniques.

5. Network Evolution

The network's evolution follows a natural progression toward a comprehensive AI verification and generation platform that will fundamentally reshape how AI systems operate. Our vision extends beyond simple verification to the creation of a new class of foundation models where verification is intrinsic to generation—a fundamental breakthrough required for AI to achieve its transformative potential.

Initially focusing on domains where factual accuracy is critical and bias risks are minimal, such as healthcare, law, and finance, the network will progressively expand to handle increasingly complex content types including code, structured data, and multimedia content. The network will extend verification capabilities to private data and additional context through data availability layers and complementary technologies, enabling secure and efficient verification without bloating the base network. Each expansion represents not just broader coverage, but a step toward more sophisticated and reliable AI systems.

The verification capabilities advance from simple validity checks to comprehensive reconstruction of invalid content, ultimately culminating in direct generation of verified outputs. This progression eliminates the traditional trade-off between generation speed and accuracy, approaching real-time performance while maintaining rigorous verification standards.

Beyond direct verification, the accumulation of economically secured facts on the blockchain enables powerful derivative applications. This verified knowledge base can support deterministic fact-checking systems and oracle services that inherit the network's security guarantees. More fundamentally, by creating economic incentives for truth verification, the network establishes a new model for converting raw data into value-backed facts—a crucial building block for reliable AI systems.

Through continuous evolution of both technical capabilities and economic incentives, the network will enable a new generation of AI applications that operate with unprecedented reliability. This represents more than an incremental improvement in AI systems—it establishes a new paradigm where error-free operation without human oversight enables AI to finally operate autonomously.

6. Conclusion

AI systems today face a fundamental challenge: while they excel at generating creative and plausible outputs, they cannot reliably provide error-free outputs, requiring human oversight. Our decentralized verification network addresses this challenge through a novel combination of content transformation and distributed consensus enabled by crypto-economic incentives, making manipulation both technically and economically impractical. Unlike traditional PoW solving arbitrary puzzles, the Mira network requires meaningful inference computations backed by staked value to ensure honest operation.

Beyond verification, our vision is a synthetic foundation model that integrates verification directly into the generation process. This streamlined approach eliminates the distinction between generation and verification, delivering error-free outputs. By distributing verification across a decentralized network of incentivized operators, we create infrastructure inherently resistant to centralized control. This represents a fundamental advancement: by enabling AI systems to operate without human oversight, we establish the foundation for actual artificial intelligence—a crucial step toward unlocking AI's transformative potential across society.

References

[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