Chandar Rathi | AI Engineer

Los Angeles, California | chandarrathi2@gmail.com | **(+1) 7149003485** | LinkedIn | Portfolio

SUMMARY:

Results-driven **AI/ML Engineer** with **4+ years** of experience designing, developing, and deploying scalable **Machine Learning** and **Generative AI** solutions across **AWS** and **Microsoft Azure**. Proficient in building **LLM-powered** applications, **RAG** pipelines, predictive forecasting models, and cloud-native ML platforms using **Python**, **LangChain**, **Scikit-learn**, **LightGBM**, **FastAPI**, **Docker**, and **MLflow**. Experienced in **MLOps**, model deployment, data engineering, and delivering enterprise AI solutions that improve operational efficiency and business outcomes across fintech and healthcare domains.

TECHNICAL SKILLS:

- **Programming:** Python, SQL
- **Generative AI:** LLMs, RAG, LangChain, Prompt Engineering, Hybrid Retrieval, Semantic Search, MCP
- **Machine Learning:** Scikit-learn, LightGBM, Prophet, Feature Engineering, Predictive Modeling, Time-Series Forecasting
- **Cloud Platforms:** AWS (Amazon Bedrock, EKS, S3, CloudWatch), Microsoft Azure (Azure ML, Databricks, ADF, ADLS Gen2, Azure SQL)
- **MLOps & DevOps:** Docker, MLflow, GitHub Actions, CI/CD, Model Deployment
- **Backend:** FastAPI, REST APIs, Microservices
- **Databases:** PostgreSQL, Redis, pgvector, Azure SQL Database
- **Tools:** Git, LangSmith, Ragas, Agile/Scrum

WORK EXPERIENCE:

AI/ML Engineer | PayPal, CA **Mar 2025 – Present**

Project: *Enterprise AI Payment Support Assistant (Generative AI Platform)*

- Designed a **Retrieval-Augmented Generation (RAG)** architecture using **LangChain**, **multi-provider LLMs (OpenAI, Anthropic, Amazon Bedrock)**, and **Prompt Engineering** to automate payment, refund, dispute, and chargeback support.
- Enhanced the **RAG** pipeline using **Hybrid Retrieval**, **Reranking**, and **Embedding Optimization** with **LangChain** and **pgvector**, improving retrieval relevance and answer accuracy by **28%**.
- Developed scalable **FastAPI** microservices and deployed AI inference workloads on **Amazon EKS** using **Docker**, enabling low-latency, highly available model serving.
- Implemented **Semantic Search** using **pgvector** and **Embedding-based Retrieval** to index payment policies, merchant documentation, and support knowledge for contextual responses.
- Integrated **Tool Calling** and **Model Context Protocol (MCP)** to securely connect the assistant with internal payment APIs and knowledge sources, automating multi-step support workflows.
- Evaluated **LLM** quality using **Ragas** and **LangSmith**, optimizing **Prompt Engineering** and retrieval strategies to improve response faithfulness and reduce hallucinations.
- Integrated **Redis** caching with **PostgreSQL** to optimize conversational context management and improve API response performance by **25%**.
- Established **CI/CD** pipelines using **GitHub Actions** and monitored application health with **Amazon CloudWatch**, ensuring reliable deployments with **99.9%** service availability.

ML Engineer | Deloitte, India (end-client – CVS Health) **Feb 2021 – Dec 2023**

Project: *Enterprise Pharmacy Demand Forecasting Platform*

- Designed batch data ingestion workflows using **Azure Data Factory (ADF)** to consolidate prescription, inventory, supplier, and pharmacy sales data into **Azure Data Lake Storage Gen2 (ADLS Gen2)**, improving data availability by **35%**.
- Processed multi-million pharmacy records using **Azure Databricks** and **PySpark** to perform data cleansing, feature engineering, and seasonal trend analysis for demand forecasting.
- Developed medication demand forecasting models in **Azure Machine Learning** using **Prophet** and **LightGBM**, improving forecast accuracy by **22%** and reducing inventory shortages across pharmacy locations.
- Optimized forecasting models with **Scikit-learn** by performing feature selection, hyperparameter tuning, and cross-validation to improve prediction performance.
- Managed model lifecycle, experiment tracking, and versioning using **MLflow**, improving model reproducibility and accelerating production releases.
- Published batch forecasting outputs to **Azure SQL Database** and archived historical datasets in **Amazon S3**, supporting business reporting, auditing, and long-term trend analysis.
- Containerized ML inference workloads using **Docker** and deployed scheduled batch prediction services on **Amazon EKS**, enabling scalable and reliable inventory forecasting.
- Configured production monitoring and automated alerts using **Amazon CloudWatch**, proactively identifying performance issues and maintaining **99.9%** service availability.

PROJECTS:

RAG Ops — Production RAG Assistant · LangChain · pgvector · FastAPI · React

- Built and open-sourced a from-scratch RAG assistant to benchmark retrieval strategies, publishing a reproducible naive-vs-hybrid-plus-reranking comparison that moved NDCG@5 from 0.64 to 0.89 on a 500-question eval set.
- Served quantized open models (Llama 3 / Mistral) on a single GPU with vLLM, keeping per-query cost near zero and P95 latency under 800ms.
- Shipped a public demo with a FastAPI backend, a streaming React/Next.js UI, and a Ragas dashboard exposing faithfulness and recall per query.

LLM Observability & Cost Platform — Self-Hosted Tracing for AI Apps · OpenTelemetry · ClickHouse · Grafana · FastAPI

- Built a self-hostable observability layer that traces every LLM request end to end — prompt, retrieved context, tool calls, tokens, latency, and cost — so teams can debug and budget AI systems instead of guessing.
- Cut blind spots in production AI by storing traces in ClickHouse and surfacing per-route cost and latency in Grafana, with automated alerts when spend or quality drifts.
- Shipped it as drop-in middleware that instruments any LangChain or LangGraph app in a few lines, with a public demo and docs.

EDUCATION:

Master of Computer Science – California state university, USA ***Jan 2024 – Dec 2025***

Bachelor of Technology in Computer Science & Engineering – Gujrat Technological University, India ***Jun 2019 – May 2023***