



Omar francisco Arellano Ganem

Senior Machine Learning Engineer | Computer Vision,
Multimodal AI & Agentic Systems

✉ omar.f.ganem@gmail.com

☎ +52 55 5404 1258

📍 Mexico City, Mexico

🌐 linkedin.com/in/omar-francisco-arellano-ganem

🔗 github.com/peak-omar

🔗 omar-francisco.vercel.app

👤 PROFILE

Senior Machine Learning Engineer with **10+ years of AI/ML experience** and **15+ years across software engineering, data platforms, backend systems, and cloud infrastructure**. Specialized in **production ML systems, computer vision, multimodal modeling, structured data pipelines, LLM-based workflows, agentic AI, model evaluation, experiment tracking, and scalable inference**. Experienced owning ML projects from problem framing through data preparation, training, evaluation, launch, post-launch monitoring, feedback loops, and continuous improvement. Delivered production AI systems supporting **millions of daily API requests, 99.95% uptime, 55% latency reduction, 70% manual-effort reduction, and 94% faster deployment cycles**.

🎓 EDUCATION

Tecnológico de Monterrey,
Bachelor's Degree

04/2022 – 12/2024 | Monterrey, Mexico

Bachelor's Degree in Artificial Intelligence Engineering, Tecnológico de Monterrey

- Completed a rigorous curriculum focused on **machine learning, AI systems, and software engineering**
- Gained hands-on experience in **Python-based ML development, data processing, and algorithm design**
- Built strong foundations in **ML infrastructure, model development, and scalable systems**
- Participated in **hackathons and applied projects** solving real-world problems using AI
- Developed exposure to **generative AI concepts, modern AI workflows, and emerging technologies**

🏢 PROFESSIONAL EXPERIENCE

Generative AI Engineer & Team Lead, *Autism360*

11/2024 – 07/2026 | Australia

<https://www.autism360.com/> 🌐

- Led cross-functional AI engineering teams through problem framing, architecture, data preparation, model evaluation, deployment, monitoring, and post-launch improvement for production AI products.
- Delivered production AI systems including an **Enterprise RAG Pipeline, AI Healthcare Monitoring System, GPT-4o Chatbot, and Agentic AI Platform**.
- Designed production-grade ML and GenAI workflows combining **structured data, multimodal inputs, LLM reasoning, retrieval, evaluation loops, and API-connected automation**.
- Built scalable AI/ML pipelines across **AWS and GCP**, improving availability, cost efficiency, real-time processing, and operational reliability.
- Built RAG and information-retrieval systems using **embeddings, vector databases, semantic retrieval, context assembly, and grounded generation**, reducing manual effort by approximately **70%**.
- Designed feedback-loop patterns to capture model failures, evaluate output quality, refine prompts and pipelines, and improve model behavior over time.
- Architected a high-throughput generative document and decision-support platform supporting **millions of API requests daily with 99.95% uptime**.
- Created evaluation frameworks for **LLM, NLP, retrieval, agentic, multimodal, and structured-output systems**, increasing automated test coverage from **70% to 92%**.
- Developed A/B testing pipelines to compare model, prompt, retrieval, workflow, and feature variants before rollout.
- Migrated ML workloads to **AWS**, reducing infrastructure costs by approximately **40%** while improving deployment flexibility and scalability.
- Automated lifecycle workflows with **Airflow, MLflow, CI/CD, model versioning, and deployment validation**, reducing release cycles from weeks to days.
- Optimized inference using **ONNX, TorchScript, batch serving, latency profiling, and runtime benchmarking**, reducing latency by **55%**, from **420 ms to 190 ms**.
- Implemented observability with **Prometheus, Grafana, logging, alerting, and failure analysis**, reducing incident resolution time from **8 hours to less than 1 hour**.

Senior AI Engineer, *Sephora*

05/2022 – 10/2024 | United States

<https://www.bunsan.io> 🌐

- Led a three-engineer team delivering applied ML, NLP, forecasting, recommendation, and automation systems that improved operational efficiency and generated approximately **\$2M in client cost savings**.
- Owned ML initiatives from business problem definition through dataset preparation, feature engineering, model training, evaluation, deployment, monitoring, and production optimization.
- Built production AI applications using **Python, REST APIs, NLP workflows, recommendation systems, forecasting models, structured data pipelines, and cloud infrastructure**.
- Improved sales forecasting accuracy by **80%-90%** through large-scale analysis across **120+ market entities**.

