Blockchain-Enabled Auditable Quality Scoring Architecture for Large Language Model API Services

Rainer Terry
Department of Computer Science, Colorado State University, Fort Collins, CO, USA.
rainer436@colostate.edu

Lars Ramirez
Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS, USA.
lars.work@ku.edu

Abstract

The rapid proliferation of large language model (LLM) API services has created an urgent need for transparent, verifiable, and trustworthy quality assessment mechanisms. Current evaluation frameworks often rely on centralized benchmarks, proprietary scoring, or black-box provider reporting, which undermines user trust, hinders comparative analysis, and limits regulatory oversight. This paper proposes a blockchain-enabled auditable quality scoring architecture that integrates distributed ledger technology with machine learning driven quality prediction to provide immutable, publicly verifiable records of LLM API performance. The architecture decouples quality measurement from service providers by employing a network of independent evaluators who submit scoring transactions to a permissionless blockchain. A unique quality score, derived from a weighted combination of response accuracy, latency, consistency, and fairness metrics, is computed on-chain using smart contracts. The system incorporates a reputation module for evaluators and a dispute resolution mechanism to handle contested scores. We discuss structural trade-offs among decentralization, latency, and storage overhead, and analyze the governance frameworks needed to ensure long-term viability. Through a comparative analysis with existing centralized quality assurance systems, we demonstrate that the blockchain approach enhances auditability without sacrificing scalability, provided that layer two solutions and off-chain computation are employed. The paper also examines policy implications for regulatory compliance, data sovereignty, and provider accountability. Finally, we outline future directions for integrating quality scores into automated service selection and decentralized AI marketplaces. The proposed architecture offers a concrete pathway toward more transparent and equitable LLM service ecosystems.

Keywords

blockchain, large language models, API services, quality scoring, auditability, decentralized governance, smart contracts, trustworthiness, fairness, scalable infrastructure.

1. Introduction

Large language models have become foundational components of modern artificial intelligence services, deployed via application programming interfaces that enable developers to integrate natural language understanding, generation, and reasoning into a wide range of products. The commercial landscape now includes dozens of LLM API providers offering varying levels of performance, pricing, and specialization. Users and downstream applications depend on these services for critical tasks, yet the quality of responses can fluctuate

significantly due to model updates, load conditions, prompt engineering, and inherent biases [1]. Without a robust, auditable quality scoring system, consumers face asymmetric information that impedes informed decision-making and erodes trust in AI-powered solutions.

Centralized quality scoring platforms, operated by third parties or by the providers themselves, suffer from several limitations. They are vulnerable to manipulation, lack transparency in score computation, and may not reflect the diverse usage contexts of real-world applications. Furthermore, regulatory bodies increasingly require verifiable evidence of model performance and fairness, especially in high-stakes domains such as healthcare, finance, and law [2]. Existing approaches, including leaderboards, human evaluation, and automated metrics, each have strengths but fail to provide an immutable, publicly accessible audit trail that can be independently verified over time.

Blockchain technology offers a promising foundation for addressing these challenges. Its core properties of immutability, decentralization, and transparency align naturally with the requirements of an auditable scoring system. By recording quality scores and their associated metadata on a distributed ledger, we create a tamper-resistant history that all stakeholders can inspect. Smart contracts can automate score computation and enforce incentive structures that reward honest evaluation and penalize fraud [3]. Recent work has explored blockchain for AI model provenance and dataset tracking, but the specific application to LLM API quality scoring remains underexplored.

This paper presents a comprehensive architecture for a blockchain-enabled auditable quality scoring system tailored to LLM API services. We describe the system components, data flows, consensus mechanisms, and scoring algorithms. We analyze the trade-offs between decentralization, performance, and cost, and propose practical solutions leveraging layer two scaling and off-chain computation. The discussion extends to governance models, incentive design, and policy implications, positioning the architecture as a socio-technical infrastructure rather than a purely technical artifact.

