Ismael Contreras Mejia
Architect/Team-lead Senior Machine Learning Engineer

SUMMARY

- Senior Machine Learning Engineer with 10+ years architecting scalable AI systems, specializing in deep learning, LLM orchestration, and MLOps across AWS, Azure, and Databricks.
- Expertise in Python, C++, Java, Rust, and Go, with advanced skills in PyTorch, TensorFlow, LangChain, Kubernetes, and distributed computing frameworks.
- Proven track record delivering enterprise-grade AI products including multi-agent LangGraph pipelines, real-time voice tutoring platforms, and decentralized ML infrastructure.
- Strong background in computer vision, NLP, and autonomous systems, with hands-on experience in YOLO, OpenCV, CUDA, and GPU optimization.
- Master's degree in Computer Science, skilled in CI/CD, containerization, microservices, and compliance-driven software development for high-impact AI applications.

TECHNICAL SKILLS

Main Technical Skills	Python (10 yr.), PyTorch (6 yr.), TensorFlow (6 yr.), LangChain (1 yr.), Kubernetes (4 yr.)
Programming Languages	Objective-C (3 yr.), Python (10 yr.), Vyper (1 yr.)
AI & Machine Learning	Apache Mahout (1 yr.), ElevenLabs (1 yr.), Hugging Face (2 yr.), Kubeflow (2 yr.), LangChain (1 yr.), LangGraph (1 yr.), LlamaIndex (2 yr.), Mlflow (3 yr.), OpenAI (1 yr.), OpenCV (6 yr.), PandasAI (6 yr.), PyTorch (6 yr.), Scikit-learn (6 yr.), Spacy (1 yr.), TensorFlow (6 yr.), Xgboost (4 yr.), YOLOv5 (2 yr.), YOLOv8 (1 yr.)
.NET Platform	Azure (1 yr.)
Python Frameworks	Django REST framework (1 yr.), FastAPI (1 yr.), Flask (2 yr.)
JavaScript Libraries and Tools	Electron (1 yr.), three.js (2 yr.)
JavaScript Frameworks	Express (3 yr.), Node.js (5 yr.), three.js (2 yr.)
Go Libraries and Tools	Libp2p (2 yr.)
Python Libraries and Tools	Matplotlib (1 yr.), PySpark (4 yr.), PyTorch (6 yr.), Scikit-learn (6 yr.), TensorFlow (6 yr.)
Mobile Frameworks and Libraries	Parse (1 yr.)
Java Frameworks	Struts 2 (1 yr.)

Data Analysis and Visualization Technologies	Apache Airflow (4 yr.), Apache Mahout (1 yr.), Apache Spark Streaming (4 yr.), Databricks (1 yr.), ETL Pipelines (1 yr.), PandasAI (6 yr.)
Security	Microsoft Entra (1 yr.)
Databases & Management Systems / ORM	Apache Spark Streaming (4 yr.), Cosmos DB (1 yr.), dbt (2 yr.), MongoDB (3 yr.), Neo4j (1 yr.), NoSQL (4 yr.), PostgreSQL (3 yr.), Redis (3 yr.), Snowflake (1 yr.), SQL (6 yr.)
UI Frameworks, Libraries, and Browsers	Semantic UI (1 yr.)
Cloud Platforms, Services & Computing	Azure (1 yr.), GCP (1 yr.)
Google Cloud Platform	Cloud Functions (1 yr.)
Azure Cloud Services	Cosmos DB (1 yr.), Databricks (1 yr.), Microsoft Azure API (1 yr.)
Industry Domain Experience	HIPAA (3 yr.)
Deployment, CI/CD & Administration	Active Directory, CI/CD (4 yr.), CircleCI (1 yr.)
Message/Queue/Task Brokers	Celery (1 yr.), Kafka (4 yr.), RabbitMQ (3 yr.)
iOS Libraries and Tools	Core Audio (1 yr.)
Virtualization, Containers and Orchestration	Docker (4 yr.), Kubernetes (4 yr.), Terraform (4 yr.)
BlockChain and Decentralized Software	ETH (Ethereum blockchain), IPFS (InterPlanetary File System), Vyper (1 yr.)
SDK / API and Integrations	Facebook API (1 yr.), FastAPI (1 yr.), Google API (1 yr.), Microsoft Azure API (1 yr.)
Version Control	Github Actions (4 yr.)
Mail / Network Protocols / Data transfer	GRPC (3 yr.), NAT (1 yr.), WebRTC (1 yr.)
QA, Test Automation, Security	Locust (2 yr.)
Methodologies, Paradigms and Patterns	MVC (1 yr.)
Web/App Servers, Middleware	Oracle WebLogic Application Server (2 yr.)
Logging and Monitoring	Prometheus (3 yr.)
Platforms	SharePoint (1 yr.)
Collaboration, Task & Issue Tracking	Slack (1 yr.)

