Hardware-Enabled Mechanisms for Verifying Responsible AI Development

Aidan O'Gara [*1] **Gabriel Kulp** [*2 3] **Will Hodgkins** [*4] **James Petrie** [5] **Vincent Immler** [3] **Aydin Aysu** [6] **Kanad Basu** [7] **Shivam Bhasin** [8] **Stjepan Picek** [9] **Ankur Srivastava** [10]

Abstract

This paper surveys the emerging field of hardware-enabled mechanisms (HEMs). We describe four kinds of HEMs: location verification for enforcing export controls, offline licensing to permit hardware usage, network verification regarding the configuration of AI clusters and real-time communications between AI chips, and workload verification of the details of AI training and inference workloads. We discuss open research questions regarding each mechanism and cross-cutting techical and governance challenges for the field. Overall, we aim to provide a research roadmap for the development of hardware-enabled mechanisms (HEMs) for verifying responsible AI development.

1. Introduction

AI offers enormous potential benefits but also poses significant risks, including misuse by malicious actors, international destabilization through AI weaponization, and even loss of human control over powerful AI systems.

Verifying the responsible use of AI hardware is a promising path to reducing AI risks, for several reasons. First, hardware is critical for AI development. Scaling laws demonstrate that exponential increases in computational resources consistently lead to improved AI performance (Kaplan et al., 2020). Motivated by scaling laws, the amount of compute used to train frontier AI systems has grown by roughly 4× annually over the last decade (Epoch AI, 2024b). Second, hardware is inherently physical and excludable, unlike data and algorithms which can be easily copied. Third, the AI hardware supply chain is highly concentrated—TSMC for chip fabrication, Nvidia for chip design, and ASML for manufacturing equipment—with facilities largely located in nations allied with the United States (Sastry et al., 2024). Fourth, high-end AI chips constitute less than 0.00025% of global semiconductor production (Heim & Pilz, 2024), enabling targeted regulation without broad economic disruption. Finally, hardware can be used to help verify claims about its own use—a critical insight which motivates research on hardware-enabled mechanisms (HEMs) for verifying responsible AI development.

This paper lays out a research agenda for the emerging field of HEMs. Based on O'Gara et al. (2025) and building on prior research including Shavit (2023), Aarne et al. (2024), Kulp et al. (2024), Brass & Aarne (2024), Petrie et al. (2025), Scher & Thiergart (2024), and others, this paper describes four promising kinds of HEMs, provides an appendix with open research questions about each one, and discusses broader technical and governance challenges to successfully reducing AI risks by verifying the responsible use of AI hardware with HEMs.

2. Mechanisms

This section discusses four major kinds of HEMs: location verification, offline licensing, network verification, and workload verification. For each, we describe the mechanism, prior work, and open questions for future research.

2.1. Location Verification

Hardware-enabled location verification would allow regulatory authorities to reliably determine the locations of AI chips, supporting the enforcement of United States export controls on AI hardware and other location-based AI policies (Brass & Aarne, 2024).

The "ping-based" approach verifies a chip's location by measuring the time it takes for signals to travel between the chip and trusted landmark servers in known positions (Abdou et al., 2015; Gueye et al., 2004; Arif et al., 2010). In this protocol, landmark servers send cryptographic challenges to the chip and measure the amount of time it takes for the chip to respond. Since signals travel at predictable speeds (bounded by the speed of light), the time measurements can

[*] Equal contribution [1] Oxford University [2] RAND Corporation [3] Oregon State University [4] Center for AI Safety [5] Future of Life Institute [6] North Carolina State University [7] University of Texas at Dallas [8] Nanyang Technological University [9] Radboud University [10] University of Maryland. Correspondence to: Aidan O'Gara <aidanogara623@gmail.com>, Gabriel Kulp <gkulp@rand.org>, Will Hodgkins <will@safe.ai>.

Workshop on Technical AI Governance (TAIG) at ICML 2025, Vancouver, Canada. Copyright 2025 by the author(s).

be converted to distance estimates, with a chip's position triangulated using multiple landmarks. Accuracy can be enhanced with methods such as traceroute information, DNS records, or satellite-based communications.

Implementation would need to address several security considerations, including preventing response time manipulation, protecting landmark servers from attacks, and securing chips against tampering that would extract private keys. Adversaries might attempt to artificially speed up responses using premium network connections, subject landmarks to DDoS attacks, or extract cryptographic keys to run verification protocols on devices in approved locations while using chips elsewhere. These threats necessitate rigorous security measures, including robust tamper-resistance for chips, redundant landmarks, and protocols resilient to partial landmark compromise.

