

Stephen Bottos

778-966-8141 | bottos.steve@gmail.com | linkedin.com/in/stephen-bottos | github.com/stevebottos

EXPERIENCE

Lead Machine Learning Engineer, Kibeam, Inc., Oakland, CA (Remote) Aug. 2023 – Present

- Architected multimodal agentic systems integrating vision, text, and audio (Langchain, MLFlow, Huggingface, OpenAI, Gemini, ElevenLabs) with custom spatial reasoning layers and RAG-enabled reflection patterns, reducing content generation turnaround time from days to under an hour across production workflows.
- Built end-to-end MLOps infrastructure (AWS ECS, Airflow, MLflow, Redshift) supporting 100+ models/week iteration and deployment of 1,000+ production models across tens of thousands of deployed units. Contributed to Unity-based synthetic data generation achieving production-grade performance.
- Enhanced ML data labeling efforts with Vision Transformers, enabling automated soft-labeling, eliminating the need for tedious and time consuming manual annotation. Optimized lightweight CNNs for ESP32 edge inference through int8 quantization and custom, hardware-specific layers.
- Led cross-functional ML strategy, defining technical roadmaps and research priorities to align agentic automation initiatives with product vision across engineering and product teams.

Senior Machine Learning Engineer, Plainsight, Inc., San Diego, CA Aug. 2021 – Aug. 2023

- Designed, trained, and deployed custom PyTorch architectures for multimodal applications (vision, video, text, LiDAR) supporting accounts up to \$10M. Scaled cloud workflows in GCP using VertexAI and custom Docker containers for training and RestAPI deployment.
- Developed internal platform using Vector Databases (Milvus, BigQuery, ClickHouse), and ANN algorithms, enabling rapid, automated train/validation/test set splitting with guaranteed quality data through deduplication and similarity inspection. Extended this technology to catalog and auto-annotate millions of open-source images from [LAION](#), allowing near-instant dataset supplementation, and improving model performance metrics across classification and object detection tasks by 10-20%
- Built Kubeflow and Apache Beam ETL pipelines handling large-scale data ingestion from client data lakes to internal warehouses.

Machine Learning Engineer, alwaysAI, Inc., San Diego, CA June 2020 – Aug. 2021

- Designed and deployed production Computer Vision models (TensorFlow, PyTorch) with quantization and pruning for real-time edge inference. Built AWS ECS-based training/inference platform handling ETL pipelines and model deployment.
- Invented patented object tracking and re-identification system ([US11915434B2](#)) enabling cross-camera tracking for physical analytics, deployed in \$1M+ contracts still in production today.

Machine Learning Engineer Qimia, Inc., San Diego, CA Sep. 2019 – June 2020

- Led data engineering and ML for large-scale advertising optimization across billions of records using Apache Spark. Developed spill detection and inventory analysis systems for robotics deployments in retail environments.
- Built PySpark/Scala data pipelines and predictive models on AWS Redshift, delivering insights through custom dashboards and Docker-based applications.

Machine Learning Research Engineer, University of Windsor, Windsor, ON Jan. 2018 – Aug. 2019

- First-authored three publications on eye-gaze tracking using Hidden Markov Models and Kalman Filters. Developed statistical models including PCA, Neural Networks, GMMs, SVMs, and Decision Trees for behavioral analysis from eye-tracking data.
- Thesis: *Statistical Methods to Measure Reading Progression Using Eye-Gaze Fixation Points* ([link](#)).

EDUCATION

University of Windsor *M.Sc, Electrical and Computer Engineering* Jan. 2018 – Aug. 2019 Windsor, ON

University of Windsor *B.Eng, Mechanical Engineering* Sep. 2012 – Aug. 2016 Windsor, ON

TECHNICAL SKILLS

ML/AI: LLMs (RAG, prompt engineering, LoRA/QLoRA, prefix tuning), Multimodal Models (CLIP, BLIP, VideoLLava, custom implementations), Vision Models (CNNs, ViTs, VJepa2), Agentic Systems, Vector Databases, Hierarchical Reasoning

Frameworks: PyTorch, TensorFlow, Langchain, MLflow, Huggingface, PySpark, Scikit-Learn, OpenCV, TensorRT, ONNX

Infrastructure: AWS (ECS, MWAA, ECR, Redshift), GCP, Azure, Docker, Airflow, Git/GitHub CI/CD

Languages: Python, C/C++, SQL, Scala, MATLAB