Near-Term Feasibility of Monitoring and Verification Methods in the Compute Governance of Artificial Intelligence

Nikhil Narayanan*†, Radhika Chandra*, Swati Dhiman*, Jonas Kgomo, Jamie Johnson, Joel Christoph

1 Introduction

Frontier AI models capabilities have grown rapidly, garnering the attention of policymakers and governments due to the multitude of risks this technology poses. The advancement of these models is largely attributed to three factors: the scaling of data, the scaling of compute, and algorithmic advancements (Sastry et al., 2024). While regulation around data is mature, regulation on algorithms tends to be challenging, whereas compute regulation (or the use of hardware) is far more tractable. Compute governance makes it possible for governments to understand and track the capabilities of frontier models, allowing them to examine potentially dangerous models pre-deployment and ease race dynamics (Allison and Herzog, 2025). However, robust regulation requires detectable non-compliance.

Monitoring is the real-time continuous observation of parties with the capacity to develop risky AI models. Verification confirms the results of monitoring; it involves auditing parties with this capacity to ensure their compliance with regulations, without infringing on privacy and security.

This paper explores six different monitoring and verification methods proposed in compute governance literature today and scores their technical and institutional feasibility (near, medium and long term), summarised in Table 1. Tradeoffs such as accuracy and vulnerability to evasion are considered when determining technical feasibility, whereas cost and enforceability help ascertain institutional feasibility. We define near term as methods that are currently deployable with today's institutions and technologies, even if imperfect. Medium term methods are those made plausible within a few (2-5) years with moderate policy, engineering or coordination work. Long term are those we consider too immature, underspecified, controversial or assumption-heavy to treat as realistically implementable yet.

* Equal Contribution † Correspondence Author (nikhil.narayanan20@alumni.imperial.ac.uk)

In Section 3 we shortlist methods with near term technical and/or institutional feasibility and specify what's required for a method to become deployable. It's important to note these methods all rely on the assumption of continued reliance on large-scale centralised compute.

We hope to help policymakers and researchers evaluate monitoring and verification methods in compute governance more critically and holistically across both technical and institutional feasibility by offering up-to-date and balanced views.

2. Methods

2.1 Hardware-Level Auditing

Hardware-level auditing uses on-chip security features to detect and report constraint violations while preserving privacy. Examples include FLOPs metering for compute caps (Ansari, 2026), workload attestation, and secure-boot enforcing signed, up-to-date firmware (Shavit et al, 2023). A staggered approach to implementation proposes leveraging existing mechanisms (performance counters, trust execution environments, hardware roots of trust) for governance, before employing hardware solutions to make physical tampering near impossible (Aarne et al, 2024), a key weakness of this method. Potential solutions include physical unclonable function (PUF) enclosures, which use slight variations in the manufacturing process to create cryptographic keys for attestation that fail if physically disturbed (O'Gara et al, 2024).

Hardware-level auditing is effective for policy measures such as usage verification and hardware-enforced licenses (Aarne et al, 2024). It assumes coordination with chip manufacturers responsible for implementing features, and trust from chip owners regarding their data security. R&D is necessary for chips to be tamper-evident and eventually tamper-proof, requiring investments in hardware and software security. However, precedent for its technical feasibility exists in NVIDIA's Lite Rate Hash GPUs , which detect crypto mining workloads and throttle performance using cryptographically signed firmware that can not be easily modified (Gartenberg, 2021). The involvement of chip manufacturers influences institutional feasibility. It is possible manufacturers will try to minimize security and reporting features that have a

negative impact on them - a similar response was seen when NVIDIA created the H800 chip to get around export controls (Peters, 2023). Existing chips without these mechanisms would need to be tackled, whether through phase-out strategies or retrofitting. Additionally, tamper-evident chips require facility-level inspections, which are considered intrusive.

