Building an Oracle Network for Ethereum to Leverage Large Language Models

Madhukara S. Holla

December 10, 2024

1. Context

Blockchain applications increasingly require off-chain data for smart contracts, but existing oracle solutions are limited to deterministic data sources. With the rise of generative AI models like Large Language Models (LLMs), smart contracts could access nuanced, context-aware insights. However, challenges such as non-determinism in LLM responses, security, and verifiability hinder their direct integration. A decentralized oracle capable of interfacing with LLMs while ensuring consensus and tamper-proof operation can address these gaps.

2. Problem

Smart contracts cannot directly access external systems like LLMs. Existing oracles lack mechanisms to handle the non-deterministic outputs of LLMs or to ensure consensus on such outputs. Moreover, logging operations in a verifiable and immutable way is critical for auditing but often underexplored. The problem becomes more complex with the need for privacy-preserving LLM interactions.

3. Proposed Work

We will implement a novel decentralized oracle system with the following components:

Core Architecture:

- Ethereum-compatible oracle network using PBFT consensus
- Smart contract interface for query submission and response retrieval
- Log storage on IPFS with on-chain hash references
- Response validation and semantic similarity scoring

Implementation Scope:

- Development of oracle node network in Python
  - Integration with LangChain/OpenAI for robust LLM handling

  - Response normalization using NLP libraries
  - Semantic similarity scoring for consensus
- Smart contract system in Solidity for request handling and hash storage
- PBFT consensus implementation with semantic response clustering
- IPFS integration for response log storage

4. Risks

- **Consensus Convergence**
  - Risk: Non-deterministic LLM outputs may prevent consensus
  - Contingency: Implement deterministic sampling with fixed parameters and semantic clustering
- **Cost Efficiency**
  - Risk: High gas costs for on-chain operations
  - Contingency: Use IPFS for storage with minimal on-chain footprint
- **Network Reliability**
  - Risk: Node failures or network partitions
  - Contingency: Implement robust node selection and fallback mechanisms

5. Success

Project success will be measured by:

- Successful consensus achievement with $f = \lfloor (n-1)/3 \rfloor$ Byzantine nodes
- Efficient storage and retrieval of LLM responses via IPFS
- Verifiable response tracking through on-chain hash references
- Sub-30 second response time for LLM queries
- Successful handling of concurrent requests

6. Path to Research Paper

To evolve this project into a research paper:
Additional Work Required:

- Comprehensive performance analysis across multiple chains
- Development of novel consensus mechanisms for non-deterministic data

- Formal security proofs for the oracle network
- Comparative analysis with existing oracle solutions

Research Contributions:

- New consensus protocols for AI-oracle systems
- Security framework for decentralized AI integration
- Performance optimization techniques for cross-chain oracle operations