Aditya Dey

Backend Developer · AI / ML Engineer

+91 79085 31553 | adey9217x@gmail.com | linkedin.com/in/aditya-dey | github.com/AditHash | adithash.in

Professional Summary

Backend and AI/ML Engineer with **2.5 years** of experience building and shipping production RAG systems, document intelligence pipelines, and agentic AI applications on AWS — deployed to production across ECS and EKS. Hands-on across LLMs, SLMs, and VLMs including fine-tuning and inference deployment. Focused on end-to-end delivery from LLM integration and vector search to scalable microservices. AWS-certified in Machine Learning; leads GenAI development and mentors junior engineers.

Experience & Projects

Backend Developer — AI **Dec 2023 — Present**

Workmates Core2Cloud Solutions Limited · Kolkata, India

- **High-Volume Insurance Document Intelligence:** Architected async FastAPI microservice ingesting **2,025,300+ documents** and processing **422,800+ claims**. Integrated AWS Textract with Surya and Paddle OCR into MongoDB. Delivered RAG-based QA chatbot on AWS Bedrock + Qdrant; containerised and deployed on AWS ECS/ECR.
- **Multi-Template Invoice & Bill Extraction Pipeline:** Built async FastAPI service extracting structured fields using AWS Textract AnalyzeExpense with Bedrock-hosted LLM normalisation layer. Persisted validated output to MongoDB with confidence-scored human-in-the-loop review, replacing manual data entry.
- **Scalable Policy Recommendation Engine:** Engineered **sub-100 ms** retrieval across **11,000+ policy records** using PostgreSQL + pgvector. Implemented hybrid search with cross-encoder re-ranking, secured with RBAC and rate-limiting on AWS ECS.
- **NBFC Loan Advisory Chatbot — WhatsApp:** Built customer-facing conversational AI over Gupshup Business API for real-time loan and account queries. RAG pipeline on AWS Bedrock with multi-turn session management; served via FastAPI on EC2 (PM2 managed).
- **AI-Driven Infrastructure Monitoring Agent:** Autonomous incident-response system using LangGraph, AWS Lambda, and Boto3 monitoring CloudWatch alerts on AWS EKS. Full observability via LangSmith, Langfuse, and Grafana for tracing and bottleneck detection.
- **Video Intelligence Chatbot:** Multi-modal platform generating full and timestamp-segmented video summaries, extracting insights, and supporting natural-language QnA over video. VLM-backed frame analysis with vector-indexed transcripts for time-anchored retrieval; ffmpeg for automated highlight and summary generation.

Skills

Languages: Python, C++, SQL

AI & ML: RAG Systems, Agentic Workflows, Vector Embeddings, LangChain, LangGraph, Scikit-learn, PyTorch, TensorFlow, LLMs, SLMs, VLMs, LoRA / QLoRA Fine-Tuning, Model Deployment & Inference

Agentic Tooling: Model Context Protocol (MCP), Claude Code, AI Agent Automation, Prompt Engineering

MLOps & Observability: Docker, AWS ECS, AWS EKS, Kubernetes, LangSmith, Langfuse, Grafana, MLflow

Databases: PostgreSQL, pgvector, MongoDB, Qdrant, ChromaDB, OpenSearch

Cloud (AWS): Bedrock, SageMaker, Textract, ECS, EKS, Lambda, S3, CloudWatch

Backend: FastAPI, Flask, Pydantic, Apache Kafka, REST APIs, Microservices

Engineering: GitHub Issues & Milestones, Conventional Commits, PR Workflows, Code Review, Semantic Versioning

Education

Techno India University, Kolkata *B.Tech — Computer Science & Engineering, 2024*

GPA: **8.46 / 10**

Certifications

AWS Certified Machine Learning Engineer – Associate	Dec 2024
AWS Certified AI Practitioner	Sep 2024
MongoDB SI Associate	Sep 2025
AWS Certified Cloud Practitioner	Feb 2024

Awards & Recognition

Rising Star — AI & **AI Innovation Spark** *Workmates Annual Business Summit, 2026*