CAYLENT

LLMOps on AWS

Powered by NVIDIA

Amplifying Impact Through Model Mobility

AI Workloads Stalled by Integration Challenges

Organizations managing large-scale AI workloads face mounting challenges in achieving and scaling value from their AI initiatives. *With 74% of companies struggling to extract meaningful outcomes*, barriers such as inefficient workflows, regulatory complexity, and disconnected systems stifle innovation. Without a scalable and optimized LLMOps approach, enterprises risk missing out on transformative insights, operational efficiency, and competitive advantage, underscoring the urgent need for an integrated AI solution.

Bringing Clarity and Simplicity to AI Operations

Our solution streamlines AI workload management by leveraging NVIDIA's NeMo Framework, NIMs, and Amazon SageMaker. This multi-model approach empowers enterprises to scale efficiently across industries, optimize performance, and maintain compliance. By integrating industry-specific tools with AWS capabilities, organizations unlock transformative insights, accelerate innovation, and ensure long-term success

Source: BCG AI Adoption Survey 2024

Industry NIMS

Financial Services

- Enable multi modal workstreams (*neva-22b*)
- Accelerate training with synthetic data *(Nemotron)*
- Support branch digitization *(Llama-3.2)*
- ***Blueprint*** *- Multimodal PDF Data Extraction*

Healthcare & Life Sciences

- Accelerate drug discovery (diffdock)
- Run real-time AI apps like medical diagnostics to hands-free clinical assistance (palmyra-med-70b)
- ***Blueprint*** *- Build A Generative Virtual Screening Pipeline*

Media & Entertainment

- Enhance digital experiences with precise audio-to-face mapping (Audio2face-3d)
- Ensure forward-facing presentations with seamless media editing (Eyecontact)
- Transform text into audio (Radtts-hifigan-tts)
- ***Blueprint*** *- Build a Video Search and Summarization Agent*

Learn More

Highlights

- Manage multiple models with NeMo and SageMaker for optimized AI operations
- Leverage NIMs to create scalable solutions for Financial Services, HCLS, and Media & Entertainment customers
- Achieve sustained innovation through continuous evaluation and cost-performance optimization
- Maintain regulatory compliance with secure, efficient workflows on AWS infrastructure

Deliverables

- Workshops to align business goals and define LLMOps requirements
- Roadmap design to outline the strategic steps for implementation and scaling
- Architecture diagrams and documentation, including workflows and testing plans
- Deployment and support for multi-model ecosystems

aws PARTNER Generative AI Competency

aws PARTNER Premier Tier Services

www.caylent.com sales@caylent.com