Naveed Asghar

Senior AI Engineer

Location: California, United States | **Phone**: +1 202 992 1720 | **Email**: itsnaveed.asghar@gmail.com
LinkedIn: https://www.linkedin.com/in/naveed-asghar42/

Senior AI Engineer with 6+ years of experience designing and deploying production-grade AI/ML systems and scalable backend architectures. Specialized in building LLM-powered applications using Retrieval-Augmented Generation (RAG), prompt engineering, and vector search. Experienced in developing high-performance AI microservices using Python, FastAPI, and distributed systems, and deploying AI platforms on AWS and Kubernetes. Proven track record of integrating OpenAI, Hugging Face, and multimodal models into enterprise applications that automate workflows, improve user engagement, and deliver measurable business impact.

Technical Skills

- **Programming Languages**
  Python, JavaScript, TypeScript
- **AI / Machine Learning**
  LLMs, RAG, Prompt Engineering, LangChain, LangGraph
  LlamaIndex Transformers, RLHF, Fine-Tuning, AI Agents, NLP
  Multi-Agent Systems, Tool Calling, Function Calling
- **Backend Development**
  FastAPI, Node.js, Django, REST APIs, GraphQL, Microservices Architecture
  WebSockets, Event-Driven Systems
- **Frontend Development**
  React, Next.js, TypeScript, TailwindCSS, Angular, Vue.js
- **Data & ML Infrastructure**
  ML Pipelines, MLflow, Vector Databases, Embeddings, Semantic Search
- **Cloud & DevOps**
  AWS (EC2, S3, RDS, ECS, ECR), Docker, Kubernetes
  CI/CD, GitHub Actions, Jenkins, Apache Kafka
- **Automation & AI Workflows:**
  n8n, Zapier, Workflow Automation
  AI Orchestration, Event-Based Systems
- **Databases**
  PostgreSQL, MongoDB, MySQL, Redis
- **Tools**
  Git, GitHub, Bitbucket, OpenAI APIs, Agile/Scrum

Professional Experience

Senior AI Engineer

Turing | Dec 2024 – Present

- Designed and deployed LLM-powered applications using Retrieval-Augmented Generation (RAG) to improve response accuracy and contextual understanding.
- Built scalable AI microservices using Python and FastAPI to support real-time inference and high-throughput systems.
- Integrated OpenAI, Hugging Face, and Llama models into enterprise platforms for conversational AI and workflow automation.
- Implemented vector search systems using Pinecone and ChromaDB to enable semantic search and enterprise knowledge retrieval.
- Collaborated with product and engineering teams to deliver AI-driven automation workflows that increased operational efficiency by 30%.

AI Engineer

Devbridge | Jun 2021 - Nov 2024

- Built full-stack applications combining AI-powered backend systems with modern frontend frameworks.
- Developed LLM-based features including intelligent search, conversational assistants, and recommendation systems.
- Designed scalable backend APIs using Python (FastAPI) and Node.js within microservices architectures.
- Implemented Retrieval-Augmented Generation pipelines connecting LLMs with enterprise data sources for contextual AI responses.
- Improved user interaction and engagement by 30% through AI-powered search and recommendation capabilities.

Full Stack Developer

Onfleet | Mar 2020 - May 2021

- Developed full-stack web applications using Python (FastAPI) and React/TypeScript for scalable and maintainable systems.
- Built high-performance REST APIs supporting real-time data processing and complex business workflows.
- Architected backend services capable of handling high-volume data pipelines and concurrent user requests.
- Optimized PostgreSQL database queries and indexing strategies, reducing query execution time by 25%.
- Deployed and maintained applications on AWS infrastructure to ensure reliable and scalable cloud services.

Projects

Amelia – Enterprise Conversational AI Platform

Technologies: Python, FastAPI, RAG, LangGraph, Vector Databases, AWS

- Built conversational AI system for automated customer support and enterprise workflows.
- Implemented RAG-based knowledge retrieval for accurate context-aware responses.
- Developed multi-agent orchestration pipelines for automation and decision support.

Design.ai – AI Content Generation Platform

Technologies: Python, Node.js, LangChain, LLM APIs, AWS

- Developed an AI platform capable of generating text, images, video, and audio from a single interface.
- Integrated multiple generative AI models for multimodal content creation.
- Built scalable backend services and cloud-based deployment architecture.

Sabermine – AI Document Processing System

Technologies: Python, FastAPI, OCR, NLP, AWS

- Built an AI system for automated document parsing and structured data extraction.
- Processed PDFs and scanned documents using NLP pipelines and OCR.
- Designed scalable processing pipelines for high-volume document workflows.

Glamezy – Social Content Platform

Technologies: Python, FastAPI, React, PostgreSQL

- Built full-stack social platform with scalable backend and responsive frontend.
- Implemented recommendation features to improve content discovery and engagement.

SeekSocial – Social Learning Platform

Technologies: React, Node.js, PostgreSQL

- Developed social learning platform supporting collaborative learning and content sharing.
- Implemented scalable backend services for real-time user interaction.

Certifications

- ChatGPT Prompt Engineering for Developers – DeepLearning.AI
- LLM Applications Development with LangChain and OpenAI
- AWS Certified Developer – Associate

EDUCATION

Bachelor's degree (Computer Science)

San Francisco State University, San Francisco, Ca