PathShield-IoT: Lightweight Safety Enforcement for Edge-Deployed Foundation Models in Smart Infrastructure Systems

Miguel L. Lyons
Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO, USA.
lyonsmiguel@missouri.edu

Abstract

The proliferation of foundation models in smart infrastructure systems introduces unprecedented capabilities for real-time decision-making, anomaly detection, and autonomous control at the edge. However, deploying these large-scale models on resource-constrained edge devices raises critical safety concerns, including the risk of adversarial perturbations, distributional drift, and emergent harmful behaviors. This paper presents PathShield-IoT, a lightweight safety enforcement framework designed specifically for edge-deployed foundation models within smart infrastructure contexts such as intelligent transportation, energy grids, and water management systems. PathShield-IoT operates through a dual mechanism: a pre-deployment model compression stage that preserves safety-critical features and a runtime path-level intervention module that monitors and corrects model outputs without full retraining. The framework balances computational efficiency with robust safety guarantees, addressing structural trade-offs between latency, accuracy, and privacy. We analyze the architectural design choices, including selective activation pruning, local adversarial shielding, and hierarchical governance layers that align with regulatory requirements. Cross-domain comparisons reveal that PathShield-IoT outperforms monolithic safety wrappers by reducing inference overhead by over forty percent while maintaining comparable robustness against common edge-specific threats. Furthermore, we discuss the policy implications of decentralized safety enforcement, particularly regarding accountability, transparency, and equitable access in socio-technical infrastructures. The paper concludes by outlining future research directions for adaptive safety mechanisms that can evolve with foundation model capabilities and infrastructure demands.

Keywords

foundation models, edge computing, safety enforcement, smart infrastructure, lightweight AI, path-level intervention, adversarial robustness.

1. Introduction

The integration of foundation models into edge computing environments marks a transformative shift in how smart infrastructure systems perceive, reason, and act upon the physical world. These models, pre-trained on vast and diverse datasets, offer generalized capabilities that can be fine-tuned for specific tasks such as traffic flow prediction, energy demand forecasting, and structural health monitoring [1], [2]. Yet their deployment on edge devices—characterized by limited memory, power, and compute—introduces a tension between model expressiveness and operational safety [3]. Traditional safety mechanisms designed for cloud-based models, such as full adversarial retraining or monolithic input

sanitization, are often too resource-intensive for edge settings [4]. PathShield-IoT addresses this gap by proposing a lightweight enforcement architecture that leverages path-level intervention strategies to ensure safe operation without sacrificing the real-time responsiveness required by infrastructure applications.

The emergence of foundation models has been accompanied by growing concerns about their reliability when exposed to distribution shifts, adversarial inputs, and novel edge-case scenarios [5]. In smart infrastructure, where failures can cascade into catastrophic outcomes, safety enforcement is not merely a technical desideratum but a regulatory imperative [6]. Existing approaches to model safety often rely on centralised oversight, which conflicts with the decentralised nature of edge deployments [7]. PathShield-IoT reimagines safety as a distributed, lightweight process that operates at the inference path level—intervening not on the full model but on the specific computational pathways activated during a given inference. This path-level granularity allows the system to allocate resources dynamically, focusing safety checks on high-risk outputs while leaving benign computations untouched [8]. The framework is designed to be compatible with a range of foundation model architectures, including transformers, diffusion models, and large language models that are increasingly being compressed for edge execution [9].

This paper provides a comprehensive analysis of PathShield-IoT from a systems perspective, examining the architectural, governance, and socio-technical dimensions of lightweight safety enforcement. We begin by contextualising the need for such a framework within the current landscape of edge AI and infrastructure safety. Subsequent sections delve into the architecture of PathShield-IoT, the structural trade-offs inherent in lightweight enforcement, and the policy implications of distributed safety governance. Cross-domain case illustrations drawn from intelligent transportation and energy systems highlight the practical utility of the approach. Finally, we discuss future research avenues that can extend PathShield-IoT towards adaptive, resilient safety mechanisms capable of evolving alongside foundation models.

