NIKHIL PALETI

510-935-8895 ◇ nikhilpaleti23@gmail.com ◇ linkedin.com/in/nikhil-paleti ◇ github.com/Nikhil-Paleti

EDUCATION

University of California San Diego — Sep 2024 - Dec 2025 (Expected)
Master of Science in Data Science (Artificial Intelligence & Machine Learning) — GPA: 4.0/4.0

Amrita Vishwa Vidyapeetham University — Oct 2020 - June 2024
Bachelor of Technology in Computer Science and Engineering (Artificial Intelligence) — GPA: 4.0/4.0

EXPERIENCE

Waymo | Google — Jun 2025 – Sep 2025
Software Engineer Intern – ML Infrastructure (Frameworks & Efficiency) — *Mountain View, CA*

- Developed a **model surgery toolkit** that automates tensor debugging and prevents silent checkpoint restoration failures across training and evaluation pipelines, reducing debugging time by over **90%** (from days to hours).
- Extended the toolkit to automate foundation model conversion and loading between **Waymo** and **Google DeepMind (Gemini)** training infrastructures.
- Implemented **dataset checkpointing** for **Waymo's foundation model pipelines** on `Grain`, enabling fault-tolerant and resumable training, and profiled `tf.data` & Python backends to guide throughput optimizations.

Hao AI Lab — Mar 2025 - Present
Research Assistant - Machine Learning Systems — *La Jolla, CA*

- Collaborating with **NVIDIA** on building `NeMo/Gym` with diverse game environments to build standardized interfaces and rollout pipelines for large-scale **LLM and RL agent** training and evaluation.
- Developing an **agent for automated profiling trace analysis**, detecting performance bottlenecks (memory, network) and suggesting optimizations for ML systems.

Tech Profuse Pvt Ltd — Jan 2024 – Jun 2024
Machine Learning Engineer Intern — *Hyderabad, India*

- Developed an **unstructured data extraction API** using **Gemini**, processing over **50K bill of lading documents** in 15 hours and reducing manual data entry effort by **98%**.
- Built a multimodal data extraction prototype by fine-tuning a **LLaVA** model with distributed training (`FSDP/ZeRO`) across 8 GPUs, improving visual-text alignment accuracy.
- Engineered a **RAG-based support system** leveraging **Cohere LLMs** for natural language issue querying, automated classification, and summarization, increasing support throughput by **130%**.

PROJECTS

Kernel Forge – Custom GPU Kernels in Triton & CUDA — *github.com/Nikhil-Paleti/kernel-forge*

- Designed and implemented high-performance GPU kernels (`matmul`, `attention` etc.) using **Triton** and **CUDA**.
- Ranked in the **top 0.2% globally (8K+ participants)** on **LeetGPU** for Triton kernel performance leaderboard.

Mini-Collectives – MPI Communication Benchmarks — *github.com/Nikhil-Paleti/mini-collectives*

- Implemented **MPI collectives** (e.g., `Allreduce`, `gather`) from scratch to analyze latency–bandwidth trade-offs.
- Built benchmarking scripts to generate throughput and latency plots visualizing scaling efficiency.

Mini-Trainer – Distributed Transformer Training from Scratch — *github.com/Nikhil-Paleti/mini-trainer*

- Built a lightweight distributed training framework inspired by **MiniGPT** and **Picotron**, implementing **data** (with gradient bucketing), **tensor**, and **pipeline parallel** Transformer training.

SKILLS

ML Systems: Distributed Training, LLM Training & Inference Infrastructure, Model Parallelism, Checkpointing, Model Surgery
Frameworks: PyTorch, JAX/Flax, TensorFlow, Ray, DeepSpeed, FSDP, Hugging Face, vLLM, TensorRT, Orbax, Grain, NumPy, Pandas, Scikit-Learn, LangChain, LangGraph, TensorBoard
GPU & Systems: CUDA C/C++, Triton, XLA, NCCL, MPI, Nsight Systems/Compute, Kernel Fusion, Memory Optimization
Programming: Python, C++