JINENDRA MALEKAR

504 S. Beltline Blvd. Apt E22 | jinu98.github.io | 803-633-5670 | jmalekar@email.sc.edu

EDUCATION

University of South Carolina — *Columbia, SC USA*
Ph.D. in Computer Engineering — **January 2024 - Present**
College of Engineering and Computing

University of South Carolina — *Columbia, SC USA*
MSc. in Computer Science — **May 2023 – Present**
College of Engineering and Computing

International Institute of Informational Technology — *Naya Raipur, India*
Bachelor of Technology in Computer Science — **August 2017 – May 2021**

EXPERIENCE

Research Assistant — *Columbia, SC*
ICAS lab, University of South Carolina — **May 2024 – Present**

- **Optimization on Raspberry pi for LLM's**
  - Deployed up to 8-Billion parameters LLM's on raspberry pi for real-time conversational AI system (> 3 tokens/sec).
  - Did Post Training Quantization (2-bit, 4-bit, 6-bit and 8-bit) and used Quantization Aware Trained (ternary weights) Models for achieving real-time performance.
- **Heterogeneous Architectures for LLM Acceleration**
  - Developed heterogeneous architectures combining TPU and PIM architecture for LLM acceleration.
  - Showed PIM-LLM, a hybrid analog-digital architecture that achieves considerable performance and throughput gains compared to conventional LLM accelerators.
- **Research on 1-Bit LLM's**
  - Evaluated how partial enhancements in 1-bit LLMs impact overall model performance.
  - Found that due to the model's computational and memory demands, focusing on custom hardware development to maximize the impact of extreme quantization is worthwhile.

.

Research Assistant — *Columbia, SC*
AI Institute, University of South Carolina — **August 2023 – May 2024**

- **WellDunn**
  - Provided two benchmark datasets for mental health and pointed out the insufficiency of the language model for medical applications.
- **CPR**
  - Created a web-based real-time application for monitoring and annotating high-risk suicidality-related posts on Reddit.
  - Leveraged language models to diagnose depression and provide clinical-friendly explanations for mental health use cases.

Research Intern — *Columbia, SC*
AI Institute, University of South Carolina — ***August* 2022 – August 2023**

- **MDiabates App**
  - Developed a comprehensive mobile app for Prisma Health to aid in monitoring type-1 diabetic patients.
  - Designed an intuitive interface enabling patients to effectively manage insulin levels by tracking daily carbohydrate intake, to support patients.
- **Disaster Record**
  - Utilized real-time Twitter data to understand better people's need for shelter, food, and healthcare to enhance disaster response.
  - Built an interactive dashboard that improved shelter and medical assistance access during calamities utilizing real-time Twitter data.

- **BMW Dashboard**
  - Collaborated with BMW Greenville/Fraunhofer plant's material planners to create a customized dashboard for inventory management.
  - Led the creation of the dashboard's front end using React.js and implemented a robust Python-based backend.

Software Developer *Delhi, India*
Krya Solutions private limited **August 2021 – August 2022**

- Handling consumer front end of the app named rippl based on nextjs framework performing bug fixes and code reviews.
- Successfully delivered new features, functionality, and capabilities to the website.
- Created a complete front-end website with react.js in collaboration with the company designer head to help improve the product reach to the consumer. https://inspire-seven.vercel.app/

Software Developer Intern Bangalore, India
MIMYK Medical Simulations Pvt. Ltd. **January 2021 – July 2021**

- Contributed to the development of the Endoarchive project, a Virtual Reality (VR) and haptics-enabled simulation platform for endoscopy procedures
- Worked on a full stack application which uses React.js, Django, and Material-UI framework to assist with the application.

AWARDS & RECOGNITIONS

- **Awards**
  - 1st Place in the University Demo at DAC 2025 (Design Automation Conference).

SELECTED PUBLICATIONS

- Mahsa Ardakani, **Jinendra Malekar**, Ramtin Zand. "LLMPi: Optimizing LLMs for High-Throughput on Raspberry Pi" 2025 CVPR Workshops
- **J Malekar**, P Chandarana, MH Amin, Elbtity, Mohammed E and Ramtin Zand. "PIM-LLM: A High-Throughput Hybrid PIM Architecture for 1-bit LLMs." *Pending Revision*
- **Jinendra Malekar**, Mohammed E. Elbtity, Ramtin Zand. "Matmul or No Matmul in the Era of 1-bit LLMs" *Pending Revision*

Full list is here: https://scholar.google.com/citations?user=4vzLrzUAAAAJ&hl=en&authuser=1

SKILLS

Technical: Python | C/C++ | Pytorch | Tensorflow | React.js | Node.js | Flask | Django | Firebase
Frameworks: MongoDB | AWS | Git | Nginx | Postman | RESTful API | FAST API