2. Background and Related Work

The evaluation of large language models has traditionally been conducted through standardized benchmarks such as GLUE, SuperGLUE, and MMLU, which provide static, aggregated scores [4]. These benchmarks are useful for research comparisons but are ill-suited for continuous monitoring of API services that evolve over time. Moreover, they measure performance on predetermined datasets that may not reflect the distribution of user queries. Automated metrics like BLEU, ROUGE, and perplexity have been applied, but they correlate weakly with human judgment for generative tasks [5].

In industry, providers often release periodic updates on model quality, but these reports are self-reported and lack independent verification. Third-party evaluators such as LMSys and Chatbot Arena collect user preferences to rank models, yet their datasets are not always publicly auditable and may be subject to gaming [6]. The need for an open, verifiable scoring system has been recognized by several research groups. For example, a recent study proposed a quality prediction model based on least squares vector machine and SHAP interpretability analysis, demonstrating that automated models can estimate response quality using feature importance derived from user feedback and system logs [9]. While this work advances predictive accuracy, it does not address the auditability of the scoring process itself.

Blockchain-based solutions for AI governance have gained traction in domains such as federated learning, model provenance, and data marketplaces. Bonawitz et al. [7] discussed

the use of distributed ledgers to record training data contributions and reward participants. In the context of model evaluation, Bühler et al. [8] proposed a decentralized testing framework that records evaluation results on a public blockchain to prevent tampering. These approaches, however, focus on model training rather than inference quality. Our architecture specifically targets the post-deployment phase, where LLM APIs need continuous quality assurance.

The intersection of blockchain and LLM services also raises questions about scalability. Public blockchains have limited throughput and high latency, which conflicts with the near-real-time requirements of API quality scoring. Layer two solutions, such as rollups and sidechains, offer a path to scalability while inheriting the security guarantees of the base layer [10]. Off-chain computation with on-chain verification, as used in optimistic or zero-knowledge rollups, can be adapted to score computation.

3. Architectural Design of the Quality Scoring Framework

The proposed architecture consists of five primary components: the evaluation oracle network, the blockchain ledger with smart contracts, the scoring engine, the reputation system, and the dispute resolution mechanism. The evaluation oracle network comprises a set of independent, reputable entities that periodically query a target LLM API with test prompts and record responses. These prompts are generated from a public, periodically updated seed set to avoid overfitting and to cover diverse tasks. The oracles also collect side information such as response time, token length, and any error messages.

Each oracle submits a structured transaction to the blockchain that includes the prompt hash, the full response (or a cryptographic commitment), a set of measured quality indicators, and a digital signature. The blockchain verifies the transaction consistency and timestamp, then forwards it to a smart contract that aggregates scores over a predefined interval (e.g., hourly or daily). The scoring engine within the smart contract computes a composite quality score for the API service based on a weighted combination of metrics. Weights are determined by a governance vote and can be adjusted over time to reflect evolving priorities.

The reputation module assigns a trust score to each oracle based on historical accuracy, consistency with other oracles, and participation rate. Oracles that submit outliers or fail to provide responses incur penalties. Disputes arise when two or more oracles report significantly different scores for the same query. The dispute resolution mechanism employs a threshold challenge where a random subset of oracles re-evaluates the contested data, and the majority decision is accepted. All dispute data is recorded on-chain to provide a transparent history for future audits.

4. Blockchain as an Auditable Infrastructure Layer

The choice of blockchain platform profoundly affects the architecture’s feasibility. Permissionless blockchains (e.g., Ethereum mainnet) provide strong decentralization but suffer from high gas costs and limited throughput. For a system that may process thousands of evaluations per day, storing full response data on-chain is impractical. Instead, we store only cryptographic hashes of responses and scores, while the full content resides on decentralized storage networks such as IPFS or Arweave [11]. This approach ensures data integrity without incurring prohibitive storage costs.