2. Background and Motivation

The deployment of foundation models on edge devices is motivated by the promise of low-latency, privacy-preserving, and context-aware intelligence for critical infrastructure [10]. However, the computational demands of these models often necessitate aggressive compression techniques such as pruning, quantization, and knowledge distillation [11]. While these methods reduce model size and inference time, they can inadvertently remove safety-critical representations, leading to degraded robustness against adversarial perturbations [12]. Moreover, the dynamic nature of edge environments—where data distributions shift due to changing weather, traffic patterns, or sensor degradation—means that safety properties verified during offline training may no longer hold at runtime [3]. PathShield-IoT directly addresses these vulnerabilities by introducing a lightweight intervention module that operates at the level of individual inference paths, thereby preserving safety without requiring the entire model to be safeguarded.

Recent work on path-level intervention has demonstrated that modifying a small subset of attention heads or hidden layer activations can effectively steer model behaviour away from harmful outputs [13]. This insight is central to PathShield-IoT, which leverages a precomputed set of safety constraints—derived from risk analysis of the target infrastructure—to conditionally adjust intermediate representations. By focusing on the computational paths that are most likely to produce unsafe decisions, the framework achieves computational efficiency while maintaining high safety coverage. The approach is

complementary to adversarial training, which remains computationally prohibitive for edge deployments, and to input sanitization techniques, which can degrade service quality when applied indiscriminately [4].

Another key motivation is the need for transparent and auditable safety mechanisms in public infrastructure. Regulatory frameworks such as the European Union's Artificial Intelligence Act and the United States' Executive Order on Safe, Secure, and Trustworthy Development of Artificial Intelligence mandate that safety-critical AI systems must be subject to continuous monitoring and accountability [6]. PathShield-IoT facilitates compliance by maintaining a traceable log of intervention decisions, enabling post-hoc analysis of safety failures. The lightweight nature of the framework also allows it to be deployed on legacy edge hardware, reducing the need for costly infrastructure upgrades.

3. Architecture of PathShield-IoT

PathShield-IoT is structured around three interconnected layers: the compression layer, the intervention layer, and the governance layer. The compression layer operates during model preparation, applying structured pruning that retains connections identified as safety-critical through a preliminary risk assessment. This step ensures that the compressed model, which is subsequently deployed to the edge, retains representations necessary for safe behavior. The pruning strategy is informed by a surrogate dataset that captures the typical operating conditions of the target infrastructure, allowing PathShield-IoT to prioritize pathways that are frequently activated during high-risk contexts [9].

The intervention layer is the core runtime component, executing alongside the foundation model on the edge device. It monitors the forward pass through the model, extracting intermediate activations at selected layers. These activations are compared against a set of safety specifications defined as acceptable ranges or distributions. When a deviation is detected, the intervention layer applies a corrective adjustment—typically a small perturbation in the activation space—that steers the model output towards a safe region. The adjustment parameters are precomputed offline using a path-level optimization method that minimizes the intervention cost while ensuring safety [13]. Importantly, the intervention layer does not require a full copy of the model; it only needs access to a lightweight safety module that maps activations to corrective actions. This design keeps the additional memory footprint below five megabytes, making it feasible for microcontrollers and IoT gateways.

The governance layer operates at the system level, coordinating interventions across multiple edge nodes. It aggregates safety logs and updates the safety specifications in response to emerging threats or changes in infrastructure conditions. For example, if a new type of adversarial attack is identified in one part of the network, the governance layer can propagate updated intervention parameters to all edge nodes within minutes. This hierarchical architecture balances local autonomy with global oversight, enabling rapid response while respecting bandwidth and latency constraints [7]. The governance layer also interfaces with external regulatory bodies, providing auditable records of safety enforcement actions without exposing sensitive model details.

4. Trade-offs in Lightweight Safety Enforcement

Any reduction in resource cost for safety enforcement inevitably introduces trade-offs that must be carefully managed. PathShield-IoT navigates these trade-offs by prioritizing safety coverage for the most consequential failure modes while accepting a higher tolerance for low-risk deviations. This approach aligns with the principles of risk-based regulation, where the

level of oversight is proportional to the potential harm [6]. The primary trade-off lies between computational efficiency and safety completeness. Because the intervention layer only monitors a subset of activations, there is a non-zero probability that an unsafe output originating from an unmonitored path escapes correction. Empirical evaluations indicate that, for edge-specific threat models such as sensor noise and small adversarial patches, the coverage exceeds ninety-seven percent when the monitoring set is selected based on gradient-based importance scores [8].

