Nicolas Basile

nico.basile24@gmail.com • linkedin.com/in/nicobasile • nicobasile.com
(239) 290-4145 • Mountain View, California

SUMMARY

Machine learning engineer focused on learning algorithms and efficient execution, with experience in post-training, reinforcement learning, compilers, and inference optimization.

PROFESSIONAL EXPERIENCE

Software Engineer | Google *Sunnyvale, CA | Mar 2025 - Present*

- Developing GNN-based compiler optimizations for scheduling, memory assignment, and replacement of an expanding range of compiler passes and decisions.
- Building graph encodings, GNN architectures, pre-training data, and hardware-in-the-loop RL infrastructure, to optimize TPU runtime and memory consumption.
- Added compiler lowering support for float4/float8 execution, enabling native accelerator execution paths and up to 2x throughput on supported workloads.
- Built value provenance and numerical-debugging infrastructure across JAX operations, HLO transformations, fusion, and device kernels, enabling root-cause analysis of accelerator correctness issues.

Machine Learning Engineer | VMware *Palo Alto, CA | Feb 2022 - Feb 2024*

- Contributed CUDA kernel optimizations to vLLM that reduced memory I/O overhead; co-authored award-winning research on distributed training and inference efficiency.
- Replaced a legacy inference stack with a custom vLLM-based runtime, increasing throughput approximately 50x through continuous batching and distributed serving optimizations.
- Post-trained LLaMA-style models via supervised fine-tuning (SFT) for multi-step enterprise tool use; generated and execution-validated synthetic conversation and tool-call trajectories.
- Built an API registry with runtime tool-call validation and an LLM-based evaluation pipeline to assess task completion and step-level failures from execution traces.

RESEARCH EXPERIENCE

Robotics Foundation Model Research | Portfolio *Oct 2025 - Feb 2026*

- Improved real-world pick-and-place success from 8% to 84% through evaluation of six training-data mixtures, with 25 trials per model with held-out lighting, background, and poses.
- Fine-tuned SmolVLA for the SO-101 arm with Conditional Flow Matching and receding horizon control, achieving 30 Hz robot control on an RTX 3060 Ti.
- Built a training dataset combining 50 real teleoperation trajectories, 150 augmented trajectories, and 1,000 MuJoCo/SAPIEN rollouts; trained with weighted over-sampling.

Embodied Agentic AI Research | Portfolio *Oct 2023 - May 2024*

- Extended a Voyager-style framework with structured hierarchical memory and a DSPy-driven optimization loop, improving success on long-horizon tasks by approximately 20%.

KEY SKILLS

Languages and frameworks:	Python, C++, JAX, PyTorch, TensorFlow, XLA, CUDA, Triton
ML Systems:	vLLM, distributed inference, profiling, Ray, MLflow, GCP, AWS
Robotics:	MuJoCo, SAPIEN, ROS 2

EDUCATION

University of Southern California *Los Angeles, CA | Aug 2018 - Dec 2021*

B.S. in Computer Science, B.S. in Business Administration