Rahul Kumar

Sr. AI Engineer

Bangalore

+91 8789850819

rahulkr31.01@gmail.com

www.linkedin.com/in/kr-rahul

About Me

I am deeply intrigued by the concept of intelligence and the possibility of building systems that capture even fragments of it. My work spans both **healthcare and e-commerce AI**, from developing a generative AI-based onboarding assistant at GE HealthCare to building large-scale retrieval and personalization systems at Makro Pro. I specialize in **LLMs, RAG, and ranking systems**, having fine-tuned PEFT-based LLaMA models for real-world multilingual deployments. Earlier, I contributed to deep learning solutions for **cardiac risk stratification**, and designed secure, adaptive systems for **AI-powered network security**. With **7+ publications and 400+ citations**, I stay actively engaged with the research community—reading papers, attending conferences, and exploring side projects. I thrive in collaborative environments where ideas evolve into elegant, empathetic solutions. If you're looking for an AI engineer who pairs strong technical depth with a love for innovation, I'd be excited to connect and shape impactful systems—one neural network at a time.

Technical Skills

Programming Languages

Python

C++,JavaScript

Software

AI (LLMs, Embeeding Models, U-Net, Deep RL), Container Technology (Docker, Kubernetes, helm, Kubeflow), Python Backend

Education

2016 – 2020 **Indian Institute of Technology (IIT), Bhubaneswar** **9.28/10**
B.Tech (Hons.) Computer Science and Engineering

2014 – 2016 **MK Convent School, Delhi (CBSE)** **94.4/100**
All India Senior Secondary Certificate Examination (AISSCE)

Working Experience

Aug 2024 – Present **Sr. AI Engineer, Makro Pro**

- Led development of a multilingual two-stage reranking system (LLM-LightRanker) using PEFT-tuned LLaMA and LightGBM, achieving **0.95 ndcg@5**, while reducing exposure bias by **42%** and optimized search/carousel models delivering a **31% revenue uplift** and **56% increase in basket adds**.
- Built anomaly detection pipelines identifying **300+ customer** and **90+ product anomalies**,
- Trained GenAI models for Thai-English product content generation with **91% accuracy**, enabling automated onboarding for SKUs.

(Python, LightGBM, LLMs, MLFlow, PEFT, GenAI, Vertex AI, AWS Sagemaker)

Apr 2024 – Aug 2024 **Sr. AI Engineer, GE HealthCare** Bangalore, India

- Led a team of 2 to build a GenAI-based Digital Health Services (DHS) Onboarding Assistant, handling over **50%** of user queries.
- Designed and implemented a RAG + LLM stack integrated with internal documentation and knowledge bases. Fine-tuned **LLaMA-3** using **5k+ GPT-4 QA pairs**, achieving a **6x** improvement in response efficiency.
- Collaborated with the GenAI Discover team for full-scale integration into GEHC's enterprise chatbot.

(Python, RAG, LLM, Chatbot, MLOps, MLFlow, AWS [Sagemaker, EC2, API Gateway], Vector Database [Mongo Vector DB, Chroma])

Oct 2022 – Mar 2024 **Software Engineer, GE HealthCare** Bangalore, India

- Onboarded Next Gen CT smart subscription to the Edison platform, delivering **50% improvement** in DL Recon CT workflow throughput via GPU optimization.
- Completed **70% of onboarding tasks**, including solution integration, feature enhancement, and risk mitigation across platform modules.
- Delivered a customer-centric solution achieving a **9/10 Net Promoter Score** on onboarding satisfaction surveys.

(Kubernetes, OpenCV, GPU Optimization, Cloud-Native Security)

Oct 2020 – Sep 2022 **Edison Engineering Development Program (EEDP), GE HealthCare** Bangalore, India

- Developed a deep learning model for pericardium segmentation from non-contrast CT scans, achieving a **clinically acceptable dice score of 0.94**; the work was accepted as a POSTER at SPIE Medical Imaging 2024. Also assisted in DOE for a fat quantification model.
- Created an ultra-minimal Edison Edge platform reducing **50%** resource usage (RAM, VM, Disk, vCPUs) while retaining system scalability and serviceability.
- Designed and built ***4 POCs*** for a universal identity framework integrating runtime security, incident response, and forensics; laid the foundation for a zero-trust architecture and future service mesh integration.

