Ronak Haresh Chhatbar

+1 716-507-2419 | ronakchhatbar@gmail.com | LinkedIn | GitHub | alphapibeta.com | Buffalo, New York

Professional Summary

Production AI systems engineer with 6 years designing and shipping inference systems end to end — GPU-aware compute scheduling, heterogeneous hardware orchestration, multimodal VLM pipelines, and agentic systems — from Jetson edge to enterprise Kubernetes deployments at scale. Architected Centific's Verity platform in production and designs a personal multi-model inference cluster serving live traffic.

Experience

- **Centific** — Remote
  AI Engineer — Verity Multimodal Agentic Platform — *Dec 2024 – Present*
  - **Owned the AI inference stack on Verity**, Centific's enterprise multi-tenant vision platform running simultaneous detection workflows for multiple clients 24/7 — cutting **facility monitoring costs 33%** by replacing continuous human surveillance with autonomous AI-driven alerting.
  - **Engineered GPU-aware workload scheduling** across a heterogeneous compute environment — routing **VLM inference to high-memory GPUs**, real-time detection to dedicated inference GPUs, and semantic search to CPU — sustaining multiple simultaneous enterprise workflows across shared infrastructure at production scale.
  - **Built VLM-powered video Q&A** into the operator triage workflow — voice and text queries route to case footage and return timestamped answers in **under 3 seconds**, with **retrieval quality checks** and **model-evaluated answer quality** verifying query satisfaction before surfacing results to operators.
  - **Cut incident response 67%** with a **LangGraph multi-agent system** routing voice and text across all triage modalities — video Q&A, live sensor feeds, facility APIs, and compliance reports — with retrieval quality checks, **LLM output evaluation**, and confidence gating before surfacing results to operators.
  - **Productionized the full heterogeneous stack** with **Helm-managed Kubernetes**, Harbor CI/CD, and TLS across all service boundaries — enabling consistent, auditable deployments across multi-tenant enterprise client infrastructure.

- **Tensorgo Technologies** — Hyderabad, India
  Computer Vision Engineer — *Sep 2020 – Aug 2022*
  - **Improved model throughput 25% and inference latency 40%** by applying **TensorRT and DeepStream** mixed-precision tuning to eye-gaze and emotion recognition models — sustaining **30–40 FPS** on Jetson NX and Nano under production memory constraints.
  - **Achieved 8% accuracy improvement** in contactless heart rate estimation (rPPG) by training on **20,000+ images** across BP4D+, UBFC-1, and UBFC-2 — explicitly modeling skin tone and lighting variation for inclusive biometric deployment.
  - **Increased meeting analytics accuracy 16%** by integrating real-time speaker segmentation into the emYt+ compliance platform via an ASR pipeline — extending the CV stack into multimodal audio+video analysis for enterprise Zoom and Webex.
  - **Reduced manual intervention 60%** and deployment time **30%** by building automated training, evaluation, and Jetson edge deployment pipelines — enabling continuous delivery of CV models to production hardware.

- **Wavelabs Technologies** — Hyderabad, India
  Machine Learning Engineer — *May 2019 – Aug 2020*
  - **Delivered sub-2-second threat detection at 30–40 FPS** on Jetson Nano by building a real-time weapon detection system integrated with iOS/Android mobile apps — trained on **150,000+ labeled images**.
  - **Generated 15% revenue increase** by developing dynamic pricing algorithms processing **10,000+ monthly transactions** for financial services clients.
  - **Reduced resource utilization 35%** and deployment time **20%** by containerizing model training and serving pipelines with **Docker** and **AWS SageMaker** — establishing reproducible workflows across development and production environments.

Projects & Research

- **Self-Hosted Multi-Model AI Inference Stack**
  Production Agentic Infrastructure — GPU · VLM · Agents · Voice · RAG
  - Designed and operate a **4-node production cluster** serving 8 live AI services: Qwen3-Coder-30B at **196K context** (FP8, tensor-parallel across 2×RTX PRO 4000 Blackwell), open-vocabulary detection at **∼300 ms** (NanoOWL/OWL-ViT TRT on Jetson Orin Nano), and dedicated RAG + embedding hardware on Jetson NX — all on wired LAN.
  - Built an **agentic inference layer** with multi-model routing, streaming SSE, **MCP** tool-calling, and **RAG** over Qdrant (bge-base-en-v1.5 + cross-encoder reranker) — **sub-150 ms** end-to-end retrieval with full LangFuse distributed tracing across all nodes.
  - Engineered a **voice pipeline** (Whisper ASR → LLM → Kokoro GPU TTS) delivering **first audio under 4 seconds** end-to-end.
  - Automated an **EPUB → M4B audiobook pipeline** dispatching parallel TTS synthesis across 2 GPU endpoints with chapter-accurate metadata — auto-delivering completed audiobooks to a self-hosted Audiobookshelf library with zero manual steps.

- **Spatial AI & Robotics Lab** — University at Buffalo
  Graduate Research Assistant — *Dr. Chen Wang* — *May 2023 – Dec 2024*
  - **Increased visual odometry inference efficiency 33%** by developing C++ plugins integrating a custom visual navigation optimizer with a TensorRT backend, enabling real-time pose estimation on resource-constrained platforms.
  - Led backend development of robotranking.org, a robotics benchmarking platform adopted as a reference resource by the international robotics research community.
  - Improved optical flow estimation accuracy and real-time throughput by applying specialized interpolation algorithms to dense frame sequences in dynamic environments.

- **GPU Computing Research Portfolio**
  Advanced CUDA Development
  - **Hessian Matrix Inversion**: Achieved **526× GPU speedup** over CPU baseline via LU decomposition with cuSOLVER and Python bindings.
  - **CUDA Performance Profiler**: Automated Nsight Compute metrics collection with an interactive Streamlit dashboard for systematic kernel optimization.
  - **Convolution Optimization Analysis**: Published analysis across **18 optimization metrics** with mathematical foundations and Nsight Compute profiling visualization.

Education

- **University at Buffalo, The State University of New York** — Buffalo, NY
  *M.S. Computer Science — GPA: **3.4/4.0*** — *Aug 2022 – Jan 2024*
  Courses: Operating Systems, Analysis of Algorithms, Biometrics Image Analysis, Reinforcement Learning, Computer Vision

- **Jawaharlal Nehru Technological University Hyderabad** — Hyderabad, India
  *B.E. Computer Science — GPA: **3.6/4.0*** — *Aug 2015 – May 2019*
  Courses: Machine Learning, Cloud Computing, Data Structures & Algorithms, Computer Networks, Probability & Statistics

Skills

- **GPU Computing & Inference Optimization**: CUDA, TensorRT, TensorRT-LLM, NVIDIA DeepStream, Nsight Compute, Mixed Precision (FP8/FP16), SGLang, Triton Inference Server, H100/A100/Blackwell GPU Clusters, Tensor Parallelism
- **Agentic AI & LLM Systems**: LangGraph, LangChain, Model Context Protocol (MCP), Multi-Agent Orchestration, RAG Pipelines, Qdrant, vLLM, LiteLLM, LangFuse, Vision-Language Models (VLM)
- **Computer Vision & Machine Learning**: PyTorch, TensorFlow, OpenCV, ONNX, Real-Time Video Processing, Object Detection, Multimodal Inference, ASR/TTS Pipelines, Biometric Systems
- **Infrastructure & MLOps**: Python, C/C++, CUDA, SQL, Docker, Kubernetes, Helm, AWS, Jetson (Orin/NX/Nano), Linux, PostgreSQL, Vector Databases, Kafka, GitOps