Sahil Padyal

Software Engineer - AI

sahilpadyal237@gmail.com | +1 (857) 230-6240 | New York, NY | www.linkedin.com/in/sahil-padyal/

SUMMARY

Software Engineer specializing in AI with 5 years of hands-on experience developing scalable systems for LLM fine-tuning, generative AI, RAG pipelines, and cloud-native backends. Demonstrated success in building AI agents, chatbots, and domain-specific models using Python, Java, LangChain, and Hugging Face Transformers, with deployments on AWS (SageMaker, Lambda) and GCP (Vertex AI). Proficient in designing microservices, REST/GraphQL APIs, distributed systems, and CI/CD pipelines with Docker and Kubernetes. Strong background in prompt engineering, NLP, 3D optimization tools, and Agile/Scrum practices.

SKILLS

Languages: Python, Java, JavaScript, TypeScript, SQL, Bash
Frameworks & Libraries: Spring Boot, ReactJS, NodeJS, NextJS, FastAPI, Flask, LangChain, LangGraph, Hugging Face Transformers, Scikit-learn, Pandas, NumPy
AI: Natural Language Processing (NLP), Large Language Models (LLMs), Prompt Engineering, RAG (Retrieval-Augmented Generation), Model Fine-tuning, Embeddings (BERT, SentenceTransformers), Vector Databases (FAISS, Pinecone, ChromaDB, Weaviate)
Cloud & DevOps: Google Cloud Platform (Vertex AI, Cloud Run, Cloud Functions, Cloud Storage, Cloud Build), AWS (S3, Lambda, EC2), Docker, Kubernetes, GitHub Actions, GitLab CI
Monitoring: MLflow, LangSmith, Ragas, Prometheus, Grafana
APIs: RESTful APIs, GraphQL, WebSockets
Databases: PostgreSQL, MySQL, Firestore, Redis, MongoDB
Frontend: ReactJS, NextJS, TypeScript, HTML5, CSS3, Redux, TailwindCSS
Tools & Platforms: Git, GitHub, GitLab, Postman, Swagger, Jupyter Notebooks, VSCode, Jira, Confluence
Operating System: Windows, Linux

WORK EXPERIENCE

Humanitarians AI, USA | Software Engineer - AI **Jan 2025 - Current**

- Contributing to open-source project for Stellis Labs to build custom fine-tuned domain-specific LLM models for reasoning and generation tasks.
- Developing a modular pipeline to fine-tune smaller LLMs like Llama3-1B and DeepSeek-1.5B for high-accuracy domain-specific reasoning.
- Designed and implemented a secure, two-click deployment workflow enabling users to launch fine-tuned models to their own AWS SageMaker instances directly from Hugging Face Hub using CloudFormation, OIDC-based role assumption, and SageMaker SDK, ensuring full customer ownership of artifacts, billing, and security boundaries.
- Designed and deployed secure APIs using Python FastAPI/Pydantic on AWS, integrating services like Lambda, API Gateway, Cognito, and CDK to support user access and metadata management.
- Implementing a Retrieval-Augmented Generation (RAG) module to synthesize a domain-specific database of 5000 entries using open-source LLMs like Llama3-70B deployed via Groq API.
- Leveraging tools such as Hugging Face Transformers, PEFT, Unsloth, and MLflow for efficient, low-latency fine-tuning and experiment tracking.
- Building RESTful APIs using Flask and managing backend storage with SQL databases to store metadata, user interactions, and embedding vectors.
- Conducted prompt engineering experiments and evaluated output with BLEU, ROUGE, and embedding-based similarity metrics for generation quality benchmarking.

Matrix Rental Solutions, USA | Software Engineer Intern (Generative AI) **May 2023 - Dec 2023**

- Engineered a high-performance, scalable AI agent-powered chatbot using LangChain and PaLM-2 LLM on GCP Vertex AI, enabling dynamic memory, tool usage, and reasoning to drive 40% higher user engagement.
- Improved context comprehension by 60% through a custom preprocessing pipeline for prompt templating, token filtering, and conversational history embedding.
- Increased model performance by fine-tuning on 10,000+ annotated prompts, improving handling of vague/ambiguous queries by 46%.

- Evaluated retrieval accuracy using Ragas, improving grounding and answer faithfulness by 35% through retriever tuning and prompt optimization.
- Integrated LangSmith to trace LLM behavior, monitor chain performance, and diagnose tool failures, reducing hallucination rates by 30%.
- Integrated moderation layers using Google Perspective API and custom filters to flag toxic or inappropriate queries, reducing offensive outputs by 90%.
- Reduced data retrieval time by 20% via optimized microservice communication using RESTful APIs (FastAPI/Flask) and GraphQL, integrated into a scalable Python backend.
- Enhanced CI/CD efficiency by 40% through automated deployment pipelines using Docker, Cloud Build, and GCP Cloud Run, ensuring continuous integration with QA gates.
- Collaborated with frontend engineers to integrate the chatbot with a ReactJS + TypeScript interface, delivering seamless UI/UX across web and mobile platforms.

Convrse.ai, India | Software Engineer Backend **Nov 2019 - June 2022**

- Spearheaded 0 to 1 development of a 3D mesh optimization tool using Quadric Error Metrics (QEM) to intelligently reduce polygon count while maintaining visual fidelity.
- Architected a distributed system with Ray and Redis for task scheduling and caching, deployed on Kubernetes (EKS) to support 100K+ daily active users, highlighting scalable backend system design.
- Optimized relational and non-relational data performance via RDS PostgreSQL and MongoDB, ensuring efficient schema design and transactional consistency.
- Developed CI/CD pipelines using AWS CodeBuild, Argo CD, and GitHub Actions, achieving 50% faster deployments and reducing manual deployment errors by 60%.
- Migrated from EC2 to AWS Fargate, cutting infrastructure costs by 15% and improving elasticity for compute-intensive workloads.
- Implemented a secure file distribution layer using AWS Lambda and S3 with signed URLs/CDN integration, ensuring integrity and access control.
- Built robust Fast APIs for client-server communication, supporting flexible queries while reducing frontend over-fetching and under-fetching problems.
- Containerized microservices with Docker, defined scalable Helm charts, and orchestrated rollouts on Kubernetes, reinforcing reliability in production.

Hexaware Technologies, India | Software Engineer **Jan 2019 - Oct 2019**

- Developed a scalable 3D SaaS inventory platform for real estate builders using Java Spring Boot, ReactJS, and WebGL, driving $500,000 in initial revenue and significantly improving 3D visualization and customer engagement.
- Ensured application reliability by developing JUnit test suites, improving backend robustness and reducing post-deployment bugs.
- Designed and implemented RESTful APIs to facilitate seamless interaction between the frontend and backend, enabling real-time inventory updates and enhanced system responsiveness.
- Utilized MySQL for relational database management, ensuring data consistency, integrity, and scalable storage for inventory records and 3D asset metadata.

EDUCATION

Master's in Information Systems
Northeastern University, Boston, MA