# Fabio Guzman

*Research Engineer | Model Efficiency & On-Device ML |*
*Associate Professor*

fabio.guzmanf@upb.edu.co
fguzman82
fguzman82.github.io
fabioguzmanfigueroa

## Summary

Research engineer with 10+ years of experience and a Ph.D. in Engineering (Explainable AI), focused on efficient intelligence: squeezing every token out of any accelerator through model optimization, on-device inference, and algorithmic efficiency across the full ML stack. Hands-on with model efficiency techniques — distillation, quantization, pruning, and reparameterization — reaching ∼1 ms inference for CLIP/SAM models on Apple Silicon, and finetuning and deploying LLMs at scale with *llama.cpp* on Nvidia GPUs. Systems-level rigor spanning algorithms, software, and hardware, from RISC processor and FPGA design — including *gateGPT*, a full Transformer synthesized to RTL on a Virtex-5 FPGA at ∼56k tokens/s — to PyTorch model optimization: the cross-stack co-design where efficiency gains actually live. Strong Python and PyTorch foundation, with production experience deploying ML workloads on AWS with Docker and CI/CD, and building observability for non-deterministic systems (latency, token consumption, output consistency). Driven by research that proves itself through working products rather than paper count, with additional hands-on experience in agent orchestration, MCP servers, and LLM-powered developer tooling.

## Experience

**Stealth Startup** *October 2024 – Present*
*Senior AI / ML Engineer, Miami, US (Remote)*

- *On-device Model Efficiency:* Co-designed model and serving runtime as one system to reach ∼1 ms inference for CLIP and SAM on Apple Silicon (A17) Neural Engine, applying distillation, quantization, pruning, and reparameterization to enable offline, real-time adaptation on mobile.
- *Synthetic Data & LLM Finetuning:* Built finetuning and distillation pipelines (Unsloth / TorchTune stack) on multi-GPU workstations, steering LLM-generated synthetic and augmented data to cover under-represented regions of the input space and improve model behavior where real data was scarce; explored cross-tokenizer distillation and drifting-model alternatives to diffusion.
- *Cross-Stack Optimization:* Collaborated across software, hardware, and algorithmic domains for system-wide efficiency gains — quantization-aware inference, runtime model routing, and latency-driven model selection co-designed end to end.
- *Real-Time Adaptation from Product Signal:* Designed feedback loops where live product signal drove algorithmic and routing decisions, adapting model selection and tool-use behavior in real time rather than relying on static, frozen configurations.
- *Metrics that Matter:* Instrumented logging, metrics, and distributed tracing for latency, token consumption, and output consistency, tying every model and prompt revision to real-world product impact through baselines and alerting.
- *Agent Orchestration & MCP Servers:* Built AI agent infrastructure with custom MCP (Model Context Protocol) servers and tool-use pipelines orchestrated with the Vercel AI SDK, plus evaluation harnesses to regression-test non-deterministic outputs across model upgrades.
- *Cloud Infrastructure & CI/CD:* Deployed ML workloads to AWS using Docker and CI/CD pipelines, ensuring reproducible rollouts, high throughput, and low latency for production systems at scale.
- *Mentorship & Technical Leadership:* Mentor junior engineers and lead technical discussions on model efficiency, real-time adaptation, and deployment of non-deterministic systems.

**Universidad Pontificia Bolivariana** *January 2009 – Present*
*Professor and Researcher in Department of Computer Science, Medellin, CO*

- *Courses Taught:* Operating Systems, AI on the Edge, Machine Learning, Embedded Systems and Computer Architecture

**Projects and Research:** *January 2014 – Present*

- *gateGPT — Transformer-to-Silicon:* Designed a full character-level GPT (RMSNorm, multi-head causal attention, MLP, persistent KV cache) as an independent RTL implementation on a Xilinx Virtex-5 FPGA, generating

tokens at ∼56k–69k tok/s at 80 MHz with no GPU or CPU. Co-designed the algorithm, fixed-point spec (Q5.11), microcode ISA, and datapath as one system, improving throughput 28× over the first working version (2.4k → 69k tok/s) while staying bit-exact to a Python reference. Open-source project that gained wide traction in the hardware/ML community.