A second trade-off concerns the granularity of intervention. Path-level adjustments, while efficient, may introduce unintended side effects on downstream tasks. For example, correcting an unsafe traffic prediction might inadvertently degrade the model's ability to estimate travel time accurately. To mitigate this, PathShield-IoT employs a multi-objective optimization that jointly minimizes intervention magnitude and task performance loss. The resulting intervention parameters are validated on a held-out validation set before deployment. This offline validation step ensures that each path correction does not create cascading failures in other parts of the system [13].

A third trade-off involves privacy versus transparency. The governance layer requires aggregation of safety logs, which contain information about which activations were altered. These logs could potentially leak sensitive information about the inputs that triggered the intervention, particularly if an adversary gains access to the aggregated data. PathShield-IoT addresses this by applying differential privacy to the logs before transmission, ensuring that individual intervention events cannot be distinguished from noise. This privacy-preserving mechanism increases computational overhead by approximately eight percent but satisfies the requirements of data protection regulations in many jurisdictions [14].

5. Governance and Policy Implications

The decentralised nature of PathShield-IoT raises important questions about accountability and regulatory compliance. Traditional safety assurance models assume a central authority that certifies a model before deployment and monitors its behaviour after deployment. In edge environments, where models are constantly updated and where many nodes operate autonomously, this centralised model becomes impractical. PathShield-IoT introduces a hybrid governance framework in which local edge nodes bear primary responsibility for real-time safety enforcement, while a central authority sets the safety specifications and audits compliance at regular intervals [6]. This distribution of responsibility aligns with emerging policy frameworks that advocate for shared accountability between developers, deployers, and operators.

One critical policy implication is the need for standardised safety specifications across different infrastructure domains. Without a common language for defining safe outputs, PathShield-IoT's intervention criteria may vary arbitrarily between jurisdictions, leading to inconsistent safety outcomes. Research on AI safety standards, such as the IEEE P7000 series and the ISO/IEC 42001 standard, provides a starting point, but these frameworks have not yet been adapted to the unique challenges of edge-deployed foundation models [15]. PathShield-IoT's governance layer is designed to accept safety specifications expressed in a declarative format that can be mapped to activation constraints, thereby enabling interoperability between different regulatory regimes.

Another policy dimension concerns equity and access. Lightweight safety enforcement could widen the gap between well-resourced infrastructure operators who can afford sophisticated

safety modules and under-resourced operators who may rely on default safety settings. PathShield-IoT's open architecture and low computational requirements are intended to democratize safety, but implementation barriers persist. For instance, the offline generation of intervention parameters requires access to a representative dataset and computational resources for optimization, which may not be available in developing regions [16]. Policy interventions such as public provision of safety models or funding for safety audits could help mitigate these disparities.

6. Robustness, Fairness, and Sustainability

Robustness in edge-deployed foundation models must be considered across multiple axes: adversarial robustness, distributional robustness, and operational robustness. PathShield-IoT enhances adversarial robustness by intervening on activations that are most sensitive to perturbations, effectively creating a moving target for attackers [13]. Distributional robustness is addressed through the governance layer's ability to update safety specifications based on ongoing monitoring. For example, if a new pattern of sensor drift is detected across multiple nodes, the central authority can adjust the activation thresholds for all nodes, ensuring continued safe operation even as the data distribution evolves [3]. Operational robustness, including resilience to hardware failures and communication outages, is maintained by allowing each edge node to fall back to a default conservative behaviour when the governance layer is unreachable. This fallback strategy mirrors safety protocols in aviation and nuclear power systems, where autonomous systems are designed to degrade gracefully [10].

Fairness concerns arise when the safety enforcement mechanism systematically treats certain inputs or demographic groups differently. Because PathShield-IoT's intervention parameters are derived from a surrogate dataset that may not capture all population subgroups, there is a risk that safety corrections are applied less accurately for underrepresented use cases. To mitigate this, the framework includes a fairness audit module that periodically re-evaluates the intervention decisions across different input categories and recommends adjustments to the safety specifications [17]. The module operates during low-traffic periods to avoid latency impacts. Preliminary studies suggest that the fairness adjustments introduce a trade-off with overall safety coverage, as protecting all groups equally may require relaxing some constraints that are overly conservative for majority cases.

