Raghav Donakanti

US Citizen – Authorized to work in the US

raghavdonakanti@gmail.com LinkedIn github.com/Raghav010 google scholar

Education

International Institute of Information Technology - Hyderabad **2021 – 2025**

B.S. in Computer Science (Honours) *GPA : 3.41/4*

- ***Teaching Assistant*** - Involved in assignments, tutorials, and grading for courses with 150+ students **CS7.501 Advanced NLP, CS6.201 Introduction to Software Systems, CS4.301 Data and Applications**
- ***E-Cell - Head of the Events and Logistics Team*** - Hosted Hyderabad's largest hackathon, Megathon, managed schedules and cross-functional team communication, and hosted inspiring keynote speakers on overcoming failure.

Experience

Adobe **05/2024 – 07/2024**

Research Intern *LLM, Multi-Agent Architectures (AutoGen)*

- Improved agentic capabilities of LLMs for multi-step planning and API-based action orchestration at Big Data Experience Labs by designing robust multi-agent workflows
- Developed a dynamic topology-based message-passing architecture, enhancing coordination and stability of agents

Precog | LINK **2023 – 2025**

Researcher *NLP, Agentic AI, LLMs for the Edge, Responsible AI*

- ***Reimagining Self-Adaptation in the Age of Large Language Models*** | ***Published, ICSA 2024*** | LINK
  Co-developed the MSE-K framework leveraging LLMs in self-adaptive systems, improving system adaptability and resilience by enabling context-sensitive, autonomous software adaptation under variable workloads.
- ***Small Models, Big Tasks: An Exploratory Empirical Study on Small Language Models for Function Calling*** | ***Published, EASE 2025*** | LINK
  Led empirical study on Small Language Models for function calling, introducing novel semantic/syntactic evaluation metrics and optimizing models for low-latency, low-memory edge deployments.

Smart City Living Lab | **01/2023 – 05/2023**

Software Engineer Intern *Agile SDLC, ReactJS, IoT, Digital Twin*

- Built an extensible Digital Twin system for a real-world IoT water network using ReactJS and oneM2M live sensor data, enabling proactive distribution planning and fault detection.
- Scoped project requirements and co-led design decisions in collaboration with stakeholders under agile SDLC practices.

Projects

Medmini - On-Device LLM Medical Q&A System | View Here | *LangChain, Docker, RAG, Quantization, Vector DB*

- Engineered a lightweight LLM-based medical data answering system for edge devices to be deployed in underserved regions, requiring $<$ 3GB RAM, 0 vRAM, and achieving sub-3-second inference times.

Spec ++ - Accelerated Speculative Decoding | *Inference Optimization, Speculative Decoding*

- Improved upon speculative decoding, achieving a 50% inference speed up over baseline by leveraging a Mixture of Experts (MoE) like approach, in effect increasing acceptance rates.
- Developed a custom attention mask to validate multiple SLM outputs in parallel without compromising speed, despite the increased validation workload.

UpDawg - Intelligent Personal Assistant | View Here | *Software Architecture Patterns, C4 Model, RabbitMQ*

- Collaborated and led a team to develop UpDawg, a personal assistant that aggregates and summarizes user data from services like Slack and Outlook, tailored to user preferences using LLMs and RAG techniques.
- Architected the system using a monolithic and event-driven model, and conducted comparative analysis across tactics to address performance/modifiability tradeoffs.

ASAC - Assisted Spatial Audio Construction | View Here | *Optical Flow, Object Detection and Tracking*

- Worked on automatically generating spatial audio (like Dolby Atmos) for scenes in movies by learning projections of objects and speakers in 3D space via video understanding, depth estimation and cross modal alignment techniques.
- Designed a system that handles complex multi-speaker environments, identity switching and occlusions by leveraging lightweight tracking and facial recognition methods.

IsThatTrue? | View Here | *PyTorch, Foundational Models, Knowledge Grounding*

- Developed a robust Multimodal Factual Verification engine over all of wikipedia (english) using Retrieval Augmented Encoding. Utilized the BridgeTower model and improved upon the ACL work: FEVEROUS.

Skills

Programming Languages/Frameworks: Python, C, C++, Bash, SQL, MongoDB, MERN, Flask, Docker, Javascript
Machine Learning / Data Analysis: PyTorch, Pandas, Seaborn, Matplotlib, Numpy, scikit-learn