

NEIL PANCHAL

Agentic AI & Generative AI Engineering | Multi-Agent Systems | RAG & LLM Applications | Computer Vision | AI/ML Engineering

+91-9624514330/ neil12panchal@gmail.com / [LinkedIn](#) / [Leetcode](#) / [Kaggle](#)

Profile Summary

- **Senior Artificial Intelligence Engineer** with **6+ years of experience** across **Artificial Intelligence, Machine Learning, Generative AI, Agentic AI, Computer Vision, and software development**, currently working on AI solutions at Matrix Comsec.
- Experienced in developing **Agentic AI and multi-agent applications** using LangGraph, including AI copilots, workflow orchestration, Human-in-the-Loop processes, episodic memory, and self-reflection capabilities.
- Skilled in **Generative AI and RAG development**, with hands-on experience in hybrid search, vector databases, BM25, reranking, knowledge graphs, LLM applications, response grounding, and citation-based answers.
- Strong background in **Computer Vision and Deep Learning**, including palm vein recognition and low-light image enhancement; achieved **98.32% TAR@FAR 10⁻⁴** in palm vein recognition and **27.16 PSNR / 0.874 SSIM** in image enhancement.
- Proficient in **AI model development, fine-tuning, inference optimization, quantization, MLOps, and cloud deployment**, with hands-on exposure to PyTorch, HuggingFace, AWS, Docker, MLFlow, Ray, and vLLM.
- **Technical team leader** managing a **4-member team**, backed by earlier experience in full-stack development using **Python, JavaScript, MERN, ReactJS, NodeJS, FastAPI, SQL, MongoDB, GCP, and Azure** for scalable application development.

Core Competencies

- | | |
|--|--------------------------------------|
| • Agentic AI Development | • Computer Vision |
| • Generative AI & LLM Applications | • AI Model Development & Fine-Tuning |
| • Multi-Agent Systems | • AI Inference & Model Optimization |
| • Retrieval-Augmented Generation (RAG) | • MLOps & AI Deployment |
| • Machine Learning & Deep Learning | • AI Solution Architecture |

Technical Skills

- **Programming Languages:** Python, JavaScript
- **Frameworks/Libraries:** PyTorch, LangChain, LangGraph, LangSmith, HuggingFace, ReactJS, NodeJS, OpenCV, Scikit-Learn, FastAPI, Ragas, vLLM
- **Databases:** SQL, MongoDB, ChromaDB, Neo4j
- **Tools/MLOps:** GitHub, GitHub Actions, AWS, EC2, S3, Bedrock, Sagemaker, Lambda, Step Functions, Glue, MLFlow, LakeFS, Ray, Docker, LiteLLM
- **Concepts:** Computer Vision, RAG, Agentic AI, Search, Reinforcement Learning, Language Models, Reasoning Models, Generative AI Modeling, Knowledge Graphs, Multi-view Stereo, Machine Learning, ETL, Inference Engineering, Quantization, System Design, Model Gateway

Organizational Experience

Matrix Comsec, Vadodara | Senior Artificial Intelligence Engineer | Since Jan’24

Palm Vein Recognition

- Architected a multi-stage NIR-based palm vein recognition pipeline covering palm detection, hand landmarking, and recognition.
- Trained the recognition model on an in-house dataset with real-time palm acceptance, user guidance, and affine-based perspective correction.
- Enhanced vein-pattern preprocessing using B-COSFIRE and ZCA whitening, fine-tuning MobileNetV3-Large with ArcFace loss to achieve 98.32% TAR@FAR 10⁻⁴.
- Evaluated system security using Extreme Value Theory (EVT) to estimate FAR under large-scale gallery conditions.
- Optimized models through asymmetric quantization for efficient NPU deployment.

