Ravi Raj Kumar

+1-216-456-7060 Email LinkedIn GitHub Cleveland, Ohio

EDUCATION AND HONORS

Case Western Reserve University Cleveland, Ohio, Expected May 2025
Master of Science in Computer Science

- Coursework: Machine Learning, Computer Vision, Robotics, High Performant Systems for AI, Probabilistic Graphical Models, Analysis of Algorithms, Computer Networks.
- Observer: Statistical Natural Language Processing, Deep Gen Models, Quantum Computing, Reinforcement Learning, ML on Graphs.

TECHNICAL SKILLS

- **Languages & Frameworks:** Python, SQL, Java, C++, JavaScript, TensorFlow, PyTorch, HuggingFace Transformers, FastAPI, Django, Streamlit, Gradio, LangChain, LangGraph
- **Machine Learning & RL:** Classification, Regression, Clustering, SVM, Random Forest, CNN, RNN, Transformers, PPO, A2C, DQN, SAC, Multi-Agent RL
- **Generative AI:** Diffusion Models , GANs, VAEs, Fine-Tuning (LoRA, QLoRA, PEFT), Image Generation, Inpainting, Upscaling
- **RAG & LLM:** FAISS, Pinecone, VectorDBs, LangChain Agents, Embedding Models, Prompt Engineering, Top-K Retrieval, Document Chunking
- **MLOps & Infrastructure:** MLflow, Docker, Kubernetes, Helm, AWS (EC2, S3, ECR), GCP, Azure, CI/CD (GitHub Actions, GitLab, Jenkins), Monitoring (Prometheus, Grafana), ONNX Runtime, Quantization, Scalable Model Serving

PROFESSIONAL EXPERIENCE

Tata Consultancy Services Hyderabad, India [October 2019 - November 2023]
Machine Learning Engineer (NLP Model Development & MLOps Integration - Banking and Finance domain)

- Built robust and scalable end-to-end ML pipelines for a Bank Member Complaint Distribution System on cloud as-well-as on-prem with components like data ingestion, data validation, feature engineering, model training, prediction, and monitoring.
- Implemented data ingestion and data validation components in the pipelines for large-scale data sources like Hadoop, Snowflake, and MongoDB and validated the output artifacts for robustness.
- Leveraged advanced NLP tokenizers, such as **BytePair Encoding (BPE)** and **SentencePiece** for tokenization, trained and finetuned several transformer-based models like **BERT, RoBERTa, and Longformer** on tasks such as Complaint Categorization and Prioritization, Named Entity Recognition (NER) for automated information extraction, and Complaint Severity Prediction.
- Built **CI/CD** pipelines with **Jenkins** for automated ML workflows and integrated **MLflow** for experiment tracking, versioning, and model registry; deployed models using **Docker** and **Kubernetes** with Helm for streamlined, version-controlled production deployment.
- Established artifact versioning and environment reproducibility using **MLflow Models**, **Docker Image tags**, and **Conda environments**, enabling rollback-safe deployments and auditability.
- Implemented automated model validation and **drift detection** to ensure reliability, and leveraged **Prometheus** and **Grafana** with custom metrics and alerts for proactive performance and infrastructure monitoring, and efficient issue detection in production.

Data Engineer (Bank Member Complaints)

- Catalogued and registered data assets to enforce security policies and enable seamless migration to the DL2 cloud zone. Designed both General-Purpose and Subject-Area marts, and implemented a Data Quality Engine to validate extracts before loading.
- Leveraged the Hera framework, UNIX shell scripting, and advanced SQL in Hive, Snowflake, and DBT to streamline data ingestion and transformation workflows.
- Developed scripts and jobs to apply complex business logic and route outputs to UNIX filesystems, on-premises databases (Netezza, Db2, Hive), and AWS Snowflake—ensuring reliable, end-to-end data delivery.

RESEARCH EXPERIENCE

OPTIMIZING RAG WITH MULTI AGENT REINFORCEMENT LEARNING

Supervising Professor: : Dr. Soumya Ray

- Designed a Multi-Agent Reinforcement Learning framework for RAG, modeling query design, document retrieval, and answer generation as cooperative agents jointly optimized via PPO under a unified F1-based reward signal.
- Implemented FAISS search with sentence-transformers/all-MiniLM-L6-v2 embeddings to enable efficient top-K passage retrieval over a custom SQuAD corpus, all within a reproducible Conda environment.
- Implemented Warm-start for each agent by employing PEFT's LoRA by freezing 96.65% of weights, to fine-tuning LoRA adapters on 5,000 SQuAD QA pairs before any RL, which improved sample efficiency and stabilized the subsequent PPO loop.
- Implemented a PPO loop using TRL's PPOTrainer that fine-tunes only the LoRA adapters and value head by iterating query rewrite, retrieve, generate with a unified reward signal, significantly improving QA performance over SFT.

