Gabriel Fagundes

Website: gabrielft.me | LinkedIn: gabrielft-me | Github: gabrielft-me | Email: contact@gabrielft.me | Phone: (813) 485 0085

EDUCATION

University of South Florida *Tampa, FL*
Bachelor of science, Artificial Intelligence (Dean's List Fall 2024, Spring 2025, Latin America Scholarship *Exp. May 2027 | GPA:* 3.81

Stanford University
Data Structures and Algorithms with Tim Roughgarden, Summer 2025

EXPERIENCE

Lyra – Backed by Y-Combinator New York City, NY
AI/ML, Backend Engineer Intern *Oct 2025 – Nov 2025*

- Developed a backend based on edge computing to manage meetings summaries with tRPC on Vercel using Drizzle ORM.
- Deployed voice microservices with ElevenLabs voice SDK on Cloudflare using Hono, launched to 800 users, grew to 5,000+ users in 7 days.

Google Developer Group devfesttampabay.com Tampa, FL
Tech Lead *May 2025 – Nov 2025*

- Organized and led monthly workshops for 50+ students (Typescript and C++), co-led a 300+ participant hackathon and mentored 5 graduate students on developing Machine Learning projects.
- Developed a Next.js website, driving a 45% increase in registrations and closing 12+ corporate partnerships totalling $8,000.

inkPen inkpen.com.br Bahia, Brazil
Co-founder, Backend Developer, DevOPS Engineer *May 2023 – Dec 2024*

- Shipped a full-stack TypeScript/React/NestJS platform for InkPen, an EdTech serving 10K teachers with daily workflows.
- Implemented JWT auth and service orchestration with Docker on AWS EC2; Created edge functions deployed on AWS Lambda, and used AWS RDS to store relational data using PostgreSQL.
- Piloted in 3 schools and scaled to 296 schools, resulting in $11,000 in pre-seed funding to keep the app free.

ENTER International enterinternational.org Venice, Italy
Front End Developer & Web designer *Jul. 2022 – Sep. 2022*

- Frontend developer in an Italian startup, leveraging HTML, CSS, and JavaScript to establish a compelling digital presence.

Nucleo do conhecimento (scientific journal) gabrielft.me/maker.html Caetite, Brazil
Published researcher, IOT/C++ Engineer *Jun 2021 – Jan 2023*

- Led an 18-month on-field research on automation to control COVID 19 proliferation, secured 1 dedicated lab room, coordinated 2 robotics professors, and 8 grad students. Designed and soldered boards, and translated microbiology/mechatronics into C++ logic.
- Shipped a firmware stack; Used linear algebra transformations to increase memory efficiency; added serial CSV logging for R analysis and reproducible builds with modular headers.
- The thesis was published in Q2 journal of 25 M annual readers validating Arduino automation for COVID 19 control.

PROJECTS

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 analytics.
- Shipped a backend with FastAPI and WebRTC, exposing a YOLOv8 Machine Learning (ML) model trained on NVIDIA A100 and deployed for real-time inference on a laptop RTX 5070 Ti powered by CUDA with mixed-precision FP16.
- Optimized GPU kernel execution and batch pipeline, reducing inference latency from 190ms to 30ms (84% reduction).

UC Berkley 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.

Small Language Models Routers – Shell Hacks 2025 devpost.com/software/slim-ygzr9c *Sep 2025*

- Built SLiM in 36 hours, a Smart LLM Router and Incremental AI Model trainer that reduces latency and cost by auto-promoting frequent queries into fine-tuned small Machine Learning models (SLM).
- Designed CUDA accelerated LoRA fine-tuning workflow in PyTorch and ChromaDB for local training, orchestrating distributed GPU jobs via Vertex AI with TensorRT for deep learning, using NVIDIA T4 (batch size of 1, FP16), and validated SLM outputs against Gemini 2.5 using sentence-transformers.
- Kept latency at 74% on repetitive prompts while maintaining response semantic similarity above 90%. The project was selected among 1,400 hackers for an interview with Google.

SKILLS

Languages & Frameworks: TypeScript, Java, Python, SQL, C++ Node.js, React, Next.js, FastAPI, PyTorch, TensorFlow, OpenCV.
AI Systems: CUDA, cuDNN, RAG, TensorRT, FP16/INT8, Distributed Training (NCCL), GPU Acceleration.
Modeling & CV: Machine Learning, YOLOv8, Computer Vision, Deep Learning, LLM Fine-tuning (LoRA), Model Quantization.
Infra: Docker, AWS EC2 g4/g5, AWS RDS, AWS Lambda Cloudflare Workers, PostgreSQL, SQL.