

# 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

- **Relevant Coursework:** Machine Learning Systems, Advanced Data Mining, Advanced Data-Driven Text Mining

### Amrita Vishwa Vidyapeetham University, India

Oct 2020 – Jun 2024

*Bachelor of Technology in Computer Science and Engineering (Artificial Intelligence)*

GPA: 9.15/10

- **Relevant Coursework:** AI in Natural Language Processing, AI in Speech Processing, Deep Learning for Signal & Image Processing, Deep Reinforcement Learning

## EXPERIENCE

### Waymo | Google

Jun 2025 – Sep 2025

*Software Engineer Intern – ML Infrastructure*

*Mountain View, CA*

- Optimized **Waymo's model training input pipelines** during migration from `tf.data` to `Grain`, implementing **dataset checkpointing** and optimizing data transformations to achieve a **5×+ throughput speedup**.
- Built a **model surgery toolkit** that automates tensor debugging to prevent silent checkpoint restoration failures across training and evaluation pipelines, reducing debugging time by **90%+**.
- Extended the toolkit to automate checkpoint conversion and migration between **Waymo** and **Google DeepMind (Gemini)** training infrastructures for large-scale foundational models.

### UCSD Hao AI Lab

Mar 2025 – Present

*Research Assistant — Machine Learning Systems*

*San Diego, CA*

- Researching **disaggregated serving** for heterogeneous accelerators to improve performance, scalability, and flexibility in LLM serving pipelines.
- 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** with **Gemini**, processing **50k** bill of lading documents in **15 hours**, reducing manual data entry requirements by **98%**.
- Built a **data extraction prototype** by fine-tuning a **LLAVA multimodal LLM** using distributed training (FSDP/ZeRO) across 8 GPUs.
- Engineered a **RAG-based support system** with **Cohere's LLMs**, combining natural language issue querying, automated classification, and summarization, improving support throughput by **130%**.

## PROJECTS

### Optimizing Deep Learning Systems for High Performance

Jan 2025 – Mar 2025

- Achieved **1.25× GPU speedup** over PyTorch via Triton matmul kernel optimizations with shared-memory tiling, register blocking, and operator fusion.
- Developed a speculative decoding engine combining draft and target LLMs, reducing inference latency by **1.7×** with >75% draft token acceptance.

### Reinforcement Learning for Reasoning in Small LLMs

Jan 2025 – Mar 2025

- Implemented GRPO-based reinforcement learning to fine-tune small LLMs (LLaMA, Qwen, Phi) on GSM8k, using multi-signal reward functions (correctness, numeric validity, and format).
- Evaluated on 1,300+ GSM8k math problems, demonstrating improved reasoning under limited compute budgets.

### Indic Verse: Indic Language LLM System

Jan 2024 – Apr 2024

- Built an Indic language LLM pipeline for translation, transliteration, dataset curation, and model fine-tuning.
- Evaluation datasets adopted by Hugging Face engineers for assessing Telugu performance in FineWeb-2.

## TECHNICAL SKILLS

**ML Systems:** Distributed Training, LLM Training & Inference Infrastructure, Checkpointing & Model Surgery

**Frameworks & Libraries:** PyTorch, JAX/Flax, TensorFlow, DeepSpeed, FSDP, Ray, Hugging Face Transformers, vLLM, TensorRT, Orbx, Grain, NumPy, Pandas

**Systems & Optimization:** CUDA C/C++, GPU Kernels, Triton, XLA, Nsight tools, Profiling & Performance Tuning

**Programming:** Python, C++