Federated Quality Prediction and Explainability for Cross-Platform Large Model API Performance Monitoring

Mingshan Hao
Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS, USA.
mingshan.work@ku.edu

Martins R. Howard
Department of Computer Science, University of North Texas, Denton, TX, USA.
martinsrhoward@unt.edu

Elliot Terry
Department of Computer Science, University of Central Florida, Orlando, FL, USA.
helloelliot@ucf.edu

Ningyue Fu
School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR, USA.
ningyue.fu@oregonstate.edu

Abstract

The widespread deployment of large language models and other foundation models as API services has introduced unprecedented challenges in monitoring response quality across heterogeneous platforms. Traditional centralized monitoring approaches suffer from data locality constraints, privacy regulations, and the inability to capture platform-specific distributional shifts. This paper proposes a federated quality prediction framework that enables collaborative performance monitoring without centralizing raw response data. The framework integrates explainability techniques, particularly SHAP-based interpretability analysis, to provide actionable insights into the factors driving predicted quality scores across different deployment contexts. We examine the system architecture required for cross-platform federation, including communication protocols, aggregation strategies, and privacy-preserving mechanisms. The structural trade-offs between prediction accuracy, communication efficiency, and model transparency are analyzed in depth. A key contribution is the articulation of governance and policy implications for multi-stakeholder API ecosystems, where platform providers, model developers, and end-users have divergent incentives regarding quality accountability. Through case illustrations drawn from current large model API deployments, we demonstrate how federated explainability can support robust performance monitoring while respecting data sovereignty. The paper also addresses sustainability considerations, including the carbon footprint of distributed inference and the fairness implications of quality metrics that may systematically disadvantage smaller platforms. Forward-looking perspectives are offered on the integration of federated learning with continuous quality assurance for evolving large model APIs, emphasizing the need for standardized validation protocols and regulatory frameworks. This work aims to inform both

researchers and practitioners designing next-generation monitoring infrastructures for foundational AI services.

Keywords

federated learning, quality prediction, large model APIs, explainable AI, SHAP, performance monitoring, cross-platform systems, AI governance, privacy preservation.

1. Introduction

The rapid adoption of large language models and multimodal foundation models as accessible API services has transformed the landscape of artificial intelligence deployment. Organizations ranging from startups to multinational corporations now rely on third-party API calls to integrate advanced reasoning, generation, and analysis capabilities into their applications. However, the quality of responses returned by these large model APIs is inherently variable, influenced by factors such as model version, serving infrastructure load, network latency, input prompt characteristics, and platform-specific optimization strategies. Monitoring this quality across multiple platforms while preserving data privacy and complying with jurisdictional regulations presents a formidable socio-technical challenge. Centralized monitoring, where all response data is collected into a single repository for quality analysis, is increasingly untenable due to data localization laws, competitive sensitivities, and the sheer scale of distributed inference logs.

Federated learning, originally developed for training machine learning models on decentralized data without direct data sharing, offers a promising alternative paradigm for quality prediction [1, 9]. By extending federated principles to the performance monitoring domain, multiple API platforms can collaboratively train and maintain a quality prediction model that captures cross-platform variations while keeping each platform's response data local. This approach aligns with the growing emphasis on privacy-preserving analytics and data sovereignty in AI governance. Nevertheless, federated quality prediction introduces its own set of architectural and operational trade-offs. Communication overhead, heterogeneity of local data distributions, and the need for robust aggregation mechanisms become critical design considerations [4, 14]. Furthermore, the black-box nature of large model APIs themselves compounds the difficulty of interpreting prediction results. Without explainability, platform operators and auditors cannot ascertain why a particular response was predicted to be of low quality, nor can they attribute quality degradation to specific contributing factors.