2.2. Offline Licensing

AI chips could be designed to require a cryptographic license in order to operate, and these licenses could expire after the chip has performed a specified amount of computational work (Kulp et al., 2024; Petrie, 2024). Similar to how software licenses control access to applications, offline licensing would allow control over the use of AI hardware without enabling invasive monitoring of how the hardware has been used.

Proposed offline licensing systems have three components. First, cryptographically signed licenses are issued by regulatory authorities and verified by the chip using a stored public key. These licenses specify permissible usage limits and include both a license ID and chip ID. Second, a throttling mechanism such as Secure Boot (Secure Boot, 2024) checks for valid licenses, limiting or disabling chip functionality when proper authorization is absent. Third, tamper-resistant meters track computational usage (through metrics like clock cycles or operations performed), with licenses expiring once usage limits are reached.

To secure offline licensing against technical attacks, several security measures are needed. First, each license would include a sequential license ID and device-specific identifier, preventing license reuse or sharing across devices. Second, meters tracking usage would need protection against tampering that might allow continued operation beyond licensed limits. Third, the throttling mechanism must be robust against attempts to bypass restrictions. More generally, secure boot implementations must resist tampering, and licenses must only be issued by the authorized provider.

The most significant governance questions center on who would control license issuance. One nation could require its AI chipmakers to give it sole authority over licensing hardware usage, but this could irk other nations. Alternatively, a multinational body could issue licenses based on the approval of some or all member states, though powerful countries might object to ceding control over their hardware supply to other nations. Thus, while offline licensing may be an effective technical tool for authorizing AI hardware usage without compromising the privacy of AI developers, successful deployment would require addressing significant governance questions.

2.3. Network Verification

The first two mechanisms we described verify a chip's physical location and basic use of computational resources. A third group of mechanisms could verify information about the networks connecting AI chips, such as claims about cluster configurations and real-time communications.

Verifying cluster configuration involves checking which chips are allowed to communicate, at what bandwidth, and under what conditions. This verification can be implemented in two ways: logically or physically. Logical verification involves enforcing network configurations through software or firmware constraints, which could then be attested using cryptographic methods. An example is the ”fixed set” mechanism proposed by Kulp et al. (2024), in which AI chips are hard-coded with a whitelist specifying a small set of other chips they can communicate with at high bandwidth, while communication with any chip outside this set is strictly limited. Similarly, an ”adjustable set” mechanism could provide a renewable license allowing a certain amount and speed of communication among a specified set of chips. Physical verification, in contrast, involves hardware constraints such as cables or switches with bandwidth limitations (Scher & Thiergart, 2024). Inspectors could verify these limits physically through direct examination, supported by tamper-evident equipment or cameras. Each of these mechanisms could verify whether chips have been aggregated together into a high-performance cluster capable of large-scale AI training.

Beyond verifying static information about cluster configurations, real-time communications between chips could also be verified. Real-time communication verification could occur either on-chip or off-chip. On-chip verification entrusts software running on AI hardware to accurately attest to the chip's network traffic. In contrast, off-chip verification places trust in other hardware, such as existing network switches within data centers or new dedicated secure processors retrofitted onto a data center's network pathways. For instance, Petrie & Aarne (2025) propose using secure hardware integrated into network interface controllers to independently observe, authenticate, and verify real-time network communications, providing a detailed, secure, and tamper-resistant record of inter-chip data transfers.

Challenges in verifying physical cluster configurations are primarily operational, not technical. Regular data center inspections would be onerous, but if run accurately, they would leave little room for circumvention. Software-based approaches, on the other hand, have a large surface area for sophisticated technical attacks. As with virtually all complex software systems, many vulnerabilities have been discovered in hardware security systems such as trusted execution environments (TEEs) and confidential computing (Muñoz et al., 2023). Secure software-based network verification would require either securing these existing software systems, or designing and retrofitting dedicated secure processors onto the network paths between chips. These vulnerabilities pose a significant challenge to software-based network verification. Finally, the value of network verification may be limited in general by algorithmic advances that reduce the communication requirements for AI training.

2.4. Workload Verification

The final type of mechanism we will discuss verifies claims about AI training and inference workloads. There are many claims that could be verified about such workloads, and multiple mechanisms that could verify them.

