

Shubhang Dhawan

shubhangdhawan2000@gmail.com | [+91-9335219193](tel:+91-9335219193) | [Linkedin](#) | [Github](#) | [Medium](#) | [Google Scholar](#)

Profile

- AI-focused Backend & ML Engineer with ~3 years of experience building production-grade GenAI and ML systems using Python and microservices architecture.
- Hands-on expertise in RAG architecture, Agentic AI workflows, LLM orchestration (OpenAI), LangChain, LlamaIndex, and vector search (Pinecone, Azure AI Search).
- Strong foundation in embeddings, semantic retrieval, FastAPI-based REST APIs, ML deployment, CI/CD pipelines, and cloud-native systems (Azure/GCP).

Experience

Software Developer – AI Engineer, SIEC Education Pvt. Ltd. 01/2026 – Present | On-Site (Gurugram)

- Built and scaled a production RAG-based GenAI chatbot using OpenAI GPT, with MCP-style model/data integration and familiarity with Gemini and Claude.
- Developed agentic LLM workflows using LangChain/LlamaIndex, with LangGraph orchestration, guardrails, response evaluation, observability, and hallucination-mitigation techniques.

Large Language Model Engineer Intern, Infutrix Technologies 05/2025 – 11/2025 | On-Site (Mohali)

- Architected an LLM-based automated grading system using multimodal OCR and NVIDIA NeMo Megatron GPT for contextual validation.
- Built GenAI evaluation and testing pipelines with semantic-similarity scoring, validation, monitoring, and reliability-focused safety controls.

Machine Learning Engineer Intern, NXG Ventures 02/2025 – 05/2025 | Remote (Gujarat)

- Built and deployed a FastAPI-based fraud detection microservice using PyTorch and transaction embeddings in a containerized environment.
- Improved reliability through threshold calibration, model evaluation, validation testing, and performance monitoring on imbalanced datasets.

Artificial Engineer Intern, Times Network | Bennett Coleman & Co. Ltd 11/2024 – 02/2025 | On-Site (Noida)

- Built an experimental AI-driven 24/7 newsroom using LLM-based content generation and TTS, including a GPT-4o Mini editorial assistant.
- Implemented retrieval grounding, prompt validation, content-safety controls, and response evaluation using LangChain and Azure AI Search.

Large Language Model Engineer Intern, Frontera Health 03/2024 – 10/2024 | Remote (USA)

- Built a GenAI health assistant using RAG, LlamaIndex and Mistral 7B, integrating semantic search with MongoDB Atlas.
- Deployed AI services on GCP using containerized workloads, with performance tuning, monitoring, evaluation, and reliability-focused optimization.

Skills

Python | SQL | FastAPI | Rest APIs | Microservices | Machine Learning | NLP | Transformers | Git | GCP | Azure | MCP | AI Guardrails | LLM Orchestration | Agentic AI | LangGraph | Docker | Gemini | Vector Database | Cloud Deployment | Claude | Monitoring | LLM Evaluation | Safety Controls | GenAI Testing | Hallucination Mitigation | Tracing & Observability

Personal Projects

[PowerMind AI – Intelligent Q&A on The Book "48 Laws of Power"](#)

[MarketIQ – AI Agent for Marketing Data Insights](#)

Published Research

[Drug Recommendation System using Bgft-Dbi-Lstm](#), IEEE

Built a system to recommend safer, more personalized drugs by considering individual health needs.

[Ethical Implications of AI in Healthcare](#), IJSREM

Explored AI's impact on healthcare, focusing on ethical concerns around privacy, data security, and bias.

Education

Master's in Data Science, Chandigarh University 2022 – 2024

Bachelor's in Science, Chhatrapati Shahu Ji Maharaj University 2019 – 2022