Bautista García

Inference Systems Engineer

Buenos Aires, Argentina | Open to relocation

bautistagarciace@gmail.com | GitHub | X | Technical Blog | LinkedIn

EXPERIENCE

MTLQwen, independent Qwen 3.5 inference engine — May 2026 - Present

- Built a public C++/metal-cpp inference engine for Apple GPU Family 8 and newer, replacing PyTorch MPS to control allocation, kernel profiling, and concurrent command scheduling. On a 512-token prompt / 128-token generation benchmark, Qwen3.5-9B (UD-Q4_K_XL) reaches about 230 prefill tok/s and 37.2 generation tok/s, 25% above llama.cpp's 29.7 tok/s under matched quantization and sampling settings (both using DFLASH spec decoding).
- Wrote FlashAttention 2, fused dequantization and GEMV/GEMM megakernels, and chunk-parallel GatedDeltaNet prefill in Metal. With no public Metal implementation of chunked GDN, derived GPU-friendly tiled matrix multiplications, parallel scans, and reductions, validating numerics against sequential PyTorch.
- Designed a unified sequence-state contract spanning continuous batching, chunked prefill, paged attention, prefix caching, MTP, DFlash and Metal Sparse buffers. This enables concurrent multi-user sessions in a local engine while committing physical memory on demand as each sequence grows.
- Built an agent-assisted optimization loop that tests candidate kernels against a stable PyTorch reference, inspects AIR disassembly or Xcode GPU traces to verify compiler output, and keeps only correct, reproducible speedups above a predefined threshold.

tinygrad, upstream contributor — February 2026

- Made uncached PyTorch checkpoint loading 2.52x faster (709 ms to 281 ms) while removing 11 lines of core code. When CI exposed a shared-storage regression, root-caused it and landed a regression-tested fix that remains upstream. PR #14734, PR #14771.

OTHER PROJECTS

Drone Controller, university project — 2025

- Reverse-engineered the drone's communications protocol and designed a custom controller PCB in KiCad; the finished controller successfully flew the drone.

Certo, Aleph Hackathon — 2024

- Owned the frontend and backend of a survey platform in a five-person team. The team placed top 3 among 55 projects and earned more than $5,000 across sponsor tracks.

Freelance Full-Stack Projects, independent client work — 2023 - 2024

- Built seven private full-stack projects for freelance clients using Next.js, Tailwind CSS, and MongoDB, spanning landing pages and custom web applications.

EDUCATION

Universidad Nacional de La Plata, Computer Engineer — March 2022 - July 2026

- GPA 8.75/10. Completed the five-year curriculum in nine semesters; coursework included mathematics, physics, computer architecture, operating systems, distributed and parallel systems, GPUs, deep learning, and advanced processor architectures.

TECHNICAL WRITING

- Published first-principles explanations of Metal dequantization and GEMV, linear attention and GatedDeltaNet, and Llama kernel optimization, using writing to expose gaps in the underlying reasoning and share findings with the inference community.