SHUBHAM PANDEY

AI Engineer | Generative AI, Agentic AI & LLM Engineering

Bengaluru, India | +91 6201075634 | shubhampandeysp01@gmail.com | linkedin.com/in/shubham-pandey-92aa46168

AI/ML Engineer, Software Engineer, and Claude Certified Architect focused on enterprise backend systems, applied machine learning, NLP, and Generative AI. Designs and deploys end-to-end LLM solutions spanning RAG pipelines, fine-tuned domain models, agentic workflows, and conversational chatbots. Proficient in Python, PyTorch, LangGraph, FastAPI, Docker, vector databases, and inference engines including vLLM and llama.cpp. Strong focus on latency optimization, retrieval accuracy, deterministic tool execution, and production MLOps.

TECHNICAL SKILLS

Generative AI & Agentic Systems: LLMs, LangGraph, LangChain, LlamaIndex, Model Context Protocol (MCP), tool/function calling, multi-agent workflows, prompt engineering, structured outputs.

RAG & Vector Databases: Embeddings, semantic search, hybrid BM25/dense retrieval, reciprocal rank fusion (RRF), cross-encoder reranking, ChromaDB, Qdrant, Milvus.

Machine Learning & Fine-Tuning: PyTorch, TensorFlow, Scikit-learn, Hugging Face Transformers, PEFT, LoRA, QLoRA, supervised fine-tuning (SFT), NLP, model evaluation.

LLM Serving & Optimization: vLLM, llama.cpp, MLX, model quantization, GGUF, AWQ, GPTQ, Safetensors, streaming inference, latency telemetry.

Backend & API Engineering: Python, FastAPI, Rust, Java, Spring Boot, C++, SQL, PostgreSQL, SQLite, REST APIs, AsyncIO, SSE, WebSockets.

Cloud Platforms: Amazon Web Services (AWS), Microsoft Azure, Google Cloud Platform (GCP).

MLOps, Evaluation & Observability: Docker, Kubernetes, GitHub Actions CI/CD, MLflow, RAGAS, LangSmith, OpenTelemetry, Git, Linux.

Development Tools: Codex, Claude, Grok, OpenCode, Cursor, Antigravity, Visual Studio, IntelliJ IDEA, PyCharm, Jupyter Notebook, Postman, Android Studio, Xcode, unsloth.

EXPERIENCE

AI/ML Engineer | Freelance / Contract | Bengaluru

Oct 2023 - Present

- Worked on OpenLLM, an OpenAI-compatible LLM inference gateway and control plane, implementing Python/FastAPI services and SQLite model metadata for routing across MLX, llama.cpp, and vLLM.
- Integrated model discovery, quantized-weight lifecycle management, Safetensors artifacts, ASGI/SSE streaming, and latency telemetry for local model serving.
- Worked on BugBounty Vault, a local retrieval-augmented generation application for context-grounded security research and question answering using Rust, Axum, Tantivy, FastEmbed, and BGE embeddings.
- Implemented hybrid retrieval combining BM25 and dense vector search with Reciprocal Rank Fusion, assembling retrieved context for LLM generation and streaming responses through SSE.
- Worked on BrowseLM, a browser-agent execution SDK with typed DOM operations in Rust, C++, and WebAssembly, enabling structured LLM tool use and browser automation.
- Implemented event replay, capability tokens, and auditable permissions for deterministic, repeatable agent execution.
- Worked on Conductor, an on-device AI assistant using Python, small language models, speech processing, and OCR, translating voice/text commands into multi-step execution plans with tiered OS permissions.

Thales Group | Software Engineer I / II | Noida, India

Aug 2022 - Sep 2023

- Developed Java and Spring Boot microservices and REST APIs for enterprise order routing, fulfillment, and delivery workflows.
- Built Python and SQL predictive-modeling pipelines for delivery planning, with data preprocessing, feature engineering, and cross-validated evaluation.
- Containerized microservices using Docker and supported Kubernetes-based deployment pipelines for reliable, automated production releases.
- Practiced test-driven development (TDD), code reviews, and root-cause analysis across distributed backend systems.

Thales Group | Software Engineer Intern | India

Feb 2022 - Jul 2022

- Automated manual business operations using Java and shell scripting, reducing recurring processing overhead by 30%.
- Diagnosed and fixed frontend and backend defects across enterprise Java and AngularJS codebases.

Defence Research and Development Organisation (DRDO) | Software Engineer Intern | New Delhi, India

May 2021 - Jul 2021

- Built unsupervised document-clustering and NLP pipelines in Python and Scikit-learn to classify and organize large unannotated corpora.
- Conducted exploratory data analysis, text normalization, and TF-IDF/semantic feature extraction to uncover patterns across defense research datasets.

EDUCATION

Amity University | Bachelor of Technology (B.Tech), Computer Science & Engineering

2018 - 2022

CERTIFICATIONS & HONORS

- Claude Certified Architect
- Dataiku ML Practitioner | Dataiku Core Designer | Dataiku Advanced Designer
- Vulnerability Management Certification
- Thales Xtra Miler Award & Quarterly Employee Recognition