Yogesh Abnave

Generative AI Engineer

+91 7350520555
abnave.yogesh@outlook.com
https://yogesh-abnave.web.app/
https://www.linkedin.com/in/yogesh-abnave-85b7191a7/

Professional Summary

To obtain a position as a Senior Generative AI Engineer with **7.5** years of experience in the IT industry, specializing in working with AWS environments.

Core Competencies

- **Gen AI:** RAG | LLMs | AI Agent | AgentCore | Transformer | Diffusion Models | Guardrails | Ollama | Python
- **Cloud:** SageMaker AI | Bedrock | Lambda | Terraform | Docker | ECS/Fargate | EC2 | VPC | IAM/KMS | EKS
- **Deployment**: Solo-deployed Amazon Nova on AWS with guardrails for a secure, scalable architecture; built scalable data pipelines from ingestion to visualization in Amazon QuickSight.

Roles & Responsibilities

CloudAge Global IT Services LLC | Sr. Generative AI Engineer — Sept 2024 - Until dates

Project: AgentCore Ai Shopping Platform | Client: NexaCorp

- The TradeIQ Shopping Agent Platform enabled e-commerce businesses to deploy intelligent shopping assistants through multi-agent architecture. Built on AWS using Bedrock, Claude 4.5 Sonnet, React, and MCP protocol, it delivered personalized product recommendations with secure Visa payment processing, reducing customer acquisition costs by 40% and increasing conversion rates by 25% through real-time conversational AI.

Project: AI-Powered Assignment Automation | Client: Duolingo

- Architected and deployed a containerized AI application on ECS Fargate, implementing a secure and scalable infrastructure comprising ALB, VPC Endpoints, and DynamoDB. Automated CI/CD pipelines with S3, ECR, and CloudFormation, while integrating Amazon Nova Pro and Nova Canvas for intelligent assignment generation and interactive content delivery.

Project: RAG Text Generation using Ollama | Client: Defontana

- Designed and deployed TitanRAG, a production-grade, high-performance Retrieval-Augmented Generation (RAG) platform for DeltaMind Analytics using Ollama to orchestrate multiple LLMs, including Llama 3-8B, Mistral, and Phi-3, on AWS GPU based EC2 instances. Built a secure VPC architecture with DynamoDB and OpenWebUI, enabling private, low-latency, multi-model inference with enterprise-level scalability, security, and cost optimization

OTS Solutions Pvt. Ltd. | Software Engineer — Jul 2022 - Aug 2024

Project: VisionFlow | Client: Jefferies

- Built a production-grade generative AI system for Synapse Vision Labs (USA) using a secure AWS architecture integrating EC2 GPU instances, SageMaker, Lambda, and S3 for dynamic Stable Diffusion model loading, enabling scalable inference, version control, and cost-efficient, low-latency AI image generation.

Project: CloudShift: AWS to OCI Migration | Client: Astra

- This project involves migrating AWS workloads, including EC2, S3, RDS, and networking setups, to Oracle Cloud Infrastructure (OCI) with minimal downtime. The migration ensures optimized performance, secure configuration, and automated deployment using modern cloud tools

TechnoGrowth Software Solutions Pvt. Ltd. | Software Developer — Feb 2021 - Jul 2022

- Built continuous integration and deployment automation using Jenkins and Docker technologies, integrated with AWS S3 and EC2 services, resulting in accelerated and more dependable software releases
- Enhanced content delivery performance and scalability through AWS CloudFront integration, supported by FastAPI and MongoDB backend architecture
- Delivered a production ready Artificial Intelligence image recognition solution on AWS cloud platform, combining Angular and Node.js technologies to enable high performance real time image analysis.
- Implemented secure role based access control with AWS IAM and Secrets Manager, improving system reliability by 15% and ensuring compliance.

AllOps Technologies Pvt. Ltd. | Software Developer — Aug 2018 - Jan 2021

- Engineered and maintained React applications using modern features like Hooks and Context API, improving development speed and ensuring scalable, maintainable front End code.
- Created role-based access controls and optimized database solutions, contributing to improved system reliability.
- Built a food delivery platform on the MERN stack with REST APIs and MySQL, delivering a dynamic user experience and strengthening skills in application development.

Education

B.E. in Computer Engineering, 2018

Certifications

AWS Certified Solutions Architect - Professional
Oracle Cloud Infrastructure 2025 Certified Generative AI Professional
Azure (AZ-900)