Reymark Garcia

AI/ML Engineer

Trece Martires City, Cavite, Philippines
reymarkgarcia040924@gmail.com
github.com/reymark196
linkedin.com/in/reymark-garcia-75a561373/

AI/ML Engineer with 5+ years of experience specializing in Machine Learning, Deep Learning, Generative AI, Large Language Models, and Retrieval-Augmented Generation. Proven track record of designing, building, and maintaining production-grade, end-to-end intelligent systems, applying strong software engineering principles such as modular system design, API-driven architectures, version control, testing, and CI/CD. Deep expertise in prompt engineering, agentic AI workflows, scalable model serving, and performance optimization, delivering reliable, maintainable, and real-world AI solutions.

Professional Experience

AI Engineer | Technical Lead *Zenovo AI, Denver, Colorado, United States | July 2025 - December 2025*

- Designed and led the development of a multi-tenant SaaS platform delivering customizable LLM/RAG-driven services, fully branded per customer via isolated tenant subdomains

- Designed and Implemented a production-scale multi-agent AI system using LangChain, LangGraph, OpenAI APIs, and Model Context Protocol to enable standardized, extensible AI tool orchestration and context sharing across agents and services.

- Implemented MCP-based integrations to expose internal services (RFP parsers, document stores, scoring engines, retrieval APIs) as reusable, discoverable tools for LLM agents, enabling clean separation between model reasoning and platform capabilities.

- Implemented Human-in-the-Loop workflows with approval gates, audit trails, override mechanisms, and LLM-as-a-Judge evaluation layers integrated into MCP-enabled agent flows for automated quality control, compliance checks, and decision validation.

- Optimized AI and platform performance through prompt engineering, vector index tuning, document chunking strategies, request batching, caching layers, and cost-efficient LLM usage, including model selection and fallback strategies.

AI Engineer *GenAI.Labs, San Diego, California, United States | November 2023 - May 2025*

- Built an enterprise-grade, production AI performance-monitoring chatbot to automate large-scale data retrieval and accelerate real-time troubleshooting for support engineers in high-throughput environments.

- Designed and implemented a production medical imaging pipeline for echocardiogram analysis using PyTorch, Lightning, MONAI, and OpenCV, delivering automated labeling, classification, and grading.

- Developed an AI-driven automation platform for a seismic construction company, generating project documentation and compliance reports. Built React-based dashboards integrated with Elasticsearch to improve reporting accuracy, observability, and operational efficiency.

- Built and deployed predictive ML models for a SaaS platform using PyTorch, scikit-learn, XGBoost, and MLflow, integrating inference into MERN-based dashboards to surface predictive insights directly within core product workflows.

AI Backend Developer *Stratpoint Technologies, Mandaluong, Metro Manila, Philippines | January 2023 - November 2023*

- Led backend development for a strategic AI experimentation team, architecting scalable data processing and ML integration systems using Django and Django REST Framework within budget-conscious constraints.

- Built robust ETL pipelines for large-scale geospatial data aggregation and transformation, processing multi-source datasets from AWS S3, Google Earth Engine API, and climate repositories for predictive modeling applications.

- Designed and developed RESTful APIs serving real-time predictive model outputs and geospatial analytics to frontend dashboards, implementing caching strategies with Redis that reduced average response times from 1.2s to 300ms.

- Developed containerized data processing microservices using Docker and Celery for asynchronous task execution, enabling parallel computation of solar irradiance analytics and site scoring across distributed cloud environments.

AI/ML Engineer *Senti Techlabs, Inc, Makati, Metro Manila, Philippines | August 2021 - November 2022*

- Developed and maintained several ML/DL models for NLP tasks using PyTorch and early Hugging Face Transformers, improved model accuracy through systematic experimentation and hyperparameter tuning.

- Performed end-to-end fine-tuning of pre-trained transformer models, applying techniques such as learning-rate scheduling, layer freezing/unfreezing, gradient accumulation, and batch-size tuning to improve convergence, training stability, and performance under limited compute constraints.

- Collaborated with data engineers to prepare training datasets, clean noisy text data, and create evaluation benchmarks for NLP workloads.

- Performed optimization tasks such as batching, caching, and basic quantization to reduce inference cost and latency on GPU-based servers.

- Implemented model-serving pipelines using FastAPI, creating small, reliable inference APIs deployed on AWS.

- Contributed to model monitoring and version control using MLflow, maintaining reproducible experiments and deployment-ready checkpoints.

- Designed and built internal document search and recommendation workflows using Sentence Transformers for embeddings and FAISS for similarity search, enabling faster access to internal knowledge bases.

AI/ML Engineering Intern *Senti Techlabs, Inc, Makati, Metro Manila, Philippines | May 2020 - August 2021*

- Assisted in developing NLP prototypes for text classification and keyword extraction using scikit-learn, spaCy, and small transformer models; handled dataset cleaning, exploratory analysis, and evaluation reporting.

- Supported AI/ML engineering and research workflows by experimenting with multiple NLP modeling approaches, benchmarking classical ML methods against transformer-based models, and documenting findings to inform design decisions and future production development.

Key Skills

- Programming Language : Python, JavaScript, TypeScript, Rust, C++, Java
- Machine Learning & Deep Learning : PyTorch, TensorFlow, scikit-learn, XGBoost, ONNX Runtime, Triton Inference Server, Model Quantization, Time-Series Forecasting, Recommendation Systems, Anomaly Detection, Hyperparameter Optimization,
- Generative AI, LLMs & Agentic Systems : Large Language Models(GPT, Claude, Gemini, Llama), Hugging Face, LangChain, LangGraph, LangSmith, CrewAI, Multi-Agent Architectures, RAG Architectures, Function/Tool Calling , Prompt Engineering, Model Context Protocol, A2A, Vector Database(Pinecone, FAISS, Milvus), Graph Database(Neo4j, Apache AGE)
- Computer Vision & Image/Video Analysis : OpenCV, PILLOW, Object Detection (YOLO), Image Segmentation, OCR (Tesseract), Feature Extraction, Video Frame Analysis, Medical Imaging Pipelines
- Full Stack Engineering & Distributed Systems : FastAPI, Django, Node.js, Express.js, React, Next.js, PostgreSQL, MongoDB, REST API, GraphQL, gRPC, WebSockets, Event-Driven Architectures, Microservices, Caching Layers (Redis)
- Cloud, DevOps & Infrastructure : AWS (EC2, S3, Lambda, RDS, CloudWatch, SageMaker, Bedrock), GCP(Vertex AI), Azure, Docker, Docker Compose, Kubernetes, Terraform, GitHub Actions, Load Balancing, Monitoring (Prometheus, Grafana)
- MLOps, Pipelines & Automation : MLflow, Airflow, Kubeflow, Model Registry, CI/CD for ML, Monitoring Metrics, A/B Testing

Education

Bachelor's Degree in Computer Science

De La Salle University Manila, Philippines | September 2017 - June 2021

Academic foundation in Data Structures & Algorithms, Machine Learning, and Deep Learning, with emphasis on algorithmic analysis and modern neural network methods.

Languages

- English - Expert
- Filipino - Expert