Sreeram Raghammudi

sr4314 [at] columbia [dot] edu | LinkedIn | GitHub | Website

Education

Columbia University — **New York, NY**
Master of Science in Computer Science - Machine Learning Track (GPA: 4.16/4.00) — *Aug 2025 – Dec 2026*

Birla Institute of Technology and Science (BITS), Pilani — **Pilani, India**
Bachelor of Engineering in Computer Science (GPA: 3.81/4.00) — *Aug 2021 – Jul 2025*

Experience

ODeX Global — **Dubai, UAE**
AI/ML and Backend Development Intern — *Jan 2025 – Jul 2025*

- End-to-end trained YOLOv8 and LayoutLMv3 models for automated shipping document ingestion, cutting manual effort by 85%+
- Built an end-to-end LLM pipeline with multi-stage OCR and coordinate validation for robust PDF data extraction.
- Built fuzzy-matching and embedding-based systems to automate document segregation and align inconsistent Excel schemas, eliminating manual overhead

Nurish Digital Inc. — **New York Metropolitan Area, NY**
AI, Research and Backend Development Intern — *Jun 2024 – Sep 2024*

- Designed and deployed OpenAI-based nutrition and hydration pipelines in Python with production services in Node.js, integrating RAG to personalize outputs using user dietary history and preferences
- Improved model accuracy by 22% and reduced latency by 4.3× through parallel inference and user data analytics

Wipro Limited — **Hyderabad, India**
AI Project Intern — *Jun 2023 – Aug 2023*

- Implemented Vertex AI PaLM 2 (text) and Codey across multiple applications; built RAG-based Agent Assist improving support workflow efficiency by 80%.

Projects

Multi-modal Scientific Question Answering — *Repo + Writeup*

- Engineered a hybrid SQA pipeline integrating BLIP-2/ViLT with instruction-tuned LLMs and chain-of-thought prompting to enable structured reasoning across visual and textual modalities.
- Conducted extensive ablation studies on early vs. late modality fusion and rationale supervision, identifying the critical role of image-conditioned representations for complex science questions.
- Achieved a 84% through rationale-aware decoding and optimized modality fusion strategies.

Parameter-Efficient Fine-Tuning of Transformer Models — *Repo + Writeup*

- Benchmarked LoRA, QLoRA, and full fine-tuning under strict GPU limits, building low-bit quantization pipelines for large-scale training on constrained hardware.
- Showed LoRA cuts trainable parameters by 95%+ while preserving accuracy; conducted system-level profiling to map memory bottlenecks and efficiency–accuracy tradeoffs.

Research Experience

AI for Multimodal Document and Forensic Analysis (*Columbia University – GILMLab, Ongoing*)

- Developing multimodal transformer & graph-based methods to associate degraded textual, and geospatial records.

Multi-modal Sarcasm Detection (*Springer Nature Q1; 1st Author, Under Revision)*

- Engineered text, audio, and video fusion models using RoBERTa and Bi-LSTMs, 0.81 F1 score

Market Connectedness and Time series Modeling *(World Finance Conference, 2nd Author; Accepted)*

- Established market connectedness using Diebold-Yilmaz, cross-quantilogram frameworks and Hybrid CNN-LSTM architectures, improving forecasting accuracy by 40%

Technical Skills

Machine Learning & Deep Learning: PyTorch, TensorFlow, scikit-learn, Hugging Face, LangChain
MLOps & Cloud: Git, Docker, Google Cloud Platform, Azure
Data & Deployment Tools: Pandas, Numpy, Flask, FastAPI
Languages & Frameworks: Python, C/C++, Java, SQL, JavaScript, React.js, Node.js