RAKSHITH KUMAR K.N

📧 **rakshitkumarkn@gmail.com** | **LinkedIn/Rakshith** | **GitHub/Rakshith** | 📱 **+91 9008796644**

EXPERIENCE

AI Engineer **August AI** ***Aug 2024 – Present***

- Developed and deployed **AI-powered healthcare applications** including automated patient interview systems, increasing medical data collection efficiency by 65% and reducing manual documentation time by 4 hours per patient, resulting in improved clinical workflow optimization across 3 major healthcare facilities.
- Implemented multi-agent healthcare AI systems for patient triage, clinical report summarization, and automated follow-ups, reducing clinician administrative burden by **45%** and processing **2,500+ patient interactions daily**, leading to improved patient satisfaction scores by **23%**.
- Engineered NLP-driven medical documentation pipelines for automated report processing and structured data extraction, achieving **97% accuracy** across **10,000+ clinical documents monthly**, resulting in **$150K annual cost savings** through reduced manual review requirements.
- Optimized LLM memory management architecture for extended medical conversations, improving context retention and diagnostic accuracy by **75%** through advanced prompt engineering and memory optimization techniques.
- Created and validated medical reasoning models for automated patient interview systems, achieving **91% accuracy** in patient data capture across **5,000+ patient sessions**, resulting in **40% reduction** in initial consultation time and improved diagnostic consistency.
- Deployed HIPAA-compliant AI workloads serving **500+ concurrent users** on AWS and Azure infrastructure with **99.99% uptime**, implementing end-to-end encryption and role-based access controls, resulting in successful regulatory audits and enterprise client retention.

Machine Learning Engineer ***Snive (Krut AI)*** ***Jan 2024 – Aug 2024***

- Engineered end-to-end creative AI pipeline processing **50,000+ product images monthly** for automated photography and content generation using Stable Diffusion and LLMs, resulting in **80% reduction** in content creation costs for e-commerce clients.
- Implemented Fast UNet and VAE optimizations for diffusion models, reducing SDXL with ControlNet inference time to **1-2 seconds** while maintaining high-quality output through advanced model compression techniques.
- Deployed production-ready APIs handling **10,000+ requests daily** on AWS G4/G5 instances with intelligent queue-based resource allocation, achieving **99.9% uptime** and **$30K monthly infrastructure cost savings** through optimized resource utilization.
- Optimized diffusion model memory usage for production workloads, reducing GPU memory consumption by **60%** through gradient checkpointing and mixed-precision training, enabling cost-efficient scaling.

Associate Data Scientist ***Lincode Labs Inc.*** ***Mar 2022 – Jan 2024***

- Delivered computer vision solutions for **8 international manufacturing clients**, implementing OCR, image segmentation, and object detection models under challenging lighting conditions, resulting in **25% improvement** in production quality control across automotive and electronics sectors.
- Built high-accuracy OCR system for leading German electrical equipment manufacturer, achieving **92% accuracy** with **0.25s inference time** through custom preprocessing pipelines and GPU-optimized deployment strategies.
- Deployed real-time anomaly detection system processing **1,000+ surface inspections daily** for automotive manufacturers, achieving **0.5s inference time** with **96% detection accuracy**, resulting in **30% reduction** in defective products reaching market.
- Reduced API infrastructure costs by **90%** through server-sent events implementation for real-time inference container updates, eliminating redundant polling and improving system responsiveness.
- Enhanced production inspection pipelines serving **15+ manufacturing facilities** across US and global markets, achieving **1.15s batch inference** speed and processing **50,000+ components daily**, resulting in **$2M annual savings** through improved quality control efficiency.

Data Science Intern ***Lincode Labs Inc*** ***Sep 2021 – Mar 2022***

- Conducted R&D on object detection and segmentation models, improving model accuracy by **15%** and GPU memory efficiency by **30%** across **20+ experimental architectures**, contributing to 3 production deployments and establishing foundation for next- generation computer vision applications.
- Strengthened technical expertise in Python, PyTorch, TensorFlow, and data optimization frameworks, contributing to production- ready AI applications and establishing foundation for advanced computer vision development.

EDUCATION

MSc, Big Data Analytics
St. Joseph's University, Bengaluru (2020 – 2022)

BCA, Computer Applications
Seshadripuram College, Bengaluru (2016 – 2019)

PROJECTS

Water Quality Prediction (IEEE Published)

Predicted water quality index using **BPNN, SVR, LSTM**, applying **WAWQI method** on Ulsoor Lake dataset. Published in *IEEE ACAI 2022*.

Paper Link : Predicting the parameters of water quality and calculating the Water Quality Index of Ulsoor Lake, Bangalore, India using Deep Learning Techniques

Hand Gesture Recognition

Built touchless interaction system using **3D CNN + LSTM**, enabling real-time gesture-based digital control.

Project Link : Hand-gestures-regonition-using-3d-cnn-and-LSTM

KEY SKILLS

LLMs & Generative AI: OpenAI GPT-4/ChatGPT, Claude, Gemini, LLaMA, Mistral, Deepseek ,RAG, Fine-tuning, PEFT, LoRA, Prompt Engineering, Chain-of-Thought, Function Calling, Embeddings, MCP, A2A, Contextual Engineering,

AI Frameworks & Tools: LangChain, LlamaIndex, Hugging Face, Transformers, Ollama, AutoGen, CrewAI, Semantic Kernel, Vector DBs (Pinecone, Chroma, Weaviate, Qdrant), Stable Diffusion, UnSloth, LamaCPP, Olama

ML/DL & Programming: Python, PyTorch, TensorFlow, Scikit-learn, OpenCV, NumPy, Pandas, CUDA, TensorRT, Quantization, Pruning, Model Optimization

AI Agents & Orchestration: Multi-Agent Systems, Tool Use, ReAct, Planning Algorithms, Memory Systems, Agent Communication, Workflow Automation, LangGraph

Healthcare AI: Medical NLP, Clinical Decision Support, FHIR, HL7, HIPAA Compliance, Medical Imaging, Patient Data Processing, Regulatory AI (FDA)

Cloud & MLOps: AWS (SageMaker, Bedrock, Lambda, Cloudfront, Loadbalancer, Autoscaling, AppRunner, Cloudwatch, EKS, EC2), Azure (OpenAI Service, AI foundry, Container apps), GCP, Kubernetes, Docker, MLflow, Weights & Biases, CI/CD, A/B Testing, Modal (GPU Instances)

Data & APIs: Postgres, Clickhouse, Aurora RDS, FastAPI, REST/GraphQL, SQL, MongoDB, Redis, Elasticsearch, Data Pipelines, ETL, Real-time Processing, API Security , AWS WAF