Hardware-level auditing is a medium term solution when considering both institutional and technical feasibility. Companies like Lucid Computing can frame the mechanisms mentioned as enhanced security features rather than for external monitoring (Lucid Computing, 2026), allowing initial deployment without international coordination.

2.2 Cloud Reporting Requirements

Compute and cloud providers can offer a powerful mechanism for monitoring at the level of domestic regulation (Ansari, 2026). These companies and organizations have access to compute usage data while also being capable of restricting access to their services - their very billing model allows them to be able to report the scale of compute of their clients training runs (Heim et al., 2024). Moreover, a result of Biden's Executive Order 11410 (which was later repealed), the US has brought into light identity verification requirements known as KYC (Know Your Customer) for cloud providers; this is a standard used in financial regulation whereby a service provider needs to verify client identities to support public oversight of such technologies (Egan and Heim, 2023). This would allow cloud providers to conduct risk assessments on clients and require additional information from clients whose usage exceeds a particular threshold to report risky usages of GPU clusters.

Criticisms of cloud metadata based methods (where you monitor risky training runs from clients who qualify for a KYC threshold) are largely focused against the use of a FLOP based threshold. These can cause detection gaps by using multiple accounts with sequential training on frozen checkpoints, or by using domain specific models for instance (Moon et al., 2025). RAND proposed the following solutions: use non-threshold based KYC where users disclose their identity and intended usage, using energy proxies (2.6), utilize model-training disclosures (2.5), and monitor large transfers of data into accounts as well as focusing on cluster size over compute

thresholds (Moon et al., 2025). Other criticisms include the loss of privacy of those running risky workloads, however, there is much ongoing research towards the medium term which can point to privacy preserving cloud monitoring and verification schemes (Apple, 2024).

Cloud reporting requirements are feasible in the near term, and governments around the world would benefit from visibility on risky actors.

2.3 Chip Tracking and Supply Chain Controls

Chip tracking and supply chain controls are governance mechanisms for monitoring and restricting access to advanced semiconductor hardware used in AI systems. Chip tracking identifies ownership, location, and movement of high-end chips, while supply chain controls use export policies and licensing to condition access. These approaches rely on the concentration of key segments of the semiconductor value chain in a small number of firms and countries, creating leverage for intervention. They assume that compute remains centralised and observable, and that actors can be compelled to comply with reporting and access restrictions.

These mechanisms offer a technically grounded approach to governing AI development by targeting compute as a critical input. By intervening at the level of hardware access, they enable upstream control over AI capabilities before deployment, (Sastry et al., 2024). Structural chokepoints in semiconductor manufacturing create leverage for a small set of actors, such as TSMC or ASML, to influence access to frontier capabilities (Allen, 2022; Khan, 2021). Tracking chip flows improves auditability and supports subsequent verification, as large-scale compute usage may be difficult to conceal (Shavit, 2023; Kulp et al., 2024).

However, these benefits depend on assumptions that may not hold in practice. Enforcement relies on compliance, yet firms can exploit jurisdictional gaps, and weaken controls by operating through intermediaries (Allen, 2022; Aarne et al., 2024). Legacy hardware limits coverage, and concerns over corporate confidentiality may reduce incentives for full disclosure (Shavit, 2023).

Effective implementation also depends on coordination across jurisdictions, sufficient state capacity, and, in some cases, resource pooling, all of which vary significantly. As AI

development becomes more efficient and distributed, the assumption of traceable compute may weaken (Shavit, 2023; Sastry et al., 2024).

We therefore find that these mechanisms are technically feasible in the near term but institutionally constrained, with effectiveness limited in fragmented settings.

2.4 Facility Level Inspections

Facility level inspections are physical visits to declared datacentres, fabs and AI development facilities to verify their compliance with compute agreements. This can consist of periodic inspections, continuous monitoring, and challenge inspections, where inspections are called on short notice if the facility is suspected of non-compliance. Literature primarily focuses on inspections of data centres, which centre around examining the chip and its activity logs, training run evidence and compliance with FLOPs limits (Wasil et al, 2024). Verification approaches in fab inspections could extend to incorporating imaging techniques with nanometer resolution to detect unauthorised modifications (Ilhan et al, 2026).

