Piero F. Orderique

☎ 706.987.3958 in pforderique pforderique

✉ pforderique@gmail.com 🌐 pforderique.com

SWE/ML Infra engineer with experience in GPU cluster validation, cloud, and applied LLM research

Education

Massachusetts Institute of Technology

MEng Computer Science & Engineering | 2024 | GPA 5.0/5.0

Concentration: Artificial Intelligence

BS Computer Science & Engineering | 2024 | GPA 4.8/5.0

Skills

Programming: Python, C++, Go, TypeScript, SQL, Java, Assembly, Bash | Google Python Readability Certified

ML/AI: PyTorch, TensorFlow, NeMo, LangChain

Infrastructure: Kubernetes (GKE), Slurm, Docker, Terraform, Google Cloud, Azure, CI/CD, Redis, Argo Workflows

Projects

(see more at pforderique.com/projects)

Market Simulation Engine (OSS, 2025)

- Developed a market exchange with a custom order-matching engine handling book management via heaps and database persistence; parallelized per-stock execution with multithreading to reduce latency (C++, SQLite)
- Simulated the end-to-end market journey, including a TCP exchange server, a UDP transaction recorder with an HTTP API for historical prices, and a terminal UI for "real-time" quotes and chart updates (C++, Python, Docker)

Stock Screener Tool (OSS, 2024)

- Automated cron job tool to refresh daily quotes and core ratings using Morningstar API and a terminal UI stock screener equipped with action recommendations (Python)
- Implemented a robust API client with a custom rate limiter, exponential backoff/retry logic, and Redis cache for stable identifiers (performanceIDs)

Device Compatibility Classifier (Microsoft, 2022)

- Trained a binary classifier to predict Microsoft Intune compatibility from device specs, addressing a 99:1 class imbalance with oversampling (SMOTE) and achieving 97% accuracy (Azure ML)
- Deployed model as a web service (API) for automated device screening

CNN Model Compression (2023)

- Reduced VGG-16 CNN model storage by 75% with <0.1% accuracy loss after finetuning using pruning and quantization (PyTorch)

Rent Prediction CNN (2023)

- Designed and built a CNN to predict rental prices by integrating structured (location, year built) and unstructured (Google Street View images) data (PyTorch, OpenCV)

Experience

Software Engineer @ Google *09/24 – present*

- Designed and implemented an end-to-end automated software qualification pipeline for scalable GPU clusters (H100/H200/GB200), cutting engineering effort per qualification from 3.6 hours to <30 minutes (-86%) and enabling self-serve use across teams (Python)
- Maintained the hardware qualification system gating release to all A* supercomputers, qualifying 100k+ GPUs (~$100M+ capacity) for production readiness (Go)
- Operated Kubernetes-based GPU clusters at scale, proactively validating 30+ major software updates to ensure zero customer disruption and no performance regressions, upholding reliability guarantees (NeMo)
- Introduced CI/CD pipelines for cloud infrastructure to provision and maintain workflows, buckets, artifacts, permissions, and Cloud Run jobs, services, and functions

Graduate Researcher @ MIT-IBM Watson AI Lab *01/24 – 07/24*

- Designed an agent leveraging causal and prescriptive reasoning tools with natural follow-ups (LangChain)
- Trained LLMs using parameter-efficient fine-tuning techniques (prompt tuning, adapters), achieving an 11% improvement in downstream task accuracy
- Translated research on hallucination mitigation and intent recognition into production-ready pipelines (Python), generalized for enterprise-scale use cases

Software Engineering Intern @ Google *06/23 – 08/23*

- Added NLP layers (tokenizers, attention, transformer blocks), introducing in-browser LLM support to TFJS
- Implemented GPT-2, TensorFlow.js's first LLM, by adapting Keras and fundamental research papers
- Designed new tensor ops with algorithmic optimizations needed for real-time inference performance

STEP Intern @ Google *05/22 – 08/22*

- Refactored the full-stack messaging/buffer system (Java, TS) to improve caching efficiency, reducing latency by 20%
- Built Angular autocomplete components, eliminating invalid options (98% to 0%) and improving UX
- Increased test coverage by 15%+ and audited accessibility across internal tools to ensure compliance within Travel org

Robotics Software Intern @ NVIDIA *06/21 – 08/21*

- Optimized OpenCV ROS2 packages to run faster on NVIDIA hardware, utilizing GPUs (C++)
- Authored unit, integration, and benchmarking tests for robotics workloads (Python, Docker)
- Resolved synchronization issues by implementing time policy algorithms, improving reliability (C++)

Software Engineer @ MIT Off. of Sustainability *09/20 – 05/21*

- Identified $600M+ of potential flood risk damages to university property from flooding simulations
- Designed a Python package for reading and visualizing 300+ specialized geodata files (SciPy, matplotlib, NumPy)
- Engineered data filters and custom compression, reducing geodata storage by 99.99% (1.1TB to 1MB)