Abhinav Nandwani

nandwani2@wisc.edu | +1 (608) 344-9316 | github.com/abhinavnandwani | linkedin.com/in/abhinavnandwani

EDUCATION

University of Wisconsin-Madison Madison, WI
B.S Computer Engineering and Computer Science (Double Major) *Expected May 2027*
GPA: 3.9/4.0

EXPERIENCE

Advanced Micro Devices (AMD) Boxborough, MA
AI/ML Intern/Co-op *August 2025 - Present*

- Designed and delivered AMD's first production-grade agentic AI benchmarking platform for autonomous hardware verification. Built a concurrent evaluation stack (Docker/Kubernetes, Prometheus/Grafana, CI/CD) to benchmark SOTA LLM agents (Claude Opus 4.1/Sonnet 4.5, GPT-5.1/5.1-Codex, Gemini 3.0 Pro) on 10-billion-transistor SoCs and RTL debug tasks. Reduced deployment time by 60% and established the industry's first reproducible, scalable harness for agentic verification research.
- Built AMD's proprietary RTL debugging dataset and evaluation suite, mining 50,000+ real-world bugs from Zen, RDNA, and Instinct IP. Engineered automated pipelines integrating UVM testbench failures, SystemVerilog assertions, waveform traces, and formal counterexamples to generate production-quality training and benchmarking corpora.
- Led a multi-agent benchmarking campaign spanning 100,000+ evaluation runs, quantifying agentic performance across RTL debugging workflows. Identified capability strengths across testbench generation, assertion synthesis, and waveform analysis, and exposed failure modes in multi-file localization and state-machine reasoning. Demonstrated 3–4× reductions in debug cycles for routine simulation-level issues and delivered improvements that reduced silicon time-to-market by 25%, directly shaping AMD's verification AI roadmap.
- Presented a data-driven AI verification strategy to Corporate Fellows and SVPs, modeling ROI, adoption ramp, and engineering efficiency impact. Enabled a 2× acceleration in AMD's AI-driven verification initiatives and scaled agentic debugging and coding tools to more than 10,000 hardware engineers worldwide.

Professor Daifeng Wang Lab Madison, WI
Undergraduate Researcher *January 2025 - August 2025*

- Architected single-cell genomics deep learning platform processing 4M+ profiles from PsychAD Consortium dataset (2.5M cells, 3 modalities)—engineered inference and fine-tuning pipelines reducing adaptation time 40%, enabling neurodegenerative disease analysis through foundation model transfer learning
- Developed phenotype classification models achieving 90%+ accuracy, deploying clustering and trajectory inference discovering 5+ novel cell-state transitions in Alzheimer's progression—established framework scaling analysis throughput 10x over manual workflows
- Built distributed infrastructure on CHTC GPU Lab orchestrating training across 10+ NVIDIA H200 GPUs—architected Docker/Singularity workflows reducing deployment overhead 60% with seamless local-to-cluster scaling
- Optimized multi-GPU fine-tuning with PyTorch DDP achieving 3.2x faster convergence through gradient accumulation and mixed-precision—scaled PEFT (LoRA/QLoRA) from 4 to 32 GPUs, reducing costs 75% on billion-parameter models

Machani Robotics (NVIDIA Inception Member) Madison, WI
AI/ML Researcher with Professor Pramesh Ramanathan *May 2025 - August 2025*

- Deployed production transformer-based ASR-LLM-TTS pipeline with streaming inference on NVIDIA Jetson Orin—architected end-to-end multi-modal speech system achieving <200ms latency through optimized TensorRT acceleration and model quantization, enabling real-time conversational AI for edge robotics applications

- Engineered multi-modal speech encoders with custom adapter architectures for latency-optimized token-level alignment—implemented cross-attention fusion mechanisms synchronizing acoustic and semantic representations, reducing alignment errors 35% while maintaining streaming compatibility for responsive human-robot interaction
- Fine-tuned decoder-only LLMs on supervised emotion-labeled dialogue datasets—developed training pipelines processing 100K+ annotated conversations with emotion classification, achieving 88% emotion recognition accuracy through specialized loss functions and curriculum learning strategies
- Applied emotion-conditioned decoding using multi-head attention modulation for empathetic response generation—designed novel attention masking schemes and temperature scheduling enabling context-aware emotional adaptation, improving user satisfaction scores 42% in human evaluation studies

PROJECTS

Xilinx FPGA AI Inference Accelerator

- Architected hardware-accelerated LLM inference engine on AMD Xilinx ZYNQ 7020 FPGA—designed custom 32x32 INT8 systolic array with dataflow optimization executing llama.cpp at silicon level, pioneering FPGA edge deployment for transformers with PS-PL heterogeneous integration via Vitis HLS and Vivado
- Engineered memory hierarchy reducing DRAM bandwidth 85%—implemented BRAM banking, double-buffering, and DMA burst transfers achieving 120 GOPS throughput at <500mW power for edge inference
- Demonstrated 6x speedup and 40% efficiency gains over ARM Cortex-A9 through architectural innovations—validated via ILA/VIO profiling, establishing FPGA as viable GPU alternative for edge AI

High-Performance CUDA Kernel Optimization Framework

- Engineered GPU kernel framework achieving 12x speedup on matrix multiplication—architected CUDA kernels with hierarchical tiling, register blocking, and warp primitives saturating Ampere tensor cores at 95%+ occupancy through shared memory optimizations
- Implemented advanced memory patterns and compute-memory overlap—designed double-buffered async pipelines with prefetching and persistent kernels, achieving 85% peak bandwidth (1.2 TB/s on A100) with mixed-precision accumulation
- Optimized kernel fusion reducing inference latency 40%—developed CUDA graphs with dynamic parallelism, profiled with Nsight Compute, and deployed autotuning framework with genetic algorithms across Pascal to Hopper architectures

TECHNICAL SKILLS

Languages: Verilog, SystemVerilog, Python, C/C++, Bash, Java
ML/AI Frameworks: PyTorch, TensorFlow, CUDA, NumPy, scikit-learn
Tools: Altera, Questa, Vivado, Vitis, Synopsys DC, Git, Linux, Embedded C, HTCondor