

Sergio Peterson

✉ sergiopeterson.dev@gmail.com | 📞 +1 (510)822-4089 | 🌐 github.com/SergioPeterson |
🌐 [linkedin.com/in/sergio-w-peterson](https://www.linkedin.com/in/sergio-w-peterson) | 🌐 www.sergiopeterson.dev

Education

University of California Berkeley

Graduation date: Dec. 2024

B.A. Data Science with a focus on ML and Software

Academic Excellence Machine Learning, Data Mining, Computer Vision, Probability/Random Processes, Robotics, Efficient Algorithms, Dynamic Data Structures, Computer Architecture, Discrete Math, Probability Theory, Operating Systems, Systems Programming, Databases, Distributed Systems, Headless Content Management, B2B eCommerce

Experience

CV/ML Summer Intern

May. 2024 - Aug. 2024

Lexius

CV, Optimization API, AWS, GCP, PyTorch, large-scale project

- Achieved **+5% accuracy** and **+17% F1-score** improvement by integrating RGB data with various window sizes, temporal resolutions, and normalization techniques.
- Migrated infrastructure from **GCP Kubernetes** to **AWS Batch**, enhancing scalability and training performance.
- Fine-tuned models like Video **Swin Transformer** and **Conv3D**, reducing false positives to **under 3%**.
- Implemented/optimized **database schemas**, data retrieval speeds by **30%** ensuring data integrity across the system.
- Designed Infrastructure** robust backend services using **RESTful APIs** to support high-performance, data-driven applications, ensuring scalability and efficiency.

LLM Research - Under PhD Akshat Gupta

Aug. 2024 - Dec. 2024

Berkeley Speech Group (in BAIR Labs)

Python, LLM, GTO, DL, Pytorch, RAG

- Developed an **LLM** for poker strategy using **NLP** and **GTO**, achieving a **20%** increase in decision-making accuracy by integrating hyper-tuned Retrieval-Augmented Generation (**RAG**).
- Designed a **"thinking pipeline"** that streamlines complex decision-making processes. Optimized RAG models for specialized datasets, improving **context-specific** response accuracy.
- Implemented **Fine-Tuning (SFT)**, enabling real-time assessment and refining model outputs on decision patterns.
- Produced a systematic pipeline in **Rust** to automate solution generation from an open-sourced solver, integrating a mechanism to store **hand EV** for advanced analysis.

Simulation and Perception Research - P.I. Dr. Allen Yang

April. 2024 - Jan. 2025

AI Racing Tech

Infrastructure Dev, C++, Azure, DevOps, Docker, Packer, large-scale project

- Developed and managed Azure simulations using Docker and Terraform for automated VM creation and deployment, enabling efficient CI/CD workflows.
- Built **CI pipelines** with **GitHub Actions** for automated testing and deployment on **Azure VMs**, utilizing custom **Docker** scripts for resource management and SSH access.
- Optimized cloud infrastructure for AI Racing Tech's autonomy stack with **Packer** and **Terraform**, automating VM provisioning and resource deallocation post-simulation.
- Integrated tools like **OSSDC Simulator** w/ asset management **GitHub CLI** for improved dynamics and testing.

Projects

Scholaria Software – ALPHA Release

React Native, Expo, iOS, Node.js, PostgreSQL, Redis, Vercel, Render

- Onboarded **100+ users**, with 95% 30-day retention.
- Aggregated and curated **1,200+** new research papers weekly from arXiv across science and tech domains.
- Developed a **Progressive Web App** with **offline caching** using **Service Workers** for seamless reading.
- Built and deployed a scalable **React Native + Node.js**, with **Redis caching** and **PostgreSQL**, ensuring constant uptime.
- Integrated with **Clerk** for authentication, **Supabase**, and optimized deployments via **Render** and **Vercel**.

Inference Claim Classifier — Corgi YCS24

Python, OpenAI API, Vector DB, PDF Parsing, NLP

- Built an **automated pipeline** to process insurance claim and ledger entries from Smartsheet.
- Extracted key information using **OpenAI API**-powered document inference and prompt engineering.
- Designed and implemented a **vector database** for accurate retrieval and classification of claim documents.
- Achieved **79% MAPE** accuracy in document classification and claim detail extraction.

Skills

SWE: Java, PHP, SQL, Rails, JavaScript/TypeScript, Rust, Kotlin, MongoDB, Flask, AWS, React, API's, PostgreSQL, C/C++, Java, Django, Flask, MySQL, PyTest, Golang, A/B Testing, spark, Hadoop, Next.js, Angular, HTML, CSS
ML/DS: NN, Sentiment Analysis, Docker, Pandas, TensorFlow, PyTorch, SKlearn, SLAM, ROS, NLTK, RL, GCP, OpenAPI, JSON API, PIM, Abstraction Layer, ETL Processes