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 systems, and backend systems. Deep expertise in transformers, LLM fine-tuning, and orchestrating AI agentic systems. Proven track record in architecting high-availability backend systems on AWS and Kubernetes, processing millions of requests per minute with sub-50ms latency.

Projects

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

- Created 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. Trained over several days, including pre-training, fine-tuning (SFT), and RLHF.
- 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.

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 trading 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 Data Cloud** (later **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.

- Served as lead tech migrating our Yield Intelligence system to **Spark on AWS EMR**, which processed ad click metrics collected by a Kafka pipeline to generate viewability predictions which were later stored on **Redis**.
- 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 our Nados application on **Kubernetes** in Oracle Cloud Infrastructure (OCI), resulting in $700,000/month savings compared to its previous deployment in AWS ECS. Nados was a latency-sensitive application deployed in multiple regions, responding to over **12 million requests/minute** at under **50 millisecond latency**, and was the second most expensive system in the Moat division.
- Served as lead tech for the migration of Moat's largest table, a **4 TB Postgres table**, into its own **PostgreSQL** database using pglogical and later into its own **MySQL** database with minimal downtime, resulting in 70% cost reduction.
- Mentored and onboarded engineers, shaping team practices around data quality, scalable ingestion, and predictive modeling at scale.
- 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.

Skills

Languages: Python, C++, Go, SQL
ML/AI: PyTorch, TensorFlow, Transformers, deep learning, natural language processing, scikit-learn, Hugging Face, LangChain, LangGraph, QLoRA, RLHF, Retrieval-Augmented Generation (RAG), multi-modal LLMs, bandits, recommendations, ranking, NLP
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.