

Marcos Hernanz

Madrid, Spain · marcos@hernanz.io · hernanz.io · [GitHub](https://github.com/mhernanz) · [LinkedIn](https://www.linkedin.com/in/mhernanz)

EDUCATION

Universidad Autónoma de Madrid, B.S. in Computer Engineering, GPA 8.48/10

Sep. 2023 – Jun. 2027

EXPERIENCE

Vercel, Software Engineer Intern, Next.js team

San Francisco Jun. – Sep. 2026

- Built an AI maintainer for Next.js that triages, reproduces, bisects and fixes GitHub and customer reports.
- Ran it as a human-reviewed queue that [closed 1,500 GitHub issues in one month](#), cutting the open backlog from 2,244 to 995 with 3 reopens.
- Owned the agent stack behind it, including resumable workflows, model routing, evaluation tracing and a benchmark of the Eve harness on Terminal-Bench 2.1.
- Merged 22 Next.js and Turbopack PRs, including a [42× faster HMR invalidation](#) and a [React Compiler precheck](#) that cut compiler time 19.6% on real modules.
- Found a cross-tenant data leak in Vercel Sandbox's storage layer, credited by name in [Vercel's \\$1M challenge post](#).

RESEARCH

[Bachelor's thesis: how the optimizer shapes routing in mixture-of-experts LMs](#)

2026 – 2027

Advised by Daniel Hernández-Lobato, UAM.

- Built the PyTorch MoE training stack (grouped-matmul experts, batched Muon, torch.compile, bitwise-exact resume) and ran 145 pre-registered runs at 55M and 108M active parameters for under \$200.
- Found that Muon makes the load balancer work harder, and that the backbone's optimizer, not the router's, causes it through a shared direction in the router's inputs.

[Multi-Head mHC: No Gain at Equal Parameters](#), preprint

Oct. 2026

- Tested a multi-head variant of manifold-constrained hyper-connections in 629 TPU runs of 27M and 112M GPTs.

[The Optimal Rate Constant of Heavy-Ball Momentum on Smooth Strongly Convex Functions](#), preprint

Oct. 2026

- Proved matching bounds showing that optimally tuned heavy-ball is asymptotically 4× faster than gradient descent.

PROJECTS

[llm-internals](#), architecture, kernels and post-training from scratch

PyTorch, Triton Sep. 2026 – present

- **Architecture:** reimplemented DeepSeek-V4.1-Flash from its technical report: compressed sparse attention with a top-k indexer, mHC with Sinkhorn mixing, Engram n-gram memory and the DSpark draft head.
- **Kernels:** hand-wrote fused Triton RMSNorm forward and backward kernels and their autograd function, accumulating the weight gradient in per-SM partial sums.
- **Post-training:** reimplemented Qwen3-0.6B with a batched KV-cache sampler, then wrote full-parameter SFT with assistant-only loss and sequence packing that took zero-shot GSM8K from 6% to 52%.
- Fused the head into a chunked cross-entropy, so 4,096-token steps fit where the plain loss ran out of memory.

[llm-lab](#), language-model training lab

JAX, PyTorch Mar. – Sep. 2026

- Wrote a byte-level BPE tokenizer and a 10B-token FineWeb-Edu pipeline, then trained a 13.7M-parameter Transformer in JAX for 39.85B tokens at 2.63M tokens/s on 8 TPU v5e chips.
- Added one modern mechanism per run across 18 runs, including RoPE, GQA, MLA, hybrid attention, a 64-expert MoE with loss-free balancing, Kimi Delta Attention, attention residuals and multi-token prediction.

[BareTensor](#), tensor and autograd library from scratch

C++, Python Feb. – Apr. 2026

- Wrote a PyTorch-style library in C++ with strided tensors, broadcasting, dynamic autograd, neural-network modules and Python bindings, and trained language models with it up to self-attention.

SKILLS AND AWARDS

ML: PyTorch, JAX, Triton, torch.compile, data-parallel TPU training **Languages:** Python, C++, TypeScript, Rust

Awards: Codeforces Expert (1800), ICPC SWERC (2×), Spanish Olympiad in Informatics (16th; 2nd in Madrid), Meta Hacker Cup top 2,000 (3×)