Rahul Kumar

ML Engineer: LLM inference, ML performance, on-device AI and computer vision

Big models, small places, real numbers

linkedin.com/in/pctablet505 | work.rahul126@gmail.com
pctablet505.github.io case studies | github.com/pctablet505 | India

SUMMARY

ML engineer with 4+ years at Google, Qualcomm and Samsung. Makes ML fast where it has to run, from GPU inference and compiled frameworks to phone NPUs and DSPs. Builds computer-vision and sensor models that ship on phones. Works profile-first: finds where the time goes, fixes the root cause in Python, C++ or a GPU kernel, and proves it with a benchmark and a check of the model's output. Case studies: **pctablet505.github.io**.

EXPERIENCE

Google Software Engineer, Machine Learning — India · Feb 2025 – Present

ML frameworks and performance: XProf, the open-source ML profiler, now; Keras and KerasHub before

- Made compiled text generation in Keras on PyTorch **11.6x faster**: profiled the Python work repeated on every call and the graph breaks that made compiled code slower than eager, then merged **15 fixes upstream**, each for one root cause, with tests and a before-and-after benchmark.
- Built the **one-call export** from a trained Keras or KerasHub model to a phone-ready LiteRT file, on TensorFlow and PyTorch; fixed the layers that blocked export and wrote the keras.io guide.
- Built the **benchmarking system** for all Keras and KerasHub models: automated weekly runs across GPU and TPU, the TensorFlow, JAX and PyTorch backends, presets and batch sizes, to catch performance regressions.
- Implemented Llama 3.1 in KerasHub and fixed Gemma 3 bugs, including a preprocessor crash across Gemma 3, 3n and 4.
- On XProf, wrote the end-to-end UI test suite, run in CI: **107 Playwright tests** over 7 profiling journeys.
- Led the migration of a major internal customer from legacy Keras to Keras 3, with backward compatibility and performance parity.

Qualcomm ML Engineer, Computer Vision — Hyderabad, India · Sep 2024 – Jan 2025

Hardware-aware optimization of on-device AI for Snapdragon phones

- Built detection, tracking and segmentation pipelines in C++ with INT8 quantization and pruning; tuned models ran **3x faster** on the Snapdragon NPU and DSP.
- Cut the memory bandwidth of image-processing models by 75% with quantization-aware training, and redesigned generative inpainting to fit those processors' latency budgets for the Snapdragon Summit demo.
- Put ML inference inside the native camera stack for a camera assistant controlled by voice.

Samsung R&D Institute India Research and Development Engineer — Noida, India · Jul 2022 – Aug 2024

Computer vision, biometrics and motion-sensor ML in production on phones, under power, memory and latency limits

- As sole researcher and developer, built real-time multi-face recognition for live video at **30+ FPS** and 99.2% accuracy on our own test set; a k-d and ball-tree index cut face search from 1,200 to 87 ms across 10,000+ users. It powered an exam-proctoring platform used by 2,000+ employees.
- Built a synthetic-data engine of 10M+ images whose labels stay true through every transform; YOLO, Mask R-CNN and U-Net models under 1 MB each read **90%+** of a damaged-barcode test set, where a commercial SDK read 70%.
- Replaced a deep-learning phone-drop detector of millions of parameters with motion-sensor features and a statistical model of **fewer than 22**; shipped, using over 95% less power for a small accuracy cost (97% to 95.2%).
- Built behavioural identification as an Android service, from touch and motion and without deep learning: it tells which hand holds the phone, right **over 90%** of the time across 30+ seconds of use, and keeps checking who is using it.
- Detected attention and distraction from facial expression, eye movement, yawns and head pose at **87% accuracy**; confirming a signal across frames cut false alarms by 40%.

Samsung R&D Institute India Intern — Noida, India · Jan 2022 – Jun 2022

- Quantized MobileNet and EfficientNet on Android at 25+ FPS through NNAPI; trained CycleGAN image translation in PyTorch, deployed with TFLite.

SELECTED PROJECTS

GLM-5.3-Flash on one GPU github.com/pctablet505/glm53-flash-single-gpu — Sep 2026

- Served a 181 GiB mixture-of-experts LLM from one 96 GiB GPU, taking decoding from **10.8 to 15.0 tokens/s** with a Triton gather kernel (52.1 of the PCIe link's 56.4 GB/s), CUDA graphs and a cache of the most-used experts.
- Evaluation before speed: a faster setting with fewer experts per token made answers worse, so it stays off, and a sweep of five FP4 kernel backends showed the default was silently wrong. Eight bugs documented.

Servedeck servedeck.rlalphalabs.com — Aug 2026 – Present

- Serves local LLMs on one GPU workstation behind one OpenAI-compatible endpoint, with one-command model switching, a live dashboard and automatic restart of an engine that died silently.

Keras on an Android phone github.com/pctablet505/keras-ondevice-android — Jul 2026

- Runs KerasHub models on the phone with LiteRT and no server, from one export call to **six working screens**, two of them Gemma, with a check of the phone's answers against Python.

RL Alpha Labs: automated trading rlalphalabs.com — May 2025 – Present

- A **point-in-time data plane** for Indian stocks with delisted companies kept in, an exact cost and tax engine, a pure-JAX simulator for reinforcement-learning agents, and pre-registered tests. Paper trading only.

OPEN SOURCE

- **100+ merged pull requests** in 40+ outside repositories (November 2026), most in Keras and KerasHub, plus OpenCV (C++ fixes to DNN inference and optical flow), vLLM and Python tooling such as pip, pydantic and pytest plugins.

SKILLS

Inference and performance: inference optimization, profiling, benchmarking, latency and throughput, torch.compile, Triton GPU kernels, CUDA graphs, KV-cache sizing, vLLM, LLM serving, mixture of experts, transformers

On-device and edge: model deployment, quantization (INT8, NVFP4, quantization-aware training), pruning, LiteRT, TFLite, ONNX, NNAPI, Snapdragon NPU, Hexagon DSP, Exynos NPU, Android

Computer vision and sensing: object detection, tracking, segmentation, face recognition, embedding search, video, synthetic data, sensor fusion, OpenCV

Languages and frameworks: Python, C++, PyTorch, JAX, TensorFlow, Keras 3, KerasHub, Kotlin, Java, SQL, Linux, CI, pytest, Playwright

EDUCATION

MicroMasters in Statistics and Data Science MITx (MIT, online) — Feb 2021 – Jun 2023

Probability, statistics, machine learning and data analysis

B.Tech, Computer Science and Engineering Birsa Institute of Technology Sindri — Aug 2018 – Jun 2022

AWARDS

MD Appreciation Award Samsung R&D Institute India, for the exam-proctoring platform — Sep 2023

HackerRank HackFest 2020 Gold, rank 29 of 8,051 — 2020

TCS CodeVita, Season 9 International rank 242

GATE 2021, Computer Science All India Rank 5,648 — 2021