Wednesday, January 07, 2026 10:39 AM

FEDERATED LLM – AS A SERVICE

OVERVIEW

Federated LLM-as-a-Service is a privacy-first generative AI solution that allows multiple organizations to **use and improve a shared AI model together without sharing their data**. The solution is built on a **pretrained TinyLlama model**, and organizations keep all their sensitive data inside their own systems. Only small, safe model updates are shared, ensuring privacy and compliance.

PROBLEM STATEMENT

Many enterprises want to use Generative AI, but they face major challenges:

- Sensitive business data cannot be shared outside the organization
- Strict compliance and regulatory rules
- High risk of data leaks when using public AI tools

Because of these challenges, organizations are unable to safely use AI models with their internal data.

PROPOSED SOLUTION

Our solution — **Federated LLM-as-a-Service** — enables organizations to collaboratively improve a global LLM **without ever sharing raw data**. Each enterprise trains the model locally within its secure environment, sending only **encrypted, lightweight LoRA updates**. A central aggregator combines these updates to enhance the global model while guaranteeing data privacy and regulatory compliance.

This approach unlocks domain-specific AI capabilities for highly regulated sectors like banking, healthcare, telecom, and public services. It eliminates data-sharing risks, accelerates safe GenAI adoption, and empowers organizations to benefit from collective learning — all while keeping their data fully protected and in-house.

Tech-AI-Thon SOLUTION

Our solution uses a **pretrained TinyLlama model** that runs inside each organization's environment. Instead of training or fine-tuning the model, each organization:

- Uses the pretrained model locally with its own data
- Generates small, lightweight model updates (without sharing data)
- Shares only these updates with a central system

The central system combines the updates to improve the shared model, while **no raw data ever leaves the organization**.

MVP SCOPE

- Use of a single pretrained TinyLlama model
- Simulation of multiple organizations (Fog Layer)
- Local model usage with private sample data
- Sharing only lightweight, non-sensitive updates
- Central combination of updates at Globe Layer
- Simple interface to demonstrate end-to-end flow

BUSINESS VALUE

- Enables safe GenAI adoption in regulated environments
- Eliminates data-sharing and compliance risks
- Reduces cost of training large models independently
- Accelerates domain-specific AI innovation
- Promotes cross-organization collaborative learning

BUSINESS OUTCOMES

- Faster GenAI deployment cycles
- Improved model accuracy for domain-specific use cases
- Reduced legal, security, and compliance risks
- Higher employee productivity with safe AI tools
- Increased trust in enterprise AI systems

TARGET USERS

- Enterprises in regulated industries
- Data science and AI engineering teams
- Compliance and security teams
- Platform and cloud engineering teams

KEY FEATURES AND USE CASES

Key Features

- Uses a pretrained AI model
- No data sharing between organizations
- Privacy-preserving collaboration
- Central model improvement
- Simple and secure design

Use Cases

- Banking: Internal policy assistance and support queries
- Healthcare: Document summarization support
- Telecom: Customer issue understanding
- Public sector: Citizen query assistance

TECH STACK

UI: HTML, CSS, JavaScript, GSAP, Bootstrap, AJAX

Backend: Python (FastAPI), PyTorch model file, LoRA file

Model: TinyLlama is used as it is lightweight, efficient, and well-suited for federated learning in resource-constrained environments.

ARCHITECTURE AND OVERVIEW

The system is designed with two simple layers to keep the solution easy to understand and secure:

Fog Layer (Organization Level)

- Runs inside each organization's secure environment
- Uses the pretrained TinyLlama model locally
- Works only on internal, private data
- Generates lightweight updates without exposing data

Globe Layer (Central Level)

- Receives lightweight updates from multiple organizations
- Combines updates to improve the shared model
- Does not access or store any enterprise data
- Shares the improved model back to organizations

This layered approach ensures privacy, security, and collaboration.

RISKS AND MITIGATION

Risk	Mitigation
Model divergence	Controlled aggregation and learning rates
Security concerns	Encrypted updates and no raw data sharing
Limited model capacity	Use lightweight, efficient models
Integration complexity	Modular and API-driven design

SUCCESS CRITERIA

- No sensitive data shared outside organizations
- Successful use of pretrained model
- Secure sharing of lightweight updates
- Working end-to-end demo
- Clear demonstration of privacy-first AI

FUTURE ENHANCEMENTS

- Support for larger LLMs
- Differential privacy and secure multiparty computation
- Advanced monitoring and audit logs
- Multi-cloud and on-prem deployment support
- Role-based access control and governance
- Auto-scaling federated training nodes