CAYLENT

SLMs for Agentic Systems on AWS

Cost-Efficient, Domain-Tuned AI with Automated MLOps

aws PARTNER Premier Tier Services

Right-Size Your AI for Speed, Cost, and Control

Small Language Models (SLMs) can deliver domain-specific intelligence with significantly lower latency and cost than large foundation models. When operationalized correctly, they enable faster iteration cycles, tighter governance, and production-ready performance for targeted business workflows. Caylent helps you design, fine-tune, and deploy SLMs on AWS with automation built in from day one.

An SLM Strategy That Scales with Your Business

Small: Targeted SLM Deployment

Designed for focused AI use cases, this stage enables teams to select, fine-tune, and deploy a SLM for a specific workflow such as document extraction, summarization, or classification. We establish secure endpoints and infrastructure to enable you to quickly prototype. We can either enable you to train SLMs on this infrastructure or train the SLM for you.

Medium: Operational SLM Framework

For teams moving into production or expanding beyond a single use case, we implement repeatable pipelines for SLM training, fine-tuning, deployment, and monitoring. This includes CI/CD for model updates, observability for latency and token usage, and governance controls to ensure reliability across environments.

Large: Enterprise Language Platform

For organizations scaling AI across business units, we design a unified SLM platform with centralized model registry, evaluation frameworks, cost controls, and multi-region deployment support. This enables controlled growth of domain-specific models while maintaining governance, security, and performance at enterprise scale.

Our Approach

01

Discovery & Planning

Through targeted discovery workshops, we evaluate your use case, data readiness, latency requirements, and cost constraints to determine where Small Language Models deliver the greatest value. We define model selection criteria, performance benchmarks, and an SLM deployment roadmap aligned to business outcomes.

02

Design & Implementation

We design and implement an end-to-end SLM pipeline including fine-tuning, optimization, containerized deployment, and CI/CD automation. Our approach embeds evaluation frameworks, observability, and cost controls to ensure production-ready performance.

03

Launch & Enablement

We operationalize SLMs with structured rollout, monitoring, version control, and governance guardrails. This enables reliable performance, auditability, and controlled expansion of domain-specific models across your organization.

www.caylent.com | sales@caylent.com

CAYLENT

SLMs for Agentic Systems on AWS

Deliverables

SLM Strategy & Architecture Design	Define the right Small Language Model approach based on use case, performance targets, cost constraints, and security requirements.
Fine-Tuning & Deployment Automation	Implement parameter-efficient fine-tuning, quantization, containerized deployment, and CI/CD pipelines to optimize accuracy, latency, and infrastructure cost.
Monitoring, Evaluation & Governance	Embed evaluation frameworks, hallucination testing, token observability, drift detection, and version control to ensure reliable, production-grade performance.

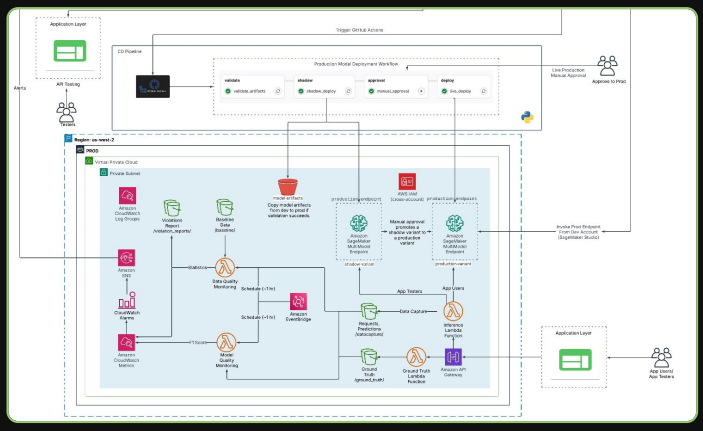
Architecture

Modular SLM architecture leveraging AWS-native AI services

Optimized, containerized inference with GPU/CPU efficiency

Integrated CI/CD for model, prompt, and version lifecycle management

Built-in observability, governance, and cost controls



Customer Stories

SYMMONS®

AI-Powered Water Anomaly Detection

Symmons partnered with Caylent to build an AI platform on AWS that analyzes IoT water system data in real time. Using automated MLOps pipelines, the solution enables predictive maintenance, improves leak detection, and helps prevent costly water damage.

Manufacturing

whatnot

Scaling Production AI on AWS

Whatnot partnered with Caylent to modernize and scale their ML operations on Amazon SageMaker, ensuring reliable model deployment, monitoring, and production support for high-growth streaming workloads.

Media & Entertainment

SmartComms Inc.

Operationalizing AI for Financial Services

A customer communication platform partnered with Caylent to modernize its AI infrastructure on AWS, enabling streamlined model deployment, improved reliability, and greater operational efficiency in regulated environments.

Financial Services

*Anonymized

Let's Get Started

No matter where you are on your journey, Caylent can help you quickly achieve your AI goals, from reinventing customer experiences and creating innovative new applications to massively improving productivity.

Whether you are looking to define a use case, create a roadmap, build a prototype, or implement a production ready solution, Caylent has a suite of offers tailored specifically to accelerate your goals.

Scan the QR code to talk with our specialists

www.caylent.com | sales@caylent.com