Kazi Hasan Ibn Arif

403 Progress St NE, Blacksburg, VA-24060, United States
☎ 5404496919 ✉ hasanarif@vt.edu LinkedIn Scholar

PhD student in Machine Learning & Systems with a peer-reviewed publications at top ML conferences. Research focuses on optimizing the training and inference of multimodal LLMs, specializing in techniques like efficient self-attention, token dropping, layer skipping, sparsity, and quantization. Prior work experience on research and development of AI-driven recommendation system at a Fortune 500 company

EDUCATION

Virginia Tech, Blacksburg, Virginia, USA *Aug 2023 – Present*
PhD Student in Computer Science Advised by Dr. Bo Ji.

Bangladesh University of Engineering and Technology, Dhaka, Bangladesh *Feb 2017 – May 2022*
Bachelor's in Computer Science and Engineering

WORK EXPERIENCE

SNAIL Lab (Virginia Tech), Blacksburg, Virginia, Graduate Research Assistant *Aug 2023 – Present*
System/Algorithmic Optimization of LLM/LMM Inference.

IQVIA, USA (Remote), Machine Learning Engineer *May 2022 – Aug 2023*
Research and Development of AI-driven recommendation engine.

PUBLICATIONS

[**AAAI 2025**] **Kazi Hasan Ibn Arif**, JinYi Yoon, Dimitrios S Nikolopoulos, Hans Vandierendonck, Deepu John, Bo Ji, "HiRED: Attention-Guided Token Dropping for Efficient Inference of High-Resolution Vision-Language Models", *Proceedings of the AAAI Conference on Artificial Intelligence* [Paper] [Code]

[**Arxiv 2024**] **Kazi Hasan Ibn Arif**, Sajib Acharjee Dip, Khizar Hussain, Lang Zhang, Chris Thomas, "Fixing Imbalanced Attention to Mitigate In-Context Hallucination of Large Vision-Language Model", *Under Review* [Paper] [Code]

[**AAAI Symposia 2024**] Sajib Acharjee Dip, **Kazi Hasan Ibn Arif**, Uddip Acharjee Shuvo, Ishtiaque Ahmed Khan, Na Meng, "Equitable Skin Disease Prediction Using Transfer Learning and Domain Adaptation", *Proceedings of the AAAI Symposium Series* [Paper]

[**INCET 2021**] Muntasir Hoq, **Kazi Hasan Ibn Arif**, Mohammed Nazim Uddin, "Local and Global Feature Based Hybrid Deep Learning Model for Bangla Parts of Speech Tagging.", *2021 2nd International Conference for Emerging Technology (INCET)* [Paper]

TECHNICAL SKILLS

Languages: Python, C, C++, Java, Shell
Machine Learning and Frameworks: PyTorch, Huggingface-transformers, vLLM, llama.cpp
Systems and Cloud: Linux, CUDA, Git (GitHub, GitLab), Docker, Kubeflow
Databases: Oracle, PostgreSQL, MongoDB

LEADERSHIP AND SERVICES

Secretary, Computer Science Graduate Council 2024-2025 at Virginia Tech
I am elected as Secretary to represent 400+ graduate students and manage active communication between students and authority within department and beyond

Reviewer, ICLR 2025
Workshop on Quantify Uncertainty and Hallucination in Foundation Models: The Next Frontier in Reliable AI

Student Scholar and Volunteer, AAAI 2025, Philadelphia, Pennsylvania, USA

AWARDS AND SCHOLARSHIPS

Cyber Innovation Scholars Program 2024: I was awarded $2000 grant from CCI SWVA Cyber Innovation Scholars Program

Fusemachines AI Fellowship 2022: I was selected for the year-long fellowship sponsored by H&M, and received best presentation award in the Machine Learning course

Dean's List Award (Senior Year): I received for achieving honors grades in consecutive semesters

Admission Test Scholarship: I was awarded for securing 72nd place (top 1%) in the 2016 undergraduate admission test at the top engineering school in Bangladesh

Bangladesh Physics Olympiad: I ranked 17th in the divisional round and qualified for the national level

PROJECTS

Full list is available here: GitHub Link

HiRED-LLaVA-Next | Link | PyTorch, Huggingface Transformer, Python **2024**

- We speed-up the inference of LLaVA-Next by 4.7x, reduce response latency by 78%, and cut the GPU memory usage by 14% on an NVIDIA TESLA P40 without sacrificing much of its multimodal tasks accuracy.

Fix In-Context Hallunication of LLaVA | Link | Python, PyTorch, Huggingface Transformer **2024**

- We mitigate in-context hallucination by 46% (CHAIR score) of Multimodal-LLM like LLaVA by intervening its self-attention and adjust the attentions of visual and text tokens in the LLM generation phase.

A C++ implementation of Rasterization and Ray Tracing Algorithm | Link | OpenGL, C++ **2021**

- Implemented Phong illumination, ray-object intersection, multi-level reflections, and texture mapping to render realistic scenes from scratch.

AI-Enabled Lines of Action Game | Link | Java, JavaFX **2020**

- Developed a heuristic for an AI-enabled board game with a GUI built using JavaFX.

CPP Compiler | Link | Yacc, Lex, C **2019**

- Developed a subset of a C compiler with Lexical, Syntax, and Semantic Analysis, including Intermediate Code Generation. Generated DAGs and TAC and converted them to 8086 assembly code.