JASIR ABDUL HAMEED

AI/ML Engineer (Fresher) | Python, Generative AI, LLMs, RAG

Malappuram, Kerala, India | Open to Kochi, Bengaluru, Chennai, remote | Available immediately

+91 8593940681 | jasirlvl@gmail.com | linkedin.com/in/jasirjru | github.com/jasirjru | jasirjru-portfolio.vercel.app

SUMMARY

Fresher AI/ML engineer who builds generative AI applications in Python. I built and deployed Jruva, a document Q&A assistant (code public) that uses retrieval-augmented generation (RAG) and failover across three LLM providers, and I measured its retrieval quality with MRR and Hit@k. Teamwork, communication and problem-solving under deadlines: NASA Space Apps 2024 Global Nominee (six-member team) and 3rd place in a 24-hour college hackathon. Looking for an entry-level AI/ML engineer role or internship.

EDUCATION

B.Tech, Computer Science and Engineering | MES College of Engineering, Kuttippuram (APJ Abdul Kalam Technological University) | 2022 - 2026

Course completed; result of one supplementary exam awaited | Class XII: 98% | Class X: 90%

SKILLS

Languages and CS: Python, SQL, Java, JavaScript, HTML, CSS; Data Structures and Algorithms, DBMS, OOP

Generative AI and LLMs: Generative AI, LLM APIs (Gemini, Groq, OpenRouter), RAG, prompt engineering, embeddings, vector databases (Qdrant), hybrid search (BM25 + RRF), reranking, LLM failover

Retrieval Evaluation: MRR, Hit@k, nDCG, development-set benchmarking, NumPy

Backend and Deployment: FastAPI, REST APIs, Server-Sent Events (SSE), PostgreSQL, pytest, GitHub Actions (CI), Git, Next.js, Render, Vercel, Qdrant Cloud

PROJECTS

Jruva - AI Assistant for Document Q&A | Sep 2026 - Present

Live: jruva.in | Code: github.com/jasirjru/jruva | Benchmark: github.com/jasirjru/jruva-retrieval-eval

- Built and deployed a document Q&A assistant: a FastAPI backend on Render and a Next.js frontend on Vercel that streams answers token by token with Server-Sent Events (SSE); on the free 512 MB server, uploads are off and a memory guard falls back to keyword search if memory runs low.
- Built a hybrid retrieval pipeline: 250-word overlapping chunks, BGE-small embeddings (384 dimensions) in Qdrant plus BM25, merged with Reciprocal Rank Fusion (RRF), with an optional TinyBERT reranker.
- Evaluated retrieval on a 100-question development set with MRR and Hit@k: vector search scored MRR 0.91 vs 0.81 for keyword search; the production hybrid (weighted RRF) scored 0.92, too small a gain to trust because its weights were tuned on the same set, so BM25 mainly covers exact terms and the low-memory fallback.
- Kept answers available when an LLM provider fails or hits its rate limit (HTTP 429), by adding automatic failover across Gemini, Groq and OpenRouter with a circuit breaker that pauses a failing provider.
- Added rule-based routing to web search (Tavily, with DuckDuckGo and Wikipedia fallback), document lookup and a calculator, and wrote system prompts for four answer styles.
- Found and fixed two code-execution vulnerabilities in a security audit of my own codebase, by replacing an unsafe eval() calculator and an unvalidated SymPy math parser with an input allow-list checked before parsing; added pytest regression tests for attack inputs, run by GitHub Actions on every change.

ACHIEVEMENTS

NASA Space Apps Challenge 2024 - Global Nominee and People's Choice | 2024

- Built the team website (HTML, CSS, JavaScript, Three.js 3D models) in the six-member Team DarkNova for "Galactic Games: Fun in a Microgravity Environment".

CodeTrek 24-hour Hackathon, MES College of Engineering - 3rd Place | 2025

- Built ExamEval with Team DarkNova in 24 hours: an AI prototype that grades handwritten answer sheets, reading them with Google Cloud Vision OCR and comparing them with the answer key using the Gemini API.