Facility level inspections are particularly useful for assessing the chips' physical condition for evidence of tampering. They rely on the assumption that key facilities will give permission for these in-depth inspections to take place. This method's failure modes include doctoring logs and concealing unauthorised hardware. Inspections are resource-intensive, requiring significant amounts of time, resources and expertise from the inspectors conducting them. Although the IAEA's on-site inspections is a commonly used point of reference to indicate this method's institutional feasibility (Baker, 2023), it is still hindered by privacy and security concerns regarding intellectual property. Its implementation may rely on permission from parties suspected of unauthorized activity, which influences its enforceability (Wasil et al, 2024). Nevertheless, if carried out well, facility-level inspection would be difficult to evade and can act as a strong supplementary mechanism for other methods made vulnerable by chip owners' physical access to the chips.

Facility level inspections are near term in terms of technical feasibility and considered to have no technological hurdles (Barnett et at, 2025), but ranks as far-term feasibility not just

institutionally, but overall. This method is not possible without international coordination, making a deployment timeline challenging to define.

2.5 Model-Training Disclosure Regimes

Model-training disclosure regimes are a class of training verification methodologies that aim to chiefly verify potential claims about the data used to train models, and safety-focused fine-tuning. See Appendix A2 for an explanation of the framework which governs these ideas.

State of the art methods include ideas such as Proof-of-Training-Data which would only add 1.3% of additional cost to a training run while verifying that a training run was completed with claims of specific training data having been used in updating the model (Choi et al., 2023), (Fang et al., 2023), (Zhang et al., 2022), (Kong et al., 2022). Proof-of-Training-Data is not provably robust and assumes it would be hard to find a random weight initialization for a model whereby fake data could result in target weights (Choi et al., 2023). See Appendix A3 for further criticisms as well as modern methods.

Overall, the low additional cost of potentially 1.3% to frontier labs make this a promising verification strategy; the main tradeoff is that these techniques need further R&D to be used in the medium term. The main assumption in this method is that it is computationally too expensive for a said covert adversary to fake a transcript and therefore are incentivized to report the truth as well as this method being available to frontier labs before they begin pretraining.

2.6 Energy or Other Infrastructure Proxies

Energy and infrastructure proxies refer to approaches that infer AI compute activity from observable physical signals rather than direct system access. These signals include electricity consumption, cooling demand, and data centre operations, which reflect the infrastructure required to support large-scale AI workloads. The approach assumes that such workloads produce observable and sufficiently distinct patterns at the facility level.

These proxies provide a technically plausible pathway for monitoring compute activity by leveraging the physical footprint of AI systems. Sustained workloads generate measurable external signals, enabling inference without relying on self-reporting or internal access (Sastry et al., 2024; Patterson, 2022). This supports non-intrusive monitoring across jurisdictions, including low-access or adversarial settings, while avoiding exposure of proprietary data (Whittlestone & Clark, 2021). Continuous observation allows detection of temporal patterns such as sustained or spiking activity, and can support cross-validation of reported activity, although only indirectly (Wasil et al., 2024). However, realising this potential requires both more data sets to be available for these problems, as well as better statistical models to regress compute from energy.

These benefits depend on assumptions that may not hold in practice. The link between energy use and compute may weaken as models become more efficient or workloads more distributed (Sastry et al., 2024; Patterson, 2022). Infrastructure signals are also noisy and difficult to attribute, as facilities host mixed workloads and energy demand fluctuates for multiple reasons (Lal & You, 2025). Current literature does not define standardised methods for implementation, and access to high-resolution data remains uneven across jurisdictions.

Energy proxies are institutionally feasible in the near term due to their non-intrusive nature. However, this feasibility depends on access to sufficiently granular infrastructure data, which may require regulatory enablement. Technical feasibility remains more limited, suggesting a medium term pathway for effective deployment.