Low-Light Image Enhancement

- Spearheaded development of a low-light enhancement model for restoring visual details in Bullet and Dome camera imagery.
- Designed a trainable HVI color space to decouple brightness and color, improving robustness across varying illumination conditions.
- Implemented CIDNet with dual branches for independent intensity and color processing in HVI space.
- Integrated Lighten Cross-Attention (LCA) to enhance image structure–content interaction while suppressing low-light noise.
- Optimized training using weighted L1, Edge, and Perceptual losses, achieving 27.16 PSNR and 0.874 SSIM.

NextWaveLogic, Location | Full Stack Developer | Feb 2020 – Aug 2023

- Architected and developed scalable applications end-to-end using the MERN stack (MongoDB, Express.js, React.js, Node.js).
- Created a scalable SaaS platform supporting C2C and B2C use cases with emphasis on modularity, performance, and extensibility.
- Applied Test-Driven Development (TDD) practices to enhance code quality and minimize production defects.
- Established CI/CD pipelines for automated testing and deployment, accelerating release cycles and improving application reliability.
- Utilized GCP and Azure for application deployment, monitoring, and cloud infrastructure management.

Projects Managed

SarthiAI | LangGraph, FastAPI, ReactJS, Neo4j, ChromaDB, Deepgram Nova-3, GPT-5.4-mini

- Architected a multi-agent AI copilot for Matrix Comsec supporting live-call assistance, post-call analysis, and always-on queries.
- Orchestrated hierarchical agents via LangGraph StateGraph with shared AgentState for centralized workflow management.
- Integrated Deepgram Nova-3 for real-time Hindi/Hinglish speech-to-text, speaker diarization, and contextual AI suggestions.
- Developed a Neo4j Knowledge Graph pipeline linking call entities with CRM companies, products, and deal opportunities.
- Unified ChromaDB vector search, Neo4j Cypher traversal, and web search to deliver contextual product, pricing, and deal insights.
- Enabled Human-in-the-Loop CRM actions through LangGraph interrupts and a React + Vite interface.

SEC Filing RAG | LangChain, ChromaDB, BM25, Cohere Rerank, GPT-5.4-nano, ReactJS, FastAPI

- Developed a production-grade hybrid RAG system for SEC 10-K/10-Q financial Q&A, achieving 35.4% FinanceBench contains-answer rate.
- Combined ChromaDB dense retrieval and BM25 sparse search using Reciprocal Rank Fusion (RRF).
- Optimized multi-ticker queries through parallel retrieval using ThreadPoolExecutor.
- Integrated configurable Cohere Rerank v3.0 and BAAI/bge-reranker-v2-m3 for two-stage reranking.
- Established input and output guardrails for query validation and answer grounding against retrieved evidence.
- Implemented citation-aware prompting with [N] source references and GPT-5.4-nano for grounded responses.

Olympiad Agents | LangGraph, USACO, Episodic Memory, Self-Reflection, HITL, GPT-5.4-nano, ReactJS, FastAPI

- Created an agentic system for USACO problem-solving using episodic memory, self-reflection, and Human-in-the-Loop workflow.
- Leveraged BM25 episodic memory to improve solve rate to 70% vs. 40% zero-shot baseline.
- Designed a self-reflection loop supporting up to three attempts with automated Python code execution and timeout handling.
- Analyzed test-case verdicts including wrong answers, runtime errors, and TLEs to iteratively refine solutions.
- Enabled Human-in-the-Loop hints through LangGraph interrupt controls after reflection attempts.
- Streamed agent progress via SSE to a React + Vite interface with attempt-wise code and test-case results.

Certifications

- | | |
|--|---|
| • Production Ready RAG workshop | • Cypher Fundamentals |
| • LLM Finetuning Workshop | • Generative AI with Large Language Models |
| • Foundation: Introduction to LangChain - Python | • Neural Networks and Deep Learning |
| • Foundation: Introduction to LangGraph - Python | • Convolutional Neural Networks |
| • Building Knowledge Graphs with LLMs | • Deep Learning with PyTorch : Image Segmentation |
| • Using Neo4j with LangChain | |

Academic Details

- **B.E. in Computer Science & Engineering** | Maharaja Sayajirao University, Gujarat | 2021