Potential claims to be verified include the number of operations or clock cycles executed by specific AI hardware; whether that hardware is performing training or inference; and if so, the weights of the model being run, as well as its performance on key benchmarks.

This information could easily be tracked and reported by AI chips, but it seems difficult for third-parties to verify the accuracy of these attestations. Ideally, on-chip hardware security features such as trusted execution environments (TEEs) would securely log and attest to this information, but they have well-documented vulnerabilities. Alternatively, Petrie & Aarne (2025) propose designing and retrofitting secure processors in AI data centers to collect and analyze information and verify claims about AI workloads.

A third possibility is that chips could be retrofitted with analog sensors to monitor side-channels such as power usage and electromagnetic radiation. Leveraging the extensive literature on side-channel attacks, verifiers could analyze side-channel information to confirm claims such as the number of clock cycles run by a chip or whether the chip was running AI training or inference. However, each of these methods may contain vulnerabilities in hardware or software, creating potential avenues for adversaries to exploit.

To avoid the risk of hardware and software vulnerabilities, workload verification could be achieved through proof-of-work systems. Shavit (2023) proposes a verification method based on periodically recording snapshots of neural network weights during training, alongside detailed information about the training data and hyperparameters. Because generating weight snapshots that precisely match these recorded training details is only feasible by actually executing the claimed training run, third parties can reliably verify that the model was trained exactly as reported. This illustrates a more general principle: If every computation in an AI data center were documented in great detail, recomputing a small number of those computations could suffice to confirm the accuracy of all reported activities in the data center. Proof-of-work verification systems face additional challenges, but they could verify claims about AI workloads without depending on hardware or software systems to remain secure against adversarial manipulation.

3. Challenges

3.1. Technical Challenges

Privacy Preservation. HEMs should protect the privacy interests of AI developers, governments, and users. Private AI developers will want to protect trade secrets about model architectures or training processes. Governments may have national security requirements to keep certain information confidential. Users of AI systems may require confidentiality for their inputs and the system's responses. Technical research can enable privacy-preserving verification, such as by minimizing the amount of information collected, stored, analyzed, and used by HEMs, or processing information locally to verify high-level claims without sharing low-level technical details.

Tampering. HEM functionality could be undermined by tampering. While perfect tamper-proofing may be unrealistic against well-resourced adversaries, deterrence may be possible by making tampering difficult, easily detected, and costly when discovered.

There are three major anti-tamper strategies. Tamper resistance raises the cost and difficulty of tampering. Tamper evidence allows inspectors to see that tampering has occurred upon physical inspection of the relevant AI hardware. Tamper response takes protective actions when tampering is detected, such as notifying a third party or self-destructing. These strategies could be combined in a defense-in-depth approach that raises the anticipated costs of tampering attempts while reducing their benefits.

Technical anti-tamper mechanisms can be complemented by non-technical deterrents such as data center inspections or legal and diplomatic consequences for detected tampering. For example, while retrofittable secure enclosures may be needed to protect HEMs from invasive tampering if an adversary has full physical control over the hardware, this could be unnecessary in some deployment scenarios, as invasive physical attacks could be effectively ruled out by security guards and cameras.

Performance, Cost, and Operational Constraints. HEMs must operate in leading AI data centers, imposing several constraints. They should minimize energy usage and heat generation, and not interfere with cooling systems. They must fit in the limited available physical space. False alarms should minimize disruptions to AI workloads, and device failures should be easily repairable. HEMs designed today must be adaptable to the hardware and software systems of tomorrow. Perhaps the most flexible part of this equation is cost: leading AI chips already cost tens of thousands of dollars, so it may be acceptable to spend up to thousands of dollars equipping each chip with HEMs.

Supply Chain Security. Hardware backdoors could undermine the functionality of HEMs, or help them perform unintended activities such as intellectual property theft.

While governments and AI developers may be willing to trust the supply chain security of hardware designers and manufacturers within their geopolitical sphere of influence, they may suspect that hardware from adversarial nations contains backdoors that would undermine HEMs. Therefore, for HEMs to help verify agreements between geopolitical rivals, we see two options. Either technical researchers must solve the perennial problem of supply chain security, or HEMs must be designed to not require international trust. As an example of the latter approach, one nation's chipmakers could design HEMs to be retrofitted to another nation's hardware. As long as each side trusts their own hardware, adversaries could verify agreements without needing to trust each other's supply chains.

