

Ravikumar M. N.

Senior AI Engineer · India

ravikumarnaduvin@gmail.com

github.com/ravikumarmn

linkedin.com/in/ravikumar-mn

ravikumarmn.github.io

Senior AI Engineer at C5i. I design multi-agent assistants, retrieval, and document intelligence on Databricks and Azure, plus the fine-tuning, evaluation, and CI/CD behind them.

EXPERIENCE

Senior AI Engineer · C5i

Jun 2026 – Present

Full-time · India

- Architecting a multi-agent assistant on LangGraph (specialized agents for retrieval, reasoning, and task execution) that holds sub-800ms average response latency across concurrent multi-tenant sessions and cut end-to-end task time 22%.
- Serve agents on Databricks Model Serving with Unity Catalog lineage and access control on every agent-facing source; Genie-based agents cut analyst query turnaround 30%.
- Built governance in from day one (model versioning, prompt and response auditing, access policies) plus Azure DevOps CI/CD that cut release cycle time 20% and deployment failures 15%.

AI Engineer · Zyoin

Nov 2025 – May 2026

Contract · India

- Designed and deployed an agentic RAG chatbot for enterprise knowledge, fed by Docling-based ingestion that turns unstructured documents into structured data.
- Built an agentic synthetic-persona generator for financial survey datasets to improve downstream model robustness and segmentation analysis.
- Ran the model lifecycle on Databricks, Azure, and MLflow: experiment tracking, versioning, and cloud deployment.

Machine Learning Engineer · Acranton

Jan 2022 – Aug 2025

Full-time · Davangere, India

- Fine-tuned transformer NLP models from 78% to 89% sentiment accuracy while cutting inference latency 30%; custom classifiers raised pipeline accuracy 18% and cut processing time 25%.
- Built computer-vision pipelines for image classification and object detection with CNN and GAN models, reaching 92% accuracy.
- Designed a Gemini API-based extractor that turns unstructured healthcare documents into structured data in under 3 seconds, raising throughput 30%.

Generative AI Engineer · ResearchBrains

Feb 2021 – Dec 2021

Internship · Remote, India

- Built AI services on FastAPI and AWS (EC2, S3, SageMaker) that cut generative-model inference response time from 5s to 2s.
- Integrated modular AI microservices with GPT-powered APIs into enterprise workflows, cutting manual content-creation time 40%.

SELECTED SYSTEMS

A governed multi-agent assistant for enterprise workflows

C5i

A LangGraph assistant that retrieves, reasons, and acts across business workflows, with sub-800ms average responses and an audit trail on every turn.

<800ms average response latency · -22% end-to-end task time · -30% analyst query turnaround

An enterprise GenAI copilot with agentic RAG and tool use

Project

One copilot over documents, tables, and analytics tools, with hybrid retrieval and query planning, answering in 1.8s end to end.

1.8s end-to-end response latency · <120ms vector search latency at scale

Agentic document intelligence, from PDFs to structured data and answers

Zyoin · Acranton · projects

Docling parsing and LLM extraction turn PDFs, forms, and reports into JSON, a knowledge graph, and grounded QA, at under 3s per page.

<3s ingestion per page · 1.5s grounded QA response · +30% healthcare form throughput

Model adaptation with LLMs, embeddings, and classifiers

Acranton · projects

Fine-tuning where prompting and retrieval run out, using SFT with LoRA, contrastive embeddings, and transformer classifiers taken from 78% to 89% accuracy.

89% sentiment accuracy, up from 78% · -30% inference latency · +10% F1 on imbalanced classes

Also built

- **Agentic research automation.** CrewAI agents that generate personas, run domain research, and compile structured reports. -50% manual research effort
- **Conversational PDFs and websites.** Turns static PDFs and sites into queryable knowledge systems. 95% query success rate
- **Multimodal chatbot.** Routes across three models to handle text and image interactions in one assistant. +35% user engagement
- **Handwritten prescription → JSON.** Structured extraction from handwritten medical prescriptions with Gemini Flash. <5s per prescription

SKILLS

LLMs & agents	LangGraph, LangChain, LlamaIndex, CrewAI, Agentic RAG, Hybrid retrieval, Tool calling & SQL agents, Databricks Genie, Guardrails & audit logging, Docling
Training & inference	PyTorch, Hugging Face Transformers, SFT, PEFT / LoRA, Sentence Transformers, Contrastive learning, scikit-learn, Optuna, Inference optimization
Data & MLOps	Databricks Model Serving, Unity Catalog, MLflow, Weights & Biases, TruLens, Azure DevOps CI/CD, Docker, SonarQube, Pinecone, Weaviate, Chroma, Elasticsearch
Languages & infra	Python, FastAPI, Flask, Pandas, Streamlit, AWS (EC2, S3, SageMaker), Azure

EDUCATION

B.Tech, Electrical Engineering · Central University of Karnataka

2018 - 2022

IEEE student member; started the Enigma AI Club.

Certifications: Finetuning Large Language Models (DeepLearning.AI); Building and Evaluating Advanced RAG Applications (DeepLearning.AI); Deep Learning Specialization (Coursera); Machine Learning Specialization (Coursera).