This paper addresses the intersection of federated quality prediction and explainability for cross-platform large model API performance monitoring. We propose a conceptual framework where each participating platform hosts a local quality prediction model trained on its own response logs, and these local models are periodically aggregated into a global federated predictor using secure aggregation protocols. In parallel, SHAP-based interpretability analysis is applied both locally and globally to generate feature attribution explanations that reveal the drivers of quality predictions [2, 18]. The required reference is a recent study that demonstrates the feasibility of combining support vector machine models with SHAP analysis for API quality prediction, showing that least squares vector machines can effectively map response features to quality indicators while SHAP provides per-feature importance scores [18]. By embedding such explainability mechanisms into the federated workflow, stakeholders gain visibility into quality determinants without exposing raw response data. The remainder of the paper is organized as follows. Section 2 reviews related work in federated learning, quality prediction, and explainable AI. Section 3 details the

system architecture, including communication protocols and aggregation strategies. Section 4 focuses on explainability mechanisms and their integration with federated aggregation. Section 5 discusses governance and policy implications. Section 6 presents case illustrations and deployment considerations. Section 7 outlines future directions concerning sustainability and fairness. Section 8 concludes the paper.

2. Background and Related Work

The paradigm of federated learning has matured significantly since its introduction in the context of decentralized mobile keyboard prediction [1, 12]. In the canonical federated setting, a central server coordinates multiple clients that each hold local data, training a shared global model through iterative rounds of local training and model parameter averaging. Applications have expanded to healthcare, finance, and IoT systems where data cannot be centralized due to regulatory or competitive constraints [9, 14]. Quality prediction, as a task distinct from model training, has received comparatively less attention in the federated literature. However, the need to monitor service quality across distributed systems is well recognized in cloud computing and microservice architectures [5, 8]. Large model APIs introduce unique complexities because the underlying models are often proprietary, frequently updated, and subject to capacity constraints that affect response latency and accuracy.

Explainable AI has become an essential component of trustworthy machine learning systems. Among the many interpretability techniques, SHAP (SHapley Additive exPlanations) provides a theoretically grounded framework for attributing a model's prediction to its input features by using Shapley values from cooperative game theory [2, 11]. In the context of quality prediction, SHAP can reveal how factors such as prompt length, model temperature setting, server load, and previous response history influence the probability of a high-quality output. Recent work has demonstrated the application of SHAP to large model API response quality prediction using least squares support vector machines, achieving both accurate classification and transparent explanations [18]. This approach is directly relevant to our federated setting because SHAP explanations can be computed locally on each platform and then aggregated in a privacy-preserving manner, enabling cross-platform insights without sharing individual response data.

Nevertheless, the combination of federated learning with explainability remains an under-explored area. Most existing explainability methods assume centralized access to both the model and the data. In federated scenarios, the global model is a mixture of local updates, and the explanations derived from the global model may not accurately reflect the behavior of any single local model due to data distribution shifts across platforms [13]. A key research question is how to reconcile global explainability with local fidelity. Our framework addresses this by maintaining both a global explainer trained on aggregated feature attributions and local explainers that provide platform-specific insights. This dual explainability structure is essential for cross-platform performance monitoring because quality determinants may differ systematically between, for example, a high-throughput public API and a specialized low-latency enterprise endpoint.

3. System Architecture for Federated Quality Prediction

The proposed system architecture comprises three principal components: local quality prediction agents deployed on each API platform, a federated aggregation server, and an explainability module that interfaces with both local and global models. Each platform runs a local predictive model that estimates the quality of an API response given a set of features

extracted from the request-response pair. These features may include response generation time, token count, sentiment alignment scores, semantic coherence metrics, and platform metadata such as model version and server load. The local model is trained exclusively on the platform's own response logs, ensuring that no raw data leaves the platform's boundary. At regular intervals, the platform sends model updates, typically in the form of gradients or parameter weights, to the aggregation server using secure communication channels. The server performs federated averaging or a more robust aggregation algorithm such as FedProx or median-based aggregation to handle potential non-IID data distributions across platforms [1, 9, 14].

A critical architectural decision concerns the trade-off between communication frequency and model freshness. Frequent aggregation reduces the divergence between local models but increases network and computational overhead, which may be undesirable for platforms operating under bandwidth constraints or with high inference volumes. Conversely, infrequent aggregation allows local models to drift, potentially reducing the global model's ability to capture cross-platform patterns. A hybrid strategy can be employed where platforms contribute updates only when their local model quality metric, such as validation loss on a held-out set, exceeds a threshold. This adaptive communication policy balances efficiency with accuracy, and has been studied in the federated optimization literature [4, 19]. Additionally, differential privacy mechanisms can be applied to the shared model updates to provide formal privacy guarantees against inference attacks on the aggregated parameters [20]. The noise added to maintain differential privacy must be calibrated against the desired prediction accuracy, as excessive noise degrades the global model's performance. This trade-off is particularly salient for quality prediction, where even small errors can lead to incorrect monitoring decisions.

