Sebastian Algharaballi-Yanow

Machine Learning Engineer & Data Scientist

San Diego, CA | sebastianalgharaballi@gmail.com | GitHub | Linkedin | Website

Technical Skills:

Programming Languages/Frameworks:

- Python, JavaScript, R, SQL. Extensive experience with FastAPI, React, Pandas, Numpy, Seaborn, Sci-kit Learn, MatPlotLib, PyTorch, Tensorflow, NLTK, Spacy, OpenCV, SciPy, Transformers, LangChain, and LLM APIs.

Machine Learning/Artificial Intelligence:

- Large Language Models (LLMs), RAG (Retrieval-Augmented Generation) architectures, Agentic frameworks, Vector embeddings, Prompt Engineering, Supervised/Unsupervised Learning, Deep Learning, Computer Vision, MLOps, Human-Centric AI.

Data Science & Enterprise Systems:

- Data Pipeline Development, ETL processes, Exploratory Data Analysis (EDA), Data Visualization, Statistical Analysis, Production-scale AI deployment, Enterprise API development.

Cloud Platforms & Tools:

- AWS Bedrock, AWS SageMaker, Google Cloud Platform, Docker, Kubernetes, CI/CD systems, PGVector, PostgreSQL, MySQL, Git/GitHub, GitHub Copilot, Tableau, Power BI.

Education:

UC Irvine – *Master of Data Science*
September 2023 - December 2024

UC San Diego – *Bachelor of Science in Cognitive Science: Machine Learning & Neural Computation*
September 2020 - June 2023

University of California, San Diego Extension – *Specialized Certificate, Machine Learning*
June 2022 - June 2023

Professional Experience:

Stealth Talent Solutions - *Lead Machine Learning/Artificial Intelligence Engineer*
November 2024 - Present

- **Built an enterprise-scale AI automation system** that replaced manual recruiter workflows, autonomously generating **200+ tailored job descriptions** daily through intelligent document creation and database integration.

- **Architected full-stack RAG solution** with Dockerized FastAPI backend, enabling recruiters to process high-volume candidate matching at enterprise scale (**15,000+ searches/hour capacity**).

- **Developed a custom transformer-based parsing engine** that automatically extracts and structures data from unstructured resumes and job descriptions, identifying 150+ successful placements that traditional keyword systems missed, **demonstrating AI's ability to assist complex human decision-making processes.**

- **Optimized vector embedding pipeline** using PGVector and AWS Bedrock, reducing similarity query latency by 30% (200ms → 140ms) while maintaining **enterprise-grade performance standards**.

- **Implemented multi-LLM orchestration system** using LangChain with dynamic switching between OpenAI, Mistral, and AWS Bedrock models, **reducing deployment cycles** from 1 hour to under 5 minutes and improving system reliability.

- **Led cost optimization initiatives** reducing monthly cloud compute spend by 30% ($8K → $5.6K) through **LLM quantization** and infrastructure optimization while maintaining SLA requirements.

- **Provided technical leadership** on prompt engineering best practices and model evaluation frameworks, fostering collaborative development of production AI systems.

Scale AI - *Generative Artificial Intelligence Prompt Engineer*
April 2024 - Present

- **Engineered advanced prompt frameworks** utilizing chain-of-thought and few-shot learning techniques for production-scale LLMs, reducing response latency from 2.3s to 1.1s while improving task completion accuracy from 88% to 95%.

- **Implemented enterprise-grade RLHF pipeline processing** 500,000+ interactions to enhance emotional intelligence in conversational AI, driving beta satisfaction scores from 3.6/5 to 4.67/5.

- **Established evaluation frameworks** for measuring LLM performance across edge cases, increasing "relevant and humanistic" response ratings from 61% to 83% through systematic testing and validation.

- Collaborated with project managers and engineering teams to integrate **prompt optimization techniques into production workflows**, enabling standardized deployment processes across multiple client applications.

Plink.bio - *Software Engineer - GenAI*
October 2024 - February 2025

- **Architected multi-modal content analysis pipeline** combining computer vision, OCR, and speech-to-text processing to automatically extract comprehensive metadata from creator content, processing videos in under 3 seconds at scale.

- **Built end-to-end LLM recommendation system** that analyzes multi-language creator content and generates personalized strategy recommendations by processing visual elements, transcripts, and engagement patterns—replacing manual content analysis workflows.

- **Developed real-time computer vision models** for automated object and brand detection, achieving 90% accuracy across 1000+ test frames for identifying monetizable product placement opportunities.

- **Integrated AI pipeline with creator platform infrastructure** through RESTful APIs, enabling content analysis capabilities for the platform's user base while maintaining sub-200ms response times.

MoodMe - *Lead Machine Learning Engineer & Co-Founder*
October 2023 - October 2024

- **Expanded AI model capabilities** across 7 major demographic groups, improving overall emotion detection accuracy from 68% to 87% and reducing bias in underrepresented populations by 62% through advanced data pipeline engineering.

- **Enhanced production emotion detection system** using transfer learning, boosting accuracy from 75% to 91.5% across 8 emotion categories for enterprise-scale deployment.

- **Created MoodMirrors wellness platform** powered by customized BERT model achieving 89% F1 score, resulting in 41% increase in user emotional self-awareness—demonstrating AI's ability to enhance human experiences beyond conversation.

Sportradar US - *Sports Data Analyst*
September 2022 - October 2024

- **Optimized data collection workflows** across **250+** NCAA and professional sporting events (basketball, baseball, volleyball, soccer) by suggesting data pipeline modifications within the basketball play-by-play workflow, **reducing** average input time per play from **8 seconds to 3 seconds** and achieving top 10% performance ratings nationwide.

Projects & Research:

Advanced NBA Referee Analysis: (Research Paper)

- Developed a comprehensive dataset (30,000+ data points) and created four neural network models to analyze referee decision-making patterns, achieving 92% test accuracy and highlighting AI's potential to support human judgment in complex, real-time scenarios.

Natural Language Financial Analytics on CEO Communication: (Presentation)

- Built an analysis pipeline using text preprocessing, TF-IDF, SVD, and sentiment analysis to investigate relationships between CEO earnings call language and financial performance, uncovering industry-specific correlation patterns between communication sentiment and financial metrics.