TraceFed: Federated Safety Intervention for Distributed Large Models with Privacy-Preserving Reasoning Monitoring

Francis Rhodes
Department of Computer Science and Engineering, University of Nevada, Reno, Reno, NV, USA.
rhodes1985@unr.edu

Kasper Fleming
Department of Computer Science, University of New Hampshire, Durham, NH, USA.
kasper.work@unh.edu

Yash Mathur
Department of Computer Science, University of North Texas, Denton, TX, USA.
yash1972@unt.edu

Abstract

The deployment of large foundation models in distributed, privacy-sensitive environments introduces unprecedented challenges for safety monitoring and intervention. While centralized safety mechanisms exist, they conflict with the distributed nature of modern machine learning infrastructures, where data and model updates reside across multiple administrative domains. This paper proposes TraceFed, a federated safety intervention framework that enables coordinated, privacy-preserving reasoning monitoring across distributed large models. TraceFed integrates cryptographic secure aggregation with path-level intervention strategies to detect and mitigate harmful outputs without exposing raw model internals or user data. The architecture leverages a federation of safety authorities that collectively maintain a global safety policy while respecting local autonomy. We examine structural trade-offs between intervention granularity, communication overhead, and privacy guarantees, and discuss governance models that balance centralized oversight with decentralized execution. The paper also addresses robustness and fairness challenges, including adversarial attacks on the monitoring infrastructure and distributional biases in safety alerts. Deployment considerations such as latency constraints, auditability, and energy sustainability are analyzed within the context of real-world large-scale systems. Policy implications for regulatory compliance and cross-jurisdictional accountability are explored. Through this work, we aim to establish a foundational framework for safe, privacy-preserving operation of distributed large models, bridging the gap between federated learning principles and modern AI safety research.

Keywords

federated safety; large foundation models; privacy-preserving monitoring; distributed AI governance; path-level intervention; secure aggregation; robust AI systems.

1. Introduction

The rapid proliferation of large foundation models across cloud-edge infrastructures has created a critical need for scalable safety intervention mechanisms that operate under strict privacy constraints. Centralized safety auditing, which requires full access to model weights, activations, and inference logs, is incompatible with many real-world deployments where data sovereignty, proprietary rights, or regulatory mandates prevent consolidation of sensitive information [1]. Federated learning frameworks have demonstrated that collaborative model training can proceed without centralizing raw data, yet analogous approaches for safety monitoring remain underdeveloped [2]. TraceFed addresses this gap by introducing a federated safety intervention system that combines privacy-preserving reasoning monitoring with coordinated, path-level intervention capabilities.

The core premise of TraceFed is that safety monitoring of distributed large models must be decentralized by design, but safety interventions require a degree of global coordination to prevent cascading failures and ensure consistency of protective measures across model replicas. Existing approaches to AI safety, such as reinforcement learning from human feedback [3] and constitutional AI [4], operate primarily at the level of individual model training and fine-tuning. They do not provide mechanisms for real-time, distributed safety enforcement across heterogeneous deployments. TraceFed extends the concept of safety intervention from a static, one-time process to a dynamic, federated service that continuously evaluates model outputs and applies corrective actions through a network of trusted safety authorities.

This paper makes several contributions. First, we present the architectural design of TraceFed, detailing the interplay between local safety agents, federated aggregators, and global policy servers. Second, we introduce a privacy-preserving reasoning monitoring protocol that leverages secure multiparty computation and differential privacy to aggregate safety signals without revealing individual data or model updates. Third, we analyze structural trade-offs inherent in federated safety systems, including the tension between intervention latency and privacy budget, the granularity of intervention actions, and the overhead of cryptographic operations. Fourth, we discuss governance and policy implications, including accountability frameworks for federated safety decisions and cross-border regulatory compliance. Fifth, we examine robustness and fairness challenges, such as adversarial manipulation of safety reports and distributional biases in alerting thresholds. Throughout, we draw on case illustrations from healthcare, finance, and public-sector deployments to ground the discussion in practical constraints.

