Taj Khunkhun

taj479505@gmail.com | +1 408 380 2726 | San Jose, California
LinkedIn: https://www.linkedin.com/in/taj-khunkhun-a46972339
GitHub: https://github.com/pauxd26

PROFESSIONAL SUMMARY

Dedicated AI Engineer with 8+ years of experience in backend systems and 2+ years specializing in Agentic Workflows, Multi-Agent Orchestration, and Autonomous Reasoners. Expert at transforming experimental prototypes into production-grade agentic AI systems using LangGraph, CrewAI, AutoGen, and MCP. Proven track record of solving the Agent Reliability Gap by architecting multi-agent systems with independent Verifier/Judge layers, MCP-based tool integration, and strict Pydantic-based tool-calling schemas. Deeply experienced in Agentic RAG, self-correcting retrieval loops that reduced hallucination rates by 40%. Passionate about bounded autonomy -- systems that reason through uncertainty but operate within deterministic guardrails.

TECHNICAL SKILLS

Ai Agents:
LangChain, LangGraph, CrewAI, AutoGen, AutoGPT, MCP, MCP Servers & Clients, Agentic AI, Multi-Agent Orchestration, RAG, Agentic RAG, LangSmith, Prompt Engineering

Ml Deep Learning:
PyTorch, TensorFlow, Keras, scikit-learn, spaCy, NLP, LoRA, QLoRA, Reinforcement Learning, Neural Networks, MLOps, LLMOps

Languages Frameworks:
Python, Golang, JavaScript, TypeScript, Django, FastAPI, Flask, Node.js, React, Angular

Data Infrastructure:
Snowflake, Databricks, Spark, Kafka, Airflow, BigQuery, Neo4j, Redis, Pinecone, Elasticsearch

Cloud Devops:
AWS, Azure, GCP, Kubernetes, Docker, Terraform, CI/CD, Microsoft Fabric, Copilot Studio, Azure Data Lake

PROFESSIONAL EXPERIENCE

Senior AI Engineer | Agentic AI Engineer
Hewlett Packard Enterprise | Spring, Texas (Remote) | Jan 2023 - Dec 2025

- Designed multi-agent Plan-and-Execute architecture using LangGraph, reducing infinite loop failures by 35%.
- Architected MCP server layer to standardize tool integration, reducing custom tool-binding code by 60%.
- Reduced inference costs by 45% on NVIDIA NIM via Router Agent triaging Llama 2 7B vs GPT-4.
- Implemented HITL checkpoint system for high-stakes financial workflows (0.85 confidence threshold).
- Built observability dashboards using LangSmith and Arize Phoenix.
- Architected LLM evaluation framework using DeepEval and LLM-as-a-Judge across 5,000+ test cases.
- Reduced LLM inference costs by 52% and p95 latency by 300ms via semantic caching with Redis.
- Eliminated catastrophic forgetting in Llama 2 with multi-objective training recipe.
- Reduced training VRAM by 65% using QLoRA 4-bit quantization for 70B models.
- Improved inference throughput 4x via multi-LoRA serving (vLLM/LoRAX).
- Led migration from monolithic Django to Microservices with Docker/Kubernetes.
- Built MCP-compliant tool registry for OpsRamp AIOps copilot.
- Architected Self-Correction loop for SQL-generating agent, reducing errors by 50%.

Data & Machine Learning Engineer

Adobe | San Jose, California | Aug 2019 - Jan 2023

- Architected ETL pipeline processing 5TB+ multi-modal data using Spark.
- Eliminated vector-relational desync via CDC (Debezium + Kafka) with 99.9% consistency.
- Engineered Blue-Green re-indexing for zero-downtime migrations across 50M+ vectors.
- Optimized semantic search by 40% via hierarchical document indexing.
- Reduced vector storage costs by $8k/month through tiered data strategy.
- Reduced inference latency by 65% via model distillation on NVIDIA A100 GPUs.
- Built synthetic data generation engine improving minority-class tasks by 18%.
- Engineered Fail-Soft orchestration saving $15k/month in compute costs.

EDUCATION

Master of Computer Science - Santa Clara University (Aug 2019 - Jun 2020)
Bachelor of Computer Science - Santa Clara University (Sep 2015 - Jun 2019)