

PHILIP SIMON DEROCK

contact@philipsimonderock.com | [+91 8300057632](tel:+918300057632) | [Linkedin](#) | [GitHub](#) | [Portfolio](#)

AI Engineer specializing in Agentic AI, RAG pipelines, Multi-Agent Systems, and LLM-powered products with hands-on experience in Fine-Tuning, AI Alignment, MCP ecosystems, and Applied AI Research. Proven in end-to-end design, deployment, and optimization of production AI systems.

EDUCATION

KARUNYA UNIVERSITY

B.TECH IN ARTIFICIAL INTELLIGENCE
AND DATA SCIENCE

May 2025 | Coimbatore, TamilNadu

SKILLS

- Python • FastAPI • REST API • Java • SQL
- System Design • Git • GitHub • Linux
- Google Cloud • Firebase • Docker • VPS
- Numpy • Pandas • scikit-learn
- Transformers • Hugging Face
- NLTK • Medical Image Processing
- Data Curation • Data Annotation
- Data Labeling • Data Augmentation
- Data Preprocessing • AI Feedback

LLM & AGENTIC AI

- LangChain • LangGraph • LangSmith
- LlamaIndex • Pydantic • OpenAI SDK
- ReAct Agent • Browser Agent
- Multi-Agent Systems • Tool Calling
- Function Calling • Prompt Engineering
- Context Engineering • Agentic RAG
- GraphRAG • Hybrid RAG • BM25
- Reranking • Knowledge Graph • RAGAS
- Embeddings • Chunking Strategies
- Dense Retrieval • Sparse Retrieval
- Hybrid Search • Vector Search • Zilliz
- Milvus • ChromaDB • Qdrant • Pinecone
- MCP • FastMCP • Ollama • vLLM
- Google AI Studio (Gemini) • Claude
- LLM API Integration • Windsurf • Cursor

AI RESEARCH

- Full Fine-Tuning • LoRA • QLORA • PEFT
- Supervised Fine-Tuning (SFT) • RLHF
- Reinforcement Learning (RL) • RLAIF • RLVR
- AI Safety • AI Alignment • Safety Evaluation
- Human-in-the-Loop Systems • Multimodal AI
- Autonomous Reasoning Systems • LLM Research

EXPERIENCE

BLACKCOAT AI | AI ENGINEER - FULL TIME

Aug 2025 - May 2026

- Engineered production-grade Agentic RAG infrastructure for legal research orchestrating LangGraph ReAct multi-agent pipelines with domain-specific chunking across legal contracts and acts, Zilliz-backed hybrid retrieval, and LangChain.
- Served the entire system via FastAPI and a dedicated FastMCP server, enabling external AI agents to invoke legal intelligence as MCP-native tools.

IISPPR | AI AGENT - INTERNSHIP

Jun 2025 - Aug 2025

- Built AI-powered workflow automations for real-time document processing & agent-based task execution.

CODINGJR | AI RESEARCH - INTERNSHIP

Apr 2025 - Aug 2025

- Worked on building an AI coding copilot using LLM-backed code generation to assist developers write cleaner, faster code.

EDUNET & AICTE | AI & DATA ANALYTICS WITH A FOCUS ON GREEN SKILLS - INTERNSHIP

Nov 2024 - Dec 2024

- Worked on Healthcare Prediction on Diabetic Patients, AI-driven feature engineering & statistical modeling to optimize predictive healthcare diagnostics.

CISCO | CYBERSECURITY - INTERNSHIP

May 2023 - Aug 2023

- Completed training in Cybersecurity Essentials and Network Security Fundamentals.

PROJECTS

XPIL - GOVERNED CLINICAL-REASONING AI | Python • Gemma 4 E4B • QDoRA

(QLoRA + DoRA) • Unsloth • GGUF • Ollama • llama.cpp • H200 • MI300X

- Fine-tuned Gemma 4 E4B via QDoRA (QLoRA + DoRA) using Unsloth on 37k truth-grounded clinical samples across 11 real-patient dataset families – building a governed clinical reasoning AI with Tamil/Tanglish support, calibrated confidence outputs, and full offline deployment on a ₹8,000 Android phone via GGUF quantization with llama.cpp and Ollama.

DEX JOBS - ZERO-COST AI JOB HUNTING AGENT | Python • Supabase • Qdrant

Jina Embeddings • Cerebras • Groq • Mistral • Cloudflare Workers AI • Telegram Bot API • Asyncio

- Engineered a framework-free asynchronous ReAct orchestration engine with a custom Think-Act-Observe loop under 200ms latency integrating 8 specialized agents, multi-tier memory (Postgres consciousness + Qdrant semantic memory), and dynamic multi-provider LLM routing with headroom-based failover across Cerebras, Mistral, and Groq ensuring 100% runtime.
- Powered by a concurrent job search pipeline querying 19 Job sources, with Qdrant vector deduplication, LLM-based relevance scoring, Jina Embeddings v3 semantic match, job card formatting, and cover letter generation all delivered via Telegram.

ADVANCED GENERATIVE AI SERVER FOR EFFICIENT AI DEPLOYMENT | Python • Ollama

LangChain • Streamlit • MySQL • Ubuntu Server • RAG • Nomic Embeddings • scikit-learn

- Transformed an idle laptop into a wireless, headless Ubuntu AI server, deploying an end-to-end RAG pipeline with LangChain, Ollama, Nomic embeddings, and MySQL-backed conversational memory via Streamlit.

MULTI-AGENT RAG CHATBOT | Python • ChromaDB • Gemini API

- Architected a decoupled multi-agent RAG pipeline, deploying a custom asynchronous message bus to orchestrate document ingestion, vector retrieval, and LLM inference.

EMAIL CLASSIFICATION API USING MACHINE LEARNING | Python • scikit-learn • FastAPI

- Built a real-time email classification system with Naive Bayes & deployed it as a FastAPI based REST service on Hugging Face Spaces for automated support ticket processing

COMPUTER VISION PLAYER TRACKER | PyTorch • YOLOv5 • OpenCV • CNNs • scikit-learn

- Architected real-time YOLOv5 multi-object tracker with Kalman filtering & Hungarian optimization for occlusion-resistant player re-identification.