The remainder of this paper is organized as follows. Section 2 reviews related work in federated learning, AI safety, and privacy-preserving systems. Section 3 describes the TraceFed architecture and design principles. Section 4 details the privacy-preserving reasoning monitoring protocol. Section 5 presents the federated safety intervention mechanisms. Section 6 analyzes structural trade-offs and governance models. Section 7 discusses deployment, sustainability, and policy implications. Section 8 addresses robustness and fairness considerations. Section 9 concludes with future research directions.

2. Background and Motivation

Federated learning has emerged as a paradigm for distributed model training where data remains on local devices and only model updates are aggregated [1, 2, 5]. The same principle motivates the need for federated safety monitoring: safety evaluations should be performed locally without exposing sensitive user interactions or proprietary model internals to a central party. However, safety is inherently a global property; a harmful output from a single node

can have widespread consequences if the model is shared or interconnected [6]. Thus, a federated approach must reconcile local autonomy with global consistency.

AI safety research has produced a range of techniques for aligning model behavior with human values, including fine-tuning with human preferences [3, 7], rule-based constraints [4], and adversarial testing [8]. These methods are typically applied during model development and are ill-suited to post-deployment monitoring where model parameters may be updated asynchronously across many nodes. Moreover, they assume access to raw model activations, which violates privacy requirements in domains such as healthcare and finance. Privacy-preserving inference techniques, such as secure enclaves and homomorphic encryption, can protect individual queries but are computationally expensive and do not scale to the real-time monitoring needs of large language models [9]. Path-level intervention, as proposed in recent work on distributed safety routing, offers a promising middle ground by allowing targeted modifications of model behavior along specific computational paths without revealing the full model [14]. TraceFed builds on this concept within a federated framework.

The motivation for TraceFed arises from concrete deployment scenarios. In a federated healthcare network where each hospital maintains a large language model for clinical decision support, safety monitoring must detect hallucinated drug interactions or biased recommendations without centralizing patient records. Similarly, in a multi-tenant cloud platform hosting foundation models for financial forecasting, safety interventions must respect proprietary trading strategies while preventing market manipulation. These scenarios require a system that can aggregate safety indicators across nodes, infer global risk levels, and issue coordinated intervention commands, all while preserving the confidentiality of local data and model parameters.

3. System Architecture and Design Principles

The TraceFed architecture comprises three hierarchical layers: local safety agents, regional aggregators, and a global policy server. Each local safety agent runs as a lightweight service co-located with the large model instance, responsible for real-time monitoring of model inputs and outputs. The agent applies a suite of safety checks, including toxicity detection, factual consistency validation, and adversarial input filtering, using locally stored reference datasets and policy rules. Results of these checks are encoded into privacy-preserving summaries that are sent to regional aggregators.

Regional aggregators employ secure aggregation protocols to combine safety summaries from multiple local agents without learning individual contributions. The aggregator processes the aggregate to compute region-level safety metrics, such as the proportion of flagged outputs, average risk scores, and temporal trends. These metrics are further aggregated at the global policy server, which maintains a holistic view of system safety across all nodes. The global policy server also houses the safety policy database, containing intervention rules, threshold parameters, and escalation procedures.

A key design principle is separation of concerns: local agents perform fast, low-latency checks for common safety violations, while global policies handle rare or high-severity events that require broader context. This hierarchical structure balances responsiveness with comprehensiveness. For instance, a local agent can immediately block an output containing hate speech without waiting for global coordination, but a subtle adversarial attack that evades local checks may be detected only through cross-node anomaly patterns identified at the global level. The architecture supports dynamic policy updates: new safety rules can be

distributed to all local agents via the global policy server, and the aggregation protocols can be reconfigured to adjust privacy parameters.

Communication between layers is encrypted and authenticated using standard PKI infrastructure. To minimize latency, local agents cache frequently used policy rules and only synchronize with aggregators periodically or upon detection of a safety event. The system is designed to be resilient to node failures and network partitions; each layer implements fallback procedures, such as defaulting to conservative safety thresholds when communication with the global server is lost.

