Ali Yassine

https://linkedin.com/in/ali-yassine-775557348 | github.com/AliYassine03 | (909) 223-1432 | ayassine2003@gmail.com

SUMMARY

NVIDIA-Certified Associate in Generative AI / LLMs, AI Engineer focused on shipping production RAG systems, fine-tuned LLMs, and inference APIs. Experience taking GenAI features from prototype to deployed service across Azure and GCP, with a track record of cutting latency, integrating retrieval pipelines, and turning research models into reliable endpoints.

EXPERIENCE

Sidereal Solutions — *AI Engineer Consultant* Aug 2024 – Present | Remote, CA

- Built and shipped production RAG pipelines for client knowledge bases, integrating LangChain, vector retrieval, and serving on Azure GPU instances; cut p95 query latency by ~50% through caching, batching, and quantized model serving.
- Fine-tuned open-source LLMs (Llama 3, Mistral) with LoRA/QLoRA for client-specific use cases, deploying behind a FastAPI service with streaming responses and structured-output guardrails.
- Migrated CPU-bound preprocessing (embedding generation, document chunking, dataframe ETL) to GPU-accelerated workflows using NVIDIA RAPIDS (cuDF), reducing batch processing time on large datasets from hours to minutes.

Product Perfect — *AI Engineer Intern* Mar 2025 – Sept 2025 | Brea, CA

- Optimized computer vision inference pipelines (Detectron2, Stable Diffusion) by moving synchronous endpoints to async FastAPI and exporting models to ONNX Runtime, cutting tail latency on image generation requests by ~70%.
- Profiled GPU workloads with Nsight Systems to identify memory and kernel bottlenecks, reducing peak VRAM usage and unblocking larger batch sizes on shared inference hardware.
- Containerized inference services with Docker and integrated them into the team's CI/CD pipeline, making model updates a one-command deploy instead of a manual handoff.

PROJECTS

Lectern — *Educational GenAI platform* **https://lectern-blond.vercel.app/**

- Built an end-to-end RAG application on a fine-tuned Llama 3 model with LangChain, vector retrieval, and a Next.js front end; supports document and video uploads streamed to an inference API for on-demand study material generation.
- Designed prompt-level guardrails and retrieval filters to keep generated content scoped to uploaded source material, with refusal behavior when relevant context wasn't found in the corpus.

Beach Finder — *Real-time AI insights app* **https://github.com/AliYassine03/Cali-Beaches**

- Shipped a full-stack app combining government-provided live weather and ocean safety data with LLM-generated summaries and ML-based condition predictions; FastAPI backend, React front end, deployed to GCP.
- Cached and parallelized external API calls to cut response time by ~50%, making generated forecasts feel real-time to the user instead of buffered.

SKILLS

Languages: Python, TypeScript / JavaScript (Node.js), Java, C/C++, Go, SQL, HTML/CSS

AI / ML: Retrieval-Augmented Generation (RAG), LangChain, Hugging Face Transformers, Llama / Mistral, fine-tuning (LoRA / QLoRA), prompt engineering, evaluation (retrieval & faithfulness)

Frameworks: PyTorch, TensorFlow, scikit-learn, ONNX Runtime, FastAPI, Next.js, React, NumPy, Pandas

Infra & MLOps: Docker, Azure AI, GCP, Git, CI/CD, Linux / Bash, NVIDIA RAPIDS (cuDF), GPU profiling (Nsight Systems)

LEADERSHIP

TableTop Titans *(Student Organization)* — *President* Aug 2023 – Jan 2025 | Fullerton, CA

- Led a 150+ member student organization at CSUF across two academic years, directly managing a 7-person officer team and running weekly events plus recurring community outreach programs.
- Grew annual operating budget by 80% by sourcing and closing local sponsorships and partnerships, reinvesting funding into expanded programming and new outreach initiatives.
- Owned end-to-end planning and execution of Titan Designs Hackathon, the organization's largest annual event, coordinating officers, volunteers, and external partners to deliver it on schedule and within budget.

EDUCATION

California State University, Fullerton — *B.S. Computer Science* Aug 2021 – May 2025

Relevant coursework: Machine Learning, Advanced Neural Networks, Parallel Computing, Cloud Computing & Distributed Systems, Advanced Algorithms, Data Structures, Operating Systems.

Certification: NVIDIA-Certified Associate — Generative AI & LLMs.
https://www.credly.com/badges/248935e2-33a5-4220-8abf-a044ac3ed710/public_url