2.7 Complementary Mechanisms

We analyse two complementary mechanisms which support the above mentioned monitoring and verification methods. These include systems that collect, standardise, and interpret data on AI development, deployment, and underlying resources such as compute and international AI agreements. See Appendix A4 for the former.

International AI agreements are another necessary complement to the mechanisms outlined here. These could be arrangements between states to coordinate the governance of AI development

and deployment through shared rules and standards. They complement chip tracking and supply chain controls, cloud reporting requirements, and facility-level inspections by aligning expectations for monitoring, reporting, and enforcement across jurisdictions. These mechanisms, in turn, provide the technical basis for verification, while agreements enable their consistent implementation across borders. This relationship is mutually reinforcing, as governance mechanisms enable verification while agreements align their implementation across jurisdictions. However, effective coordination depends on sustained alignment (Wasil et al., 2024) and access to comparable data and measurement approaches (Sastry et al., 2024). In the absence of such alignment, fragmented implementation across jurisdictions can limit the effectiveness of these mechanisms, highlighting the role of agreements as an enabling layer rather than a standalone solution.

Method Type	Mechanism	Technical Feasibility			Institutional Feasibility		
		Near term	Medium term	Long term	Near term	Medium term	Long term
Monitoring	Cloud reporting requirements	✓			✓		
	Chip tracking and Supply chain controls	✓				✓	
	Energy or other infrastructure proxies		✓			✓	
Verification	Facility level inspections	✓					✓
	Model-training disclosure regimes		✓			✓	
	Hardware-level auditing		✓			✓	

Table 1. The six mechanisms for monitoring and verification scored on technical and institutional feasibility, across near term, medium term and long term time horizons.

3. Feasibility Analysis

In a seemingly polarised world, international agreements are very challenging and time-consuming to achieve. Policy levers that depend on international coordination tend to take more time to become effective, and due to the potentially worrying pace of development of AI, it is important to prioritize the best interventions that can help immediately. The methods requiring

international coordination - namely chip tracking and facility level inspections - are both technically feasible, and as such, as soon as international coordination can be achieved, they should be prioritized. Until then, the below domestic methods may be most effective. The key assumption for these methodologies, that states are compliant or compliance can be enforced, may be challenged due to evasion risks that are amplified without international coordination. The absence of coordinated implementation may reinforce a negative feedback loop where fragmentation could contribute to geopolitical disagreements which would limit incentives for cooperation. In absence of sustained international alignment, these mechanisms, though promising, are limited when compared to domestically viable approaches.

In the near term, policy makers should allocate funding for:

> **Cloud reporting requirements:** Both KYC and sharing top-level metadata on risky users can improve monitoring and visibility on activities outside of frontier labs. Existing regulation for cloud reporting should be built on by focusing on non-threshold based KYC.

In the medium term, however, the best monitoring approach could be:

> **Energy and infrastructure proxies:** May face relatively low institutional friction. Policymakers can promote their adoption by incentivising the development of standardised measurement frameworks, infrastructure-level data protocols, and methods for signal interpretation. Early approaches suggest that energy data could be used to approximate compute activity (Desislavov et al., 2023), but these remain preliminary, reinforcing the need for further technical development.

The verification methods we researched all require some level of technical or policy work to be deployable and as such are at best achievable in the **medium term**. We consider the stronger verification approach to be:

Model-training disclosure regimes: Governments should encourage investment into research grants and startups developing model training disclosures via proof-of-training paradigms. This could be a key pillar by which we build trust with KYC counterparties and clients while also better policing activities in frontier labs (with potentially added training costs of around 2% in SoTA workflows) (Choi et al., 2023).

