Sarvesh Nand Kumar Khetan

College Park, MD | +1 (240)-726-8367 | skhetan@umd.edu | Linkedin (15k+) | Google Scholar | Website |
Book | Medium (70+ technical blogs featured at top publishers) | Github (1k+ commits, 20+ repos)

EDUCATION

University of Maryland, College Park, MD, USA **Aug 2024 - Expected May 2026**
Master of Science in Machine Learning; TA for Machine Learning (x2) *GPA : 3.95 / 4.0*

Birla Institute of Technology and Science (BITS) Pilani **Aug 2018 - Jun 2022**
*Bachelor of Technology in Mechanical Engineering (Minor : **Data Science**); TA for Linear Optimization* *GPA : 3.77 / 4.0*

TECHNICAL SKILLS

Languages and Databases : Python3, SQL, Snowflake, Pinecone, Neo4j, MySQL, PostgresSQL, MongoDb, C++, Java
Frameworks : PyTorch, Hugging Face, Langchain, LangGraph, LlamaIndex, PySpark, scikit-learn, NLTK, Spacy, Pandas, NumPy
Cloud : AWS Sagemaker, AWS Bedrock, AWS Glue, AWS Airflow, AWS EMR, AWS Step Function, AWS API Gateway
Dev Tools : VS Code, Git / Github, Docker, MLFlow, vLLM, UnSloth, W&B, Rest API (FastAPI, Flask), Jira, Airflow
IT Constructs : Large Language Models (LLMs), LLM Alignment (RLHF/DPO), Generative AI, Reinforcement Learning (RL), Deep Learning (DL), Natural Language Processing (NLP), Computer Vision (CV), Graph Neural Networks (GNN), MLOps, Machine Learning (ML), DSA, DBMS, OOPs, Distributed Training, Inference Optimization (Quantization/ Pruning/ Distillation)

WORK EXPERIENCE

AI Research Co-op (Early Stage Open Source Research Lab) | Sentient Labs, Remote, USA ***Nov 2025 - Present***

- Operated in a **startup-style research** setting, contributing to open source **recursive deep research** framework ROMA.
- Integrated **long-term agentic memory** system for Sentient Chat using DSPy, leveraging vector databases for persistent, context-aware conversational recall and real-time memory updating.

Data Scientist - AI Engineer | Piramal Capital & Housing Finance, India ***Jun 2022 - Aug 2024***

- Spearheaded a team of 5 to design a patented RAG-based **Text2SQL Graph AI agent**, leveraging **knowledge graphs**, **LLMs**, Graph Neural Networks (GNN) and Langchain, securing **$1M+** in management funding for Generative AI initiatives.
- Upgraded unimodal RAG system to a **multimodal RAG** using multimodal VLMs, improving retrieval relevance by **35%**.
- Built an Azure Blob **connector** for a RAG pipeline, enabling one-click ingestion of 1,000+ files across 10+ formats.
- **MLOps/ LLMOps:** Deployed open-source LLMs via **CI/CD** pipelines on AWS ECS using **Docker**, reducing build latency by 30%.
- Built DataMart pipelines using **PySpark** and **Airflow**, improving querying efficiency by **2x** and reducing processing time by **50%**.

Software Engineering - AI Research Intern | MicroStrategy, Virginia, USA ***June 2025 - Aug 2025***

- Optimized a multi-agent **ReAct-style Deep Research** system with **LangGraph** by integrating planners and external tools, enabling coordinated search and achieving a **40%** improvement in LLM response quality.
- Built a long-context **LLM evaluator** using LLM-as-a-Judge and checklist scoring, achieving **94%** agreement with human labels.

LLM Research Assistant | University of Maryland, USA ***Sep 2024 - Jan 2025***

- Trained a Graph Attention Network with **18%** higher accuracy versus GNNs for single cell classification on omics dataset.
- Used transfer learning, MoE, Lasso regularizer to improve AUC by 15% on EHR dataset to assess dementia risk factors.
- Finetuned AlexNet for defect classification in AM parts to achieve **95%+** accuracy, it got published in Scientific.Net.

RESEARCH WORK AND PROJECTS

DeepSeek-R1 SLM - Paper Implementation ***Feb 2025 - May 2025***

- Accelerated inference **3x** by replacing MHA / GQA with Multi Headed Latent Attention and utilizing RoPE.
- Added **Mixture of Experts (MOE)** with auxillary loss free **load balancing** and **KV Caching**, reducing inference latency by **35%**.
- Implemented RL finetuning with GRPO algorithm to enhance model reasoning capabilities without value function overhead.
- Developed multi-node GPU (4 A100s) data pipelines for large-scale LLM training using PyTorch Data Parallel (DDP).
- Achieved efficient inference by applying **quantization** and deploying with **VLLMs**, maintaining 90% of original model accuracy.

Computer Vision (CV) - Paper Implementations | *Python, PyTorch* | *Github* ***Jan 2024 - Present***

- Engineered MNIST augmentation pipeline advancing from VAE/GANs to **Diffusion Transformers** with **98.4%** accuracy gain.
- Improved multimodal image-text alignment by **27%** using Visual Language Model (VLM) architecture with **CLIP** encoder.
- Developed PyTorch DiT-based video generation model achieving **30%** faster inference than U-Net baselines.
- Implemented **YOLOv1** object detection paper from scratch and analyzed **DETR** for accuracy vs latency tradeoffs.
- Trained Deep-Q-Network (DQN) and PPO agents, optimizing decision-making via reward shaping and policy gradient methods.
- Engineered CNN from scratch and other foundation models like ResNet, MobileNet and **Vision Transformers (ViT)**.

ACHIEVEMENTS & EXTRACURRICULAR

- Received the **Piramal Delivery Excellence Award** for demonstrating out of box thinking & stellar professionalism.
- Selected for **ACM India** Summer School in NLP by IIIT Hyderabad and **Microsoft** with only **0.5% acceptance rate**.
- Accomplished **top 10** rankings in UCMAS (Mental Math) Competition for **two consecutive years**.
- Captained district cricket team while advancing through Art of Living's meditation disciplines and mindfulness practices.
- Volunteered 3-5 hours weekly, tutoring 10th-graders in computer science and math, improving grades by 20%.