Raj Krishnan **Vijayaraj**

Machine Learning Expert

159 Beverley St , Toronto, Canada, M5T 1L7

+12498760601 | raj.vijayaraj@mail.utoronto.ca | unni12345.github.io/ | unni12345 | raj-krishnan-vijayaraj/

Summary

Machine Learning Engineer with 5+ years' experience building scalable, production-grade ML systems in legaltech and healthcare. Skilled in distributed training (Spark, Ray), RAG pipelines, and LLM evaluation. Passionate about optimization and agent-based learning, with growing focus on real-time pricing, robustness, and decision modeling in complex multi-agent systems.

Skills

Languages	Python, Scala, Typescript, C, C++, Java, Bash, JavaScript, Ruby
Web frameworks	Rails, Flask, Node.js, GraphQL, REST API
Machine Learning	PyTorch, TensorFlow, LLM, LangChain, Optimization, Reinforcement Learning, Transformers, Hugging Face
Cloud & DevOps	AWS, GCP, Azure, Docker, Kubernetes, KubeRay, Jenkins, Semaphore, AWS Secrets Manager
Developer Tools	Git, Linux, SumoLogic, Shell scripting, Kernel Configuration, IntelliJ, PyCharm, Visual Studio Code

Work Experience

LLM Engineer — *Toronto, Canada*

LexisNexis/Appfabs — *July, 2024 - now*

- Building LLM systems with RAG pipelines, scalable workflows, and automated MLOps for production-grade AI.

Machine Learning Associate — *Toronto, Canada*

Vector Institute, University of Toronto — *May, 2024 - July, 2024*

- Fine-tuning medical LLMs and developing RAG pipelines for diagnostic QA in healthcare.

Machine Learning Engineer — *Toronto, Canada*

Scribble Data — *Sep. 2022 - Oct. 2024*

- Developed AI agents and LLMs-as-Judge framework for contract analysis, enabling automated legal assessment and PII redaction.

Backend Developer — *Bengaluru, India*

G2 — *July 2019 - Nov 2021*

- Built GraphQL APIs, data pipelines, and backend systems for a SaaS platform, supporting real-time data processing.

Projects

Machine Learning

Clinical Tiger: Agentic RAG for Multimodal Clinical QA — *Toronto, Canada*

Clinical Tiger — *April 2025*

- Built an agentic Retrieval-Augmented Generation (RAG) system for clinical question answering with multimodal support for tables and images.
- Implemented tools for intent classification, verified answer synthesis, and section-aware hybrid retrieval using ClinicalBERT embeddings.
- Used GPT-4-Vision for image captioning and structured citations, including chain-of-verification and Vancouver-style reference tracking.

Embeddings to Diagnosis: Geometry-Aware Evaluation of Clinical LLM Robustness — *Toronto, Canada*

Independent Research — *Feb. 2025 – Present*

- Developed a perturbation- and geometry-based framework to evaluate diagnostic robustness of clinical LLMs using synthetic notes and PCA-driven latent analysis.
- Showed that surface metrics miss instability, while boundary crossings in latent space reveal diagnosis fragility, emphasizing need for geometry-aware evaluation.

Development of RAG Pipelines and SQL Generation Systems — *Toronto, Canada*

LexisNexis/Appfabs — *July. 2024 - now*

- Developing RAG and SQL generation pipelines with LLMs, Python, and SQL for automated query generation.
- Building distributed data pipelines using Apache Spark, Scala, and distributed computing frameworks.
- Leveraging Jenkins, KubeRay, and AWS Secrets Manager for orchestration, automation, and secure credential management.

Clinical Panda: A Large Language Model for Diagnostic Explanation

Toronto, Canada

University of Toronto / Prof Mark Chignell — *Jan. 2023 - Jan.2024*

- Developed a novel framework using Large Language Models (LLMs) to generate synthetic clinical notes from patient question-answer pairs.
- Built Clinical Panda, an LLM trained to generate evidence-based explanations for diagnoses.
- Proposed a lightweight perturbation analysis technique to evaluate model robustness against missing information in real-world data.

PIIguardLLM: Enhanced Data Privacy with PII Masking

Toronto, Canada

Scribble Data — *Sep. 2023 - April. 2024*

- Advancing a proprietary Large Language Model (Masking and PII Data Privacy) project, emphasizing advanced data masking techniques.
- Curating a Specialized Dataset for Model Optimization, Prioritizing PII Protection.
- Significant experimental investigation on properties like quantization, model size and model optimization.

Multi-modal Language Modelling in Medical Domain

Toronto, Canada

Wang Lab, University of Toronto — *May 2023 - Oct. 2023*

- Utilizing the BLIP-2 architecture, combining multiple modalities to construct a multi-modal LLM tailored for the medical domain.
- Implementing distributed training of the model using the Fully Sharded Data Parallel (FSDP) technique to enhance efficiency.

Neuro-inspired Sparse Visual Explanation for Convolutional Neural Networks

Toronto, Canada

University of Toronto/ Independent Project — *Jan. 2023 - April. 2023*

- Innovated a neuro-inspired sparse visual explanation algorithm for CNNs, merging LRP and Convolutional Sparse Coding..
- Validated its superiority over GradCam and GradCam++ through experiments.
- Showcased its applicability in high-dimensional fields like medical and 3-D imaging.

Ontology-Based Information Extraction

Trichy, India

Undergraduate Thesis — *Dec 2018 – May 2019*

- Built a custom NER system using spaCy and a domain ontology (Protégé) to extract and enrich technical data from OCR-processed documents.

Software Development

Automation of Data Processing and Export

Bengaluru, India

G2 — *June. 2021 - July. 2021*

- Developed a new module for end-to-end automated processing and export of data replacing multiple manual processes written in R.
- Scheduled the task monthly using sidekiq (job scheduler).
- Brought down the completion time to 20 hours from 3 days.

Sycomore: a blockchain that adjusts to transaction demand

Rennes, France

IRISA Lab, Inria — *May. 2018 - July. 2018*

- Conducted literature survey to compare and contrast various approaches for develop a blockchain that self-adjusts to the transaction rate.
- Implemented algorithms for a bitcoin blockchain data structure using Java and Akka programming.

Education

MEng Electrical and Computer Engineering

Toronto, Canada

University of Toronto, GPA 3.85 / 4.00 — *Jan 2022 - Dec 2023*

B.Tech Electrical and Electronics Engineering

Trichy, India

National Institute of Technology - Tiruchirappalli, GPA 8.00 / 10.00 — *July 2015 - July 2019*

Publications

Embeddings to Diagnosis: Robustness Evaluation of Clinical Reasoning with Agentic Synthetic Notes
Raj Krishnan Vijayaraj
Agentic & GenAI Evaluation Workshop @KDD2025

ARMatron — A wearable gesture recognition glove: For control of robotic devices in disaster management and human Rehabilitation
Anand Asokan; Allan Joseph Pothen; **Raj Krishnan Vijayaraj**
RAHA 2016: IEEE Robotics & Automation Society Conference

Honors & Awards

2022	**Recipients**, Mitacs Business Strategy Internship with Scribble Data	*Toronto, Canada*
2020	**Recipient**, PEAK Giver Award from G2 for performance at work	*Bengaluru, India*
2019	**Selected**, for travel grant offered by Embassy of France	*Bengaluru, India*
2015	**National Topper**, in Mathematics for AISSCE ’15 attended by 1.4 million students	*Kerala, India*