- *CLIP Finder:* Developed an iOS app for semantic offline searches of gallery photos using natural language descriptions or the camera, employing Core ML models optimized for the Neural Engine and MPSGraph for pre- and post-processing on the GPU.
- *CoreMLProfiler:* Developed a visual profiling tool for macOS to analyze Core ML models, providing estimation times and validation messages for the Neural Engine.
- *PhD Thesis - Explainable AI:* Developed attribution methodologies to accurately identify concepts in convolutional neural networks (CNNs), generating reliable explanations to enhance confidence, transparency, and security in decisions made by CNNs.
- *RISC Processor Design:* Designed and simulated a RISC processor suitable for tape-out using SKY 130nm technology, including a hand-coded assembler-level compiler tool — full algorithm-to-silicon co-design for efficient, high-throughput execution.
- *LLM Models Deployment & Finetuning:* Led the deployment of locally finetuned LLM models using an engineering question bank, applied to generating exams for students. Inference was performed with *llama.cpp* on Nvidia GPUs, optimizing for throughput and memory footprint.
- *Adversarial Samples Mitigation:* Developed a methodology for concept retrieval in CNNs, using reliable explanations to mitigate inaccuracies in predictions produced by adversarial samples.
- *Early Identification of Developmental Disorders:* Advised on projects applying LLMs for the early identification of developmental disorders in infants through the analysis of babbling.

**GREEMSY** *December 2014 – June 2017*
*Co-founder & Embedded System Engineer, Medellin, CO*

- *DSP Acceleration Platforms:* Developed digital signal processing acceleration platforms for software-defined radio systems using Nvidia Jetson embedded architecture.
- *Parallella Libraries:* Developed libraries for the 16-core Epiphany coprocessor on Parallella, enabling the deployment of software-defined radio applications in ultra-low-power systems.
- *LTE Schedulers:* Developed schedulers for the deployment of LTE submodules using a hybrid FPGA and Embedded-GPU architecture.

## Skills

**Languages:** Python, JavaScript, TypeScript, Swift, SwiftUI, C, C++, Verilog
**Model Optimization & Efficiency:** Distillation, quantization, pruning, reparameterization, finetuning, RLHF, cross-tokenizer distillation, on-device / Neural Engine inference, llama.cpp
**ML & Deep Learning:** PyTorch, JAX, Core ML, MPSGraph, Unsloth, TorchTune
**AI Agents & LLMs:** MCP (Model Context Protocol) servers, Vercel AI SDK, tool schema design, prompt engineering, evaluation harnesses
**Cloud & Infrastructure:** AWS, Docker, CI/CD pipelines, observability (logs, metrics, tracing for ML workloads: tokens, latency, output consistency)
**Backend & Frameworks:** Node.js, Next.js, Restify, Fastify
**Tools:** Xcode, VS Code, Git, Linux shell scripting

## Education

- **PhD in Engineering, Artificial Intelligence** *2017 – 2022*
  *Universidad Pontificia Bolivariana, Medellin, CO*
  Cum Laude, Thesis: Explainable AI: Methodologies for Identification and Retrieval of Concepts Applied to Convolutional Neural Networks.
  Conducted significant research in the field of AI interpretability, focusing on the development of concept attribution techniques to make deep learning models more transparent and trustworthy.
- **Master of Advanced Studies in Embedded System Design** *2012 – 2013*
  *Università della Svizzera italiana | USI, Lugano, Switzerland*
  Diploma is awarded by USI in collaboration with ETH and Politecnico di Milano
  Thesis: Wireless Communications with FPGA.

- **M.Sc. in Engineering (Emphasis in Telecommunications)** *2009 – 2011*
  *Universidad Pontificia Bolivariana, Medellin, CO*
  Thesis: Evaluation and implementation of techniques for symbol timing synchronization and carrier phase synchronization in digital receivers.
- **B.Sc in Electronics Engineering** *1999 – 2005*
  *Universidad Pontificia Bolivariana, Medellin, CO*
  Thesis: 32 bits RISC Processor.

## Fellowship Awards

- **Minciencias, Colombia**
  Government research grant and full scholarship, 2017-2022
  *PhD. Degree*
- **Università della Svizzera italiana | USI, Lugano, Switzerland**
  Full ALaRI scholarship, 2012-2013
  *M. Sc. in Embedded Systems Design*

## Licenses and Certifications

- **Deep Learning Specialization (5 courses)**
  Certificate awarded by Coursera, October 2018
- **Building Transformer-Based Natural Language Processing Applications**
  Certificate awarded by Nvidia, June 2022
- **Fundamentals of Accelerated Data Science with RAPIDS**
  Certificate awarded by Nvidia, June 2022