

Stephen Bottos

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

EXPERIENCE

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

- 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.
- Enhanced ML data labeling with Vision Transformers for automated soft-labeling, eliminating manual annotation. Optimized lightweight CNNs for ESP32 edge inference through int8 quantization and custom 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. Aug. 2021 – Aug. 2023
San Diego, CA

- Designed, trained, and deployed custom PyTorch model 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) for automated train/validation/test set splitting through deduplication and similarity inspection. Extended this to catalog and auto-annotate millions of images, improving model performance by 10-20%.

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

- 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. Sep. 2019 – June 2020
San Diego, CA

- 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 Researcher — University of Windsor Jan. 2018 – Aug. 2019
Windsor, ON

- First-authored three publications related to eye-gaze tracking and signal processing ([Researchgate](#)). Thesis: *Statistical Methods to Measure Reading Progression Using Eye-Gaze Fixation Points* ([link](#)).

EDUCATION

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

B.Eng, Mechanical Engineering — University of Windsor 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