Alan Casallas

Machine Learning Engineer
(347) 813-2260 | alancasallas@gmail.com | https://alancasallas.com/ | www.linkedin.com/in/alancasallas

Professional Summary

Machine Learning Engineer (Master’s in Computer Science, MIT) with 6 years of experience designing ML systems, predictive models, and large-scale data pipelines (Kafka/Airflow/Spark). Proven track record in architecting high QPS and low-latency systems on AWS and Kubernetes. Deep expertise in transformers, LLM fine-tuning, and multi-modal architectures.

Professional Experience

East Summit Capital

Founder & Lead Engineer — October 2024 – June 2025

- Built and productionized an intraday trading pipeline handling **$500K intraday AUM**, with real-time feature extraction, model inference, and automated order execution via WebSockets with Interactive Brokers and Alpaca; implemented model validation, backtests, and risk controls that improved fill/slippage predictions in live investment strategies.
- Led feature engineering and hyperparameter tuning pipelines (NumPy, Pandas, scikit-learn, TensorFlow) for forecasting and execution models, including **XGBoost, GRU RNNs, and Random Forests**, applied to historical market data to optimize trade timing and execution.

Oracle

Senior Software Engineer — July 2021 - September 2024
Software Engineer — August 2019 - June 2021

- Worked in Oracle’s Moat division, part of **Oracle Advertising.**
- Maintained a feature ingestion pipeline that fed our bot detection IVT system, which processed user-agent, device, user behavior, and other features to flag requests as bots with high accuracy.
- Led migration of Moat’s Yield Intelligence system, a publisher-side ad yield optimizer, to **Spark** on **AWS EMR**, processing billions of impression and click events from **Kafka** pipelines to generate real-time viewability predictions.
- Designed and managed **Elasticsearch**/**Lucene** indexes to store and query large-scale prediction metrics, supporting ad campaign forecasting and model monitoring.
- Migrated our labeling system, which ingested 200 GB of data per day, to **Apache Airflow**, allowing us to shut down a fleet of always-on EC2 instances and saving 60% in costs.
- Served as lead tech designing and deploying Nados**, a real-time inference platform**, on Kubernetes, resulting in $700,000/month savings compared to its previous deployment in AWS ECS. Nados was Moat’s central inference platform, supporting bot detection, pre-bid scoring, yield intelligence, and NLP-based brand safety scoring across multiple regions, handling **12M+ predictions/minute** at less than **50ms latency**.
- Mentored engineers, shaping team practices around data quality and scalable ingestion.

- Served as Scrum Master during several sprints, monitoring and unblocking the progress of team members to achieve an average of 90% ticket completion during sprints I monitored.
- Worked with Oracle Security team to ensure systems handling IP address and user agent data complied with security and PII requirements.

Projects

CasaLLM - An LLM Created From Scratch July 2025 - August 2025

- Trained a 350M parameter LLM from scratch in PyTorch using the transformer GPT architecture, optimized with TF32, BF16, and Flash Attention. Implemented RoPE and kv caching. Performed pretraining, fine tuning (SFT) and DPO over several days.
- Live demo: https://huggingface.co/spaces/alancasallas/casallm-ui

Custom CLIP Implementation July 2025 - August 2025

- Trained a 36M-parameter multi-modal CLIP network from scratch in PyTorch on 3M image-text pairs, leveraging contrastive learning and experimenting with an RNN text encoder.

Skills

Languages: Python, C++, Go, SQL
ML/AI: PyTorch, TensorFlow, Transformers, Scikit-learn, Hugging Face, LangChain, QLoRA, RLHF reward modeling/alignment, Retrieval-Augmented Generation (RAG), multi-modal LLMs, Bandits, Recommendations, Personalization
Infra & Systems: Spark, Kafka, Airflow, PostgreSQL, Elasticsearch, Weights & Biases (wandb), AWS (S3, EC2, EMR, SQS), Kubernetes, Prometheus, Elasticsearch, Grafana

Education

Massachusetts Institute of Technology (MIT) — Cambridge, MA
Master of Engineering (M.Eng.), Electrical Engineering & Computer Science (EECS) • GPA: 5.0/5.0 • Sep 2017 – Aug 2019

- **Thesis:** Applied AI/ML thesis on current estimation using point magnetic-field measurements; applied **signal processing**, **linear regression**, **autoencoders**, and **generalized least squares (GLS)** for sensor replacement with machine learning. US Patent no. US12085591B2.
- Selected Coursework: Statistical Learning, Computer Vision, Feedback Control, Distributed Systems.

Massachusetts Institute of Technology (MIT) — Cambridge, MA
Bachelor of Science, Computer Science • GPA: 4.9/5.0 • Sep 2013 – May 2017

- Selected Coursework: Computer Architecture, Advanced Algorithms.