(Cordex) Decentralized LLM Compute Network: Technical and Economic Overview

April 16, 2025

1 Introduction

This document outlines the technical architecture, incentive structures, and blockchain integration model for a decentralized compute network specialized in Large Language Model (LLM) inference (e.g., LLaMA, DeepSeek). The network leverages blockchain technology, smart contracts, and token-based incentives to provide cheaper and decentralized LLM inference APIs.

2 System Overview

The platform consists of three primary actors:

- **Providers (Nodes):** Supply compute resources hosting specific LLM models.
- **Users:** Consume inference APIs specifying model and performance requirements.
- **Blockchain Layer:** Manages payments, reputation scores, escrow accounts, and resource marketplace logic via smart contracts.

3 Provider Workflow

Providers follow these steps to participate:

1. Setup inference node by hosting standardized containers of supported LLMs.
2. Register nodes on-chain with metadata (models supported, hardware specs, pricing).
3. Periodically submit cryptographic performance proofs to update reputation scores.

4 User Workflow

Users request inference through:

1. Specifying desired model (e.g., "LLaMA 2 13B") and performance criteria (latency, availability).
2. System automatically querying blockchain provider registry for qualified providers.
3. Selecting the most cost-effective provider meeting user requirements.
4. Executing payments via escrow smart contracts.
5. Receiving inference result and providing cryptographic confirmation to release escrow payment.

5 Blockchain and Smart Contracts

The following Ethereum-compatible smart contracts facilitate marketplace operations:

- **ProviderRegistry.sol:** Maintains provider metadata, pricing, and reputation data.

- **InferenceRequestEscrow.sol:** Handles escrow payments from users to providers; manages disputes and settlements.
- **ReputationManager.sol:** Aggregates off-chain oracle data to regularly update provider reputation scores.
- **Token (APICoin ERC20):** Native utility token for payments, staking, and governance.

6 Tokenomics and Economic Incentives

The **APICoin** token serves multiple roles:

- Users pay APICoin tokens for inference requests.
- Providers earn APICoin tokens by delivering inference results.
- Providers stake APICoin tokens to ensure reliability and honesty.
- Token holders govern network parameters and reputation mechanisms through decentralized governance.

7 Off-chain Components

Due to latency considerations, the system integrates off-chain infrastructure:

- **Orchestrator/API Gateway:** Routes inference requests in real-time using periodically updated blockchain data.
- **Decentralized Oracle Network:** Supplies aggregated provider performance data for on-chain updates.

8 Security and Trust Model

Key security mechanisms include:

- Escrow-based smart contract payments minimizing need for user-provider trust.
- Provider staking mechanisms incentivizing honest behavior.
- On-chain reputation system incentivizing high-quality service provision.

9 Implementation Roadmap

Proposed phased development steps:

1. Develop standard node software for provider deployment of LLMs.
2. Deploy initial smart contracts (Provider Registry, Escrow, Reputation Manager) on Ethereum-compatible blockchain or Layer-2 solution.
3. Integrate Orchestrator/API Gateway off-chain service for low-latency inference requests.
4. Implement APICoin token distribution and staking mechanics.
5. Launch decentralized oracle network integration for reputation tracking.
6. Provide developer-friendly SDKs/APIs targeting early adopter communities.

10 Conclusion

This decentralized compute architecture represents an innovative approach to drastically reducing costs and centralization risks associated with current LLM inference solutions. Leveraging blockchain-based economic incentives, smart contracts, and decentralized reputation systems enables a robust marketplace tailored specifically for the rapidly growing AI inference market.