Yogesh Abnave

Generative AI Engineer

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

Professional Summary

To obtain a position as a Senior Generative AI Engineer with **7.1** years of experience in the IT industry, specializing in working with AWS environments.

Core Competencies

- **Gen AI:** Transformers | Hugging Face | Diffusion Models | RAG | LangChain | AgentCore | Ollama
- **Cloud:** SageMaker AI | Bedrock | Lambda | CloudFormation | Docker | ECS/Fargate | VPC | IAM/KMS
- **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

AI-Powered Assignment Automation | 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, CodeBuild, ECR, and CloudFormation, while integrating Amazon Nova Pro and Nova Canvas for intelligent assignment generation and interactive content delivery.

Fan Engagement | Deloitte

- Engineered real-time, interactive dashboards (Combine IQ, Draft IQ) serving over 1M fans by orchestrating data pipelines from S3 through Lambda, with visualizations powered by QuickSight. Integrated Amazon Q Business with GenAI chatbots, enabling conversational fan queries using football-specific language with governance guardrails for secure, scalable engagement.

Brain Knowledge Hub | Deloitte

- Developed an AI platform on AWS SageMaker AI to train and deploy models like scVI, GNNs, BioBERT, and ResNet for brain research, generating insights from large biological datasets, visualized in QuickSight and accessible via an Amazon Q chatbot.

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

RAG Text Generation using Ollama | Jefferies

- Pioneered a bare-metal RAG solution on AWS EC2 GPUs to deliver high-fidelity text generation, reducing inference latency to <100 ms and improving answer accuracy by 35% through fine-tuning Llama 3-8B via Hugging Face. Orchestrated the semantic search and generation pipeline with LangChain and ChromaDB, containerized with Docker for portability. Managed infrastructure securely at scale using VPC, IAM, and CloudFormation templates, enabling reliable one-click deployments

CloudShift: AWS to OCI Migration | 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

- Automated CI/CD automation streams using Jenkins and Docker, integrated with AWS S3 and EC2, enabling faster and more reliable deployments.
- Implemented secure role-based access control with AWS IAM and Secrets Manager, improving system reliability by 15% and ensuring compliance.
- Optimized content delivery and scalability using AWS CloudFront, integrated with FastAPI and MongoDB for backend services.
- Developed and deployed an AI-powered image recognition system on AWS (EC2, S3) leveraging Angular and Node.js, enabling real-time image processing at scale.

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

Certification

Azure (AZ-900)