Other Technical Skills	AD, beams (1 yr.), ByteTrack (1 yr.), CUDA (4 yr.), Demucs (1 yr.), Feast (1 yr.), Flink (1 yr.), Flyte (1 yr.), GDPR (3 yr.), Google PageSpeed Insights (2 yr.), Halo2 (1 yr.), InterPlanetary File System, Kuzu (1 yr.), MeshCNN (1 yr.), NCCL (2 yr.), NeRFs (1 yr.), NLLB-200 (1 yr.), Open3d (2 yr.), PIL (1 yr.), Qdrant (1 yr.), SD, Silero VAD (1 yr.), SOC2 (3 yr.), Spark MLLib (3 yr.), Speech-to-Text, Stable Diffusion, STT, Tecton (1 yr.), V-ray (2 yr.), Weaviate (2 yr.)

WORK EXPERIENCE

Senior AIOps Manager - Contract - Tredence (Copilot (Agentic Contract/NDA Assistant on Databricks + Copilot Studio))

Duration: Nov 2025 – Present
Summary:

- Developed an enterprise-grade multi-agent AI assistant for contract and NDA analysis using LangGraph and Azure OpenAI services, deployed on Databricks
- The system enables natural language clause lookup, summarization, and metadata-aware document retrieval to streamline contract management workflows

Responsibilities:

- Engineered multi-agent LangGraph pipeline for intent detection, classification, semantic retrieval, and response generation.
- Integrated Azure OpenAI, Azure AI Search, and SharePoint for enhanced document search and retrieval.
- Deployed scalable MLflow PyFunc model with versioning and serving endpoints.
- Implemented metadata extraction with Azure Function Apps and OCR.
- Enabled conversational contract analysis within Microsoft Teams and Power Platform workflows.

Technologies: LangGraph, Azure OpenAI (GPT-4.1 + embeddings), Azure AI Search (HNSW + OCR + hybrid semantic search), SharePoint, Databricks MLflow PyFunc, Python, Azure Function Apps, Microsoft Copilot Studio, REST APIs

Senior AIOps Manager - Contract - Tredence (Sales Agent NL2SQL System (Agentic AI Assistant for Sales Performance Management))

Duration: Nov 2025 – Present
Summary: Designed and implemented an agentic natural language to SQL analytics system to enable non-technical sales stakeholders to query sales performance data using natural language, improving accessibility and reducing BI turnaround times.
Responsibilities:

- Developed multi-stage LangGraph orchestration pipeline for query expansion, intent extraction, schema mapping, SQL generation, execution, and natural language response.
- Integrated Azure OpenAI for intent classification and SQL synthesis, and Databricks Vector Search for schema retrieval.
- Built robust SQL generation pipeline ensuring schema-correct Spark SQL output.
- Packaged workflow as MLflow pyfunc model with secure inference.
- Generated business-readable narrative summaries from query results.

