Synechron

Synechron

LLM Sandbox

Accelerating Intelligent Workflows via LLM Sandbox

Empowers teams to deploy and manage LLM solutions for higher productivity

The **Background**

Organizations need a scalable, secure, and flexible platform to structure, enrich, and converse with their local knowledge using advanced LLMs.
Existing solutions lack unified access to multiple models and seamless integration with enterprise data sources.

The **Challenge**

Key challenges includes managing large-scale data pipelines, ensuring low-latency model transitions, and maintaining prompt and model standardization across distributed teams. Integrating multiple cloud environments—AWS, Azure, and GCP—along with on-premise infrastructure, demands advanced orchestration, unified governance, and enterprise-grade security to ensure reliability at scale.

The **Solution**

LLMSandbox delivers a robust, API-first framework that integrates seamlessly with AWS Bedrock (Claude, Titan, Nova) & other frontier models to enable unified model access. It supports rapid experimentation, fine-tuning, and real-time observability, empowering teams to build and optimize AI solutions faster. The platform leverages AWS services for embeddings, storage, and scalable compute, while also extending support for Azure and local model deployments to ensure flexibility across environments.

The **Outcome**

LLMSandbox enables unified LLM/SLM access, enhances risk assessment accuracy, and drives operational efficiency across enterprise workflows. AWS Bedrock integration allows seamless scaling and cost optimization, while built-in feedback loops and real-time monitoring ensure continuous improvement in user experience and model performance.

Accelerating Intelligent Workflows via LLM Sandbox

Key Features

Unified Model Access via One API

The platform seamlessly connects to multiple LLM providers, including AWS Bedrock (Claude, Titan, Nova), through a single unified API. It enables rapid switching and comparison of models based on cost, latency, and quality, and supports integration with Azure, Google Cloud, and on-premise models for maximum flexibility across environments.

Scalable Experimentation & Fine-Tuning

The platform empowers teams to A/B test prompts and models, optimizing user experience and productivity. No-code fine-tuning pipelines enable rapid adaptation using user-validated data, while AWS Bedrock managed services streamline model deployment and scalable operations.

Real-Time Observability & Monitoring

The platform provides dashboards for live tracking of token usage, latency, errors, and moderation events. It includes widgets for cache savings, paragraph-level attribution, and retrieval inspection, while AWS cloud infrastructure ensures reliable, scalable monitoring across deployments.

Centralized Prompt Hub & Knowledge Reuse

The platform offers a marketplace for managing, sharing, and reusing prompts across teams and instances, integrated directly into bot creation and instance setup for standardized workflows. It supports prompt versioning and internal knowledge reuse, leveraging AWS-managed storage and compute to ensure scalability and reliability.

Accelerating Intelligent Workflows via LLM Sandbox

Solution Design Approach

Modular Microservices Architecture

LLMSandbox is built on a microservices-based architecture that orchestrates core components including bot creation, context services, and an admin console. AWS Bedrock powers model hosting and inference, while AWS Lambda and Amazon API Gateway enable scalable, event-driven integrationsacross the platform.

Advanced Embedding & Retrieval Pipeline

Documents are split, embedded, and indexed using AWS Bedrock embedding models and Amazon OpenSearch (VectorDB), enabling high-performance Retrieval-Augmented Generation (RAG) for context-aware responses and efficient enterprise knowledge search.

Secure Data Persistence & Storage

The platform leverages AWS-managed storage services such as Amazon S3 and Amazon DynamoDB to manage document libraries, embeddings, and chat histories at scale. Integration with AWS Identity and Access Management (IAM) ensures secure access control, data integrity, and enterprise-grade compliance across all data pipelines.

Automated Deployment & Monitoring

Continuous integration and deployment are orchestrated through Jenkins pipelines and AWS CodePipeline, enabling automated, reliable release cycles.
Real-time analytics and monitoring, powered by AWS CloudWatch, provide end-to-end visibility into model performance, latency, and usage metrics for proactive optimization.

Synechron

Thank you

For more information, please contact:

Partnerships@Synechron.com