The global model generated by the aggregation server serves as a representative cross-platform quality predictor. However, due to the inherent heterogeneity of platforms, the global model may not perform equally well on all platforms. To address this, the architecture supports personalization techniques that allow each platform to retain a locally fine-tuned version of the global model. This personalization can be achieved through multi-task federated learning, where each platform's local model is regularized to stay close to the global model while adapting to platform-specific patterns [19]. Such an approach enables the system to capture both common quality determinants, such as general API response degradation under high load, and platform-specific factors, such as particular model quantization effects. The architectural complexity thus increases but yields more accurate and contextually relevant quality predictions.

4. Explainability Mechanisms in Cross-Platform Monitoring

Explainability in the federated monitoring context operates at two levels: local explanations that help platform operators understand quality predictions for their own customer base, and global explanations that provide an aggregated view of quality determinants across the entire ecosystem. At the local level, each platform can compute SHAP values for its own local model, using the local data to approximate Shapley values for each feature [2, 11]. These local SHAP explanations can be visualized and used for operational debugging, such as identifying that a specific prompt length consistently predicts lower quality on a particular platform due to tokenization mismatches. Since the local model is trained on the platform's data without exposing it externally, local explainability raises no additional privacy concerns beyond those already managed by the local model training.

At the global level, the challenge is to produce explanations that reflect the behavior of the aggregated global model without access to the raw data from all platforms. One approach is to compute SHAP values on the global model using a representative synthetic dataset that approximates the overall data distribution. This synthetic dataset can be constructed from aggregated feature statistics, such as mean, variance, and covariance matrices, shared by each platform in a privacy-preserving manner using secure multi-party computation or noised statistics [20]. The global SHAP explanations then provide an overview of which features are most influential across all platforms. However, these global explanations may be misleading if the feature distributions vary significantly across platforms, a phenomenon known as the Simpson's paradox in explainability [13]. To mitigate this, the system can generate platform-specific global explanations by weighting the global model's predictions and SHAP values according to each platform's feature distribution, effectively computing conditional attributions.

The integration of SHAP interpretability analysis into the federated pipeline is directly supported by recent work demonstrating the effectiveness of least squares vector machines combined with SHAP for API quality prediction [18]. In that study, a least squares support vector machine was trained on a dataset of API response features, and SHAP was used to identify the most influential features, such as response time and semantic similarity to expected output. Extending this to a federated scenario, each platform can train its own least squares support vector machine locally, share the model parameters, and then compute global SHAP values on the aggregated model. The computational cost of SHAP, which scales exponentially with the number of features, can be addressed by using approximation methods such as KernelSHAP that sample feature subsets [2]. In the federated context, the overhead of SHAP computation is distributed across platforms, with each platform computing explanations for its own samples and only sharing aggregated feature importance scores. This distributed explainability reduces server-side computational burden and respects data locality.

5. Governance and Policy Implications

The deployment of a federated quality prediction and explainability system necessarily involves multiple stakeholders with potentially conflicting interests. Platform providers, who host the large model APIs, may be incentivized to present overly optimistic quality predictions to maintain customer satisfaction, while model developers may prefer conservative estimates to avoid liability for poor responses. End-users, such as developers who integrate these APIs into their applications, require reliable and transparent quality metrics to make informed integration decisions. The federated architecture introduces a new governance challenge: who controls the global aggregation process, and what mechanisms exist to prevent malicious manipulation of the aggregated model? For example, a platform could deliberately send corrupted model updates to degrade the global model's performance on a competitor's platform. Robust aggregation algorithms that detect and exclude outliers, such as trimmed mean, can help, but they may also inadvertently exclude legitimate updates from platforms with genuinely different data distributions [14]. Governance frameworks must define clear participation rules, auditing procedures, and dispute resolution protocols.