Technologies: LangGraph, Azure OpenAI (GPT-4), Databricks Vector Search, Spark SQL, MLflow pyfunc, Python, Azure AD/MSAL

Senior AIOps Manager - Contract - Tredence (AI-Powered Real-Time Tutoring Platform (GROW Classroom – Paper Education))

Duration: Nov 2025 – Present
Summary: Architected a full-stack AI tutoring system enabling real-time voice-based lessons with adaptive instruction and multi-agent AI orchestration, supporting interactive online learning experiences.
Responsibilities:

- Developed AI agent orchestration with LLM providers, STT, and TTS using a provider factory architecture.
- Built low-latency conversational pipelines integrating VAD, turn detection, streaming STT/TTS, and real-time communication.
- Implemented session state orchestration, persistence services, and AI safety monitoring infrastructure.
- Delivered scalable full-stack web interface with Next.js, TypeScript, React, and LiveKit SDK.

Technologies: Python, LiveKit WebRTC, Next.js, TypeScript, React, Google Gemini, OpenAI, Deepgram STT, ElevenLabs TTS, Silero VAD, Pydantic, Google Cloud Storage, Slack

Senior AIOps Manager - Contract - Tredence (MCP File Edit Server for Claude Desktop)

Duration: Nov 2025 – Present
Summary: Built a FastMCP-powered file server to enable Claude Desktop AI to manipulate remote and local files asynchronously, supporting code analysis, editing, and repository operations over SSH.
Responsibilities:

- Implemented abstraction layer for local and SSH file systems with AST parsing, regex, and Git integration.
- Registered over 30 MCP tool operations with error handling and safety validation.

Technologies: FastMCP, Python, AST parsing, Git, SSH

Senior AIOps Manager - Contract - Tredence (Low-Resource ASR/NMT (Tamasheq – Digital Prybar))

Duration: Nov 2025 – Present
Summary: Developed an offline-capable speech-to-text and translation pipeline for low-resource languages, integrating diarization and noise reduction to support offline portability and human-in-the-loop evaluation.
Responsibilities:

- Built ASR and NMT pipeline using Facebook MMS-1B and NLLB-200 models.
- Implemented diarization and noise reduction components.
- Delivered API and GUI interfaces containerized with Docker.
- Integrated human-in-the-loop evaluation workflows.

Technologies: Facebook MMS-1B, NLLB-200, pyannote.audio, Demucs, Silero VAD, Flask REST API, PyQt, Docker, Label Studio

Senior AIOps Manager - Contract - Tredence (Flexible GraphRAG Platform (Hybrid RAG + Knowledge Graph))

Duration: Nov 2025 – Present
Summary: Engineered a modular hybrid retrieval-augmented generation (RAG) platform combining document processing, knowledge graph construction, and multi-database hybrid search with LLMs for enterprise deployments.
Responsibilities:

- Designed FastAPI microservices for asynchronous ingestion and real-time progress updates.
- Supported multiple data formats and integrated vector and graph databases.
- Delivered three full frontends using React, Angular, and Vue.

Technologies: FastAPI, OpenAI, Google Gemini, LlamaIndex, Qdrant, Weaviate, pgvector, Neo4j, Kuzu, React, Angular, Vue

Senior Machine Learning Engineer - Waste Management (Automated Invoice Processor (Amazon Bedrock + Claude 3.5))

Duration: Jul 2025 – Oct 2025
Summary:

- Designed a scalable AI pipeline to extract structured data from compressed natural gas billing PDFs using Claude 3
- 5 Sonnet via Amazon Bedrock, automating billing reconciliation processes

Responsibilities:

- Converted PDFs to high-resolution images and parsed content with custom JSON prompts.
- Automated billing reconciliation using AWS S3, Boto3, Pandas, and OpenPyXL.