4. Conclusion

Ultimately, we determine cloud reporting to be the most near term feasible form of monitoring, and model training disclosure regimes to be the most near term feasible verification method. International coordination is the key dependency to leverage the true power of chip tracking and facility level inspection which would both be powerful complements to the aforementioned methods

Policymakers should prioritise the deployment of near term feasible mechanisms while investing in the development of technically underdeveloped approaches in the near term to enable more robust, multi-layered compute governance over time. Monitoring and verification are key levers towards governing compute, thus enabling safer, aligned, and trustworthy artificial intelligence.

5. Limitations

This research was limited in scope, as we have focused on near term mechanisms, but this does not mean methods excluded from this are not worthy of further development, as they may prove to be more promising in the long term. Moreover, while examining the institutional feasibility of cloud reporting requirements, we gave civilian data privacy considerations less significance because we deemed it to not be a key factor in the method longevity.

6. Future Work / Recommendations

We recommend the following calls-to-action for entrepreneurial readers as startups tackling the below problems may offer high impact contributions to the field of monitoring and verification:

- Build a high quality training data set working with energy companies to help model how energy data can be used to measure the scale (in FLOPs) of a training run
- Implement model training disclosures such as Proof-of-Training at scale and convince a frontier lab or sovereign AI lab submit a training transcript for your verification

Disclaimer

The views expressed are those of the authors and do not necessarily reflect the views of their employers or affiliated institutions.

References

Aarne, O., Fist, T., & Withers, C. (2024, January 8). Using On-Chip Mechanisms to Manage National Security Risks from AI & Advanced Computing. *Center for New American Security*. https://www.cnas.org/publications/reports/secure-governable-chips

Akgul, H., Borg, D., Berisha, A., Rahimova, A., Novak, A., & Petrov, M. (2025). *Verifiable Fine-Tuning for LLMs: Zero-Knowledge Training Proofs Bound to Data Provenance and Policy* (arXiv:2510.16830). arXiv. https://doi.org/10.48550/arXiv.2510.16830

Allen, G. C. (2022). *Choking off China's Access to the Future of AI*. https://www.csis.org/analysis/choking-chinas-access-future-ai

Allison, D. M., & Herzog, S. (2025). Artificial Intelligence and Nuclear Weapons Proliferation: The Technological Arms Race for (In)visibility. *Risk Analysis*, *45*(11), 3839–3859. https://doi.org/10.1111/risa.70105

Ansari, S. (2026). *Hardware-Level Governance of AI Compute: A Feasibility Taxonomy for Regulatory Compliance and Treaty Verification* (arXiv:2604.04712). arXiv. https://doi.org/10.48550/arXiv.2604.04712

Baker, M. (2023). *Nuclear Arms Control Verification and Lessons for AI Treaties* (arXiv:2304.04123). arXiv. https://doi.org/10.48550/arXiv.2304.04123

Balan, K., Learney, R., & Wood, T. (2025). *A Framework for Cryptographic Verifiability of End-to-End AI Pipelines* (arXiv:2503.22573). arXiv. https://doi.org/10.48550/arXiv.2503.22573

Barnett, P., Scher, A., & Abecassis, D. (2025). *Technical Requirements for Halting Dangerous AI Activities* (arXiv:2507.09801). arXiv. https://doi.org/10.48550/arXiv.2507.09801

Choi, D., Shavit, Y., & Duvenaud, D. (2023). *Tools for Verifying Neural Models' Training Data* (arXiv:2307.00682). arXiv. https://doi.org/10.48550/arXiv.2307.00682

Desislavov, R., Martínez-Plumed, F., & Hernández-Orallo, J. (2023). Compute and Energy Consumption Trends in Deep Learning Inference. *Sustainable Computing: Informatics and Systems*, *38*, 100857. https://doi.org/10.1016/j.suscom.2023.100857

Egan, J., & Heim, L. (2023). *Oversight for Frontier AI through a Know-Your-Customer Scheme for Compute Providers* (arXiv:2310.13625). arXiv. https://doi.org/10.48550/arXiv.2310.13625