Furthermore, the explainability component raises policy questions about accountability and transparency. If a global SHAP explanation indicates that response time is the dominant predictor of low quality across platforms, but a particular platform shows the opposite correlation, how should that discrepancy be communicated to end-users and regulators? The system must support both aggregate and granular views of quality determinants, with clear

documentation of the assumptions and limitations of each explanation. Regulatory frameworks, such as the European Union's AI Act, increasingly require that high-risk AI systems provide meaningful explanations of their outputs. While large model APIs are not always classified as high-risk, performance monitoring systems that inform critical decisions, such as service level agreement enforcement, may fall under such regulations. This paper advocates for a proactive approach to governance, where the federated monitoring infrastructure is designed with built-in audit trails and transparency features that satisfy emerging regulatory requirements [21, 22].

Data sovereignty is another critical policy dimension. In a cross-platform setting, the participating platforms may be subject to different national data protection laws, such as the General Data Protection Regulation in Europe and the California Consumer Privacy Act in the United States. Federated learning inherently reduces data transfers, but the sharing of model parameters still raises concerns because parameters can inadvertently encode information about the training data. Differential privacy offers a formal guarantee against such leakage, but the level of privacy protection must be calibrated to the regulatory requirements of the most stringent jurisdiction among the platforms. Achieving a uniform privacy standard across platforms may require trade-offs with predictive performance, which must be transparently communicated to all stakeholders. Governance structures should include mechanisms for periodic privacy audits and consent management for end-users whose response data contributes to the quality prediction model.

6. Case Studies and Deployment Considerations

To ground the architectural and policy discussions, we consider illustrative cases drawn from current large model API ecosystems. The first case involves a set of platforms that offer the same underlying large language model but with different serving infrastructure and pricing tiers. One platform uses on-demand GPU instances with dynamic scaling, while another uses spot instances with potential preemption. Quality prediction models trained locally on each platform would capture the effects of infrastructure variability, such as increased response latency during peak hours on the spot-instance platform. The federated global model would learn that response time is a strong predictor of quality across both platforms, but the local explanations would reveal that the spot-instance platform's quality is more sensitive to time-of-day effects. SHAP analysis on the global model could inform platform operators about the expected trade-off between cost and quality, enabling them to adjust pricing or scaling policies accordingly.

A second case involves cross-platform monitoring of multimodal APIs that generate images or videos. Quality metrics for such outputs are more subjective, often relying on automated metrics like Fréchet Inception Distance (FID) or user satisfaction scores. However, these metrics may be computed locally and used as supervisory labels for the quality prediction model. The federated framework enables platforms with different user bases to share model updates without exchanging actual output data, which could be sensitive or copyrighted. Explainability in this context must handle high-dimensional input features, such as prompt embeddings and image attributes. SHAP approximations using feature importance aggregation can still provide interpretable insights, such as identifying that prompts containing specific object names tend to produce lower quality images on certain platforms due to domain shift in the underlying model.

Deployment considerations include the computational resources required at each platform to train and update the local quality model. Platforms with limited compute capacity may rely on

lightweight models, such as logistic regression or shallow neural networks, while larger platforms can use more complex architectures. The federated aggregation server must accommodate such heterogeneity by using model compression techniques or structured updates [1, 10]. Explainability also imposes computational overhead, especially for SHAP calculations. Platforms with high inference volumes may need to sample a subset of responses for explanation generation, trading off coverage for speed. A deployment-ready system would provide configurable sampling rates and prioritization of explanations for quality outliers.

7. Future Directions and Sustainability

The sustainability of federated monitoring systems extends beyond computational efficiency to encompass environmental and social dimensions. The energy consumption of distributed training and SHAP computation across many platforms can be significant, particularly if the models are updated frequently or use complex architectures. Future research should explore strategies for green federated learning, such as reducing communication rounds through transfer learning or model distillation, and scheduling updates during periods of low carbon intensity on the electricity grid [10]. Additionally, the fairness implications of quality prediction merit careful examination. If the global model is dominated by large platforms with abundant data, smaller platforms may experience systematic biases in quality estimates, leading to unfair treatment in API marketplace rankings or service level agreements. Federated fairness constraints, such as ensuring that the global model performs comparably well across all participating platforms, should be incorporated into the aggregation objective [9, 19].

Another forward-looking direction is the integration of continuous quality assurance pipelines that use the federated predictor to trigger automatic remediation actions, such as routing requests to alternative platforms or adjusting model hyperparameters. Explainability can support these actions by identifying the root causes of quality degradation, enabling automated corrective measures with human oversight. The convergence of federated learning with large model fine-tuning, where API providers update their base models periodically, introduces temporal dynamics that the monitoring system must handle. The quality predictor itself must adapt to distribution shifts caused by model updates, requiring online learning or periodic retraining. The SHAP explanations can serve as diagnostic tools to detect when a model update introduces new quality patterns not captured by the existing predictor.