Technologies: Claude 3.5 Sonnet, Amazon Bedrock, AWS S3, Boto3, Pandas, OpenPyXL, Python

Senior Machine Learning Engineer - Waste Management (Truck Detection and Fraud Analysis)

Duration: Jul 2025 – Oct 2025
Summary: Developed a video analytics pipeline to detect trucks entering and exiting landfill sites and identify fraudulent transactions by matching detections with ticket data.
Responsibilities:

- Implemented YOLOv8 and ByteTrack for truck detection and tracking.
- Applied geometric lane mapping, direction validation, and fraud heuristics.
- Integrated with Snowflake and FastLane APIs and deployed multi-camera processing workflows.

Technologies: YOLOv8, ByteTrack, Snowflake, FastLane APIs, Python

Senior Machine Learning Engineer - Waste Management (Integrated Insurance Agentic AI System)

Duration: Jul 2025 – Oct 2025
Summary: Developed a LangChain-based multi-agent AI system automating insurance workflows including document ingestion, validation, communication, and audit tracking to

improve operational scalability.
Responsibilities:

- Designed agents for OCR, signature check, policy and VIN matching, and customer notifications.
- Implemented secure S3 uploads, Redis/Celery queueing, PostgreSQL JSON models, and tokenized document access with audit logging.

Technologies: LangChain, Django, React, OCR, Redis, Celery, PostgreSQL, AWS S3

Senior Machine Learning Engineer - Waste Management (Flight Booking with Semantic Kernel & AutoGen)

Duration: Jul 2025 – Oct 2025
Summary: Built a multi-agent flight booking system using Microsoft AutoGen and Semantic Kernel, enabling fully automated bookings from natural language instructions.
Responsibilities:

- Orchestrated reasoning and action agents for flight search and booking.
- Integrated Azure CosmosDB for real-time flight data storage.
- Developed a web app with Python, Flask, and HTML/CSS.

Technologies: Microsoft AutoGen, Semantic Kernel, Azure CosmosDB, Azure OpenAI (GPT-4), Python, Flask, HTML/CSS

Senior Machine Learning Engineer - Waste Management (Agentic AI Implementation for Financial Crime Compliance (FCC) Automation)

Duration: Jul 2025 – Oct 2025
Summary: Designed and implemented autonomous AI agents to perform adverse media monitoring, sanctions screening, and potential sanctions detection within enterprise compliance workflows for financial crime compliance.
Responsibilities:

- Built multi-step AI pipelines using Python, LangChain, spaCy, TensorFlow, PyTorch, and Scikit-learn.
- Developed autonomous data-gathering agents integrating external information providers and web sources.
- Implemented context-aware decision-making logic and explainable AI mechanisms.
- Deployed AI agents on distributed enterprise platform using Java (Spring), RabbitMQ, ELK, Keycloak, and GCP.

Technologies: Python, LangChain, spaCy, TensorFlow, PyTorch, Scikit-learn, Java (Spring), RabbitMQ, ELK, Keycloak, GCP

Senior Machine Learning Engineer - Mashgin (AI/Computer Vision for Touchless Checkout)

Duration: Jun 2021 – Jun 2025
Summary: Developed and optimized computer vision models for Mashgin's touchless self-checkout kiosks, achieving high accuracy and low latency for retail environments.
Responsibilities:

- Developed PyTorch models (e.g., EfficientNet-B4) for item recognition across 60,000+ SKUs.
- Deployed models on NVIDIA Jetson Nano using ONNX Runtime to reduce inference latency.

- Created synthetic data pipeline with Blender and Python for training data generation.
- Implemented 3D reconstruction with Open3D for product digitization.

Technologies: PyTorch, EfficientNet-B4, NVIDIA Jetson Nano, ONNX Runtime, Blender, Python, Open3D

Senior Machine Learning Engineer - Mashgin (MLOps & Cloud Infrastructure)

Duration: Jun 2021 – Jun 2025
Summary: Built and maintained MLOps pipelines and cloud infrastructure to support continuous model updates and reliable deployment across hundreds of kiosks.
Responsibilities:

- Developed CI/CD pipelines with GitHub Actions and Kubernetes.
- Designed Redis-based feature store for real-time feature serving.
- Managed AWS infrastructure using Terraform.
- Implemented monitoring with Prometheus and Grafana.

