

SAJID SHAIKH

(213) 756-9690 | sajidsha@usc.edu | in/connectwithsajid | connectwithsajid.github.io

SUMMARY

Software Engineer specializing in AI infrastructure, distributed systems, and production backend services. Experienced in architecting scalable LLM applications, retrieval pipelines, cloud-native APIs, and data platforms with a focus on reliability, performance, observability, and operational excellence.

EXPERIENCE

Visic (Pre-Seed Startup)

Jan 2026 – Present

Founding AI Engineer

San Francisco, CA

- Architected an end to end production AI platform by decoupling document ingestion, summarization, and conversational inference into independently deployable services using FastAPI, Docker, PostgreSQL, and Azure OpenAI.
- Engineered hybrid RAG system powering municipal AI chatbot by combining PostgreSQL results with ChromaDB vector search, implementing asynchronous indexing, content-hash deduplication, and low-latency semantic retrieval for context-aware responses.
- Operationalized cloud-native AI services on Azure by integrating Docker, GitHub Container Registry, Key Vault, and environment driven deployments while designing versioned APIs, session management, and secure authentication for production-ready AI applications.

Kognitic, Inc

Jul 2025 – Aug 2025

Data Pipeline Analyst - AI

New Jersey, USA

- Engineered production-grade data ingestion and transformation pipelines in Python and SQL, implementing schema validation, metadata-driven workflows, and automated processing for 500K+ clinical records across multiple healthcare datasets.
- Designed scalable entity resolution and deduplication pipelines using fuzzy matching and statistical similarity techniques, improving dataset integrity by 30% while increasing downstream analytics reliability across large-scale clinical data.
- Collaborated with cross-functional engineering and data quality teams to validate ingestion workflows, establish governance rules, and document pipeline architecture, achieving 99.9% reliable data flow, reducing onboarding time by 20%, and ensuring compliant, reproducible data operations.

ZS Associates

July 2021 – May 2024

Business Technology Solutions Associate Consultant

Pune, India

- Architected and deployed serverless AI services using FastAPI, AWS Lambda, and distributed retrieval pipelines, enabling real-time inference APIs and reducing manual clinical decision workflows by 40% for 50+ production users.
- Led the design of microservice-based backend systems across a 17-engineer cross-functional team, integrating external services, distributed storage, and Blue-Green AWS CI/CD deployments to improve query performance by 35% while reducing deployment downtime by 30%.
- Engineered automated AWS Glue ingestion pipelines supporting ML training workflows through schema validation, feature engineering, & workflow orchestration, reducing manual data movement by 70% .

TECHNICAL SKILLS

- **AI Systems:** PyTorch, TensorFlow, Scikit-learn, Prompt Engineering, Vector/Semantic Search, Embedding Models, RLHF, LoRA, MoE
- **LLM Infrastructure:** Context Window Management, Token Budgeting, Prompt Caching, Model Routing, Provider Fallback, Rate Limiting, TPM/RPM Management, Structured Output Validation, Guardrails, Retry Policies
- **AI Platform:** FastAPI, LangChain, LangGraph, Langfuse, Azure OpenAI, Hugging Face, ChromaDB, PostgreSQL, n8n
- **Backend Engineering:** FastAPI, Flask, REST APIs, Microservices, JWT Authentication
- **Distributed Systems:** Apache Spark, PySpark, Celery, Redis, Asynchronous Processing, Event-Driven Architecture
- **DevOps & Observability:** Git, AWS CI/CD, Linux VM, Structured Logging, Monitoring, Azure OpenAI, Docker, GitHub Actions

PROJECTS

AI Resume Optimizer | Skills: Python, FastAPI, LangChain, Groq, OpenAI, Anthropic, Langfuse, LaTeX

- Architected provider-agnostic AI resume optimization platform using FastAPI and LangChain, implementing capability-aware structured output with automatic JSON fallbacks across Groq, OpenAI, Anthropic, and multiple LLM providers.
- Engineered an end-to-end document processing pipeline that transformed resumes into structured JSON, independently optimized resume sections against job descriptions, and generated production-ready PDFs through automated compilation.
- Designed resume retrieval system that indexed historical resume versions, selected contextually relevant experience for each job description, and computed deterministic resume-job match scores using weighted evidence retrieval instead of model-generated estimates.

NightPulse - A CRM platform | Skills: Python, FastAPI, Celery, Redis, Docker, PostgreSQL, JWT, Meta Developer toolkit, Meta Graph

- Architected an event-driven backend platform using FastAPI, Celery, and Redis, decoupling business workflows from asynchronous messaging, automation, and background processing through scalable domain events and worker queues.
- Engineered a provider-agnostic messaging framework integrating Meta Graph APIs (Instagram, Facebook, WhatsApp) with secure webhook ingestion, JWT-based authentication, RBAC, and extensible service abstractions supporting future communication channels.
- Built a production-ready deployment foundation with Docker, structured logging, idempotent retry policies, workflow automation via n8n, and environment-driven configuration, enabling reliable integrations, observability, and scalable service operations.

EDUCATION

University of Southern California, Los Angeles

Aug 2024 – May 2026

Master of Science in Computer Science, Courses: Analysis of Algorithms, LLM(RAG), NLP, Machine Learning

GPA 3.67/4