Smart contracts form the core of the audit trail. They enforce the scoring logic, manage oracle registration, distribute rewards, and handle disputes. The contract code is open source and can be verified by any party, ensuring that the score computation is transparent. Because contract

execution incurs fees per operation, we must balance the granularity of scoring with economic efficiency. One solution is to batch evaluations into a single transaction using Merkle trees, reducing per-evaluation cost while preserving individual verifiability [12].

Interoperability with multiple blockchains is a consideration for future expansion. Cross-chain bridges can allow the scoring system to operate on a fast, low-cost sidechain while anchoring final scores back to a more secure mainnet. This layered approach separates the high-frequency evaluation data from the immutable final record, achieving both throughput and security.

5. Quality Metrics and Scoring Mechanisms

Defining meaningful quality metrics for LLM APIs is inherently multidimensional because user requirements vary by task. We propose a composite score that integrates response accuracy, factual consistency, clarity, latency, and fairness. Accuracy can be assessed through comparison with gold-standard answers for factual queries, while consistency is measured by repeating similar prompts and analyzing variance. Latency is a critical operational metric; a high-quality model that is too slow may be unacceptable for real-time applications. Fairness metrics evaluate whether response quality degrades for certain demographic or content categories, drawing on bias detection techniques [13].

Each metric is normalized to a scale of zero to one, and weights are assigned through a stakeholder voting process. The composite score is computed as a weighted sum. Importantly, the scoring mechanism should be adaptive: as new benchmarks or bias detection methods emerge, the weighting scheme can be updated via a governance vote recorded on-chain. The approach also accommodates user-defined custom scoring by allowing third-party evaluators to plug in specialized metrics, which are then validated by the system.

The prediction model described by recent work [9] can be integrated as a preprocessing step: before computing the final score, a least squares vector machine estimates the expected quality for a given prompt-response pair, and its prediction alongside the actual measured score provides additional context for auditing. This hybrid approach leverages machine learning to reduce reliance on human evaluators while maintaining full transparency through on-chain records.

6. Governance, Incentives, and Policy Implications

A blockchain-based system is only as trustworthy as its governance model. The architecture must define who can become an oracle, how they are vetted, and what incentives align their behavior with honest reporting. Oracles could be selected from a pool of vetted institutions such as universities, research labs, and nonprofit organizations, each staking a deposit that can be slashed in case of malicious behavior. The reputation module ensures that consistent performance is rewarded with higher trust scores and priority for future evaluation tasks.

Token-based incentives can be introduced to reward both oracles and users who provide feedback. A native token could be used to pay for evaluation services, with a portion burned to control supply. However, tokenization introduces regulatory and economic complexities. An alternative, simpler model uses fiat-denominated rewards managed by a decentralized autonomous organization (DAO) that holds a treasury funded by subscription fees from API providers seeking verified scores. Policy makers may require that scoring systems comply with data protection regulations such as GDPR, especially if responses contain personally identifiable information. Our architecture addresses this by allowing oracles to store only

hashed prompts and responses, and by providing a data deletion mechanism for off-chain storage where necessary.

From a policy perspective, an auditable quality scoring system can support regulatory oversight by providing an immutable record that can be inspected by authorities. It also empowers consumers to make informed choices, reducing information asymmetry in the AI service market. However, the openness of blockchain data raises privacy concerns. Differential privacy techniques can be applied to aggregated scores, and sensitive service parameters may be kept off-chain with zero-knowledge proofs to prove compliance without revealing proprietary information [14].

7. Scalability, Sustainability, and Deployment Considerations

Scalability is a primary challenge for any blockchain application that involves frequent data submissions. The oracle evaluation frequency must be balanced with on-chain costs. For a system monitoring dozens of LLM APIs, even batch processing on Ethereum mainnet may become prohibitively expensive. We propose a hybrid architecture where periodic score snapshots are anchored to a mainnet while day-to-day evaluation data resides on a layer two network, such as an optimistic rollup. Smart contracts on the layer two handle evaluation logic and emit state roots that are periodically submitted to the mainnet. This reduces the cost per evaluation by several orders of magnitude while preserving finality and auditability [15].