Fang, C., Jia, H., Thudi, A., Yaghini, M., Choquette-Choo, C. A., Dullerud, N., Chandrasekaran, V., & Papernot, N. (2023). *Proof-of-Learning is Currently More Broken Than You Think* (arXiv:2208.03567). arXiv. https://doi.org/10.48550/arXiv.2208.03567

Gartenberg, C. (2021, April 29). *Nvidia has reinstated its RTX 3060 Ethereum cryptocurrency mining limit*. The Verge. https://www.theverge.com/2021/4/29/22409838/nvidia-rtx-3060-etherium-cryptocurrency-mining-limit-back-driver-update

Heim, L., Fist, T., Egan, J., Huang, S., Zekany, S., Trager, R., Osborne, M. A., & Zilberman, N. (2024). *Governing Through the Cloud: The Intermediary Role of Compute Providers in AI Regulation* (arXiv:2403.08501). arXiv. https://doi.org/10.48550/arXiv.2403.08501

Ilhan, A., Withers, C., Gietz, H., & Harack, B. (2026, April 14). Verifiable Semiconductor Manufacturing. *Oxford Martin AIGI*. https://aigi.ox.ac.uk/publications/verifiable-semiconductor-manufacturing/

Jia, H., Yaghini, M., Choquette-Choo, C. A., Dullerud, N., Thudi, A., Chandrasekaran, V., & Papernot, N. (2021). *Proof-of-Learning: Definitions and Practice* (arXiv:2103.05633). arXiv. https://doi.org/10.48550/arXiv.2103.05633

Khan, S., Peterson, D., & Mann, A. (n.d.). The Semiconductor Supply Chain. *Center for Security and Emerging Technology*. Retrieved 26 April 2026, from https://cset.georgetown.edu/publication/the-semiconductor-supply-chain/

Kong, Z., Roy Chowdhury, A., & Chaudhuri, K. (2022). Forgeability and Membership Inference Attacks. *Proceedings of the 15th ACM Workshop on Artificial Intelligence and Security*, 25–31. https://doi.org/10.1145/3560830.3563731

Lal, A., & You, F. (2025). Advances and challenges in energy and climate alignment of AI infrastructure expansion. *Advances in Applied Energy*, *20*, 100243. https://doi.org/10.1016/j.adapen.2025.100243

Lucid Computing. (n.d.-a). *Confidential Computing—Lucid Developer Platform*. Retrieved 26 April 2026, from https://lucidcomputing.ai/site/concepts/tee

Lucid Computing. (n.d.-b). *EU AI Act Conformity Auditor*. Lucid Computing. Retrieved 26 April 2026, from https://lucidcomputing.ai/solutions/eu-ai-act-conformity.html

Moon, A., Vedula, P., Geneson, J., & Bar-on, S. (2025). *Strategies and Detection Gaps in a Game-Theoretic Model of Compute Governance*. https://www.rand.org/pubs/research_reports/RRA3686-1.html

O'Gara, A., Kulp, G., Hodgkins, W., Petrie, J., Immler, V., Aysu, A., Basu, K., Bhasin, S., Picek, S., & Srivastava, A. (2025). *Hardware-Enabled Mechanisms for Verifying Responsible AI Development* (arXiv:2505.03742). arXiv. https://doi.org/10.48550/arXiv.2505.03742

Patterson, D. (2022, March 12). *Reducing the carbon emissions of AI*. https://oecd.ai/en/wonk/reducing-ai-carbon-emissions

Peters, J. (2023, October 17). *Nvidia's H800 AI chip for China is blocked by new export rules*. The Verge. https://www.theverge.com/2023/10/17/23921131/us-china-restrictions-ai-chip-sales-nvidia

Private Cloud Compute: A new frontier for AI privacy in the cloud - Apple Security Research. (n.d.). Private Cloud Compute: A New Frontier for AI Privacy in the Cloud - Apple Security Research. Retrieved 26 April 2026, from https://security.apple.com/blog/private-cloud-compute/