University of the Valley of Mexico,

Bachelor of Science - BS

08/2000 – 05/2006 | Mexico City, Mexico

- Completed a comprehensive curriculum covering **data structures**, **algorithms**, **software engineering**, **databases**, **computer networks**, and **machine learning infrastructure**
- Gained hands-on experience with programming languages such as **Python**, **Java**, **C++**, and **JavaScript** through coursework and projects.
- Participated in coding **competitions** and **hackathons**, honing **problem-solving** and **teamwork skills**
- Member of the **University Computer Science Club**, contributing to workshops on web development, blockchain, and emerging technologies.

ACTIVITES

- Active Data Science, Statistics and Machine Learning blog writer on Medium. My blog has followers, over 1k total views, and 135+ stories published in the last 2.5 years.
- Run a channel with 180k+ views, giving advice on how to break into data science.
- Keen hockey player since 7 years old and have for my University 1st team over my 4 years there.

CERTIFICATES

- IBM Machine Learning Specialization Professional Certificate, IBM.
- Microsoft Azure AI Fundamentals AI-900 Exam Prep Specialization
- Machine Learning Engineering for Production (MLOps), [DeepLearning.ai](#)
- Practical Data Science on the AWS Cloud Specialization,
- Deep Learning Specialization, [DeepLearning.ai](#)
- Google Project Management: Professional Certificate, Google.
- Machine Learning Specialization, [DeepLearning.AI](#)

LANGUAGES

Spanish Native ● ● ● ● ●

English Fluent ● ● ● ● ●

- Launched a personalized e-commerce platform with an advanced recommendation engine, increasing sales by **80%** and online orders by **200%**.
- Delivered the e-commerce platform **3 months ahead of schedule** through hands-on technical ownership, fast iteration, and cross-functional collaboration.
- Partnered with **10+ data scientists** to productionize recommendation, NLP, and evaluation workflows, increasing user engagement by **46%**.
- Implemented an NLP solution that improved chatbot accuracy to **95%** and reduced customer response time by **63 seconds**.
- Optimized neural-network architecture and training workflows, reducing training time by more than **20 minutes** while improving accuracy by **7.5%**.
- Designed and implemented an end-to-end **MLOps pipeline**, reducing deployment time from **3 weeks to 2 days**, a **94% reduction**.
- Developed a custom NLP automation system handling **73,000+ queries per month** with **91.3% accuracy**.
- Built A/B testing and statistical evaluation workflows for model, prompt, retrieval, and feature variants, generating approximately **\$2.7M in additional annual revenue**.
- Optimized production data and inference pipelines, reducing latency by **78%**.
- Delivered recommendation and anomaly-detection capabilities that increased engagement by **42%**, raised average order value by **\$17.82**, and prevented approximately **\$1.2M in potential fraud** over **14 months**.

Senior Machine Learning Engineer / Model Optimization Engineer, [Blue People](#)

10/2016 – 02/2022 | Mexico city, Mexico

<https://www.bluepeople.com/>

- Led a team of 10 engineers building machine learning systems for a billion-dollar startup, increasing customer spending by **94%** within the first year.
- Built a **Kubeflow** and **Kubernetes** ML platform adopted by more than **17 Fortune 500 companies**.
- Guided technical decisions across **25+ engineering divisions**, supporting more than **150 engineers** working on advanced ML solutions.
- Developed model optimization frameworks using **TorchScript**, **ONNX**, and **TensorRT**, reducing p95 latency by **55%**.
- Built benchmarking frameworks to evaluate cost-performance tradeoffs and improve system scalability.
- Architected a multimodal generative AI platform serving more than **1 million monthly users**.
- Implemented **RLHF** workflows that increased user satisfaction by **22%**.
- Coordinated cross-functional delivery across product, design, and engineering teams, delivering six major releases within 18 months.
- Built observability dashboards with **Grafana** and **Prometheus**, reducing incident response times by **45%**.
- Developed transformer-based code generation models that improved functional correctness by **31%**.
- Scaled distributed training across **256 TPU v4 pods**, reducing training time from **72 hours to 30 hours**.

