

Gabriel Fagundes

gabrielft.me | linkedin.com/in/gabrielft-me | github.com/gabrielft-me | contact@gabrielft.me | (813)-485-0085

EDUCATION

University of South Florida

Bachelor of Science, Artificial Intelligence Dean's List Fall 2024, Spring 2025, Latin America Scholarship

Tampa, FL

Exp. May 2027 | GPA: 3.81

EXPERIENCE

Lyra – Y Combinator X25

AI/ML & Product Engineer Intern

New York City, NY

Oct 2025 – Dec 2025

- Built an edge-native AI voice agent on Cloudflare using tRPC, WebRTC streams, Drizzle ORM, Hono, and the ElevenLabs SDK.
- Engineered a custom RAG pipeline with Pinecone and a knowledge graph built from entity-relation triplets to improve retrieval quality before reranking, reducing latency for related-information retrieval.
- Built a trajectory evaluation framework scoring agent actions and tool-selection accuracy, achieving over 95% outcome accuracy (human-in-the-loop) and sub-200 ms p90 evaluation latency. Launched to 10 startups and scaled to 500+ startups in one week.

Google Developer Group

Tech Lead

Tampa, FL

May 2025 – Nov 2025

- Organized and led monthly workshops in full-stack development, co-led two hackathons with 300+ participants, raised \$10,000+ through direct 1:1 negotiations with local companies, and mentored 20 students exploring the startup ecosystem.

inkPen

Co-founder, Product Engineer, DevOps

Bahia, Brazil

Dec 2022 – Dec 2024

- Co-built a school platform end-to-end, owning backend architecture, frontend delivery, and product iteration through direct customer feedback.
- Traveled to 20+ cities interviewing public schools, created customer feedback channels with 800+ teachers across all states, led a UX research team analyzing national databases, and designed 50+ Figma pages from research insights.
- Built a platform with 150+ API endpoint, using NestJS for the backend, deployed to Amazon EKS, AWS Lambda for serverless workloads, and AWS RDS for data storage, while shipping the React/Tailwind/Typescript web app and React Native mobile app.
- Helped scale the product from a school project into a real company through direct customer exposure, growing from pilots in 3 schools to adoption in 250+ schools and 10,000 teachers, securing \$11,000 in pre-seed funding, and ending in a successful exit.

Nucleo do conhecimento (scientific journal)

Published researcher, IoT/C++ Engineer

Caetite, Brazil

Jun 2021 – Dec 2022

- Led an 18-month field study on hands-free automation to limit COVID-19 transmission; secured a sponsored lab room, coordinated 2 robotics professors, and collaborated with 8 graduate students.
- Coded a C++ library, designed PCB layouts in EasyEDA, soldered custom hardware boards, and developed computer vision algorithms in R for data analysis. Co-authored a peer-reviewed thesis published in a Q2 journal with 25M annual readers.

PROJECTS

Harper – Y Combinator W25 (5-day onsite) Engineering Trial

Feb 2026

- Deployed a self-hosted Qwen3-32B retrieval assistant on a Google Cloud Confidential VM (A3, 1× H100 80 GB), keeping client history inside the trusted environment while giving internal teams instant recaps of prior interactions at call time.
- Served the model with vLLM against a pgvector-backed RAG index, instrumenting throughput, KV-cache utilization, and CPU load in Grafana to size serving capacity against observed saturation.
- Buffered inbound requests through Apache Kafka so traffic bursts could not overwhelm single-GPU inference capacity, and built the orchestration layer in FastAPI to route concurrent memory-retrieval requests across internal workflows.

Eyrie – Harvard Hackathon 2025 Winner Top 5% devpost.com/software/eyrie-idxhj8

Oct 2025

- Built Eyrie in 36 hours, a computer-vision monitoring system that predicts crowd crush incidents using live video from drones.
- Shipped a backend with FastAPI and WebRTC, exposing a YOLOv8 model trained on NVIDIA A100 and deployed locally.
- Optimized GPU kernel execution and batch pipeline, reducing inference latency from 190ms to 30ms (84% reduction).

SLiM – Small Language Models Orchestration – Shell Hacks 2025 devpost.com/software/slim-yggr9c

Sep 2025

- Built SLiM in 36 hours, an adaptive ML orchestration loop that routed queries across fine-tuned distilled small language models, a retrieval layer, and Gemini 2.5 based on query intent, semantic similarity, latency, and inference cost.
- Implemented a fallback pipeline escalating low-confidence queries to Gemini 2.5, and a continuous-learning pipeline backed by a CUDA-accelerated LoRA fine-tuning loop in PyTorch, distributed jobs on Vertex AI, and ChromaDB for embedding retrieval.
- Built an automated evaluation layer with sentence-transformers, holding 90%+ semantic similarity to Gemini 2.5 while reducing latency by 26% on similar semantic requests. Selected from 1,400+ hackers for a technical interview with Google.

AI Auth – UC Berkeley CalHacks 2025 Winner Top 1% devpost.com/software/auth-agent

Oct 2025

- Built an npm package (AI Auth) for JWT authentication enabling autonomous agents to securely authenticate across platforms.
- Deployed on Cloudflare Workers using TypeScript with Hono as an edge framework. Selected among 200+ projects.

SKILLS

Tools & Inference: vLLM, self-hosted LLM, TensorRT, YOLOv8, LoRA, PyTorch, TypeScript, Node.js, Python, C++, React, Next.js

Infrastructure: Kubernetes (Amazon EKS), Docker, AWS (EKS, EC2, Lambda, RDS, S3, DynamoDB), Google Cloud (Confidential VM A3/H100, Vertex AI), Cloudflare Workers, Grafana

Backend & Realtime: FastAPI, NestJS, tRPC, WebRTC, Apache Kafka, PostgreSQL/pgvector, Pinecone, ChromaDB