

HARSH KAMAT

AI Engineer · Generative AI · LLM Applications

Bengaluru, India | +91 8660843889 | harshkamat.2307@gmail.com

linkedin.com/in/harshkamat | github.com/datapiratepy | harshkamat.vercel.app

SUMMARY

Final-year Computer Science Engineering student who builds LLM applications end to end — retrieval pipelines, agent routing, and the deterministic logic and tests that make them trustworthy. Shipped a RAG assistant on IBM watsonx.ai (Granite) with 156 automated tests at 88% coverage in CI, where citations are constructed in code so the model cannot fabricate a source. Seeking AI / Generative AI / Applied AI / AI Automation internships and entry-level roles.

TECHNICAL SKILLS

Generative AI & LLMs: Retrieval-Augmented Generation (RAG), IBM watsonx.ai, Granite models, Embeddings, Vector databases (ChromaDB), Similarity search & thresholding, Chunking, Agent routing & orchestration, Prompt engineering, Structured (JSON) outputs, Grounding & hallucination mitigation

Languages: Python, SQL, JavaScript (basic), HTML/CSS

Backend & Engineering: Flask, REST APIs, Server-Sent Events (SSE), SQLAlchemy, pytest, CI/CD (GitHub Actions), Git

ML & Data: scikit-learn, Pandas, NumPy, EDA, Classification, Model evaluation

Data & Tools: Google BigQuery, SQLite, Power BI, Matplotlib, IBM Cloud, Jupyter, Google Colab

PROJECTS

NutriMind AI — Retrieval-Grounded Nutrition Assistant | Python · IBM watsonx.ai (Granite) · RAG · ChromaDB · Flask · pytest

- Built a RAG pipeline over user-uploaded PDFs — page-bounded overlapping chunking, ChromaDB cosine search, and a similarity threshold below which answers are labelled general knowledge instead of presented as evidence. **Citations are constructed in Python from chunk metadata, so the model cannot invent a source.**
- Designed a three-tier embedding stack — Granite 768-dim when credentials are present, sentence-transformers or a deterministic hash provider offline — each isolated in its own vector collection, keeping the full pipeline testable with zero credentials.
- Implemented a Coordinator that routes each request to one of four specialist agents: deterministic rules first at zero token cost, Granite JSON classification only for ambiguous messages, with two intents answered directly in Python. Every response exposes the routing method and reason.
- Shipped a Flask app with SSE token streaming, a deterministic nutrition engine (every displayed figure computed in Python, never by the LLM), and **156 automated tests at 88% coverage** in GitHub Actions CI requiring no credentials.

Redrob Ranker — Interpretable Deterministic Candidate-Ranking Engine | Python · Algorithm Design · Feature Engineering

- Ranked the top 100 candidates from a 100,000-profile pool against a Senior AI Engineer job description in under a minute on CPU alone — streaming a 465 MB dataset in under 2 GB against constraints of 5 minutes wall-clock, no network and 16 GB RAM.
- **Evaluated an embedding-based semantic layer and rejected it with evidence:** the corpus collapsed to 44 unique text templates, so exact hash classification beat approximate similarity on accuracy, latency and reproducibility. Output is deterministic and byte-identical across runs.

Customer Churn Prediction & Retention Analytics | Python · scikit-learn · Pandas · Power BI

- Raised churn-class recall from 0.52 to **0.79** on 7,032 Telco records via class balancing, accepting the precision trade-off (0.50, F1 0.61) because a missed churner costs more than a false alarm; audited the pipeline and fixed an encoding-order defect that had inflated the feature matrix from 30 to 7,061 columns.

EXPERIENCE

AI & Cloud Intern — IBM SkillsBuild · Edunet Foundation · AICTE | Jun 2026 – Jul 2026

- Selected for the IBM SkillsBuild AI & Cloud internship to build a Generative-AI product under industry mentorship; designed and built NutriMind AI on IBM watsonx.ai Granite models.
- Applied prompt engineering, retrieval systems and software-engineering practice — automated testing, CI, environment-driven configuration and fail-fast validation — to ship a documented application that runs with or without cloud credentials.

EDUCATION & CERTIFICATION

B.E., Computer Science Engineering — AMC Engineering College, Bengaluru — 2023 – 2027

Google Data Analytics Professional Certificate — Google / Coursera — 2025

Hackathon: Redrob × Hack2Skill INDIA.RUNS Data & AI Challenge — 2026