RAJAT PATEL

(443) 253-7424 raj55.patel@gmail.com linkedin.com/in/rajat-patel github.com/rajathpatel23 rajathpatel23.github.io

SKILLS

Languages: Python, Go, JavaScript/TypeScript, SQL, Shell Scripting
ML & Frameworks: TensorFlow, PyTorch, HuggingFace (TRL, PEFT, Transformers), LoRA, FastAPI, Flask, Django, AWS SageMaker, Kubeflow, MLflow, Ray
Data & Streaming: Apache Flink, AWS Glue, Kafka, Prefect, Snowflake
Tools & Databases: Milvus, DynamoDB, Postgres, MongoDB, Kubernetes, Docker, Arize, DataDog

EXPERIENCE

Senior Machine Learning Engineer | *Chime Financial Inc.* Nov 2023 - Present

- Fine-tuned Mistral 7B for merchant name normalization using supervised fine-tuning (SFT) and LoRA via HuggingFace TRL; improved accuracy from 79.8% to 91.2% (+11.4pp). Built end-to-end SFT pipeline covering dataset curation, training, and evaluation — validated across diverse merchant categories.
- Designed and built a three-agent self-healing system (Diagnosis, Remediation, Validation) for automated ML pipeline maintenance: agents identify rule failures, generate and apply code fixes, and validate changes against 10M+ real datapoints before issuing a safety recommendation — eliminating manual intervention from the debugging loop.
- Built a scalable LLM-based labeling system (LangGraph + HITL) for ML training data annotation: 150% improvement in throughput, turnaround reduced from weeks to hours, 90% reduction in manual review load, and 99% reduction in annotation cost across model iteration cycles.
- Improved ML infrastructure across training and production: integrated MLflow for experiment traceability, built near-real-time feature pipelines (Apache Flink, AWS Glue, DynamoDB-backed feature store) serving 10M+ daily transactions at $<$30ms latency, and used Arize for model observability.
- Prototyped graph-based (GNN) and sequential transformer approaches for fraud and payment integrity, exploring foundation-model-inspired behavioral signal modeling beyond traditional rule and feature pipelines.

Senior Machine Learning Engineer | *Interos Inc.* Jan 2020 - Nov 2023

- Built and owned a full-stack logo image search engine (scraping, active learning, ImageNet-based visual search) to automate website logo detection, reducing manual identification time by $>$80%.
- Shipped a deep-learning entity dedup/linking pipeline (contrastive candidate generation, cross-encoder reranking, Milvus vector search) achieving $\sim$90% top-10 recall and materially reducing false matches in supply-chain search.
- Co-led deployment and operation of large-scale Milvus vector search over $\sim$400M entities, tuning indexing and HNSW-based retrieval to meet latency/availability requirements across multiple ML products.
- Developed and productionized Transformer-based NER + event extraction pipelines, and built human-in-the-loop feedback/retraining loops (SageMaker GroundTruth) to handle concept drift and continuously enrich the Interos knowledge graph.

EDUCATION

University of Maryland Baltimore County May 2020
Master of Science in Computer Science
Shivaji University May 2014
Bachelors of Engineering in Electronics and Telecomm. Engg.

PUBLICATIONS

- On the Complementary Nature of Knowledge Graph Embeddings, Fine Grain Entity Types and Language Modeling. InProceedings, EMNLP Workshop on Deep Learning Inside Out, November 2020
- InterosML @ Causal News Corpus 2023: Understanding Causal Relationships: Supervised Contrastive Learning for Event Classification. In Proceedings, RANLP CASE 2023 Workshop, September 2023

PROJECTS

LabPilot — R&D Experiment Optimization Platform | *Python, FastAPI, React* | github 2026

- AI copilot for lab experiment optimization using surrogate models (RandomForest) and adaptive decision policies (UCB, LinUCB); LLM-powered explanations and literature retrieval close the interpretability loop. Full decision-support cycle: ingest → train → recommend → submit result → adapt.

Emergency Guidance Agent — CPR Copilot | *Python, TypeScript, Google Gemini Live* | github 2026

- Real-time voice-and-video CPR coaching assistant built on Google Gemini Live (multimodal audio + video); strict FSM controls step transitions — app enforces safety state, model provides guidance within each state.

Evidence-Driven Deep Research Agent | *Python, Custom Agentic Graph* | github | slides 2026

- Built a research agent with step-level reward signals (PRM proxy, 5 outcome signals) and an explicit typed evidence state — guided planner prioritizes uncovered sub-questions, challenges high-confidence claims, and stops on diminishing returns; ablations show higher coverage and grounding rate vs. naive baseline.