LTM

Federated Inferencing with BUD

LTM Solution, 2026

Solution Offering

A Larsen & Toubro Group Company

Federated Hybrid LLM Inferencing

"Cloud-Grade AI at the Edge — Without Cloud-Scale Cost."

Enterprise AI, Financial Services, Healthcare, Automotive, Industrial IoT

Business Problem

- High cloud inference cost for enterprise-scale LLM deployments
- Data sovereignty constraints preventing full cloud centralization
- Edge devices unable to deliver cloud-grade reasoning quality
- Network latency impacting real-time AI experiences
- No mechanism to reuse cloud reasoning across distributed devices

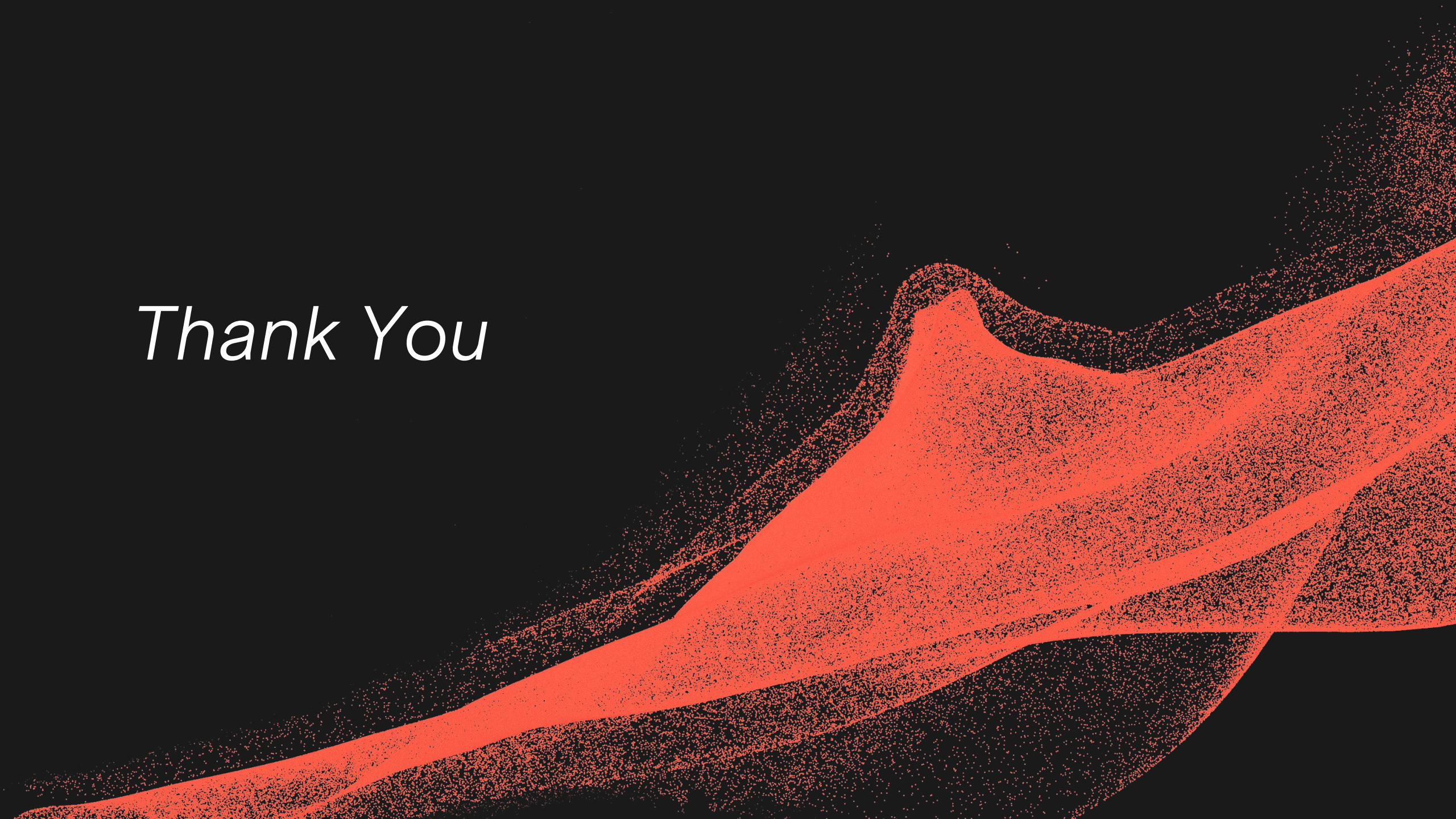
Solution Approach

- Deployed Small Language Models (1M–4B) across edge and on-prem systems.
- Implemented token-level routing using a trained reward model to validate edge-generated tokens.
- Routed only uncertain tokens to a 70B+ cloud LLM running on H200-class GPUs.
- Introduced an actively learning N-Gram cache to eliminate repeat cloud calls.
- Enabled tunable quality–cost control without retraining or redeployment.

NVIDIA + BUD STACK

- **NVIDIA H200 / H100 / B200** – 70B+ cloud LLM inference tier
- **DGX Spark / RTX Workstations** – 600M–4B subject-matter SLM execution
- **RTX A-Series / Jetson Orin** – Context & task expert models at edge
- **BUD Inferencing** – Real-time token verification & inference scaling

Distributed AI environments where latency, sovereignty, and inference cost directly influence user experience, compliance, and long-term AI scalability.


Thank You