Given the potential for hardware backdoors to undermine verification systems, researchers should explicitly consider supply chain security requirements when proposing methods for verifying responsible AI development.

3.2. Governance Challenges

Authority and Control. Deploying HEMs requires determining who controls critical functions like receiving attestations, issuing licenses, or accessing monitoring data. Failing to adequately answer these questions could prevent HEMs from being deployed even if they are technically sound.

Perhaps the simplest path to HEM adoption would be a single country mandating HEMs for certain hardware. For example, the United States Congress is currently considering a bill to require location verification systems on AI accelerators facing export controls. In the future, perhaps they would consider mandating offline licensing or workload monitoring devices on certain chips. This still creates some questions about control—Who will operate the landmark servers? Who will have access to the location data? What if these are compromised by hackers?—but if these questions can be answered to the satisfaction of the United States government, then the HEMs can be rolled out.

More difficult could be an international agreement involving HEMs. Adversaries might not trust one another with access to information about their AI development, or the ability to block the functionality of their hardware. Further exploration of the kinds of international agreements that might be incentive-compatible would be useful for guiding technical research on HEMs towards realistic solutions.

Coverage of relevant workloads. To mitigate AI risks, HEMs must be deployed on relevant AI hardware. But covering the relevant hardware will be challenging for several reasons.

Timing. Hazardous AI workloads may be run in the near future, or may already be possible (Brent & Jr, 2025). If HEMs require redesigning chips and manufacturing processes, rollout may be too slow to mitigate key risks. Firmware updates and retrofitting hardware may be faster paths to HEM deployment.

Legacy compute. Today's AI hardware supply contains the equivalent of roughly 10 million Nvidia H100s, and it is projected to grow to the equivalent of 100 million H100s by the end of 2027 (Dean, 2025). Perhaps HEMs will be retrofitted to a small number of the world's largest data centers, but for most of the existing compute supply, it may be impractical to deploy HEMs. Instead, HEMs might only be applied to newly manufactured chips, thus limiting their reach.

AI diffusion. Over time, improvements in hardware and algorithm efficiency reduce the cost of training AI models with a given capability level (Pilz et al., 2024). Consequently, hardware unable to support hazardous AI workloads today may become capable of doing so in the future. Deploying HEMs on all hardware capable of hazardous workloads is therefore an ever-receding horizon, perhaps neither feasible nor desirable. A more practical goal may be to govern the highest risk frontier AI workloads, while recognizing that AI capabilities will eventually diffuse more broadly.

Ensuring deployment. Technical solutions alone cannot guarantee that HEMs will be installed and activated on relevant AI hardware. Effective deployment requires complementary non-technical measures, such as identifying semiconductor fabs capable of producing relevant AI hardware and verifying that they install and activate HEMs. Fortunately, leading-edge fabs are large, expensive, and few in number, making them relatively easy targets for oversight. Ensuring widespread deployment of HEMs may require new kinds of monitoring at these facilities.

Together, these challenges suggest HEMs can only be a partial solution to mitigating risks from AI development.

4. Conclusion

Hardware-enabled mechanisms (HEMs) are a promising tool for verifying responsible AI development. Location verification supports export controls by ensuring AI hardware remains in trusted jurisdictions. Offline licensing ensures that AI hardware can only be used with a valid license. Network verification mechanisms can attest to the configuration of AI clusters and the real-time communications between AI chips, potentially yielding insights about large-scale AI training runs. Finally, key facts about AI workloads such as the number of operations or the capabilities of a model in training or inference could be verified through multiple mechanisms, including on-chip software, off-chip secure processors or analog side-channel monitors, and proof-of-work systems.

Yet there are substantial challenges to mitigating AI risks with HEMs. Technical challenges include privacy preservation, tamper deterrence, performance constraints, and supply chain security. Governance challenges include deploying HEMs to relevant workloads and deciding who controls them. Given these challenges, HEMs cannot be a silver bullet solution to AI risk.

We call for targeted research into the design, prototyping, and red-teaming of specific hardware-enabled mechanisms and their core components. This research agenda should involve collaboration between hardware security experts, AI researchers, and policy specialists to ensure that solutions are technically sound, operationally feasible, and aligned with responsible governance objectives. By developing these capabilities now, we can create options for policymakers to address emerging risks as AI capabilities continue to advance.