Sastry, G., Heim, L., Belfield, H., Anderljung, M., Brundage, M., Hazell, J., O'Keefe, C., Hadfield, G. K., Ngo, R., Pilz, K., Gor, G., Bluemke, E., Shoker, S., Egan, J., Trager, R. F., Avin, S., Weller, A., Bengio, Y., & Coyle, D. (2024). *Computing Power and the Governance of Artificial Intelligence* (arXiv:2402.08797). arXiv. https://doi.org/10.48550/arXiv.2402.08797

Srivastava, M., Arora, S., & Boneh, D. (2024). *Optimistic Verifiable Training by Controlling Hardware Nondeterminism* (arXiv:2403.09603). arXiv. https://doi.org/10.48550/arXiv.2403.09603

Wasil, A. R., Reed, T., Miller, J. W., & Barnett, P. (2024). *Verification methods for international AI agreements* (arXiv:2408.16074). arXiv. https://doi.org/10.48550/arXiv.2408.16074

Whittlestone, J., & Clark, J. (2021). *Why and How Governments Should Monitor AI Development* (arXiv:2108.12427). arXiv. https://doi.org/10.48550/arXiv.2108.12427

Zhang, R., Liu, J., Ding, Y., Wang, Z., Wu, Q., & Ren, K. (n.d.). *"Adversarial Examples" for Proof-of-Learning*. Retrieved 26 April 2026, from https://www.computer.org/csdl/proceedings-article/sp/2022/131600b542/1FlQEN9cor6

Appendix A1

Startups such as Lucid Computing have been working on auditing training pipelines with presumably similar techniques marketed towards the EU AI Act (Lucid Computing, 2026). Overall, some methods such as Proof-of-Training-Data may be feasible and cheap and a valuable solution as part of a suite of solutions, as well as Akgul et al's work on verifying fine tuning which may be especially valuable in experiments around fine tuning / post-training pertinent to safety at CBRN risks. While these methods are not a catch-all for verification, they could be imposed on frontier labs with startups like Lucid, or government institutes like UK AISI as trusted verifiers with further investment into R&D for these methods

Appendix A2

Model training disclosures originated from a concept known as 'Proof-of-Learning' that relies on a 'Prover' (a frontier lab) and a 'Verifier' (a governing body). The 'Prover' generates a transcript with some sample data of a subset of a training run as well as checkpoints of the model weights before and after the training (Jia et al., 2021). Probabilistic and cryptographic algorithms are then used to verify that a training did occur with model weights updating as expected due to the exposure of some shared training data. Jia's original method required intensive memory and computational costs, as well as being susceptible to data reordering, and weight initialization attacks from the 'Prover' (who would be a 'covert adversary' who may cheat so long as they appear compliant).

Appendix A3

Aforementioned model training disclosure methods also had some issues in accuracy due to nondeterminism of GPUs whereby floating point errors can make it harder to reproduce a training run. This was addressed by some researchers by running audit / proofs with higher precisions than what is used for training (Srivastava et al., 2024) but such a tradeoff would cost 20% of a training run. Recent solutions have popularly been utilizing zero knowledge proofs for cryptographic verifiability (Balan et al., 2025), and some promising methods can verify fine tuned models which used LoRA (Low Rank Adapters) for Parameter Efficient Fine Tuning (but

unfortunately costs would scale with the number of trainable parameters at the current state of the art) (Akgul et al., 2025).

Appendix A4

One complementary method would be an integration layer that aggregates inputs from cloud reporting, chip tracking, and energy or infrastructure proxies, and synthesizes this information together to be used more effectively. By structuring how these signals are captured and compared, it can enable cross-validation and reduce information asymmetry across actors and jurisdictions (Wasil et al., 2024; Whittlestone & Clark, 2021). In practice, data is often incomplete or inconsistently reported, and the absence of operational integration methods constrains its effectiveness and limits its immediate deployment.