MULTIMODAL TRANSFORMER FOR IMAGE-CONDITIONED TEXT GENERATION [Github Link]

Independent Research Project

- Designed and implemented a multimodal transformer integrating a SigLIP-style Vision Transformer with a Gemma-based causal decoder for image-grounded generation tasks such as captioning and visual question answering.
- Engineered autoregressive decoding with KV caching and Rotary Positional Embeddings, reducing inference latency by **40%** per token on long sequences.
- Built a PaLI-inspired tokenizer pipeline with image-token prefixing and robust image normalization, improving input alignment and reducing token mismatch errors by **23%**.
- Enabled diverse generation using temperature-scaled top-p sampling; achieved a **+3.7 BLEU-4 score improvement** over greedy decoding on benchmark prompts.

Latent Diffusion Transformer for Text-Conditioned Image Synthesis [GitHub]
Implemented from scratch using PyTorch, inspired by Stable Diffusion architecture

- Developed a complete multimodal pipeline integrating CLIP-based language encoder, variational autoencoder (VAE), and U-Net-based denoising diffusion model for high-fidelity text-to-image generation.
- Achieved **4× faster inference** over pixel-space models by operating in latent space (64×64 vs. 512×512), reducing memory and compute by over **85%**.
- Implemented **Classifier-Free Guidance** to enhance prompt alignment, improving semantic accuracy of generated images by **23%** (measured via CLIPScore).
- Reproduced Stable Diffusion v1.5 results with **FID: 12.4** and **CLIPScore: 0.32** on the COCO validation set, matching baseline model accuracy.
- Enabled multimodal capabilities: **text-to-image**, **image-to-image translation**, and **inpainting**, through unified conditional diffusion framework.
- Built and optimized 12-layer Transformer encoder for text representation using PyTorch, achieving token encoding throughput of **4000+ tokens/sec** on NVIDIA A100.
- Used custom DDPM scheduler with dynamic timestep control for guided denoising, reducing inference steps from 100 to **50 steps with no perceptual loss**.

DEEP GENERATIVE MODELS TO ENHANCE SEMI SUPERVISED LEARNING [Github Link]
Supervising Professor: : Dr. Soumya Ray

- Conducted an in-depth survey of semi-supervised learning with deep generative models and implemented Kingma et al.'s M2 approach on MNIST and CIFAR-10. Matched paper results: 94.8% test accuracy on MNIST (1k labels) and 63.1% on CIFAR-10 (4k labels), establishing a reliable baseline.
- Conducted a rigorous mathematical and experimental analysis of the Evidence Lower Bound (ELBO) within a variational inference framework, identifying and resolving three critical issues in the paper: entropy penalization for sharper decision boundaries; mutual information maximization to strengthen input–label coupling; and smoothed-label integration into the classification loss.
- Delivered a 4% accuracy improvement over the original paper's baseline on MNIST and a 2.5% gain on CIFAR-10 and 15% reduction in classifier entropy.

3D POINT CLOUD SEGMENTATION USING 2D IMAGE SEGMENTATION [Github Link]
Supervising Professor: : Dr. Yu Yin

- Developed a novel 3D point cloud segmentation framework leveraging state-of-the-art 2D image segmentation models (OneFormer) and a voting-based approach to project 2D semantic and panoptic labels onto 3D point clouds, achieving real-time segmentation with reduced computational overhead.
- Utilized RGB images, depth maps, and LiDAR data captured with the iPhone 13 Pro, integrating segmentation masks generated by OneFormer with a voting mechanism to accurately transfer semantic labels to 3D point clouds.
- Achieved segmentation accuracy of 96.5%, matching PointFormer, while significantly reducing computational overhead and memory usage demonstrating the model's scalability and efficiency.

PROJECTS

MULTI-AGENT MEDICAL APPOINTMENT SYSTEM [Github Link]

- **Built a modular multi-agent AI system using LangChain and LangGraph**, implementing a supervisor-agent architecture to route user queries to specialized agents for doctor information and appointment management.
- **Engineered a full-stack solution with FastAPI for scalable backend services and Streamlit for real-time, interactive user interfaces**, enabling seamless doctor appointment management via natural language queries.
- **mproved query-to-response turnaround time by $\tilde{6}0\%$** through automated slot filtering and decision routing, reducing manual filtering from $\tilde{3}0$ seconds to under 12 seconds across 4,000+ simulated booking records.

RAG-POWERED CUSTOMER SUPPORT AGENT [Github Link]

- Built an end-to-end ETL pipeline with Pandas to parse Flipkart reviews into LangChain documents, then ingested them into AstraDB Vector Store for high-performance, semantic top-k retrieval.
- Integrated Google Gemini-1.5-Pro embeddings and LangChain's ChatPromptTemplate to orchestrate context-aware retrieval and natural-language answer generation for real-time product recommendations.
- Developed a RESTful FastAPI backend with environment-driven configuration with PyYAML, python-dotenv and integrated secure secret management, paired with a modular AJAX chat UI using Jinja2, Bootstrap 4 and jQuery.