Deval Patel | Software Engineer

devalrocket@gmail.com • (647)-470-7505 • deval-patel • Deval Patel

Experience

Microsoft

SDE II — Graphics Kernel (Core OS) 2025 - Present

- Investigated GPU/OS performance across **NVIDIA, AMD, Intel, and Qualcomm** hardware using **GPUView**, **WPA**, and kernel trace analysis, identifying vendor driver issues and OS scheduling inefficiencies.
- Diagnosed and fixed a critical scheduling bug where workloads were incorrectly placed into a **penalty box** after power-transition sleep states, restoring correct QoS behavior and improving GPU responsiveness.
- Led design for the next-generation **Hardware Scheduler Logging System**, enabling packet-level start/stop telemetry, VM utilization metrics, and significantly improved debugging scope for GPUView.
- Enhanced **TDR (Timeout Detection & Recovery)** logic to better identify true GPU hangs and reduce false positives during hardware bring-up and partner validation.
- Built and flashed Windows OS test builds, performed low-level validation on internal hardware, and debugged crashes/hangs using **WinDBG** and GPU scheduling traces.

Qualcomm

Embedded ML Software Engineer 2022 - 2025

- Designed eNPU (embedded neural processing unit) driver in **C** for Qualcomm's LPAI automotive chip, enabling multi-master DSP support and advancing pre-silicon schedules by 3 months.
- Optimized eNPU firmware for **2–3µs** latency on RTOS, achieving industry-leading real-time ML inference performance.
- Delivered cross-platform driver support across automotive, mobile, IoT, and XR/VR devices through large-scale refactoring and system unification.
- Performed pre-silicon verification on FPGA platforms, identifying and resolving critical bring-up issues for next-generation SoCs.
- Built automated **eAI model profiling** pipeline (Python + ADB) comparing HW vs. SW scheduling to diagnose model latency bottlenecks for internal customers.

Projects

CareOverflow — Disease Diagnosis with BioBERT

Lead ML Engineer 2025 - Present

- Built a complete ML pipeline using **PyTorch**, **BioBERT**, and custom transformers to classify symptoms into 41 diseases (85–95% accuracy depending on configuration).
- Implemented dataset preprocessing, structured vs. natural-language formatting, and train/val/test splitting with custom PyTorch Dataset classes.
- Designed a modular configuration-driven training framework supporting AdamW optimization, label smoothing, gradient accumulation, and per-class evaluation metrics.
- Developed inference engine for both structured and natural-language inputs, supporting top-k predictions and GPU/CPU dispatch.
- Architected a pluggable interface allowing additional transformer models and trainers with minimal integration overhead.

Education

Georgia Institute of Technology

MS in Computer Science (Machine Learning Specialization) — GPA: 4.0 2025 – Present

University of Toronto

HBSc. Computer Science Specialist — GPA: 3.52 2018 – 2022

Skills and Technologies

Systems/OS: C, C++, RTOS, Linux, Drivers

ML/AI: PyTorch, Transformers, BioBERT, CUDA

Tools: Make/CMake, Docker, ADB, Git, GDB

Other: Python, SQL, Bash, Web