Kasuru Kodanda Rama Raju

Jersey City,NJ | kkasuru@stevens.edu |+1(551)228-9910 |linkedin.com/in/kasuru/ |github.com/Ramkasuru

EDUCATION

Stevens Institute Of Technology — **New Jersey, US**
Master of Science in Applied AI — *Graduation Date: Dec 2026*

Dayananda Sagar University — **Bangalore, IN**
Bachelor of Technology CSE (AI & ML) — *Graduation Date: May 2024*

SKILLS

- **Machine Learning & AI**: Large Language Models (LLMs), Natural Language Processing (NLP), Computer Vision, Retrieval-Augmented Generation (RAG), Agentic AI, Tool-Use Pipelines, Representation Learning, Causal Inference, Optimization, Model Evaluation, Reasoning Systems, Reinforcement Learning
- **Programming & Frameworks:** Python, C++, SQL, CUDA, NumPy Systems & ML Engineering: Linux, Git, Docker, PyTorch, TensorFlow, Hugging Face, scikit-learn
- **Research & Data Science:** Statistical Analysis, Experimental Design, Data Processing, Hypothesis Testing, Reproducibility Pipelines

WORK EXPERIENCE

Outlier AI — **Remote**
Prompt Engineer — *Jul 2024 - Nov 2024*
- Improved LLM evaluation accuracy by 12% across 10k+ prompts using reinforcement-based prompt tuning and experimental refinement.
- Developed automated benchmarking pipelines that reduced evaluation latency by 18% and ensured reproducibility across 5+ LLM families.
- Delivered enterprise-grade evaluation suites (Goldfish V2, Gemini Nexus), directly influencing modeling and product roadmap decisions.

Ziberr Communications — **Bangalore, IN**
Product Operations & AI Solutions Intern — *Jan 2024 - Apr 2024*
- Built scalable Shopify–Facebook ML automation systems used by 200+ clients, improving engagement by 40%.
- Designed and deployed APIs and automated data pipelines, increasing operational efficiency by 28%.
- Improved data consistency and reliability through integration debugging, monitoring, and version-controlled workflows.

RESEARCH EXPERIENCE

Master's Thesis — *Aug 2025 - Current*
From Many Small to One Large: Aggregation of SLMs
- Conducting comparative analysis of SLMs and LLMs on reasoning and QA benchmarks (MMLU-CF, HumanEval).
- Implementing RAG pipelines to evaluate trade-offs between accuracy, compute efficiency, and scalability.
- Building a multi-metric benchmarking suite (latency, FLOPs, robustness, memory) to compare real-world deployment efficiency.

Stevens Institute of Technology
Summer Research — *May 2025 - Jul 2025*
- Built and optimized CUDA-based matrix multiplication and reduction kernels for AI performance benchmarking on NVIDIA H100 GPUs.
- Compared GPU vs. CPU metrics using event timing, Nsight profiling, and reliability analysis to improve reproducibility.
- Explored LangChain pipelines and fine-tuned LLMs for reasoning and summarization tasks within applied AI safety research.

Viola
Multimodal AI Assistant — *Aug 2023 - Feb 2024*
- Built an end-to-end multimodal assistant integrating YOLOv8 (CV), Whisper (ASR), and GPT-4 (NLP).
- Achieved 90% object detection and 98% ASR accuracy through fine-tuning and multimodal model fusion.
- Deployed and tested across 5+ users with disabilities; awarded Best AI for Social Impact at Tech Spark (DSU).