Impact Statement

This paper presents work with the goal of advancing the field of responsible AI development through hardware-based governance mechanisms. The societal impacts of this work include potential improvements to AI safety through verifiable compliance with governance frameworks, but also risks of restricting legitimate AI research if mechanisms are poorly designed or implemented. We emphasize the importance of balancing security risks with privacy and innovation.

References

Aarne, O., Tim, F., and Caleb, W. Secure, governable chips. *Center for a New American Security. https://www. cnas. org/publications/reports/secure-governable-chips*, 2024. Using On-Chip Mechanisms to Manage National Security Risks from AI & Advanced Computing.

Abdou, A., Ashraf, M., and C, V. O. P. Cpv: Delay-based location verification for the internet. *IEEE Transactions on Dependable and Secure Computing*, 14(2):130–144, 2015.

Arif, M. J., Shanika, K., and Santosh, K. Geoweight: internet host geolocation based on a probability model for latency measurements. In *ACSC*, pp. 89–98, 2010.

Brass, A. and Aarne, O. Location Verification for AI Chips, 2024. URL https://static1.squarespace.com/static/64edf8e7f2b10d716b5ba0e1/t/6670467ebe2a477eb1554f40/1718634112482/Location%2BVerification%2Bfor%2BAI%2BChips.pdf.

Brent, R. and Jr, T. G. M. Contemporary ai foundation models increase biological weapons risk, 2025. URL https://arxiv.org/abs/2506.13798.

Dean, R. Ai 2027 compute forecast. Technical report, AI Futures Project, April 2025. URL https://ai-2027.com/research/compute-forecast.

Epoch AI. Data on Notable AI Models, 2024a. URL https://epoch.ai/data/notable-ai-models. Accessed: 2025-01-24.

Epoch AI. Key trends and figures in machine learning, 2024b. URL https://epoch.ai/trends. Accessed: 2025-01-24.

Gueye, B., Artur, Z., Mark, C., and Serge, F. Constraint-based geolocation of internet hosts. In *Proceedings of the 4th ACM SIGCOMM conference on Internet measurement*, pp. 288–293, 2004.

Heim, L. and Pilz, K. What share of all chips are high-end data center AI chips?, 2024. URL https://blog.heim.xyz/share-of-ai-chips/.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models, 2020.

Kulp, G., Gonzales, D., Smith, E., Heim, L., Puri, P., Vermeer, M. J. D., and Winkelman, Z. Hardware-Enabled Governance Mechanisms: Developing Technical Solutions to Exempt Items Otherwise Classified Under Export Control Classification Numbers 3A090 and 4A090. 2024. doi: 10.7249/WRA3056-1.

Muñoz, A., Ríos, R., Román, R., and López, J. A survey on the (in)security of trusted execution environments. *Computers & Security*, 129: 103180, 2023. doi: 10.1016/j.cose.2023.103180. URL https://www.sciencedirect.com/science/article/pii/S0167404823000901.

O'Gara, A., Kulp, G., Hodgkins, W., Petrie, J., Immler, V., Aysu, A., Basu, K., Bhasin, S., Picek, S., and Srivastava, A. Hardware-enabled mechanisms for verifying responsible ai development, 2025. URL https://arxiv.org/abs/2505.03742.

Petrie, J. Near-term enforcement of ai chip export controls using a minimal firmware-based design for offline licensing. *arXiv preprint arXiv:2404.18308*, 2024. URL https://arxiv.org/abs/2404.18308.

Petrie, J. and Aarne, O. Flexible hardware-enabled guarantees, part ii: Technical options for flexible hardware-enabled guarantees. April 2025. URL https://www.flexheg.com/report-2.pdf.

Petrie, J., Aarne, O., Ammann, N., and 'davidad' Dalrymple, D. Flexible hardware-enabled guarantees, 2025. URL https://flexheg.com/report-1.pdf.

Pilz, K., Heim, L., and Brown, N. Increased compute efficiency and the diffusion of ai capabilities, 2024. URL https://arxiv.org/abs/2311.15377.

Sastry, G., Lennart, H., Haydn, B., Markus, A., Miles, B., Julian, H., Cullen, O., K, H. G., Richard, N., Konstantin, P., et al. Computing Power and the Governance of Artificial Intelligence. *arXiv preprint arXiv:2402.08797*, 2024. URL https://cdn.governance.ai/Computing_Power_and_the_Governance_of_AI.pdf.

