Guadalupe Higuera

Software Engineer · GenAI Platform & LLM Infrastructure · Agentic AI
Guadalupe.Higuera@protonmail.com

SKILLS

GenAI / LLM	Large Language Models, Prompt Engineering, RAG Systems, Agentic AI, LLM Orchestration (Semantic Kernel, AWS Bedrock Agents, LangGraph), Vector Databases (Pinecone, ChromaDB), QLoRA Fine-Tuning, Hugging Face, LLM Evaluation, MCP
Cloud & Infra	AWS (Bedrock, Lambda, DynamoDB, API Gateway, S3, IAM), Azure (AZ-900 Certified), Terraform, CI/CD, LLMOps
Languages	Python, FastAPI, SQL, C#, JavaScript / TypeScript, React, ASP.NET Core
Testing & Monitoring	Playwright, Splunk, LLM Observability

PROFESSIONAL EXPERIENCE

Wells Fargo
Chandler, AZ

Software Engineer — GenAI Platform Infrastructure
July 2023 – Present

- Built and operated production platform powering agentic AI systems for 10,000+ users across 5 business use cases (policy, finance, investments) — agents leverage MCP server integrations, multi-step reasoning, single- and multi-agent orchestration, tool-calling, and database connectivity.
- Designed agent deployment and configuration infrastructure enabling non-technical business users to independently deploy and iterate on AI agents, reducing engineering bottlenecks and accelerating pilot-to-production timelines.
- Architected and own automated regression testing framework validating multi-step agent execution paths in a drag-and-drop AI agent builder during platform migration.
- Debugged and optimized AI agent testing pipelines, improving deployment reliability and reducing CI/CD time.

Sandhills Global
Scottsdale, AZ

Software Development Intern
December 2021 – February 2023

- Developed .NET backend services and React front-ends for internal sales tools and customer-facing applications; shipped production code across multiple releases.

PROJECTS

iRacing AI Race Strategist

- Fine-tuned Llama 3.1 8B with QLoRA on 9,269 synthetic examples; eval framework on 55 cases showed score improvement 29.7 → 78.5 and hallucination elimination (48/55 → 0/55) via 12-metric weighted scorer and LLM-as-judge blind A/B comparison.
- Designed event-driven LLM orchestration with priority-based dispatch, per-event cooldowns, and state-machine transitions for race events — replaces naive per-tick inference. Generated training data using Claude API with category-balanced distribution across key racing scenarios.
- Engineered async telemetry → LLM → TTS pipeline (800–1,100ms latency) with GGUF q4_k_m quantization for local inference and deterministic fallback when LLM fails.

AWS Bedrock: Agentic Shopping Assistant

- Built autonomous shopping agent using AWS Bedrock Agent with custom action groups for natural language search and tool-calling.
- Provisioned full AWS infrastructure with Terraform: DynamoDB, Lambda, API Gateway, S3, IAM Roles.
- Developed FastAPI backend integrating Bedrock Agent runtime; Lambda functions query DynamoDB based on LLM-extracted parameters.

Titanic Historical RAG

- Built a retrieval-augmented search engine over the 1912 Titanic inquiry transcripts that surfaces contradictions between witnesses instead of hiding them.
- OpenAI text-embedding-3-large + Pinecone serverless for retrieval; Claude Haiku 4.5 with structured JSON output for pairwise contradiction verdicts with 0–1 confidence scores.
- Indexed over 3,300 pages of US Senate Inquiry and British transcripts across 68 witnesses (~40,000 chunks).

EDUCATION

Arizona State University

Tempe, AZ

Master of Science in Computer Science

August 2023 – December 2027

Relevant Coursework: Knowledge Representation & Reasoning, Artificial Intelligence, Bio-Inspired Computing

Arizona State University

Tempe, AZ

Bachelor of Science in Computer Science

August 2020 – May 2023