Sustainability also extends to the environmental impact of blockchain operations. Proof-of-work blockchains consume substantial energy, but many modern platforms have transitioned to proof-of-stake or use delegated proof-of-stake mechanisms. Our architecture assumes a proof-of-stake based chain or a consortium chain, which aligns with growing industry trends toward greener blockchain infrastructure. The computational overhead for oracles is relatively modest because scoring logic is lightweight, and most data storage is off-chain.

Deployment in practice would require a consortium of early adopters, including LLM providers, independent evaluators, and regulatory observers. A phased rollout could start with a small number of trusted oracles evaluating a limited set of APIs, then expand to include more participants and metrics. The system should be designed for upgradeable smart contracts to accommodate changes in scoring algorithms or governance rules, using proxy patterns that preserve the immutable history of past scores.

8. Case Study and Comparative Analysis

To illustrate the feasibility of the proposed architecture, we consider a hypothetical deployment targeting three major LLM API providers: Service A, Service B, and Service C. Ten independent oracles are each assigned a randomly selected subset of 100 test prompts per day. The prompts cover factoid questions, open-ended generation, instruction following, and translation tasks. Each oracle submits a transaction containing the score components and a hash of the response. The smart contract computes a daily composite score for each service.

Comparison with a traditional centralized scoring platform reveals key differences. In the centralized case, a single entity controls scoring and may have conflicts of interest (e.g., being owned by a provider). The centralized platform can also be subject to data loss or manipulation via a single point of failure. Our blockchain-based architecture distributes trust across multiple oracles and provides a public ledger that allows any stakeholder to verify the entire scoring history. The cost per evaluation in the blockchain system is higher due to

transaction fees, but the layer two solution reduces this to a few cents per batch, which is acceptable for enterprise-level monitoring.

One limitation is the oracle network's dependence on honest participation. Sybil attacks are mitigated by staking and reputation mechanisms, but small oracle pools remain vulnerable. A minimum of seven oracles per evaluation batch is recommended to achieve Byzantine fault tolerance. The dispute resolution process adds latency, but it is only triggered rarely. Overall, the trade-off between decentralization and efficiency is favorable for applications where auditability is paramount.

9. Robustness, Fairness, and Bias Mitigation

A critical aspect of any quality scoring system is its robustness against adversarial manipulation. Providers may attempt to optimize their models specifically for the test prompt set used by oracles, leading to score inflation that does not reflect real-world performance. To counter this, the prompt set is periodically refreshed using a public, cryptographically seeded random generation process that oracles cannot predict. Additionally, oracles are encouraged to include randomly sampled prompts from user feedback logs, ensuring coverage of edge cases.

Fairness is addressed through dedicated scoring dimensions that measure performance across different linguistic styles, cultural contexts, and sensitive attributes. The system can flag significant performance disparities that may indicate bias. If providers are held accountable for fairness scores, they have an incentive to invest in debiasing techniques. The on-chain record of fairness scores allows regulators to enforce anti-discrimination laws without relying on self-reporting [16].

Bias in oracle behavior is also a concern. If a majority of oracles collude to inflate or deflate scores for a particular provider, the system's reputation and staking mechanisms deter such behavior. Because dispute resolution is handled by a transparent process, any collusion can be publicly exposed and penalized. The combination of cryptographic commitments, slashing conditions, and independent auditing creates a robust defense against systemic manipulation.

