Stephen Bottos

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

Summary

Senior Machine Learning Engineer with 7+ years building production ML systems across vision, text, audio, and LiDAR. Expertise in efficient inference under real constraints—from int8 quantization on edge devices to cost-optimized LLM pipelines. Builds and maintains MLOps infrastructure supporting 1,000+ models across 10k+ devices. Architects multimodal agentic systems integrating vision, text, and audio with foundation models. Leads cross-functional ML initiatives (20+ employees). Patent holder.

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

Experience

Lead Machine Learning Engineer — Kibeam, Inc. Aug. 2023 – Dec. 2025
Oakland, CA (Remote)

- Architected multimodal agentic systems integrating vision, text, and audio (Langchain, MLFlow for GenAI, Huggingface, OpenAI, Gemini, ElevenLabs). For multimodal implementations, custom architectural layers for pixel-level localization were required to address LLM spatial reasoning gaps. For internal content-generation workflows, RAG-enabled reflection patterns spanning dozens of agents reduced content generation from days to under an hour.
- Built end-to-end MLOps infrastructure from scratch (AWS ECS, Airflow, MLflow, Redshift), which supports 100+ model training runs per week, and manages deployment and tracking of 1,000+ unique production models across 10k+ edge devices in the wild today. Contributed to Unity-based synthetic data generation as part of this system, achieving production-grade generalization across all models without manual real data collection.
- Designed a bespoke, lightweight, hardware-optimized CNNs architecture for real-time edge inference on ESP32 through int8 quantization and custom layers, which is our default production model on all edge devices.
- Trained and deployed fine-tuned Vision Transformers and VLMs to analyze databases of internally collected images and propose soft-labels for regular re-training cycles as a form of offline knowledge distillation, ensuring continuous model improvement and eliminating manual annotation entirely.
- Led cross-team Machine Learning initiatives spanning 20+ employees, defining technical roadmaps, research direction, and development priorities across engineering and product teams. Directed advanced modeling and data curation efforts toward agentic automation to ensure direct alignment with key product vision and user needs.
- Supported the CEO directly in high stakes technical demonstrations in front of audiences both large and small, fully owning any and all ML/AI system components.

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

- Responsible for designing, implementing, deploying, and training custom PyTorch model architectures driving business logic for customer accounts worth up to $10M. While mostly Computer Vision models, many other modalities were leveraged on a per-use-case basis including video, text, and LiDAR pointcloud data.
- Pioneered use of foundation models (CLIP, Owl-ViT) for fine-tuning, zero-shot detection and auto-labeling. Leveraged early agentic systems like ViperGPT for compositional visual reasoning.
- Built and scaled cloud workflows in GCP, using custom Docker containers along with VertexAI to train and deploy models behind RestAPI endpoints for inference. Implemented CI/CD pipelines for automated model testing and deployment across multiple client environments.
- Improved upon internal data management software to support de-duplication, clustering, and exploration of millions of images/videos, accelerating data aquisition/QC workflows by 80% by leveraging embeddings and vector databases (Milvus, BigQuery, ClickHouse).
- Built Kubeflow and Apache Beam ETL pipelines handling large-scale data ingestion from client data lakes.

- Obtained Google Cloud Certified Professional Machine Learning Engineer certification.

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 platform handling ETL pipelines, model training, and serving infrastructure.
- Sole inventor of patented object tracking and re-identification system (US11915434B2). System consists of many deployed models across edge devices in a network which identify and track objects (usually people) across many cameras in a store in order to identify interactions in regions of interest, producing digital-like metrics in a physical environment. During employment, two contracts yielding upwards of $1M used this technology, and still do.

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

- Gained hands-on exposure to a broad range of ML applications, solely responsible for data engineering, data science, and machine learning demands of a large advertisement agency whose mission was to optimize campaigns regionally, with billions of data points relating to individuals across the United States. Also worked with other sub-contractors to develop early-stage spill detection and inventory analysis software commonly running on robots deployed to grocery/big-box stores.
- Utilized Apache Spark to facilitate data science workflows including analytics/BI, data manipulation, feature engineering, and predictive modeling at scale. Built PySpark/Scala pipelines and models on AWS Redshift, delivering insights upon Big Data through custom dashboards and Docker-based applications.

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

- First-authored three publications on eye-gaze tracking using Hidden Markov Models and Kalman Filters (Researchgate). 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).
- Primary area of research was, generally, developing and optimizing application-specific algorithms which utilize various machine learning/statistical modeling and predictive analysis techniques to extract information about an individual, given eye-gaze fixation data.
- Regularly developed models and techniques which include but are not limited to Principal Component Analysis, Neural Networks, Hidden Markov Models, Kalman Filters, K-Means Clustering, Gaussian Mixture Models, Support Vector Machines (SVM), and Decision Trees.

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