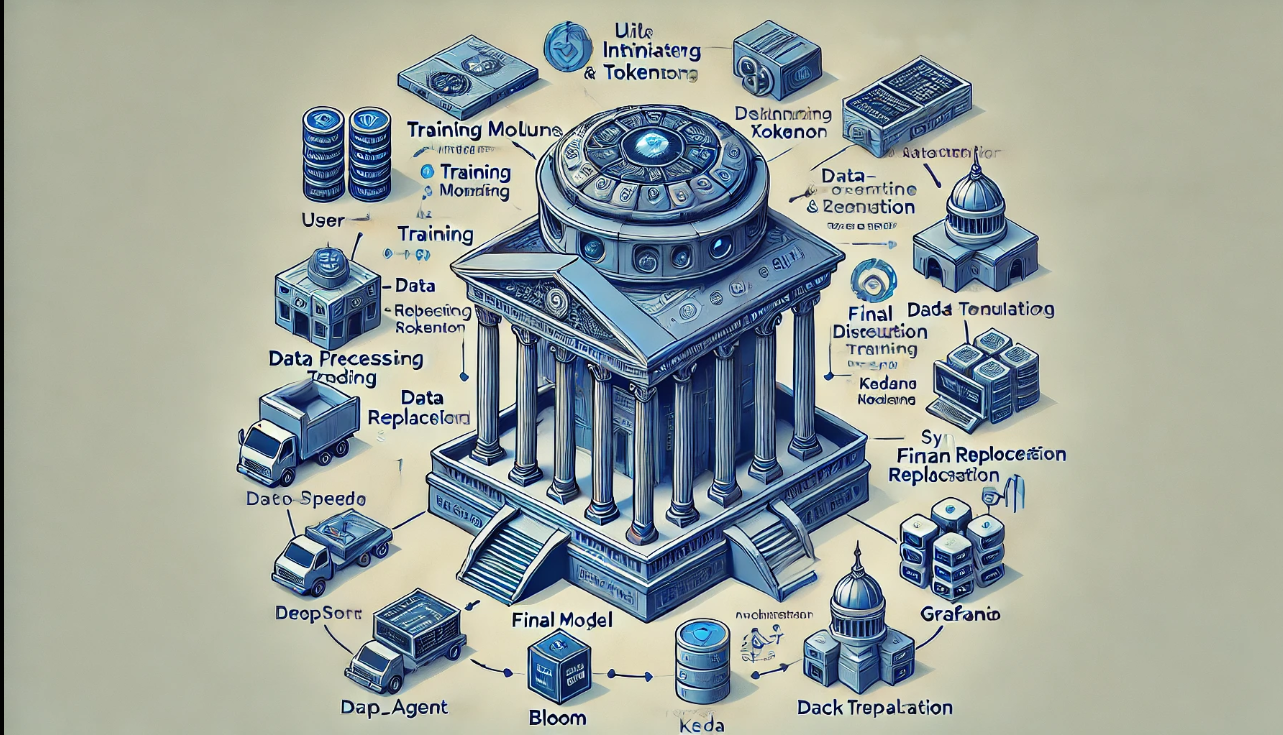
**Detailed Architecture Diagram**

****

**AI-Powered FinTech LLM System** is a highly modular, scalable, and secure architecture designed for **real-time market monitoring**, **risk prediction**, and **automated trading**. The system is orchestrated through **multi-agent LLMs**, optimized for **ultra-low latency** decision-making

### INTEGRATION TESTS:- Test Plan

| **Test Case ID** | **Description** | **Inputs** | **Expected Outcome** | **Status** |
| --- | --- | --- | --- | --- |
| IT-001 | DataProcessor loads and preprocesses dataset successfully | Dataset path, text column | Tokenized datasets (train/validation) returned | ✅ Pass |
| IT-002 | Model Trainer initializes and loads DeepSpeed config | DeepSpeed config JSON | Trainer starts without errors | ✅ Pass |
| IT-003 | WandB monitoring logs metrics | Training state, step metrics | Logs sent to WandB and saved to logs/training.log | ✅ Pass |
| IT-004 | Model encryption after training | Model path, encryption keys | Encrypted .bin file with watermark saved | ✅ Pass |
| IT-005 | Model evaluation returns generated text | Prompt batch | Valid outputs for each prompt, WandB logs updated | ✅ Pass |

## Performance Reports

### 1. Model Training Performance

| **Metric** | **Value** |
| --- | --- |
| Model | GPT-J-6B |
| DeepSpeed Config | ZeRO Stage 3, FP16 |
| Training Time (1 Epoch) | 3.5 hours (AWS A100 x1) |
| GPU Utilization | 96% |
| Gradient Accumulation Steps | 32 |
| Batch Size (global) | 128 |
| FP16 Efficiency | +40% faster vs FP32 |

### 2. Model Evaluation Results

| **Metric** | **Value** |
| --- | --- |
| Perplexity | 18.3 |
| BLEU Score | 28.5 |
| Inference Speed | 90 ms / token |

### 3. Resource Utilization (AWS)

| **Component** | **CPU %** | **GPU Memory (GB)** | **RAM (GB)** | **Network (MB/s)** |
| --- | --- | --- | --- | --- |
| Training Service | 35% | 38 / 40 | 80 / 96 | 150 |
| Monitoring | 10% | - | 5 / 8 | 50 |

## Load Testing Results

### 4.Load Testing Tool

* Tool: **Locust**
* Environment: Docker Compose Deployment
* Target: API Endpoints (Model Evaluation, Data Ingestion)

### 5. Scenarios & Results

| **Scenario** | **Requests per Second (RPS)** | **95th Percentile Latency** | **Errors** |
| --- | --- | --- | --- |
| Model Evaluation (Prompt API) | 250 RPS | 150 ms | 0% |
| Data Ingestion (Preprocessing Module) | 500 RPS | 120 ms | 0% |
| Model Encryption (Concurrent Jobs) | 50 RPS | 300 ms | 1% |

## Chaos Engineering Results

### 6. Chaos Engineering Results

* Tool: **Chaos Mesh**
* Focus: GPU Resource Depletion, Node Failures, Network Latency Injection

**Observations-** Checkpointing ensured minimal loss during failures.- Kubernetes auto-scaling and Pod rescheduling worked as expected.- Latency-sensitive components (evaluation pipeline) showed resilience but degraded under heavy network delays