

ROYDON SEQUEIRA

Udupi, Karnataka, India | +91-6362261716 | roydNSEQUEIRA@gmail.com
[linkedin.com/in/roydonsequeira](https://www.linkedin.com/in/roydonsequeira) | github.com/roydonsequeira | roydonsequeira.com

PROFESSIONAL SUMMARY

GenAI Engineer and AI Agent Architect with 2+ years of production experience taking autonomous AI systems from architecture to deployment. Specialized in Multi-Agent Systems, LLM Orchestration, RAG, and Semantic Retrieval. Core stack: Python, LangChain, Qdrant, FastAPI, Docker.

EXPERIENCE

AI Agent Developer

Code Crew Studio

Feb 2026 - Present

Mumbai, India (Remote)

- Automate customer onboarding by building and maintaining a production AI agent on the official WhatsApp Business API that runs structured intake conversations, classifies user problems, and returns analyzed feedback.
- Preserve conversation continuity across sessions by engineering a multi-turn query system with persistent context, using LangChain for orchestration-heavy flows and direct LLM API calls where lower latency is required.
- Persist agent state, conversation history, and structured customer records by developing FastAPI backend services on PostgreSQL.
- Technologies:** WhatsApp Business API, Python, LangChain, FastAPI, PostgreSQL

GenAI Engineer

ReinHealth (Stealth-Mode AI Healthcare Startup)

Jul 2024 - Feb 2026

Colorado, USA (Remote)

- Reduced manual patient-intake workload by an estimated 70% and cut average intake time from 15 minutes to under 5, by leading end-to-end delivery of a production autonomous medical AI agent conducting structured intake over text and voice.
- Eliminated third-party exposure of patient data, with zero records sent to external APIs, by architecting an LLM orchestration layer running multi-step reasoning (intent analysis, follow-up selection, tool execution, grounded synthesis) on local Ollama inference.
- Enabled grounded clinical responses by building a privacy-first RAG pipeline on Qdrant pairing dense embeddings with PostgreSQL, and automating speech-to-text clinical notes and TTS summaries in n8n with conflict-aware scheduling.
- Hardened the platform for clinical use by engineering input validation, emergency-symptom escalation, and audit logging, and unified agents, automation workflows, and frontend behind a single REST API layer.
- Technologies:** Python, LangChain, Flask, Qdrant, PostgreSQL, n8n, Ollama, Docker, STT/TTS

Machine Learning Intern

Igeeks Technologies

Jun 2023 - Jul 2023

Bengaluru, India

- Built CNN, AlexNet, and MLP image-classification pipelines on custom datasets with hyperparameter tuning and OpenCV preprocessing. **Technologies:** Python, OpenCV, NumPy, CNN, MLP

PROJECTS

AI Resume Analyzer | FastAPI, Qdrant, BGE Embeddings, LangChain, Claude/Gemini API, Docker

2025

- Delivered auditable, reproducible resume scoring by building a deterministic 5-dimension model weighting keyword alignment (30%), semantic similarity (25%), experience relevance (20%), formatting (15%), and impact metrics (10%).
- Achieved sub-400ms end-to-end pipeline execution excluding LLM calls, benchmarked across 5 runs, by architecting a three-agent system (Analyzer, Optimizer, Formatter) with SSE streaming.
- Safeguarded release reliability with 39 unit and integration tests, CI/CD on GitHub Actions, pre-commit hooks (ruff, mypy), and Docker Compose; integrated BGE local embeddings with Qdrant for semantic retrieval.

CORTEX: Local AI Agent Framework | Python, LATS Orchestration, OpenTelemetry

2026

- Enabled persistent context retention across long-horizon tasks by designing a local-first agent framework on a four-tier memory architecture spanning short-term, episodic, semantic, and procedural stores.
- Extended agent reasoning beyond single-shot chains by implementing LATS (Language Agent Tree Search) orchestration supporting tree-search planning, multi-step reasoning, and self-correction.
- Made agent failures traceable by instrumenting end-to-end OpenTelemetry tracing for latency monitoring and behavioral debugging.

TECHNICAL SKILLS

Generative AI / LLMs: Generative AI, Large Language Models (LLMs), LLM Orchestration, Multi-Agent Systems, LangChain, RAG Pipelines, Semantic Retrieval, Prompt Engineering, Ollama, BGE Embeddings, STT/TTS

Vector Databases: Qdrant, ChromaDB, Vector Search, Embedding Pipelines, Semantic Search

Languages & Backend: Python, SQL, FastAPI, Flask, REST APIs, PostgreSQL, Redis, Pydantic, Unicorn

ML & Computer Vision: PyTorch, TensorFlow, scikit-learn, CNN, AlexNet, MLP, OpenCV, NumPy, Pandas, Hyperparameter Tuning

Infrastructure & Tools: Docker, Git, GitHub Actions (CI/CD), n8n, Linux, Twilio, Hugging Face, Vercel, ruff, mypy

EDUCATION

NMAM Institute of Technology

B.E. in Artificial Intelligence and Machine Learning

2020 - 2024

Karnataka, India

iHUB DivyaSampark, IIT Roorkee

Executive Post Graduate Certification in Data Science & Artificial Intelligence

2024 - 2026

Online