Kristen Pereira

kristenp1123@gmail.com p-kris10.github.io linkedin.com/in/pkris10/ github.com/p-kris10

Education

Georgia Institute of Technology, Atlanta, GA **August 2023 - May 2025**
Master of Science in Computer Science **GPA: 4.0/4.0**
Coursework: Conversational AI, Efficient ML, Social Computing, Grad Algorithms, ML, Big Data Systems, HRI

Sardar Patel Institute of Technology, Mumbai **August 2019 - May 2023**
B.Tech in Information Technology **GPA: 9.5/10.0**
Coursework: AI, Computer Vision, Advanced Databases, Distributed Systems, OS

Skills

Frameworks & Libraries: React, Node, Express, Django, Flask, FastAPI, Redis, PyTorch, TensorFlow, Scikit-learn
Tools & Languages: Python, C++, Java, JavaScript, TypeScript, Git, AWS, Docker, Google Cloud, Apache Spark

Experience

Member of Technical Staff 2, Nutanix AI, San Jose, CA June 2025 – Present

- Designed and built an end-to-end **agentic code-review system** for Nutanix's on-prem Gerrit workflow, automating structured reviews across large internal repositories.
- Implemented **MCP-driven toolcalling** to fetch relevant code context via semantic indexing, improving review accuracy and reducing manual effort.
- Created a comprehensive **internal code-review evaluation dataset** to benchmark agent performance and drive iterative improvements.
- Scaled the system to production with **Kubernetes-based deployment**, enabling autoscaling and observability; processed **10,000+ PRs** to date.
- Developed an enterprise-grade **RAG offering** ("NAI - Talk to My Data"), including ingestion, indexing, and secure on-prem retrieval workflows for customer datasets.
- Collaborated across AI and platform teams to deliver high-availability on-prem inference and optimized model-serving performance.
- **Technologies:** Kubernetes, MCP, FastAPI, Python, CI/CD, Arize AI, Grafana, Prometheus.

Machine Learning Engineer Intern, Skinzy Software Solutions, Mumbai October 2021 - June 2022

- Designed APIs for PyTorch-based vision models, handling image data preprocessing and inference
- Optimized ML models to have **40% less storage** size and **60% less response time** using **pruning, quantization and custom CUDA kernels with Pytorch bindings**
- Led the migration of key backend services to **AWS Lambda, cutting infrastructure costs by 20%** while ensuring scalability and high availability
- Integrated **AWS CloudWatch** for real-time performance monitoring and automated alerting, ensuring system reliability and prompt issue resolution
- Reduced deployment time via **GitLab CI/CD** optimization and enhanced test automation, improving system reliability
- **Technologies:** PyTorch,CUDA C, ReactJS, AWS Lambda, S3, CloudWatch, ONNX, Docker, Git, Postman, Jira.

Projects

Dynamic Resolution Input for DeIT in HuggingFace Transformers ↗

- Enhanced Vision AI models in HuggingFace library **(150k stars and 25k forks)** through open source contribution

Token Compression in RAGs for Inference Cost Reduction ↗

- Architected a Python implementation of TCRA-LLM using **LLamaIndex**, **HuggingFace**, and **Tonic**, achieving a 30% token reduction in RAG systems while maintaining model accuracy and optimizing operational costs for paid LLMs by reducing retrieved context

Dynamic Quantization of Large Language Model ↗

- Extended Meta's Fairseq library to support CPT ↗ and implemented both post-training quantization and **quantization-aware fine-tuning** on RoBERTa model
- Optimized multi-GPU communication and pipeline parallelism, improving training throughput and minimizing memory overhead. Technologies : **Meta's Fairseq, PyTorch Profiler, NVIDIA Visual Profiler**

Smart Healthcare Diagnostics Using Federated Learning

- Engineered a **full-stack web application** that enables healthcare institutions to securely collaborate on CNN model training via **federated learning**, preserving sensitive data privacy
- Built support for real-time inference and progress visualization across worker nodes. Tools used: **Flask, React, Flower, TensorFlow, WebSockets, AWS S3, AWS EC2**

Multilingual Text-based Image Search

- Developed a web application for stock image search using content-specific text queries, enabling precise cross-modal retrieval through multilingual knowledge distillation and cosine similarity. Tech stack : **React, TensorFlow, ONNX, Flask, Heroku, Docker**