HACKXTREME'26

CUBEAISOLUTIONS Tech Pvt Ltd

The "Nano-Voice" Agent for Smart Commerce

Context: Conventional LLM (Large Language Models) are often too slow and expensive for high-volume voice applications. In the fast-paced world of restaurant reservations and tele-marketing, a 2-second delay in response feels like an eternity to a human caller. Furthermore, businesses need AI that doesn't just "talk," but actually "acts"—syncing instantly with internal databases to prevent double-bookings or outdated sales leads.

The Challenge: Build a high-performance, **Low-Latency Voice Agent** using **Small Language Models (SLMs)**. This agent must handle both **Inbound** (Restaurant Table Reservations) and **Outbound** (Marketing/Lead Gen) calls across multiple regional languages.

Key Objectives for Participants:

- **SLM Optimization:** Utilize a Small Language Model (e.g., Phi-3.5, Mistral, or Qwen) to ensure sub-second response times (latency < 800ms).
- **Real-time Database Synchronization:** The agent must be "State-Aware." If a user books a table for 8:00 PM via voice, the database must update instantly, and the agent must not offer that same slot to a second caller 10 seconds later.
- **Multilingual Fluidity:** The agent should detect the caller's language (e.g., English, Hindi, Tamil, Spanish) from the first sentence and switch its persona and logic accordingly.
- **Agentic Actions:** For outbound calls, the agent must be able to "read" a lead's status from a CRM, attempt the call, and "write" the outcome (e.g., Interested, Not Interested, Follow-up scheduled) back to the system.

Target Impact:

- **Cost Efficiency:** Using SLMs allows for 10x cheaper inference compared to massive models like GPT-4.
- **Zero Missed Revenue:** 24/7 handling of every single inbound call ensures no restaurant reservation is lost to a busy signal.
- **Human-Parity Experience:** Eliminating the "robotic delay" makes automated marketing calls feel like natural conversations, significantly increasing conversion rates.