(Python, U-Net, Image Segmentation, Zero-Trust Security)

Rahul Kumar

Sr. AI Engineer

Framework

Huggingface, Langchain, Pytorch, TensorFlow, MLflow

Domain

Deep Learning, NLP, Reinforcement Learning, RAG, Cyber-Security,Computer Vision (CV)

Hardware

Raspi-3, Arduino,
LoRa module, small Sensor devices

Soft Skills

- Communication
- Leadership
- Customer Focus
- Team Work & collaboration
- Planning & Prioritisation
- Strategy & Analysis

Relevant Courses

- Cloud Computing, Software Engineering
- Terraform Basics: Automate Provisioning of AWS EC2 Instances (Coursera) **(Certificate)**
- Data Science and Machine Learning Courses (Udemy, Coursera courses) **(Certificate)**

My Interest Bubble

Extra-Curricular activities

- Painting
  (3-year degree course)
- Football
  (Basic football course)

May 2019 - July 2019 — **Summer Intern, GE Healthcare** — Bangalore, India
Integrated LDAP/LDAPS server with the identity platform along with healthcheck and telemetry and created nearly **100-page documentation** for precise usage of the platform as a reference to the clients.
(Windows Server, SSL, X86 certificates, spinrg boot, WSO2 IDAM)

May 2017 - May 2018 — **Founding Engineer: Vasitars Pvt. Ltd. (Project Prajjawala)** — Bhubaneswar, India
Developed two generation prototypes for IoT based LPG Distribution System along with a five technical members and demonstrated the idea to the representatives from major Oil and Manufacturing Companies (OMCs) as potential customer and **raised 7 lakhs** initial funding.
(Python, Raspberry-pi, IoT, Data Analysis)

Major Publications

Published 8+ papers in various AI/ML, Cyber Security Journals having 400+ citations and 10000+ reads. Some of the pertinent ones are mentioned below:- (Research Profile)

Jul. 2025 **(Under Review)** — **LLM-LightRanker (LLM-LR): Early Results from Multi-Objective Reranking via LLM Distillation into LightGBM for E-Commerce.** **ACM RecSys '25:** R. Verma, M. Bhala, **Rahul Kumar**, et al. *Patent filed*

Jul. 2025 **(Under Review)** — **A Lightweight, Language-Aware Phonetic BK-Tree with Deep Distance Refinement for Real-Time Multilingual Query Correction in E-Commerce.** **ACM RecSys '25:** R. Verma, M. Bhala, R. Verma, M. Bhala,**Rahul Kumar**, et al. *Patent filed*

Feb 2023 **(Published)** — **Automated pericardium-segmentation using an attention-based convolutional neural network:** **SPIE Medical Imaging 2024:** A. Suntwal, **Rahul Kumar**, S. Kumar et. al.

Sep. 2021 **(Published)** — **Attention based multi-agent intrusion detection systems using reinforcement learning:** **JISA:** K. Sethi, Y. V. Madhav, **Rahul Kumar**, P. Bera

Dec. 2020 **(Published)** — **Robust Adaptive Cloud Intrusion Detection System Using Advanced Deep Reinforcement Learning:** **11th SPACE 2020:** K. Sethi, **Rahul Kumar**, D. Mohanty, P. Bera

Dec. 2019 **(Published)** — **Deep Reinforcement Learning based Intrusion Detection System for Cloud Infrastructure:** **COMSNETS 2020 Cybersecurity and Blockchain Workshop, 2020** K. Sethi; Rahul Kumar; N. Prajapati; P. Bera

Sep. 2018 **(Published)** — **A Context-Aware Robust Intrusion Detection System: A Reinforcement Learning Based Approach:** **IJIS:** K.Sethi, E. S. Rupesh, **Rahul Kumar**, P. Bera Y.V. Madhav

Awards and Achievements

Aug. 2022 — **Awarded** for collaboration efforts in Lotus NPI by Mohan B., Sr. Program Manager, GEHC MIC

Mar. 2020 — **Selected** in Youth India Delegate by **Min. of Youth Affairs and sports, Govt. Of India**

Dec. 2019 — **Awarded** Gold prize at **Inter-IIT Techmeet, IIT Roorkee, India**

Mar. 2019 — **2nd runner up** in SIH (Smart India Hackathon), MHRD, Govt. of India

Mar. 2019 — **Recipient** of **NRDC** Budding Innovators award 2019.

Dec. 2018 — **Winner** of Grand India IoT Innovation Challenge (GIIOTIC)

Dec. 2017 — **Received Honorary mention:** at ACM ICPC Onsite Contest amongst 250+ teams from over the country. **(Certificate)**

Volunteer Experience

June 2020 – Sept 202 — **MoSahay (NGO)** — Bhubaneswar, India
Developed a web portal (along with a 3 member team) as a part of social initiative to connect job seekers and employees, especially for migrants labourers. corporate portal *(MEAN Stack)*