Rajesh Marudhachalam

Toronto, ON rajesh.marudhachalam@gmail.com +1-999-999-9999
linkedin.com/in/rajesh1804 github.com/rajesh1804 rajesh1804.github.io

Summary

AI/ML Engineer with 4+ years of experience building low-latency AI systems in production — from real-time inference and drift-aware forecasting to agentic LLM pipelines and semantic search. Proven track record delivering measurable impact across enterprise and consumer-scale use cases at BlueCat and J.P. Morgan. Now seeking to build next-gen agentic or real-time AI at frontier orgs.

Experience

BlueCat Networks — Toronto, Canada
ML Engineer - Production AI Systems Jan 2024 – Present

- Architected an internal **RAG assistant for infra diagnostics** — handling **1500+ daily queries**; cut support ticket time by **40%**.
- Built a real-time **anomaly detection system** with drift-aware alerts — **92% precision**, **40% fewer false positives**.
- Led integration of **ETS + ARIMA forecasting models** into planning pipelines, improved infra forecasts by **30%**. Scaled across **500+ distributed edge devices**.

ML Research Intern — Security AI May 2023 – Dec 2023

- Shipped transformer-based **pairwise relation model** model for DNS tunnel detection; beat baseline statistical model by **23%** F1.

J.P. Morgan Chase — Bengaluru, India
Software Engineer II — Wealth Management (Top Performer Q1 2024) Jan 2022 – Aug 2022

- Led migration of mission-critical **ETL pipelines** to **AWS EKS + Snowflake** — reduced infrastructure costs by **20%**.
- Architected a **15TB HDFS data lake** used by **10+ downstream teams** — improved query latency by **60%**.

Software Engineer I — Wealth Management (Top Performer Q4 2020) Aug 2020 – Jan 2022

- Designed and deployed a **PySpark ingestion framework** handling **10M+ records/day** — achieved **11x** throughput over legacy ETL.
- Led automation of **schema validation, alerting, and failover workflows** — cut pipeline failures by **80%**.

Software Engineer Intern — Asset Management (Top 6 APAC Intern) Jan 2020 – Jul 2020

- Shipped a **React-based portfolio dashboard** to replace Excel workflows — saved analysts **2 hours/day**.

Projects

RideCastAI 2.0 — Real-Time Fare Predictor (Uber-style) [Demo] [Medium] [GitHub]
Built a latency-aware system with **ONNX**, **joblib**, and **river** for fare prediction with real-time ingestion, dual drift detectors (KSWIN, HST), and online learning. Reduced MAE by **27%** vs XGBoost. Tracked prediction errors live to simulate post-deployment feedback loops.

ThreadNavigatorAI 2.0 — Reddit Analyzer (Agentic AI) [Demo] [Medium] [GitHub]
Multi-agent pipeline with summarizer (Kimi), fact-checker (DeepSeek), and evaluator (Mistral) using LangGraph-style orchestration. Fact-check latency managed via async+cache. Improved LLM summarization faithfulness by **25%** using RAG + LLM-as-a-Judge loop.

StreamWiseAI — Movie Recommender + Retention Coach (Netflix-style) [Demo] [Medium] [GitHub]
Hybrid system combining Sentence-BERT RAG, fuzzy search, and session-state LLM agent. Delivered **19%** boost in simulated retention scores with model retries, state-aware personalization, and real-world watch history traces.

GroceryGPT+ — Semantic Grocery Search (Instacart-style) [Demo] [Medium] [GitHub]
RAG-style grocery engine using Sentence-BERT vectors, Weaviate DB, and reranking via OpenRouter LLM. Handled typos, cold-starts, and semantic overlap with **99% recall@5** at **¡200ms** latency under free-tier constraints.

Core Skills

ML/AI: Real-Time ML, Drift Detection, Forecasting, Online Learning, Semantic Search
LLMs: LangGraph, OpenRouter, LLM-as-a-Judge, Multi-Agent Reasoning, CoT
Tools: PyTorch, Tensorflow, Hugging Face, ONNX, river, FAISS, Airflow
Infra: Docker, AWS (EKS, Lambda), Snowflake, Kafka
Languages: Python, SQL, JS, Shell

Education

University of Toronto — MSc, Applied Computing (AI) 2022 – 2023
Focused on production ML systems, forecasting, and research-driven LLM applications.
Vellore Institute of Technology — B.Tech, Computer Science 2016 – 2020
Focused on Algorithms, Distributed Systems, and Data Engineering.