Christopher Millones

Irvine, CA • 949-842-8097 • ChristopherMillones@hotmail.com • linkedin.com/in/christophermillones

SUMMARY

AI/ML Engineer with 4+ years building production-grade LLM systems, MLOps pipelines, and GenAI applications in a HIPAA-regulated healthcare environment. Expertise spans multimodal document understanding, RAG architecture, prompt engineering, and cloud-native deployments on AWS and the use of GCP with focus in VertexAI to handle millions of annual invocations.

EXPERIENCE

Glidewell Dental Inc. — *Software Engineer – MLOps & AI* Nov 2021 – Present

Dental Rx Form Extraction (Multimodal LLM • Document Understanding)

- Designed and deployed a production pipeline processing complex RX dental prescription image into structured JSON using a hierarchical LLM fallback architecture, serving ~1M orders annually through an AWS-orchestrated Step Functions workflow.
- Led iterative prompt engineering across 8+ versions covering material mapping, checkbox interpretation, tooth notation logic, and validation rules; achieved an 84% context cache hit rate, significantly reducing per-call token costs.

GlidewellGPT (RAG • GenAI Chatbot)

- Built a Modular RAG chatbot enabling call center agents to retrieve accurate answers from internal documentation, reducing search time and improving response consistency across the team.
- Implemented advanced retrieval techniques including semantic chunking, query rewriting, two-stage retrieval with LLM-based reranking, and metadata filtering.

Order Comment Automation (High-Throughput LLM • NLP • Workflow Routing)

- Engineered an NLP pipeline extracting structured routing decisions from free-text doctor comments, processing ~1M annual orders via serverless invocation.

LLM Evaluation Framework (Model Assessment • MLOps • A/B Benchmarking)

- Developed an automated evaluation system comparing LLM predictions against ground truth across relational and data warehouse databases, enabling data-driven prompt iteration that drove overall extraction accuracy from 70.2% → 78% → ~90%.

Additional Engineering

- Orchestrated MLOps pipelines for CV, OCR, and tabular models on cloud infrastructure; built model monitoring system, production error alerting (Teams, SMS, Email), and a facial recognition access control system using cloud vision services.
- Drove LLM provider migration POC from Bedrock to Vertex AI, delivering cost analysis, cache performance benchmarks, and architecture recommendations to stakeholders.

Star Micronics Inc. — *Market Solutions Engineer* Aug 2019 – Jan 2021

- Managed full project lifecycle for hardware/SDK integration projects; created developer training materials and provided vendor-level technical support for SDK and hardware issues.

SKILLS

AI / ML	LLMs, prompt engineering, RAG (Modular), multimodal extraction, embeddings, vector stores, reranking, fine-tuning, computer vision, NLP
Cloud	AWS (Lambda, Step Functions, SageMaker, Bedrock, Aurora, Redshift, DynamoDB, EventBridge, Rekognition, and more..) • GCP (Vertex AI, Gemini)
Languages	Python, JavaScript / TypeScript, SQL, Bash, Java
Frameworks	LangChain, Hugging Face Transformers, ReactJS, Serverless Framework, CloudFormation
MLOps / DevOps	CI/CD, Docker, model monitoring, batch & real-time inference, A/B evaluation, GitLab
Compliance	HIPAA-aware design, PHI data handling, GCP Vertex AI BAA

EDUCATION

California State University, Fullerton — *B.S. Computer Science, GPA 3.70* Aug 2021