

Ahasan Kabir | Ph.D. Student in Computer Science

Orlando, FL, USA

☎ +1 (689) 250 9987 • ✉ ahasankabir146@gmail.com • 🌐 ahasannn.github.io
in ahasankabir • 🔄 Ahasannn

Education

University of Central Florida

Ph.D. in Computer Science

Advisor: Dr. Qian Lou

Orlando, USA

Aug 2025 – Present

Chittagong University of Engineering and Technology

B.Sc. in Computer Science and Engineering

Chattogram, Bangladesh

Jan 2018 – Jul 2023

Research Interests

LLM serving, AI infrastructure, and efficient inference systems.

Work Experience

University of Central Florida

Graduate Research Assistant

Orlando, FL

Aug 2025 – Present

- Built **HW-Router**, a router that decides which LLM serves each request from live GPU signals rather than static model specs; delivered **3.4–3.9× lower end-to-end latency** and **46–48 pp higher SLO attainment** than state-of-the-art routers, with $\sim 200 \mu s$ overhead (DAC 2026, first author).
- Built **InfraMind**, a reinforcement-learning orchestrator that plans multi-agent LLM workflows around live cluster load; achieved **up to +7.6 pp accuracy**, **7× lower latency**, and **99.9% SLO compliance** under high load (arXiv 2026, first author).

Kinetik Healthcare Solutions, Inc.

Software Engineer (Remote)

New York, USA

May 2023 – Jul 2025

- Built Kinetik's medical transportation iOS app from scratch in Swift; published on the App Store (link).
- Implemented trip booking and tracking flows, core UI logic, and key app functionality.
- Contributed to backend services using TypeScript, Node.js, NestJS, and gRPC.

Publications & Manuscripts

[1]: **HW-Router: Hardware-Aware Routing for Scalable Multi-LLM Serving** (first author).

Accepted at DAC 2026.

[2]: **InfraMind: Infrastructure-Aware Multi-Agent Orchestration** (first author).

arXiv preprint arXiv:2606.11440, 2026 (link)

[3]: **BEemoLexBERT: A Hybrid Model for Multilabel Textual Emotion Classification in Bangla by Combining Transformers with Lexicon Features** (first author).

Bangla Language Processing Workshop (BLP), EMNLP 2023 (PDF)

[4]: **BEemoFusionNet: A Deep Learning Approach for Multimodal Emotion Classification in Bangla Social Media Posts.**

IEEE ICCIT 2023 (PDF)

Technical Skills

Languages: Python (research), C/C++ (competitive programming), Swift (production app)

ML & Systems: PyTorch, CUDA, Linux; familiar with Docker, vLLM

Achievements

Awards: DAC Young Fellow, ACM/IEEE Design Automation Conference (2026).

ICPC: Asia West Continent Finalist (2021 & 2022); 22nd place (2022), 25th place (2021).

Meta Hacker Cup: 38th/34,584 in Qualification Round 2021; 116th/12,692 in Round 1.

Codeforces: Candidate Master (Max Rating 2094, Top 1%); 3300+ problems solved; 330+ rated contests.

Leadership

Vice President, CUET Computer Club (Aug 2022 – Apr 2023): Organized workshops, seminars, and programming contests; coordinated annual planning and logistics.