4. Privacy-Preserving Reasoning Monitoring

The core of TraceFed's monitoring capability is a privacy-preserving reasoning protocol that enables the global safety authority to estimate the distribution of harmful outputs across the federated network without learning which specific nodes produced them. The protocol operates on privacy summaries that are constructed using a combination of local differential privacy and secure aggregation. Each local agent perturbs its safety indicators—for example, the count of toxic outputs in a sliding window—by adding calibrated noise drawn from a Laplace distribution, ensuring differential privacy guarantees. The noisy counts are then submitted to the regional aggregator, which uses a secure multiparty computation scheme inspired by Bonawitz et al. [2] to compute the sum without revealing individual contributions. The same approach can be extended to more complex safety metrics, such as embedding similarities or confidence scores, by using randomized response mechanisms.

This privacy-preserving approach introduces a trade-off between accuracy and privacy budget. Higher noise levels protect privacy but reduce the sensitivity of the global safety metrics, potentially delaying detection of emerging threats. TraceFed addresses this through adaptive privacy budgeting: the system allocates a limited privacy budget per time window across multiple safety dimensions, and dynamically adjusts the noise scale based on the severity of observed safety events. For example, during a period of low incident rates, the privacy budget can be reduced to improve accuracy, while a surge in flagged outputs triggers a switch to stronger privacy protection to shield local nodes from scrutiny. This adaptive mechanism is governed by a policy that balances privacy and safety priorities.

Additionally, TraceFed supports a dual-mode operation: a high-accuracy mode for non-sensitive deployments where privacy requirements are relaxed, and a high-privacy mode for regulated environments. In the high-privacy mode, the global policy server can only learn aggregate statistics above a threshold, preventing inference about any single node's behavior below that threshold. This is achieved through thresholded secure aggregation, where the aggregator only reveals the sum if it exceeds a minimum value. The thresholds are set based on the desired privacy level and the minimum safety signal required for intervention.

5. Federated Safety Intervention Mechanisms

Safety intervention in TraceFed occurs at multiple levels of granularity, from local blocking of individual outputs to global model rollbacks. The intervention mechanisms are designed to be minimally invasive, preserving model utility while mitigating harmful behavior. At the local level, the safety agent can apply output filtering, input sanitization, or temporary suspension of the model's inference capability. These actions are taken immediately upon detection of a high-confidence safety violation and are logged for audit purposes. However, local interventions alone may be insufficient for coordinated attacks or subtle biases that only become apparent when aggregated across nodes.

Regional and global interventions are orchestrated by the aggregator and global policy server. When regional safety metrics cross a predefined threshold, the aggregator can issue a regional alert, triggering a review of all nodes in that region. More severe events, such as a sudden spike in toxic outputs across multiple regions, may prompt a global intervention, such as distributing a patch to all local models or temporarily disabling inference on the entire network. The intervention commands are sent as cryptographically signed directives that local agents must obey, with non-compliance logged and reported.

Path-level intervention [14] is a particularly promising technique for federated safety systems. Instead of modifying the entire model, path-level intervention targets specific computational pathways that are responsible for harmful outputs. In a federated context, this requires coordinating the application of intervention masks across nodes while preserving the confidentiality of the model's internal structure. TraceFed implements a federated version of path-level intervention where the global policy server distributes a set of intervention rules encoded as gradient masks or attention block modifications. Each local agent applies these masks only to precomputed safe zones, ensuring that the intervention does not degrade overall model performance. The effect of the intervention is then assessed locally and the aggregated results are used to refine the masks in an iterative feedback loop.

6. Structural Trade-Offs and Governance

Designing a federated safety system involves navigating several fundamental trade-offs. The most prominent is the tension between intervention latency and privacy guarantees. Stronger privacy mechanisms, such as high-distortion differential privacy or secure multiparty computation with large communication overhead, introduce delays that may be unacceptable for real-time safety enforcement. Conversely, relaxing privacy to achieve low latency exposes individual nodes to potential re-identification attacks. TraceFed addresses this by employing a tiered approach: low-latency, low-privacy local checks are used for immediate safety actions, while higher-latency, high-privacy global aggregation is used for strategic oversight. The governance model must define clear thresholds for when each tier is invoked.

