Llama Stack

Llama Stack Overview

Llama Stack standardizes the core building blocks that simplify AI application development. It codifies best practices across the Llama ecosystem. More specifically, it provides

- Unified API layer for Inference, RAG, Agents, Tools, Safety, Evals, and Telemetry.
- Plugin architecture to support the rich ecosystem of different API implementations in various environments, including local development, on-premises, cloud, and mobile.
- Prepackaged verified distributions which offer a one-stop solution for developers to get started quickly and reliably in any environment.
- Multiple developer interfaces like CLI and SDKs for Python, Typescript, iOS, and Android.
- Standalone applications as examples for how to build production-grade AI applications with Llama Stack.

Llama Stack Benefits

- Flexible Options: Developers can choose their preferred infrastructure without changing APIs and enjoy flexible deployment choices.
- Consistent Experience: With its unified APIs, Llama Stack makes it easier to build, test, and deploy AI applications with consistent application behavior.
- Robust Ecosystem: Llama Stack is already integrated with distribution partners (cloud providers, hardware vendors, and AI-focused companies) that offer tailored infrastructure, software, and services for deploying Llama models.

Llama 4 Maverick

Llama 4 Maverick is a Mixture-of-Experts (MoE) model with 17 billion active parameters and 128 experts.