AFSARA BENAZIR

hys4qm@virginia.edu | 434-956-0446 | website | linked-in

EDUCATION

University of Virginia, School of Engineering and Applied Science — Started in Aug 2022
PhD. In Computer Science — Expected Graduation: Aug 2027
Supervisor: Dr. Felix Xiaozhu Lin

Bangladesh University of Engineering and Technology (BUET) — Mar 2016-Feb 2021
BSc. In Computer Science and Engineering

RESEARCH AREA

Efficient ML, On-Device AI, Systems Optimization, Speech, Privacy, GPU performance analysis, Apple Silicon

WORK EXPERIENCE

Department of Computer Science, UVa — August 2022-Present
Graduate Research Assistant — *Charlottesville, VA*

- Exploring efficient Mixture-of-Experts (MoE) architectures for on-device LLM inference; investigating scalable expert routing, quantization-aware optimizations, and dynamic execution strategies for NPUs and GPUs.
- Developed a resource-efficient framework for on-device speech understanding, leveraging cache and temporal locality with deep models and cloud offloading to enable real-time speech understanding on tiny devices.
- Advanced privacy-preserving automatic speech recognition with OpenAI Whisper through on-device execution, LoRA finetuning and foundation model adaptation to safeguard sensitive information while maintaining usability.
- Conducted large-scale CPU/GPU benchmarking of foundation models such as LLaMA, DeepSeek across 26+ quantization schemes using *llama.cpp*; profiling latency/throughput tradeoffs and uncovering hardware bottlenecks across heterogeneous hardware (Apple Silicon vs CUDA GPUs).

Graduate Teaching Assistant

- Led 4 semesters of core CS courses (Operating Systems, NLP, Signal Processing & ML)

Systems, Solutions and Development Technologies (SSD-Tech) — March 2021-July 2022
Engineer, Technology — *Dhaka, Bangladesh*

- Developed client-side features for an e-commerce website using Laravel Framework (PHP)

PUBLICATIONS

- **[SIGMETRICS’26] Benchmarking and Characterization of Large Language Model Inference on Apple Silicon** [PDF]
  Afsara Benazir, Felix Xiaozhu Lin.
  Benchmarked 8B–405B LLMs across 26 quantization schemes uncovering performance bottlenecks on Apple Silicon vs NVIDIA GPUs while revealing non-intuitive hardware bottlenecks (latency, memory b/w, compute, power).

- **[Mobisys’24] Speech Understanding on Tiny Devices with A Learning Cache** [PDF]
  Afsara Benazir, Zhiming Xu, Felix Xiaozhu Lin.
  Integrated on-device execution with cloud offloading to understand human like speech in a $5 MCU at 1.5MB memory with 75% faster latency.

- **[SEC’25] Privacy-Preserving Edge Speech Understanding with Tiny Foundation Models** [PDF]
  Afsara Benazir, Felix Xiaozhu Lin.
  Developed and edge/cloud privacy preserving speech inference engine that filters >83% sensitive entities on-device, maintaining transcription accuracy at 0.11 WER.

Poster: [MobiCom'24] Maximizing the Capabilities of Tiny Speech Foundation Models in a Privacy Preserving Manner [PDF]

- **[SOSP'25] A Journey of Modern OS Construction From boot to DOOM** [PDF]
  Wonkyo Choe, Rongxiang Wang*, Afsara Benazir*, Felix Xiaozhu Lin*.
  **Co-primary authors*
  Worked in developing an instructional OS on Raspberry Pi 3 with modern features (multicore, threading, USB, DMA, per-app address spaces, debugging, and a window manager.

- **[WI-IAT'20]** Credibility assessment of User Generated health information of the Bengali language in micro blogging sites using NLP techniques and Machine Learning. [PDF] *Afsara Benazir, Sadia Sharmin.*
  Workshop paper at the 2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology

ACADEMIC PROJECTS

Moving Cube Game with interactive sound effects (2023) video

- Designed and implemented a handheld game on the TM4C123G MCU and Booster Pack MKII fulfilling the constraints of RTOS including multi-threading and deadlock prevention, using C and Arm Keil Studio IDE

Adaptive Step Tracking using Smartwatch for Smart Health Application (2023)

- Collected and evaluated data to design a closed loop feedback system using WaDa app and WEKA classifier

BetterSound: a real time location based noise alert android application (2022)

- Notifies users to avoid historically noisy areas built using Java Frontend and Firestore Database in Backend

Implemented **Face and Hand gesture recognition system** (2024) video

- On XIAO ESP32S3 MCU with 512KB SRAM for low-resource on-device authentication and control

TECHNICAL SKILLS

Machine Learning: Deep Learning[CNN, Transformer, MoE], Quantization, Evaluation Pipeline, Named Entity Recognition
Languages: Python , C++ ,C, Java, PHP, Bash, SQL, Assembly (8086)
Framework/Lib: llama.cpp, PyTorch,lm-eval, HuggingFace, Metal, CoreML, Laravel, Django
Libraries: Pandas, Numpy, soundfile, SpaCy, NLTK
Software: PyCharm, VS code, GPU Profiling (Nsight, Instruments), Embedded (STM32CUBE IDE, Arm Keil, Atmel Studio)
Miscellaneous: RaspberryPi, STM32F7 Booster Pack, XIAO ESP32 series, xv6, Linux, Git, LaTeX.

ACHIEVEMENTS

- Student travel grant at MobiCom'24 (ACM International Conference on Mobile Computing and Networking)
- Faculty choice award at the poster presentation session of UVa CSGSG Research Symposium (2023) *poster*
- HPCI selected participant at The International Conference for High Performance Computing, Networking, Storage, and Analysis (SC'20)
- Undergraduate University Merit Scholarship in Level 4/Term 1 (March 2020)
- Undergraduate ABI student scholar at Grace Hopper Celebration of Women in Computing (GHC'19)

LEADERSHIP

- Student committee chair (lightning talk segment) at the 1st LLM workshop at UVA
- Mentored 4 Charlottesville high school students in developing a hands-on engineering capstone project in collaboration with Link Lab.
  - Conducted weekly meetings, supervised prototyping. (news article) *Fall'24 & Spring'25*
- Reviewer: AE@SIGCOMM'25, AE@PPoPP 2025