

SRAVANTH KODAVANTI

☎ +91-9550173710 ✉ ramasravanthkodavanti@gmail.com [in sravanthk27](https://www.linkedin.com/in/sravanthk27) [Google Scholar](#) [Website](#)

Education

Indian Institute of Technology, Hyderabad

2020 - 2024

Bachelor of Technology, **Major** - Computer Science and Engineering, **Minor** - Entrepreneurship

CGPA - 8.58

Research Publications

1. **Sravanth Kodavanti***, Sowmya Vajrala*, Srinivas Miriyala*, Utsav Tiwari*, *et al.*,
Unlocking the Edge Deployment and On-Device Acceleration of Multi-LoRA Enabled One-for-All Foundational LLM, *Under Review at EMNLP 2025*
2. **Sravanth Kodavanti***, Srinivas Miriyala*, Sowmya Vajrala*, Vikram N R,
On Distillation of Transformers into State-Space Models for Efficient Image Restoration,
Under Review at NeurIPS 2025
3. Srinivas Miriyala*, Sowmya Vajrala*, Hitesh Kumar, **Sravanth Kodavanti**, Vikram N R,
Mobile-friendly Image De-noising: Hardware-Conscious Optimization for Edge Application,
Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), 2025

Patents

1. **Sravanth Kodavanti***, Srinivas Miriyala*, Sowmya Vajrala*,
System and Method for Accelerating On-Device Large Language Model Inference.
Provisional Specification Filed. Proposes a novel self-speculative decoding technique to significantly improve efficiency and reduce latency in LLM inference on edge devices.

Work Experience

Samsung Research Institute Bangalore

Aug 2024 - Present

Machine Learning Research Engineer

- Developing **Samsung Neural Acceleration Platform** for AI model acceleration and deployment on mobile devices, leveraging **Neural Architecture Search (NAS)** and **Quantization** for performance optimization.
- **Commercialization & Impact:**
 1. **Accelerated on-device inference** for the **Samsung LLM Gauss - L, 3B Model** by implementing **Speculative Decoding**, achieving a **2X improvement** in tokens per second (toks/sec). Successfully integrated into the **Samsung S25 flagship series**.
 2. **Optimized low-light video de-noising model** using **NAS, Quantization**, achieving a **2.5× speedup in inference**. Successfully deployed in the **Samsung A56 device**.
- **Awards & Recognition:**
 1. **Spot Award (Q2 2025)** – For novel speculative decoding for LLM acceleration.
 2. **MD Project Incentive Award (2024–25)** – For AI model optimizations and deployment on edge devices.
 3. **Team Awesome Award (Q4 2024)** – For on-device optimization of **Samsung Gauss - L** in **S25** mobile series.

Stealth Startup - Gika.AI

June 2024 - Sep 2024

AI Researcher

- Worked on a thesis for proving **Knowledge Graphs (KG)** are better than **Vector Databases** in **Retrieval Augmented Generation (RAG)**.
- Worked on **Coreference & Entity** Resolution of the documents, which is the data used for finetuning the model. Used many LLMs such as **GPT-4,4o, LLAMA3, SpanBert & LingMess** by **Spacy** AI agent for the task.

Hexagon R & D India

Jan 2024 - June 2024

Machine Learning Intern

- Developed a website for implementing segmentation & classification based on tags of various manufacturing plant sketches. Used various segmentation algorithms such as **RANSAC, DBSCAN, K-Means**.
- Used **Azure form recognizer** models & other OCR models for tag identification.

- Developed a **flask website** for text recognition on manufacturing plant sketches using **Form, Doc Recognizer** & also used OCR models for tag identification.
- Involved in an **LLM** research project. Compared the results for text generation between **Mixtral, Mistral, LLAMA2**.

OnePlus / Oppo (OPLUS) Mobiles India R & D

Jan 2023 - June 2023

Research Intern - Device AI

- Implemented model compression techniques called **Quantization , Pruning & Distillation** on various deep learning models such as **ResNet , Yolo , ViTs , ConvNeXt , Stable Diffusion Models & Large Language Models (LLAMA)** . **Compressed all these models for deploying in mobile devices.**
- Involved in research work on **Neural Style Transfer (NST)**.
- Implemented a DL model for **LaTeX - OCR** task . The model's aim is to detect & recognize the mathematical equations present in a research paper. Compressed the model for the integration with edge devices.

Department of Computer Science , IIT Hyderabad

Nov 2021 - Apr 2024

Teaching Assistant

- I have worked as a **TA** for the courses **Operating Systems (CS3510)** under *Prof. Sathya Peri* , **Discrete Maths (CS1010)** under *Prof. Rakesh Venkat* , **DBMS (CS3550)** under *Prof. Manish Singh* & **Theory of Computation (CS2030)** under *Prof. Subramanyam Kalyanasundaram*.

Academic Service

Research Reviewer

2025 – Present

Top ML/NLP Conferences and Workshops

- Reviewed submissions for **ICML, NeurIPS**, and **ACL** workshops.
- Provided constructive feedback to maintain academic rigor and contribute to the research community.

Projects

Continual Learning for 3D Point Cloud | *Prof. P.K.Srijith*

- Implemented Continual Learning on **Pointnet** architecture by addressing the challenge of catastrophic forgetting.
- Implemented **Knowledge Distillation** approach for the Continual Learning & Model was trained on **ModelNet10** dataset.

Computer Vision & NLP | Personal Projects

- Image denoising using **Auto - Encoders** . Model trained on **MNIST** dataset.
- Human Face generation using **GANs** . Model trained on **celeba** dataset.
- Underwater Object Detection and Classification using **DGYOLO**. Model trained on **URPC 2019** dataset.

Technical Skills

Languages: C , C++ , Python , SQL , HTML , CSS

Frameworks & Tools : PyTorch, ONNX, Hugging Face, Git, LaTeX, Markdown

Operating Systems : Linux , Windows

Familiar: Tensorflow , Tensorrt , JavaScript

Achievements

- Secured AIR **1156** in Open Category & AIR **91** in EWS Category in IIT JEE Advanced 2020.
- Secured AIR **493** in Open Category & AIR **45** in EWS Category in JEE Main B Planning 2020.
- Secured **99.75** percentile in IIT JEE Main 2020.
- AP EAMCET 2020 Rank **135**.
- TS EAMCET 2020 Rank **307**.
- Solved around 700+ CP problems over multiple platforms [LeetCode](#) , [CodeChef](#) , [CodeForces](#) , [GeeksForGeeks](#).

Leadership / Extracurricular

- Worked as a Core member for **Epoch : AI-ML club of IITH**.
- Worked as an **Internship & Placement** Coordinator for Office of Career Service of IITH.
- Member of IITH **Chess** team , Represented IITH at **InterIIT** & various competitions.