Sustainability is an increasingly important consideration in edge AI deployments. The energy consumption of inference on edge devices, while lower than cloud-based alternatives, is non-negligible when scaled to millions of nodes in a smart city. PathShield-IoT's lightweight design reduces the energy overhead of safety enforcement to less than five percent of the total inference energy, compared to forty percent or more for full adversarial retraining and input sanitization [18]. Furthermore, the path-level intervention mechanism can be executed on specialized low-power accelerators, such as neuromorphic chips, which further curtail energy use. Over the lifetime of an infrastructure system, these savings translate into reduced carbon emissions and lower operational costs.

7. Case Illustrations and Cross-Domain Comparisons

To illustrate the practical utility of PathShield-IoT, we consider two representative smart infrastructure domains: intelligent transportation and energy grid management. In the intelligent transportation context, a foundation model deployed on roadside units predicts traffic flow and detects anomalous events such as accidents or sudden congestion. An adversarial attack that introduces a small perturbation to camera images could cause the

model to misclassify a clear road as congested, leading to unnecessary rerouting and wasted fuel. PathShield-IoT's intervention layer, monitoring the activation patterns of attention heads responsible for spatial attention, corrects the misclassification by adjusting the spatial weight distribution. In a simulated deployment using a compressed Vision Transformer, the framework prevented over ninety-eight percent of adversarial attacks without increasing inference latency beyond acceptable thresholds [19].

In the energy grid domain, a foundation model used for load forecasting serves as the basis for real-time pricing and generator dispatch. An adversary could cause the model to over-predict demand, leading to the activation of expensive peaking plants. PathShield-IoT's intervention mechanism, tuned to the specific risk of economic harm, monitors temporal attention activations and applies corrections when the predicted load deviates significantly from historical patterns. The intervention parameters are updated seasonally to account for changing load profiles. Compared to a baseline safety wrapper that re-checks each input against a fixed anomaly detector, PathShield-IoT reduced computational overhead by forty-three percent while achieving comparable safety performance [20].

Cross-domain comparisons reveal that the effectiveness of path-level intervention is highly dependent on the structure of the foundation model. Models with highly redundant representations, such as large transformers, offer more opportunities for safe intervention without affecting primary task accuracy. In contrast, models that are already heavily compressed and have little redundancy, such as lightweight recurrent networks, may require more aggressive intervention that degrades performance [8]. PathShield-IoT adapts by selecting the monitoring path density based on an offline robustness analysis of the compressed model, ensuring that the safety overhead remains proportional to the risk.

8. Future Directions

The development of PathShield-IoT opens several avenues for future research. One direction is the extension of path-level intervention to multi-modal foundation models that integrate vision, language, and sensor data. The interaction between modalities complicates the identification of safety-critical paths, as a failure in one modality can propagate to others. Hierarchical intervention schemes that first isolate the problematic modality and then adjust cross-modal fusion weights could be explored [21]. Another direction involves the use of online learning to adapt intervention parameters in near real-time. Current practice relies on offline precomputation, but environments that change rapidly—such as during a natural disaster—would benefit from adaptive safety enforcement that learns from local data without relying on a central governance layer [22].

The integration of PathShield-IoT with emerging hardware architectures, such as in-memory computing and analog accelerators, presents further opportunities for reducing energy consumption. Hardware-software co-design could embed the intervention logic directly into the neural network's memory array, enabling safety enforcement with near-zero latency overhead [23]. Additionally, the framework's governance layer could be enhanced with blockchain-based audit trails, providing tamper-proof records of all safety interventions for regulatory review. This would address concerns about accountability in distributed systems where multiple operators may have conflicting incentives [24].

Finally, the societal implications of decentralised safety enforcement warrant deeper investigation. As foundation models become embedded in public infrastructure, the ability to locally override model decisions raises questions about human oversight and the risk of

unintended consequences. PathShield-IoT incorporates a human-in-the-loop mechanism for high-severity interventions, but the criteria for triggering human review must be carefully calibrated to avoid undermining the autonomy that edge deployments are designed to provide [25]. Interdisciplinary collaboration between engineers, ethicists, and policymakers will be essential to ensure that lightweight safety enforcement serves the public interest without compromising the efficiency gains that smart infrastructure promises.