Another trade-off exists between central coordination and local autonomy. A fully centralized governance model ensures consistency but creates a single point of failure and may violate jurisdictional data sovereignty. A fully decentralized model enhances autonomy but risks fragmentation of safety policies across nodes, leading to inconsistent protections. TraceFed adopts a hybrid governance structure where a global policy server defines minimum safety standards and escalation protocols, while regional aggregators and local agents retain discretion to implement stricter local policies. This federal governance model aligns with practices in multi-national organizations and is designed to accommodate variations in regulatory regimes, such as the General Data Protection Regulation in Europe and the Health Insurance Portability and Accountability Act in the United States.

Accountability is a critical governance concern. When a safety incident occurs, it must be possible to trace the sequence of decisions without revealing sensitive information. TraceFed incorporates an immutable audit log that records the hashes of all privacy summaries, aggregation steps, and intervention commands, but not the underlying raw data. The audit log is maintained by a consortium of independent auditors that are distributed across jurisdictions. This approach provides transparency without compromising privacy, enabling post-hoc forensic analysis of safety failures.

7. Deployment, Sustainability, and Policy Implications

Deploying TraceFed in production environments requires careful consideration of infrastructure requirements. The system must support heterogeneous hardware ranging from cloud servers to edge devices, each with differing computational and bandwidth capacities. Local safety agents are designed to be lightweight, with memory and compute footprints similar to standard inference accelerators. Regional aggregators require moderate computational resources for secure aggregation, which can be offloaded to trusted execution environments like Intel SGX or AWS Nitro Enclaves to further protect the aggregation process. The global policy server is the most resource-intensive component, but its workload can be distributed across multiple data centers for fault tolerance.

Energy sustainability is an emerging concern, as the cryptographic operations required for privacy-preserving aggregation consume significant power. TraceFed addresses this by optimizing the frequency of aggregation: instead of continuous real-time aggregation, the system uses a batched approach where privacy summaries are accumulated locally and sent in periodic bundles. The batch interval is adjusted dynamically based on the safety risk level, with higher-risk periods triggering more frequent but shorter batches. This adaptive batching reduces the overall energy footprint without compromising safety responsiveness.

Policy implications extend beyond technical design to regulatory compliance. Many jurisdictions are considering or have enacted laws requiring algorithmic accountability and transparency. TraceFed's privacy-preserving design supports compliance by demonstrating that safety monitoring can be achieved without unauthorized access to personal data. However, the federated governance model also raises questions about liability: when a safety failure occurs across multiple nodes, which entity is responsible? TraceFed's audit trail provides a basis for attribution, but legal frameworks must evolve to recognize federated accountability structures. Additionally, the system must ensure that safety interventions do not disproportionately impact underrepresented groups, which requires careful calibration of thresholds and fairness-aware aggregation.

8. Robustness and Fairness Considerations

The robustness of TraceFed against adversarial manipulation is paramount, as attackers may attempt to subvert safety monitoring by crafting inputs that evade local checks or by poisoning safety summaries. The combination of local differential privacy and secure aggregation provides inherent resilience against summary poisoning: an adversary controlling a minority of nodes cannot significantly distort the aggregate without detection, because each node's contribution is bounded and noised. However, a coordinated collusion of many nodes could still bias the aggregate. TraceFed mitigates this through anomaly detection on the aggregate distribution, flagging sudden deviations that exceed expected statistical bounds. If an anomaly is detected, the system can escalate to manual review or temporarily disable the affected nodes.

Fairness considerations arise in the design of safety thresholds and intervention criteria. If thresholds are set uniformly across all nodes, they may fail to account for differences in user demographics, language varieties, or deployment contexts. For example, a toxicity detector trained on standard English may over-flag non-standard dialects, leading to unfair censorship. TraceFed allows per-node customization of safety thresholds within a globally defined range, but this customization must be auditable to prevent intentional discrimination. The system incorporates fairness metrics into the global policy server's monitoring dashboard, enabling administrators to track disparities in flag rates across nodes and adjust policies accordingly.

Another robustness concern is the potential for intervention itself to cause unintended harm. An overly aggressive intervention may suppress legitimate outputs, reduce model utility, or even trigger cascading effects where multiple nodes simultaneously degrade performance. TraceFed introduces a gradual intervention escalation protocol: the first intervention step is always observation with minimal action, followed by targeted path-level modifications, and only as a last resort, full model suspension. Each intervention is preceded by a predictive simulation that estimates its impact on utility, using a lightweight surrogate model. This cautious approach minimizes the risk of over-correction.

