Mayur Kishor Kumar

AI Engineer
+91 7760842842, mayurakash1999@gmail.com

Professional summary

AI/ML Engineer with 1+ year of experience building and deploying LLM agents, reasoning systems, and data pipelines. Proven results across healthcare, fintech, and platform teams through scalable AI tools, CI/CD workflows, and cost-optimized model deployment.

Links

LinkedIn: www.linkedin.com, Github: github.com, Portfolio: mayurkishorkumar.github.io.

Skills

Python, SQL, Java, JavaScript, React, NodeJS, C++, HTML, CSS, PyTorch, TensorFlow, Scikit-Learn, XGBoost, NLTK, SpaCy, OpenCV, AWS, GCP, Docker, Kubernetes, Git, FastAPI, Flask, Django, Apache Spark, Airflow, MLflow, Snowflake, Tableau.

Employment history

AI Engineer – Applied Research, Sep 2024 - May 2025

HUMANITARIANS AI, Boston, MA, USA

- Co-developed 10+ LLM-powered agents, including Multi-Hop Reasoning, Chain-of-Thought, and Patient-Triage-RAG, used by teams at MIT, Harvard, and Northeastern, supporting 2000+ real or simulated interactions across education and healthcare use cases.
- Integrated LangChain, ChromaDB, and LLaMA 3 into agent pipelines to handle 4–6 subqueries per prompt with memory support, improving multi-step reasoning depth by 2× compared to baseline.
- Built MCP-style chatbots via FastAPI to deliver 500+ educational and healthcare sessions with 30% lower dropout rate and 20% faster average response time.
- Benchmarked Claude, Mixtral, Dolphin, and Mistral models to guide architecture selection, reducing inference latency by 25% and API cost by ~18% across deployments.
- Deployed modular agent frameworks with containerized CI/CD pipelines on GCP and AWS, achieving 95% test pass rates and <3 min container build time, reducing environment bugs by 40%.

Software Engineer – Platform Reliability & Tooling, Mar 2025 - May 2025

AFTERQUERY EXPERTS, San Francisco, CA, USA

- Led debugging of platform-critical issues across data frameworks (NumPy, PyTorch, Pandas), improving stability and reducing runtime failures by 40% in pre-deployment stages.
- Built and integrated end-to-end test suites that raised CI coverage from 68% to 92%, enabling smoother deployments and catching regressions before release.
- Containerized internal dev environments using Docker, cutting onboarding time for new engineers by 2 days and standardizing local builds across OS platforms.
- Partnered with platform architects to deploy low-regression performance patches, reducing job execution time by 15% in critical data pipelines.
- Authored reusable triage templates and debugging guides, shortening issue resolution cycle by 50% and improving internal contributor velocity.

Internships

Machine Learning Intern – FinTech Risk & Insights, Aug 2021 - Nov 2021

IG GROUP, Bangalore, KA, India

- Designed and deployed fraud detection pipelines using XGBoost on high-frequency trading data, reducing false negatives and saving ₹12L/month in fraud-related losses.
- Engineered a real-time trade analytics engine ingesting 500+ events/min, improving actionable insight delivery speed by 25% for portfolio managers.
- Enhanced internal tool discoverability by 45% through transformer-based semantic search, improving analyst efficiency in report retrieval workflows.
- Automated ETL refresh cycles with Airflow DAGs, eliminating 10+ manual hours/week and increasing data pipeline reliability across compliance reporting.

Project Trainee – Predictive Analytics & Systems Automation, Mar 2021 - Jun 2021

INDIAN SPACE RESEARCH ORGANIZATION(ISRO), Bangalore, KA, India

- Developed predictive models on 10+ years of telemetry data to forecast server overheating risks, helping ensure zero-downtime operations during mission launches.
- Built a D3.js-powered EMS dashboard that reduced metric access latency from 120s to under 30s, boosting real-time situational awareness across ops teams.
- Deployed an alerting system in Node.js to automate fault detection in remote systems, accelerating incident response by 30%.
- Collaborated with systems engineers to embed ML-driven monitoring into legacy infrastructure without disrupting mission-critical workflows.

Education

Master of Science in Artificial Intelligence, May 2024

Northeastern University, Boston, MA, USA

GPA : 3.42/5

Bachelor of Engineering in Computer Science, Jun 2022

Dayananda Sagar College of Engineering, Bangalore, KA, India

CGPA : 8.77/10

Projects

Automated Stock Forecasting Pipeline with AWS and Airflow

- Predicted stock prices with 93% accuracy by training a CNN-LSTM model on historical data using AWS Glue and S3 pipelines.
- Eliminated 10+ hours/week of manual work by automating 8+ tasks with Apache Airflow and QuickSight.

Brain Tumor Classification and Segmentation Pipeline

- Processed 5K+ MRI scans to classify tumors with ResNet50 and segment regions using ResUNet with 98% pixel accuracy.
- Enhanced segmentation precision by optimizing Tversky loss and added visual interpretability using Grad-CAM heatmaps.

AI-Powered Document Query System with Vector Database Integration

- Designed an LLM-based semantic search tool using Dolphin-Mistral, GenAI embeddings, and Chroma across 1,200+ documents.
- Accelerated query speed from 3.5s to <1s by optimizing embeddings and text splitting in Streamlit.

Multi-Hop Reasoning Agent for Business Intelligence

- Engineered a reasoning agent with Groq API + Ollama to analyze SQL, PDFs, and APIs in parallel using LLaMA 3-70B.
- Extracted insights across 4+ data streams by chaining LLM outputs into subqueries, reducing BI response time to <10s.