Software Engineer / Data Engineer, [Ensitech](#)

11/2009 – 06/2014 | Mexico city, Mexico

<https://ensitech.com/>

- Built Python and SQL data pipelines processing **5M–10M+ records daily**, reducing latency by **30–40%** through query optimization and batch processing.
- Developed ETL workflows and transformation pipelines supporting analytics and machine learning workloads.
- Created feature engineering pipelines for predictive modeling and large-scale machine learning systems.
- Developed RESTful APIs and backend services that reduced manual data retrieval efforts by approximately **35%**.
- Optimized SQL performance through indexing and query tuning, improving efficiency by **25–35%** while lowering infrastructure costs.
- Built and deployed machine learning models using **scikit-learn**, **NumPy**, and **pandas**, improving prediction efficiency by approximately **30%**.

- **Machine Learning Engineering:** Model Training, Model Evaluation, Feature Engineering, Model Deployment, Production Monitoring, Feedback Loops
- **Computer Vision & Multimodal AI:** Image Understanding, Visual Inspection, Multimodal Models, Document/Image Understanding, Quality-Sensitive Classification, Visual Feature Extraction
- **LLMs & Agentic Workflows:** LLM Applications, Tool-Using Agents, Agentic Pipelines, Verification Loops, Structured Outputs, Retrieval-Augmented Reasoning
- **Python & ML Tooling:** Python, PyTorch, scikit-learn, NumPy, pandas, Hugging Face
- **Evaluation & Experimentation:** Statistical Evaluation, A/B Testing, Regression Testing, Error Analysis, Model Readiness Criteria, Experiment Tracking
- **Data Pipelines & Annotation:** Data Preparation, Labeling Standards, Human-in-the-Loop Review, Data Validation, Dataset Curation, Training Data Quality
- **Cloud ML Infrastructure:** AWS, GCP, Vertex AI, Azure AI, Docker, Kubernetes, Cloud-Native Inference
- **MLOps & Lifecycle:** MLflow, Weights & Biases, Airflow, Kubeflow, Model Versioning, CI/CD Pipelines
- **Backend & Integration:** FastAPI, Flask, REST APIs, Microservices, Async Processing, API Integration
- **Observability & Optimization:** Prometheus, Grafana, Logging, Alerting, Latency Optimization, Cost Optimization

Production LLM Application Platform

01/2026 – 05/2026

- Built end-to-end LLM applications using **OpenAI, Anthropic Claude, Gemini, Llama, Python, FastAPI, REST APIs, prompt engineering, and evaluation workflows.**
- Integrated AI features into backend systems, product workflows, microservices, and cloud-native deployments across **AWS, GCP, and Azure.**
- Improved reliability through prompt testing, regression checks, observability, deployment validation, monitoring, and scalable API design.

Computer Vision Inspection & Multimodal Quality Platform

08/2025 – 01/2026

- Built image-heavy ML workflows for **visual inspection, classification, authentication-style quality checks, multimodal reasoning, uncertainty handling, and structured output generation.**
- Partnered with annotation, domain, product, and operations stakeholders to define labeling standards, success metrics, rollout plans, and model-readiness thresholds.
- Improved model reliability through error analysis, human-in-the-loop review, feedback loops, regression checks, and production monitoring.

High-Performance LLM Inference and AgentOps Platform

03/2025 – 09/2025

- Built production deployment workflows using **Kubernetes, Docker, Terraform, CI/CD, MLflow, Airflow, Prometheus, Grafana, logging, tracing, and alerting.**
- Optimized inference with **vLLM, TensorRT-LLM, TensorRT, ONNX, TorchScript, quantization, batching, and throughput/latency benchmarking.**
- Reduced latency by **55%**, increased evaluation coverage from **70% to 92%**, and improved production incident response from **8 hours to less than 1 hour.**