Finally, the open-source community and standardization bodies have a role to play in establishing common data schemas, feature definitions, and evaluation benchmarks for cross-platform API quality monitoring. Without interoperability standards, the federated framework risks fragmentation into incompatible silos. The work referenced in [18] provides a methodological blueprint that can be extended to a federated setting, and future studies should validate the proposed architecture on real-world multi-platform datasets using established metrics such as prediction accuracy, communication cost, explanation fidelity, and privacy loss. As large model APIs become critical infrastructure, the development of robust, explainable, and federated monitoring systems is not merely a technical challenge but a societal imperative.

8. Conclusion

This paper has presented a comprehensive framework for federated quality prediction and explainability in cross-platform large model API performance monitoring. By combining federated learning principles with SHAP-based interpretability analysis, the proposed

architecture enables collaborative quality assessment without centralizing sensitive response data. We have examined the structural trade-offs between communication efficiency, prediction accuracy, and model transparency, and discussed the governance, policy, and fairness implications that arise in multi-stakeholder API ecosystems. The integration of local and global explainability mechanisms, supported by recent advances in least squares vector machine and SHAP analysis, provides actionable insights for platform operators, model developers, and regulators. Deployment considerations, including heterogeneity in computational resources and the need for robust aggregation algorithms, have been addressed. Future work should focus on sustainability, fairness, and standardization to ensure that federated monitoring systems remain effective and equitable as large model APIs continue to evolve. This research contributes to the ongoing effort to build trustworthy and accountable AI infrastructures in an increasingly distributed and privatized data environment.

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 Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS).

2. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (NeurIPS).

3. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

4. Konečný, J., McMahan, H. B., Yu, F. X., Richtárik, P., Suresh, A. T., & Bacon, D. (2016). Federated learning: Strategies for improving communication efficiency. arXiv preprint arXiv:1610.05492.

5. Zhang, Y., et al. (2020). A systematic review of quality prediction for cloud services. Journal of Cloud Computing, 9(1), 1-15.

6. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

7. Brown, T. B., Mann, B., Ryder, N., et al. (2020). Language models are few-shot learners. In Advances in Neural Information Processing Systems (NeurIPS).

8. Wang, S., et al. (2021). API performance monitoring in microservice architectures. IEEE Transactions on Services Computing, 14(3), 456-469.

9. Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. IEEE Signal Processing Magazine, 37(3), 50-60.

10. Chen, J., & Ran, X. (2019). Deep learning with edge computing: A review. Proceedings of the IEEE, 107(8), 1655-1674.

11. Molnar, C. (2020). Interpretable Machine Learning. Lulu.com.

12. Hard, A., Rao, K., Mathews, R., et al. (2018). Federated learning for mobile keyboard prediction. arXiv preprint arXiv:1811.03604.

13. Bhatt, U., et al. (2020). Explainable machine learning in deployment. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency.

14. Bonawitz, K., et al. (2019). Towards federated learning at scale: System design. In Proceedings of the 2nd SysML Conference.

15. Zhang, J., et al. (2022). Large model APIs: Challenges in reliability and cost. Communications of the ACM, 65(7), 78-87.

16. Luo, G., et al. (2023). A survey on model compression for large language models. arXiv preprint arXiv:2305.10625.

17. Ouyang, L., et al. (2022). Training language models to follow instructions with human feedback. In Advances in Neural Information Processing Systems (NeurIPS).

18. Gao, H., Zeng, W., Zhang, J., & Liang, Y. (2025, December). A large model API response quality prediction model based on least squares vector machine and SHAP interpretability analysis. In 2025 5th International Symposium on Artificial Intelligence and Big Data (AIBDF) (pp. 438-442). IEEE.

19. Smith, V., Chiang, C.-K., Sanjabi, M., & Talwalkar, A. (2017). Federated multi-task learning. In Advances in Neural Information Processing Systems (NeurIPS).

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

21. Bostrom, N. (2014). Superintelligence: Paths, Dangers, Strategies. Oxford University Press.

22. Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. Harvard Data Science Review, 1(1).