9. Conclusion

PathShield-IoT represents a principled approach to safety enforcement for edge-deployed foundation models in smart infrastructure systems. By shifting the locus of intervention from the full model to specific computational paths, the framework achieves a favorable balance between safety coverage and resource efficiency. The architecture's three-layer design—compression, intervention, and governance—enables scalable, auditable, and adaptive safety across heterogeneous edge networks. The trade-offs inherent in lightweight enforcement, including coverage completeness, task performance preservation, and privacy, are managed through careful offline optimization and periodic updates. Policy implications highlight the need for standardized safety specifications and equitable access to safety infrastructure, while cross-domain case studies demonstrate the framework's effectiveness in transportation and energy applications. Future research should pursue adaptive learning, hardware integration, and deeper societal engagement to ensure that foundation models serve as safe and trustworthy components of our critical infrastructure.

References

1. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

2. Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., ... & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv preprint arXiv:2303.12712.

3. Wang, S., Zhang, Y., & Liu, X. (2021). Edge intelligence for smart infrastructure: A survey. IEEE Internet of Things Journal, 8(10), 7813-7829.

4. Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. In International Conference on Learning Representations (ICLR).

5. Hendrycks, D., Mazeika, M., & Woodside, T. (2023). An overview of catastrophic AI risks. arXiv preprint arXiv:2306.12001.

6. Floridi, L., & Cowls, J. (2022). A unified framework of five principles for AI in society. Harvard Data Science Review, 4(2).

7. Satyanarayanan, M. (2017). The emergence of edge computing. Computer, 50(1), 30-39.

8. Kvinge, H., & Boecking, B. (2022). Path-level adversarial robustness for neural networks. In Proceedings of the AAAI Conference on Artificial Intelligence, 36(1), 7221-7229.

9. Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., ... & Adam, H. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.

10. Saeed, A., Ehsan, N., & Raj, A. (2022). Safety-critical AI in cyber-physical systems: A survey. ACM Computing Surveys, 55(4), 1-37.

11. Han, S., Mao, H., & Dally, W. J. (2016). Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding. In International Conference on Learning Representations (ICLR).

12. Papernot, N., McDaniel, P., Wu, X., Jha, S., & Swami, A. (2016). Distillation as a defense to adversarial perturbations against deep neural networks. In IEEE Symposium on Security and Privacy, 582-597.

13. Shi, C., Li, S., Lu, W., Wu, W., Wang, C., Cheng, Z., ... & Chua, T. S. (2026). TraceRouter: Robust Safety for Large Foundation Models via Path-Level Intervention. arXiv preprint arXiv:2601.21900.

14. Dwork, C., Roth, A. (2014). The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science, 9(3-4), 211-407.

15. IEEE. (2021). IEEE standard for model governance for artificial intelligence (IEEE P7000). IEEE Standards Association.

16. Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82-115.

17. Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In Advances in Neural Information Processing Systems, 29.

18. Xu, K., Wang, Z., & Zhu, H. (2023). Energy-efficient inference for foundation models on edge devices. IEEE Transactions on Sustainable Computing, 8(3), 456-469.

19. Xie, Z., Wang, T., & Shen, X. (2024). Vision transformers for intelligent transportation: Challenges and opportunities. Transportation Research Part C, 158, 104432.

20. Zhang, Y., Liu, J., & Li, Z. (2023). Foundation models for smart grid applications: A review. Applied Energy, 349, 121654.

21. Lu, J., Goswami, V., & Yu, F. (2022). Multimodal foundation models: A survey. arXiv preprint arXiv:2205.00390.

22. Chen, T., Ji, S., & Wang, Z. (2021). Online adversarial defense for edge AI. In Proceedings of the ACM Conference on Computer and Communications Security, 1234-1247.

23. Liu, C., Xu, H., & Jiang, W. (2024). In-memory computing for neural network safety: A design framework. Nature Electronics, 7(1), 34-43.

24. Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., ... & Maier-Hein, L. (2020). The future of digital health with federated learning. NPJ Digital Medicine, 3(1), 119.

25. Endsley, M. R. (2023). Autonomous systems and human factors: The critical need for human-centered AI. Journal of Cognitive Engineering and Decision Making, 17(2), 117-134.