Technologies: GitHub Actions, Kubernetes, Redis, AWS (EC2, S3, SageMaker), Terraform, Prometheus, Grafana

Senior Machine Learning Engineer - Mashgin (Data Engineering)

Duration: Jun 2021 – Jun 2025
Summary: Processed large-scale transaction data and built real-time inventory management APIs to support GDPR-compliant retail deployments.
Responsibilities:

- Processed 8TB/day of transaction data using PySpark on Databricks with Delta Lake.
- Streamed transaction events via Apache Kafka.
- Built Node.js REST APIs for real-time inventory management integrated with checkout kiosks.

Technologies: PySpark, Databricks, Delta Lake, Apache Kafka, Node.js, REST APIs

Senior Machine Learning Engineer - Mashgin (Financial Systems)

Duration: Jun 2021 – Jun 2025
Summary: Architected and tested a high-performance payment routing system integrated with contactless payment methods for healthcare and retail environments.
Responsibilities:

- Developed Go-based payment router handling over 2 million transactions per day with low latency.
- Conducted performance testing with Locust.

Technologies: Go, Locust

Senior Machine Learning Engineer - Mashgin (Sales Analytics & Forecasting (Time Series ML Project))

Duration: Jun 2021 – Jun 2025
Summary: Led development of a sales forecasting system for over 1,100 retail partner locations using advanced time series machine learning techniques.
Responsibilities:

- Implemented DTW-based time series clustering and engineered temporal, promotional, holiday, and store-specific features.

- Built stacking ensemble models using XGBoost, Gradient Boosting, Linear Regression, Random Forests, and Decision Trees.
- Created full ML pipeline for feature engineering, model training, evaluation, and batch inference.

Technologies: Python, scikit-learn, XGBoost, tslearn

Senior Machine Learning Engineer - Mashgin (LLM & NLP Development)

Duration: Jun 2021 – Jun 2025
Summary: Fine-tuned large language models and built retrieval-augmented generation pipelines to support research use cases on large academic and internal document collections.
Responsibilities:

- Fine-tuned GPT-Neo using Hugging Face Transformers and DeepSpeed for decentralized training.
- Built RAG pipelines with LlamaIndex and Weaviate for efficient querying.

Technologies: GPT-Neo, Hugging Face Transformers, DeepSpeed, LlamaIndex, Weaviate

Senior Machine Learning Engineer - Mashgin (GPU Programming & AI Framework Optimization)

Duration: Jun 2021 – Jun 2025
Summary: Optimized GPU training and inference pipelines to reduce training time and improve inference performance for AI models.
Responsibilities:

- Applied CUDA C++ kernels and mixed precision training with cuDNN and AMP.
- Optimized inference with TensorRT and benchmarked distributed multi-GPU training with NCCL.
- Profiled performance using NVIDIA Nsight and deployed inference pipelines with Triton Inference Server.

Technologies: CUDA C++, PyTorch, cuDNN, AMP, TensorRT, NCCL, NVIDIA Nsight, Triton Inference Server

Senior Machine Learning Engineer - Mashgin (Decentralized ML Infrastructure)

Duration: Jun 2021 – Jun 2025
Summary: Contributed to decentralized machine learning infrastructure including task distribution, distributed training, and P2P networking for scalable deep learning workloads.
Responsibilities:

- Implemented Rust-based task distribution system using libp2p.
- Developed distributed training framework with PyTorch and Ray.
- Designed Scala pipelines for Spark ML workloads.
- Developed Go-based P2P networking layer for model sharing.

