Krishnam Gupta

San Francisco, CA | kayjee1994@gmail.com | krishnam-gupta | krishnam94 | Google Scholar

Summary

Applied AI Engineer with 8+ years shipping production ML systems in healthcare. Architected LLM systems serving **36,000+ users** across clinical programs in South Africa and Zimbabwe, extracting **120+ clinical fields** at scale with evaluation frameworks cutting inference costs by **over 50%**. Built CV infrastructure behind **1.3M+ diagnostic tests** across 15+ countries. 7 peer-reviewed publications, 2 patents, h-index 5.

Experience

Senior Software Engineer, Applied AI 2022 – Present
Audere – AI-powered diagnostics for global health *Seattle, WA (Remote from SF)*

- Led design of **structured clinical data extraction** from multi-turn WhatsApp conversations – **120+ validated fields** across 4 HIV/TB programs serving **36K+ users**; iterative prompt refinement reduced extraction failures to **zero** across core clinical categories
- Built **LLM evaluation framework** with LLM-as-judge scoring on 6 locally grounded clinical metrics; benchmarked **11+ models**, enabling task-specific selection that cut inference costs **50–63%** while maintaining clinical accuracy
- Improved extraction stability that enabled reliable downstream **risk scoring, care encounter classification, and conversation routing** – directly supporting adoption at scale across programs
- Architected CV pipeline behind **1.3M+ rapid diagnostic tests** (malaria, HIV, COVID-19) – phone camera to result in **<3 seconds** on low-end Android devices across **15+ countries**
- Led rule-based to deep learning migration for test interpretation: sensitivity **82% → 94%**, specificity held at **>99%**, validated across diverse lighting and device conditions
- Designed **multi-provider fallback** with open-weight models at **80%+ lower cost**; production guardrails including content moderation, safety rules, and clinician escalation pathways

Lead Data Scientist 2019 – 2022
Mfine – AI-first telehealth platform, 100K+ monthly consultations *Bangalore, India*

- Designed **AI-powered differential diagnosis system**: symptoms, vitals, and lab results → ranked differential diagnoses for clinician review; served **100K+ consultations/month**
- Built **contactless vital signs pipeline** – PPG signal extraction from phone camera video for heart rate measurement; enabled remote patient monitoring without specialized hardware
- Led team of 4 data scientists across computer vision and NLP workstreams with automated quality monitoring
- Built **skin disease recognition model** from consumer-grade smartphone images; published at IEEE ICPR 2020

Research Engineer Oct 2017 – 2019
Microsoft Bing *Hyderabad, India*

- Built search relevance and ranking models for Bing Local Search serving billions of daily queries
- Developed ML features for query-document matching and click-through rate prediction at web scale

Publications & Patents

7 peer-reviewed publications | 63 citations | h-index 5 | **2 US patents** (1 granted, 1 pending)

Patents

- US20220084677A1 – System and Method for Generating Differential Diagnosis in Healthcare
- Patent Pending – System and Method for Extracting PPG Signals and Body Vitals Prediction

Healthcare AI

- Transforming Rapid Diagnostic Tests into Trusted Diagnostic Tools Using AI in LMICs – **IEEE CAI '23**
- Task-Specific CV vs Large Multi-Modal Models for Diagnostic Test Interpretation – **VeriXiv '25**
- HealthPulse AI: Diagnostic Trust and Accessibility in Under-Resourced Settings – **VeriXiv '25**
- Dual Stream Network with Selective Optimization for Skin Disease Recognition – **IEEE ICPR '20**

Robotics & Vision

- MergeNet: A Deep Net Architecture for Small Obstacle Discovery – **IEEE ICRA '18**
- Small Obstacle Detection Using Stereo Vision for Autonomous Ground Vehicle – **ACM AIR '17**
- Mobile Robot Navigation Amidst Humans with Intents and Uncertainties – **IEEE CDC '15**

Technical Skills

ML/AI: LLM evaluation (LLM-as-judge, benchmark design), structured data extraction, RAG, prompt engineering, production guardrails, computer vision, on-device ML
Languages: Python, JavaScript/TypeScript, SQL
Infrastructure: PyTorch, HuggingFace, Claude/OpenAI APIs, LangChain, FastAPI, React, Docker, GCP, AWS

Education

MS + BTech, Computer Science 2012 – 2017
IIIT Hyderabad *Hyderabad, India*