Andres Vasquez-Restrepo

(+57)3005258250 | anvasquezre@gmail.com | linkedin.com/in/anvasquezre | github.com/anvasquezre | Medellin, Colombia

Senior Machine Learning Engineer with over five years (5+) of experience specializing in NLP, Document AI, speech analytics, and unstructured data analysis. I have designed and deployed more than 8 AI solutions in production, optimizing pipelines that process thousands of requests per hour and hundreds of documents and audio hours with 95% accuracy. My expertise spans PyTorch, OpenCV, YOLO, and LangChain, with a strong focus on MLOps to ensure scalable, reliable AI systems.

I have led projects that analyze large-scale text and speech data, enabling 75% faster decision-making and automating hundreds of hours of manual work per week. I thrive on bridging the gap between AI research and business impact, translating complex models into real-world solutions that drive measurable results.
Beyond coding, I enjoy simplifying technical concepts for engineers, business teams, and clients—because clear communication is as vital as well-optimized algorithms.

Education

Master of Science in Artificial Intelligence — La Rioja, España
Universidad Internacional de la Rioja — *Sept. 2024 – Present*

Master of Science in Computational Biology and Bioinformatics — Medellín, Colombia
Universidad Nacional de Colombia — *Aug. 2019 – Dec. 2021*

Bachelor of Engineering in Biotechnology — Medellín, Colombia
Universidad Nacional de Colombia — *Jan. 2014 – Aug . 2019*

Experience

Senior Machine Learning Engineer (L4) — Aug. 2024 – Present
Provectus — *SF, USA, Remote*

- Led the development of machine learning workflows for data processing, model training, evaluation, and deployment in "The Sphere" project.
- Designed and built an automatic speaker recognition pipeline from videos using a 40-GPU cluster. The pipeline integrated text-to-speech, speaker diarization, silence detection, and face matching.
- Developed a hybrid-search engine combining keyword and semantic search with LLM-driven metadata augmentation, enabling efficient retrieval of video content.
- Reduced batch processing costs for video analysis from $5,000 to $100 while achieving 95% accuracy in identifying members present in over 10,000 video hours.
- Achieved 80% MAP@12 performance for the search engine, significantly improving video content discoverability.
- Led the Data and ML team, managing 2 mid-level ML engineers and 2 data engineers, ensuring alignment with project goals and delivering high-impact solutions.
- Implemented MLOps best practices, including data and model versioning, experiment tracking, and reproducibility for large-scale LLM deployments.
- Developed and optimized distributed training pipelines using multi-GPU/TPU architectures for large-scale NLP models.
- Designed and deployed cloud-native ML solutions on AWS, leveraging serverless architectures, cost optimization strategies, and multi-cloud integrations.
- Integrated large-scale text data pipelines with cloud-based data lakes and warehouses, ensuring efficient storage, retrieval, and processing.
- Collaborated with cross-functional teams to align ML models with business objectives, ensuring high-impact AI solutions.

Lead Machine Learning and Artificial Intelligence Engineer — Jul. 2023 – Aug. 2024
Tenant Evaluation — *FL, USA, Remote*

- Designed, developed, and deployed over 5 AI/ML solutions while mentoring junior developers.
- Engineered a production-grade AI-driven virtual assistant serving 20,000 users monthly, reducing support ticket load by 50%.
- Architected a microservices framework with 6 components integrating REST APIs and WebSockets.

- Implemented a Document AI platform automating 75% of internal document processing, cutting costs from $9,000 to $236 per month.

Machine Learning Engineer Sep. 2022 – Jul. 2023
AnyoneAI *SF, USA, Remote*

- Developed and deployed 3+ end-to-end ML solutions for global clients, covering recommendation systems and classification models.
- Built and maintained ML models for NLP, Computer Vision, and Business Intelligence applications.
- Automated e-commerce reporting with Apache Airflow, reducing report generation time from hours to minutes.

Data Scientist Sep. 2021 – Sep. 2022
Grinsup *Medellín, Colombia*

- Developed pricing algorithms that maintained a 40% gross profit margin, boosting net profit by 10%.
- Created nutritional estimation models using multi-objective optimization for efficient product development.
- Recognized as a leader in Colombian innovation and selected for the Young Leaders of the Americas initiative.

Data Analyst Sep. 2019 – Sep. 2021
Centricol Industries *Medellín, Colombia*

- Led market analysis to develop new biotechnology product lines, increasing revenue by 50% in two years.
- Improved data accessibility by 30% through centralized data repository implementation.
- Generated sales reports that enabled a 50% increase in revenue through targeted product strategies.

Projects

Job Recommendation System for Hunty Mar 2023

- Developed an end-to-end job recommendation API integrating daily web scraping using Ada-02 embeddings.
- Implemented FastAPI with MongoDB and MariaDB for scalable data storage.
- Automated job ingestion using DAGs in Apache Airflow, reducing manual work and improving efficiency.
- Enabled the platform to post over 1,000 new jobs daily and rank them based on user profiles.

ChatFLOW Jan 2023

- Developed an open-source, drag-and-drop chatbot design platform similar to Voiceflow.
- Implemented state-of-the-art NLP techniques and Generative AI models to enhance user interactions.
- Built modular components enabling users to craft detailed conversational flows with minimal coding.

Technical Skills

Programming Languages: Python, Javascript, Java
Machine Learning: SQL, TensorFlow, Pytorch, Scikit-learn, XGBoost, LightGBM, LLM, SML, Finetunning, HuggingFace, LangChain, OpenCV, YOLO, Llama.cpp
Data Processing: Pandas, NumPy, NLTK, OpenCV, NetworkX, Apache Airflow
Cloud & MLOps: AWS (EC2, EKS, Lambda, S3, SageMaker, CloudWatch, RDS, DocumentDB, SQS), Kubernetes, Docker, MLflow, ArgoCD, Jenkins, Rundeck, Grafana, Bash
Web Frameworks: FastAPI, Django
Databases: PostgreSQL, MongoDB, Redis
Developer Tools: Git, Bitbucket, Jira

Certifications

AWS AI Practitioner: AWS
English C1 Proficiency: Duolingo English Test 120/160
Applied Machine Learning: University of Michigan
Machine Learning with Python: IBM
Leadership and Team management: University of Florida
Leader of the Americas: Department of State, United States