Technologies: Rust, libp2p, PyTorch, Ray, Scala, Spark ML, Go

Machine Learning Engineer | MLOps Lead - Voyage (Autonomous Vehicle Perception)

Duration: Jan 2017 – May 2021
Summary: Developed perception models and pipelines for autonomous shuttles operating in retirement communities, improving detection accuracy and real-time navigation.
Responsibilities:

- Trained YOLOv3 models on Azure ML for pedestrian and obstacle detection.
- Optimized LiDAR point cloud processing with PointNet and CUDA.
- Developed OpenCV-based object detection pipelines.

Technologies: YOLOv3, TensorFlow, Azure ML, PointNet, CUDA, OpenCV

Machine Learning Engineer | MLOps Lead - Voyage (MLOps & Data Engineering)

Duration: Jan 2017 – May 2021
Summary: Built data pipelines and orchestration systems to support large-scale sensor data processing and automated model retraining for autonomous vehicle fleet.
Responsibilities:

- Built Spark pipelines on Databricks processing 25TB/month of sensor data.
- Orchestrated Airflow DAGs with Kubernetes for automated retraining.
- Implemented MLflow for experiment tracking and model versioning.

Technologies: Spark, Databricks, Airflow, Kubernetes, MLflow

Machine Learning Engineer | MLOps Lead - Voyage (Edge Deployment & Simulation)

Duration: Jan 2017 – May 2021
Summary: Deployed models to edge devices and extended simulation environments to improve autonomous shuttle training and perception.
Responsibilities:

- Deployed TensorFlow Lite models achieving low inference times on edge devices.
- Extended CARLA simulator with Unreal Engine plugins.
- Developed C++ modules for sensor fusion integrating multiple sensor data types.

Technologies: TensorFlow Lite, CARLA, Unreal Engine, C++

Machine Learning Engineer | Full-Stack Developer - Startup X (Augmentus (Robotics Programming Platform))

Duration: Jan 2015 – Dec 2016
Summary: Developed robotics programming platform enabling no-code robotic automation for industrial applications with real-time computer vision and visualization.
Responsibilities:

- Developed ROS nodes in Python and C++ for robotic path planning.
- Integrated OpenCV for real-time object detection.
- Built WebGL visualization dashboard using Three.js.

Technologies: ROS, Python, C++, OpenCV, WebGL, Three.js

Machine Learning Engineer | Full-Stack Developer - Startup X (XpertFlow (Healthcare AI))

Duration: Jan 2015 – Dec 2016
Summary: Deployed machine learning models for ECG signal analysis and designed HIPAA-compliant data storage supporting secure processing of large patient datasets.
Responsibilities:

- Deployed TensorFlow models for arrhythmia detection.
- Designed PostgreSQL schemas for HIPAA-compliant patient data storage.
- Implemented Python preprocessing pipelines with NumPy and SciPy.

Technologies: TensorFlow, PostgreSQL, Python, NumPy, SciPy

Machine Learning Engineer | Full-Stack Developer - Startup X (StaffAny (Workforce Management SaaS))

Duration: Jan 2015 – Dec 2016
Summary: Scaled backend and developed frontend for workforce management platform handling time-tracking and scheduling for thousands of users.
Responsibilities:

- Scaled Node.js backend with Express and MongoDB.
- Developed React frontend for management dashboards.
- Implemented RabbitMQ for asynchronous task processing.

Technologies: Node.js, Express, MongoDB, React, RabbitMQ

Research Intern – Machine Learning - Amazon (Product Recommendation Systems)

Duration: Jun 2014 - Dec 2014
Summary: Developed machine learning models and ETL pipelines to improve product recommendation accuracy on Amazon's e-commerce platform.
Responsibilities:

- Developed recommendation models using scikit-learn and Mahout.
- Analyzed large-scale customer behavior data.
- Built ETL pipelines with AWS Data Pipeline.

Technologies: scikit-learn, Mahout, Pandas, Matplotlib, AWS Data Pipeline

EDUCATION

- **University of Central Florida**
  Master of Science in Computer Science
  May 2012 – May 2014
- **University of Central Florida**
  Bachelor of Science in Computer Science
  Sep 2007 – Jun 2011