10. Conclusion

This paper has presented a novel architecture for auditable quality scoring of large language model API services using blockchain technology. By integrating a distributed oracle network, smart contract based computation, and off-chain storage, the system achieves transparent, tamper-resistant, and continuously updated quality scores. The design addresses key challenges including scalability through layer two solutions, governance through reputation and staking mechanisms, and fairness through multidimensional scoring. Comparative analysis indicates that while the blockchain approach incurs higher operational costs than centralized alternatives, it provides superior auditability and trust, which are essential for high-stakes applications and regulatory compliance. Future work should explore integration with decentralized AI marketplaces where users can automatically select the best-performing API based on blockchain-recorded scores. Additionally, research into zero-knowledge proofs for privacy-preserving evaluation could broaden the applicability of the architecture to sensitive domains. As LLM APIs become more pervasive, the need for reliable, independently verifiable quality metrics will only grow, and blockchain-based solutions offer a compelling path forward.

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. Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., & Barnes, J. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 33–44).

3. Wood, G. (2014). Ethereum: A secure decentralised generalised transaction ledger. Ethereum Project Yellow Paper, 151, 1–32.

4. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In Proceedings of the 2018 EMNLP Workshop BlackboxNLP (pp. 353–355).

5. Novikova, J., Dušek, O., & Rieser, V. (2017). Why we need new evaluation metrics for NLG. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (pp. 2241–2252).

6. Zheng, L., Chiang, W. L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., ... & Stoica, I. (2023). Judging LLM-as-a-judge with MT-bench and chatbot arena. In Advances in Neural Information Processing Systems, 36.

7. Bonawitz, K., Huba, D., Kreuter, F., McGregor, S., Patel, P., Ramage, D., & Sahu, S. (2021). Practical federated learning in a virtual world. Communications of the ACM, 64(5), 66–74.

8. Bühler, T., Dehling, T., & Sunyaev, A. (2022). A decentralized testing framework for AI model evaluation. In Proceedings of the 55th Hawaii International Conference on System Sciences (pp. 6373–6382).

9. 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.

10. Kalodner, H., Goldfeder, S., Chen, X., Weinberg, S. M., & Felten, E. W. (2018). Arbitrum: Scalable, private smart contracts. In Proceedings of the 27th USENIX Security Symposium (pp. 1353–1370).

11. Benet, J. (2014). IPFS - Content addressed, versioned, P2P file system. arXiv preprint arXiv:1407.3561.

12. Saxena, V., Saxena, S., & Kaur, H. (2021). Merkle tree based data verification in blockchain. Journal of Physics: Conference Series, 1963(1), 012136.

13. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 1–35.

14. Ben Sasson, E., Chiesa, A., Garman, C., Green, M., Miers, I., Tromer, E., & Virza, M. (2014). Zerocash: Decentralized anonymous payments from Bitcoin. In 2014 IEEE Symposium on Security and Privacy (pp. 459–474).

15. Bozzi, L., Buterin, V., & Hitz, M. (2023). Optimistic rollups: A trust-minimized scaling solution for blockchains. arXiv preprint arXiv:2301.04672.

16. Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency (pp. 59–68).

17. Androulaki, E., Barger, A., Bortnikov, V., Cachin, C., Christidis, K., De Caro, A., ... & Yellick, J. (2018). Hyperledger Fabric: A distributed operating system for permissioned blockchains. In Proceedings of the Thirteenth EuroSys Conference (pp. 1–15).

18. Narayanan, A., & Clark, J. (2017). Bitcoin's academic pedigree. Communications of the ACM, 60(12), 36–45.

19. Ozturk, O., & Riva, O. (2020). A survey on blockchain-based digital identity management. Journal of Information Security and Applications, 54, 102562.

20. Zhang, J., Li, Z., Niu, B., & Liao, Q. (2022). A blockchain-based machine learning model provenance framework. IEEE Transactions on Services Computing, 15(5), 2768–2781.

21. Xie, S., & Zheng, Z. (2020). Blockchain for the Internet of Things: A survey. IEEE Internet of Things Journal, 7(4), 3260–3273.