

Rahul Anand

FREELANCE ML & GENAI ENGINEER · RAG, AI AGENTS, PRODUCTION LLM SYSTEMS

CONTACT

+91 98348 42796
rahulanand1103@gmail.com
rahulanand.me
linkedin/rahul-anand1103
github/rahul-anand1103
Bengaluru, India · IST

AVAILABILITY

[30] hrs / week
Fixed-scope projects or monthly retainer
Typical engagement [6–12 weeks]
Daily overlap with EU and US-East mornings

SERVICES

Production RAG systems
Multi-agent & tool-use workflows
Agent tooling & API integration
Long-term memory & state design
Human-in-the-loop escalation
Agent & RAG evaluation, red-teaming
LLM application backends
Recommender systems
ML pipelines & MLOps

CORE STACK

GenAI LangChain, LangGraph, Graphiti, vector DBs, Hugging Face, GPT, LLaMA, fine-tuning, evals

ML / DL PyTorch, TensorFlow, scikit-learn, XGBoost, CatBoost, LightGBM, transformers, LSTM

MLOps Airflow, MLflow, Docker, DVC, Feature Store, AWS (S3, EC2, SageMaker), GCP

Backend Python, FastAPI, Flask, Pandas, NumPy, Dask, Git

EDUCATION

B.E., Computer Science & Engineering
Chandigarh University
2013 – 2017

PROFESSIONAL OVERVIEW

Machine Learning engineer with 7 years in production ML, now freelancing on generative-AI systems — grounded RAG, multi-agent workflows, and the evaluation harnesses that keep them honest. Author of MODE (arXiv), a retrieval method for small domain-specific corpora. I take projects from prototype to something a team can operate without me.

SELECTED FREELANCE ENGAGEMENTS

- **Grounded RAG + agent platform, logistics operations** MAY 2025 – PRESENT
 - Built the full retrieval stack — ingestion, chunking, embeddings, retrieval, citation enforcement — and red-teamed it against hallucination and injection cases. Reached ~0.88 faithfulness and ~92% citation coverage on an internal eval set of [N] questions.
 - Built multi-agent workflows in LangGraph (planning, tool use, routing, state and memory) to automate shipment-exception handling. ~70% of cases now close without human escalation.
 - Added long-term memory with Graphiti so context persists across sessions, cutting repeat questions ~30%.
 - Shipped ETA-prediction and delay-risk models into the same workflow, ~20% more accurate than the [heuristic] baseline they replaced.

RESEARCH & OPEN SOURCE

MODE: Mixture of Document Experts for RAG arXiv · github.com/rahulanand1103/mode

A cluster-and-route retrieval method for small, domain-specific corpora that removes the need for an external vector database or a reranker. Competitive with standard pipelines on HotpotQA and SQuAD at substantially lower retrieval latency.

- **rag-citation** — attaches verifiable source citations to RAG output, so answers can be audited. github.com/rahulanand1103/rag-citation
- **ytkit** — fetches subtitles from YouTube channels, playlists, and search results in a few lines of code; a corpus-building tool for RAG. github.com/rahulanand1103/ytkit
- **youtube-script-writer** — an agent that drafts video scripts to a given title, language, and tone. github.com/rahulanand1103/youtube-script-writer

PRIOR FULL-TIME EXPERIENCE

- **Machine Learning Engineer** FEB 2022 – MAY 2025
Careervira
 - Built the company's agentic search over LangChain, LangGraph, and FAISS on EC2, plus a LangChain report pipeline delivering course insights as PDFs via S3.
 - Built and deployed the course recommender — collaborative filtering in PyTorch, plus content-based course-to-job and role-to-course mapping on embeddings.
 - Predicted skills and topics for partner courses at 0.731 F1 (SVC) and 0.879 (RoBERTa), later moving the workload to LLMs on SageMaker. Owned the end-to-end pipeline: Airflow, Feature Store, MLflow, Docker, DVC.
- **R&D, Machine Learning** DEC 2019 – JAN 2022
Teknuaance Pvt. Ltd.
 - Built intent classification for the Sabhae assistant at 90% intent and 93.2% sub-intent accuracy, and custom NER models (TensorFlow, BERT, BiLSTM) at 0.78 F1.
 - Built the AI-Studio ML pipelines and parallelised workloads with Dask, cutting computation time 30%. Improved model robustness ~10% with paraphrasing and word-level augmentation.