9. Conclusion

TraceFed presents a comprehensive framework for federated safety intervention in distributed large models, addressing the critical gap between privacy-preserving operations and effective safety monitoring. By combining local safety agents, regional aggregators, and a global policy server with privacy-preserving reasoning monitoring and path-level intervention, the system enables coordinated safety enforcement without centralizing sensitive data. The analysis of structural trade-offs, governance models, deployment challenges, and fairness considerations provides a roadmap for practical implementation in domains ranging from healthcare to finance. Future research should explore the integration of TraceFed with emerging hardware-based security technologies, the development of formal verification methods for federated safety policies, and empirical evaluations across diverse deployment scenarios. As large models become increasingly embedded in critical infrastructure, the need for scalable, privacy-preserving safety systems like TraceFed will only grow, making this work a timely contribution to both the AI safety and distributed systems communities.

References

1. McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. In Artificial intelligence and statistics (pp. 1273–1282). PMLR.

2. Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., ... & Roth, E. (2017). Practical secure aggregation for privacy-preserving machine learning. In Proceedings of the 2017 ACM SIGSAC conference on computer and communications security (pp. 1175–1191).

3. Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. In Advances in neural information processing systems, 30.

4. Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., ... & Kaplan, J. (2022). Constitutional AI: Harmlessness from AI feedback. arXiv preprint arXiv:2212.08073.

5. Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., ... & Zhao, S. (2021). Advances and open problems in federated learning. Foundations and Trends in Machine Learning, 14(1–2), 1–210.

6. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. arXiv preprint arXiv:1606.06565.

7. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. In Advances in neural information processing systems, 35.

8. Gehman, S., Gururangan, S., Sap, M., Choi, Y., & Smith, N. A. (2020). RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In Proceedings of the 2020 conference on empirical methods in natural language processing (pp. 3356–3369).

9. Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In Theory of cryptography conference (pp. 265–284). Springer.

10. Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., ... & Christiano, P. (2019). Fine-tuning language models from human preferences. arXiv preprint arXiv:1909.08593.

11. Dinan, E., Abercrombie, G., Bergman, A. S., Spruijt-Metz, D., Neff, M., & Prabhumoye, S. (2021). SafetyKit: First aid for measuring safety in open-domain conversational systems. In Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (pp. 4699–4713).

12. Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., & Meger, D. (2018). Deep reinforcement learning that matters. In Proceedings of the AAAI conference on artificial intelligence, 32(1).

13. Hard, A., Rao, K., Mathews, R., Ramaswamy, S., Beaufays, F., Augenstein, S., ... & Ramage, D. (2018). Federated learning for mobile keyboard prediction. arXiv preprint arXiv:1811.03604.

14. 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.

15. Wang, J., Charles, Z., Xu, Z., Joshi, G., McMahan, H. B., & Scdoris, S. (2023). On the robustness of large language models against adversarial examples. In International conference on machine learning (pp. 36402–36418). PMLR.

16. Zhang, C., Xie, Y., Bai, H., Yu, B., Li, W., & Gao, Y. (2022). Federated learning with non-IID data: A survey. IEEE Transactions on Neural Networks and Learning Systems, 33(12), 7050–7069.

17. Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., & Smith, V. (2020). Federated optimization in heterogeneous networks. Proceedings of Machine Learning and Systems, 2, 429–450.

18. Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated machine learning: Concept and applications. ACM Transactions on Intelligent Systems and Technology, 10(2), 1–19.

19. Reddi, S. J., Charles, Z., Zaheer, M., Garrett, Z., Rush, K., Konečný, J., ... & McMahan, B. (2020). Adaptive federated optimization. arXiv preprint arXiv:2003.00295.

20. Bagdasaryan, E., Veit, A., Hua, Y., Estrin, D., & Shmatikov, V. (2020). How to backdoor federated learning. In International conference on artificial intelligence and statistics (pp. 2938–2948). PMLR.