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

Projects

Custom CLIP
Self-Directed June 2025 - August 2025

- Created a 15M parameter implementation of CLIP from scratch in Pytorch trained on 4M image-text pairs, achieving a one-shot performance of 60.6% on Imagenet. Used Weights and Biases (wandb) for experiment tracking and hyperparameter tuning.

CasaLLM - An LLM created from scratch
Self-Directed June 2025 - August 2025

- Coded a 400M parameter LLM from scratch in Pytorch using the transformer GPT architecture. Performed supervised pre-training, SFT, and RLHF on a 4090 GPU over 10 days. Used Huggingface tokenizer for BPE tokenization and FastAPI to build the serving platform.

Professional Experience

Casallas Capital
Founder & Lead Engineer November 2024 – May 2025

- Designed backend system to ingest realtime prices and place trades using websockets with Interactive Brokers and Alpaca, placing an average of 30 trades a day per symbol. Assets held during the day totaled $400,000.
- Ran feature engineering pipelines and hyperparameter tuning using numpy, pandas, scikit-learn, Keras, and Tensorflow to test random forests, xgboost, and GRU RNN's on historical data to predict price action, fill prices, and slippage.

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

- Worked in Oracle's Moat division, part of the adtech organization **Oracle Data Cloud** (later **Oracle Advertising**)

- Served as lead tech migrating our Yield Intelligence system to Spark on AWS EMR, which processed click metrics collected by a Kafka pipeline and stored as Parquet files to generate viewability predictions using Wilson Score rating and then used an AWS SQS queue to store them 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.
- Maintained a feature ingestion pipeline that fed our bot detection ivt system, which used user agent, device, and other impression information to flag requests as bots with over 90% accuracy.
- Served as lead tech designing and deploying our Nados application 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.
- Created Kubernetes deployment for the Nados system, including the Load Balancer Service, HorizontalPodAutoscaler (HPA), and IngressController k8s objects.
- Set up observability for Nados using Prometheus, Grafana, and Elasticsearch. Along with multi-region failover, this helped us achieve a 99.99% uptime.
- Migrated Nados ad blocking rules from an in-memory json to a distributed Redis server, using distributed locks to achieve 99.9% availability and eventual consistency.
- 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.
- Developed a Flask application called Enrichments that served as an API for customers to update ad campaign settings.
- Set up a failover system on Akamai to AWS S3 for our Pixie system, which was the first point of ingestion for impression information, ingesting 10 million requests/minute.
- 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, Bash

ML/AI: PyTorch, Transformers, CLIP, scikit-learn, Hugging Face, RLHF, RAG, multimodal LLMs, Fine-Tuning, AI Agents, Distributed Training (DDP), bandits, recommender systems, A/B testing.

Data & Streaming: Apache Spark, Apache Kafka, Apache Airflow, Apache Parquet, PostgreSQL, MySQL, Redis, Elasticsearch, Weights & Biases (wandb), AWS (S3, EC2, EMR, Lambda, SQS), Kubernetes, Docker, Terraform, Prometheus, Elasticsearch, Grafana, Git

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:** Contactless voltage/current estimation using point magnetic-field measurements; applied **signal processing**, **linear regression**, **autoencoders**, and **generalized least squares (GLS)** for sensor replacement with **machine learning**. Resulting product was patented as 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.