Alazer Manakelew

Boston, MA | 240-703-0593 | alazermanakelewb@gmail.com | linkedin.com/in/alazer-manakelew1 | github.com/Lazarus-931

Available for Co-op: January – June 2027

Education

Northeastern University Boston, MA

Candidate for Bachelor of Science Major: Computer Science and Mathematics GPA: 3.68/4.0 Expected Aug 2028

- Related Courses: Algorithms, Computer Systems, Theory of Computation, Fourier Series and PDEs, Linear Algebra

Technical Skills

Languages: Python, C++, Metal Shading Language, Swift, Objective-C++, Rust, Java
Libraries & Frameworks: PyTorch, JAX, MLX, CUDA, CUTLASS/CuTe DSL, NumPy, Flax, Optax, React, FastAPI
Developer Tools: Git, Docker, Linux, macOS, Xcode, FFmpeg

Work Experience

Neywa Labs Remote

Member of Technical Staff *June 2026 – October 2026*

- Engineered inference support for vision-language and text workloads in mlx-vlm (**5M+** monthly downloads), implementing the design and maintenance of model architectures for generation, cache, and decoding paths.
- Implemented speculative decoding with multi-token-prediction drafter detection and rollback for ragged batches; fixed recurrent-kernel dispatch, KV-cache reuse, chunked prefill, and end-of-sequence handling.
- First engineer hire, lead the led the inference team on key redesigns such as speculative decoding, APC, as well as kernels (GDN, sparse attention, MoE gather). Raised 5 million to scale local inference

Khoury College, Northeastern University Boston, MA

Teaching Assistant, CS2100 *Jan 2026 – Present*

- Led weekly lab sessions covering lecture review, debugging support, and reinforcement of core course topics.
- Held 1-on-1 outreach with struggling students and coordinated with faculty on course design and logistics.

Projects

mccl / mccl-jax | Author | *Metal, Obj-C++, JAX* | github.com/Lazarus-931/mccl Jun 2026 – Present

- Wrote a collective communication library for Apple-Silicon Mac clusters over Thunderbolt, modeled on NVIDIA NCCL, with automatic topology discovery and ring vs. double-binary-tree selection per message size.
- Built a JAX PJRT plugin on top that lowers StableHLO to MPSGraph (attention, conv, scan, full Transformer fwd+bwd); matched Apple's jax-metal within **2%** on large GEMM and ran **1.15**× faster on attention.
- Trained data-parallel Transformers to **2B** parameters across **5** M4 Mac minis at **1.3**–**1.5**× the steps/s of mlx.distributed, and a **1B** decoder-only Transformer on The Pile at **43%** MFU.

MetalBench | Author | *Apple Metal, C++* | github.com/Lazarus-931/MetalBench Apr 2026 – Present

- Built a kernel benchmark and authoring harness for Apple Silicon modeled on KernelBench (CUDA → Metal, PyTorch → MLX), grading **5** metrics across **3** kernel tiers from basic ops to full model blocks.
- Developed Agent Steel, a closed-loop LLM harness that writes, profiles, and verifies kernels per chip; outperformed PyTorch's KernelAgent and published results on a public leaderboard.

SuperKittens | Author | *Apple Metal, C++, C* | github.com/Lazarus-931/SuperKittens Mar 2026 – Present

- Developed a Metal kernel library for Apple Silicon inspired by ThunderKittens, shipping GEMM (simdgroup MMA), attention, norms, and activations as `.metal` sources plus C headers.
- Built a GGUF-quantized inference engine on top decoding Qwen3-4B at **40** tok/s on an M4 Mac mini; removed redundant memory barriers and added native Q4_K_M paths for a **5.6**× decode speedup.

Interests

Writing and reading on wide range of topics; adventure travel; soccer