

ABHINAV HARBOLA

AI / ML Engineer

Varanasi, India (open to remote or relocation) | +91-9415291480 | abhinavdaharbola@gmail.com
linkedin.com/in/abhinavharbola | github.com/abhinavharbola

PROFESSIONAL SUMMARY

AI/ML Engineer who builds and rigorously evaluates LLM agents, guardrailed RAG, recommender systems and MLOps pipelines. Raised agent completeness from 3.8 to 4.8 via paired ablation, doubled popularity-baseline Recall@10 with a SASRec and LightGBM pipeline, and shipped a fail-closed RAG system with multi-provider failover. Strong in statistical testing.

TECHNICAL SKILLS

LLM and Agentic AI: LangGraph, Multi-Agent Orchestration, RAG, Model Context Protocol (MCP), Prompt Engineering, LLM Guardrails, Prompt-Injection Defense, LLM-as-Judge Evaluation, Ragas, Semantic Caching, Gemini API, Groq, NVIDIA NIM
Retrieval and Recommenders: Qdrant, FAISS, Vector Search, Embeddings, Reranking, SASRec, Two-Tower, LambdaRank, NDCG
Machine Learning: PyTorch, Scikit-learn, XGBoost, LightGBM, Hugging Face, Feature Engineering, Bootstrap CI, McNemar Test
MLOps and LLMops: MLflow, DVC, Evidently (Drift Detection), Model Monitoring, CI/CD (GitHub Actions), Docker, pytest
Languages and Backend: Python, SQL, C++, FastAPI, REST APIs, PostgreSQL, SQLite, Pydantic, Streamlit, Pandas, NumPy

PROJECTS

Competitive Intelligence Agent | LangGraph, Gemini, Groq, NVIDIA NIM, Postgres, FastAPI 2026

- Raised judge-scored completeness from **3.8 to 4.8** and groundedness from **3.8 to 4.6** (0 to 5 scale, 5-company paired ablation) by adding a Critic node that flags coverage gaps and triggers up to 3 targeted replans, at 1.7x the tool calls.
- Bounded every run to **15 tool calls**, 8 steps per plan and an **8-minute wall clock**, degrading gracefully to confirmed fields, by building hard-stop guardrails, loop detection, 60s LLM timeouts and 3-attempt retry with backoff.
- Delivered fully sourced 5-field briefs at **zero API cost** by matching 4 agent roles to free-tier models by call volume, checking each field's status against the scratchpad, and caching confirmed fields in Postgres for 7 days.

Kubernetes Gated RAG | LangGraph, Qdrant, FlashRank, Ragas, Groq, Gemini, Streamlit 2026

- Blocked unsafe, off-topic and ungrounded answers by building a LangGraph pipeline with **4 fail-closed gates** (input safety, topic, rerank relevance threshold, response safety) that refuse or return "no grounded docs" instead of guessing.
- Reduced repeat retrieval and LLM calls with exact and Qdrant-backed semantic caches (cosine ≥ 0.95) and sustained availability through a **3-link failover chain** (2 Groq accounts, then NVIDIA NIM) with circuit breakers and 15s timeouts.
- Scored **0.95 faithfulness** and **0.91 answer relevancy** (6 Ragas metrics, 8-row set), judged by a different model family (Gemini) than the generator (gpt-oss-120b), with node- and provider-level Logfire tracing.

ReelBench: Multi-Stage Recommender Benchmark | PyTorch, FAISS, LightGBM, FastAPI, MLflow 2026

- Improved Recall@10 to **0.0485**, **2.0x** the popularity baseline and **29%** above ALS (NDCG@10 up 18%, Recall@20 up 28%), by training SASRec on 11.2M interactions and re-ranking FAISS candidates with a LightGBM LambdaRank model.
- Ran a leakage-tested 5-model comparison across **12.45M interactions** by building the harness first, with a global time split, a hard-failing leakage assertion and a separate temporal split for ranker labels, scored on 49,533 test interactions.
- Avoided a **~3.9GB** dense item-item matrix over 31,195 items with sparse row-blocked cosine and bounded top-K and served the pipeline on CPU-only hardware via FastAPI, with an offline Qwen3-Embedding index for cold-start items.

Continuous Credit Risk Governance Pipeline | Python, MLflow, Evidently, DVC, Postgres, GitHub Actions 2026

- Automated the model lifecycle (drift detection, retraining, gating, promotion, rollback) by streaming 63,750 rows of a 150,000-row credit dataset as **318 live batches** on a GitHub Actions schedule, with no human in the loop.
- Prevented noise-driven model swaps at roughly 13 positives per batch by gating promotion on a **2,000-resample bootstrap CI** for the AUC-PR delta and added a paired-bootstrap rollback comparing champion and predecessor on identical live rows.
- Held expected false retrain triggers near **1%** per undrifted batch via a pinned K-S test ($p < 0.05$), 0.3 drift-share threshold and 3-batch cooldown, with every decision in a Postgres audit trail covered by **105 tests** across 11 files.

EXPERIENCE

Data Science Intern | Coding Blocks Jun 2024 - Aug 2024

- Improved churn prediction accuracy by **9%** over the baseline with an XGBoost model tuned via GridSearchCV.
- Built a queryable dataset from **5,000+** scraped and cleaned competitor reviews for the product team.
- Delivered the churn model end to end, from ingestion and feature work through model selection, evaluation and handover.

EDUCATION

B.Tech in Computer Science and Engineering | Lovely Professional University, Punjab, India Jul 2026

CGPA: 8.23 / 10.0 | Dean's List (top 10% of students), 2023 to 2025

CERTIFICATIONS

- IBM DevOps and Software Engineering Certificate (Coursera, 2025); SAP Certified Data Analyst, SAP Analytics Cloud (2026)