Scher, A. and Thiergart, L. Mechanisms to verify international agreements about ai development, 2024. URL https://intelligence.org/wp-content/uploads/2024/11/Mechanisms-to-Verify-International-Agreements-About-AI-Development-27-Nov-24.pdf.

Secure Boot. Hardware Secure Boot. In Kelly, B. (ed.), *OCP Security workgroup*. Microsoft Corporation, 2024. URL https://www.opencompute.org/documents/secure-boot-2-pdf.

Shavit, Y. What does it take to catch a chinchilla? verifying rules on large-scale neural network training via compute monitoring. *arXiv preprint arXiv:2303.11341*, 2023. URL https://arxiv.org/abs/2303.11341.

A. Open Research Questions by Mechanism

This appendix compiles the key open research questions for each hardware-enabled mechanism discussed in this paper, drawing from the detailed analysis in the source literature.

A.1. Location Verification

- Given that the delay to distance relationship varies depending on local network conditions, what level of accuracy in location estimates is feasible in various regions of the world, particularly those in and around countries subject to export controls on AI chips?
- What improvements are possible on existing protocols for converting time delays to distances and distances from individual servers to absolute locations?
- How can location verification protocols be calibrated or complemented with other tools to minimize the rate of false positives, which could lead to unnecessary operational disruption for chip users?
- How scalable are the proposed location verification protocols, and what modifications might be required to enable location verification for millions of high-performance AI chips?
- Given their potential for greater accuracy, can satellite-based communications provide a sufficiently secure and economically feasible alternative to communication over the Internet for verifying locations?
- How can landmarks be secured against DDoS and other kinds of attacks?
- How can secure memory protect private keys from being read by chip owners?
- How can this protocol be made compatible with desires for strong cybersecurity at data centers that might involve airgapping?

A.2. Offline Licensing

- To what extent are existing secure boot technologies appropriate for protecting licensing (and other hardware-enabled governance mechanisms) against tampering? What new vulnerabilities could arise in this context and how can these be mitigated?
- How can the authenticity of licenses be verified in a scalable way across hundreds of thousands of chips? What other implementation challenges should be expected in deploying a licensing scheme at scale, and how can these be addressed?
- Which quantities should be metered, and how can this be done securely? What are the strengths and limitations of potential candidate metrics such as floating-point arithmetic unit uses?
- What technical and operational constraints should be considered when determining how much of each quantity should be allowed per license? Smaller limits

would require chip owners to more frequently renew their licenses, which has costs and benefits from a governance perspective.

- How should licenses be issued? This is primarily a policy question, not a technical question. However, technical researchers could enable more desirable policy choices, such as designing systems for multi-party provision of licenses that enable multilateral AI governance.

A.3. Network Verification

- What is the optimal mechanism of implementation for monitoring and enforcing chip interconnection limits? For example, how feasible is the use of cluster management software, network switches or individual AI chips for this purpose?
- How viable are secure, remote, post-manufacturing adjustments to previously specified limits on communication bandwidth? Can flexible caps or whitelists for interconnection of AI chips be implemented without creating security vulnerabilities that would make the system easy to bypass for moderately resourced actors?
- What are the appropriate technical parameters for pod size and external communication bandwidth limits, given current AI training needs and anticipated future developments in distributed AI training techniques?
- How can heterogeneous devices be identified in a secure way? How can individual device authentication or attestation mechanisms be integrated into a cohesive architecture that verifies and controls the number of interconnected AI chips?
- How can the integrity of fixed set pods be remotely attested, and what mechanisms could be developed to detect tampering with these configurations?
- If restrictions are enforced using networking equipment, how can this be done without undesired impacts on non-AI training activities that use this equipment?
- Is there a way to securely update which chips are permitted in the pod so that broken hardware can be replaced?

A.4. Workload Verification

- What are the most effective metrics and indicators for accurately classifying AI workloads and performing compute accounting, while minimizing performance overhead?
- How can adversarial robustness be improved for these metrics, ensuring resilience against intentional manipulation and evolving AI training techniques?
- What are the technical and practical challenges of implementing large-scale cryptographic proofs and secure enclaves (e.g., TEEs) for AI verification across multi-node and multi-GPU systems? What modifications to current hardware, protocols, and security features would be needed to enable widespread use of TEEs for verifying governance-relevant properties of AI training at scale?
- How can dataset verification be reliably performed in distributed systems, accounting for challenges like pipeline parallelism and data parallelism, where only subsets of GPUs interact with input data?