ANUSHRI SURESH

asures13@jh.edu anushrisuresh anushrisuresh

EDUCATION

Johns Hopkins University *May 2026*
Master of Science and Engineering in Computer Science GPA: **4.0/4.0**
Activities: Teaching Assistant - Machine Learning: Deep Learning (Spring '25)

National Institute of Technology, Tiruchirappalli *May 2022*
Bachelor of Technology in Electronics and Communication Engineering GPA: **8.43/10**

SKILLS

Programming Languages: Python, C/C++, SQL, Bash
Developer Tools: Git, Docker, Kubernetes, Linux, SLURM, Airflow
ML & AI Frameworks: PyTorch, TensorFlow, Keras, vLLM, HuggingFace Transformers, CUDA, OpenCV

EXPERIENCE

Johns Hopkins University - Center for Language and Speech Processing *Baltimore, USA*
*Summer Research Assistant — **Supervisors**: Prof. Daniel Khashabi, Prof. Eric Nalisnick* *May 2025 – Present*

- Building a PyTorch and vLLM-based inference engine for budget-aware decoding on mathematical reasoning benchmarks, maintaining performance degradation within **5%**.

Carnegie Mellon University - InfiniAI Lab *Pittsburgh, USA*
*Research Assistant — **Supervisor**: Prof. Beidi Chen* *February 2025 – Present*

- Optimizing speculative decoding, Key-Value (KV) cache compression, and continuous batching within MagicDec, achieving up to **2.5×** throughput for long-context LLMs (**32K** tokens, batch size **256**) on A100 GPUs.

Johns Hopkins University - ARCADE Laboratory *Baltimore, USA*
*Graduate Research Assistant — **Supervisor**: Prof. Mathias Unberath* *September 2024 - January 2025*

- Built a language-promptable digital twin using a multi-modal foundation model (FluoroSAM) for real-time segmentation, 3D reconstruction, and automatic collimation, reducing radiation exposure by **60%**.
- Engineered an LLM-driven interface to interpret complex verbal instructions for the Brainlab Loop-X robotic C-arm, allowing autonomous, hands-free surgical imaging with **84%** end-to-end success in cadaveric trials.

Bosch Global Software Technologies *Bangalore, India*
Senior Engineer - Machine Learning *August 2022 - July 2024*

- Spearheaded a team of 6 to develop an AI-powered test case optimizer for the Automotive Electronics Division, reducing test runtime by **66%** and increasing code coverage to **94%**.
- Delivered scalable chatbot automation for the Bosch Automation Platform, integrating Named Entity Recognition and ElasticSearch to improve search relevance, boosting customer satisfaction by **31%** across **5,000+** enterprise users.

University of Zürich - Artificial Intelligence and Machine Learning Group *Zurich, Switzerland*
*Summer Research Fellow — **Supervisor**: Prof. Manuel Günther* *June 2021 - October 2021*

- Boosted face recognition on low-quality surveillance footage by designing a Super-Resolution GAN, increasing accuracy by **37%** on downsampled and **5%** on full datasets for reliable identification in real-world security scenarios.

SELECTED PROJECTS

Don't Drop It, Compress It: Selective KV Quantization [GitHub]

- Led a team of 4 to design and implement Selective KV Quantization for LLM inference, preserving sink/window tokens in full precision while quantizing older cache entries to int8, achieving **2×** memory savings with minimal impact on perplexity **(5.56)** and ROUGE-L **(0.2073 → 0.1709)**.

Rich Teacher Features for Efficient Single-Image Haze Removal [Bachelor Thesis, Optik Journal]

- Devised a lightweight haze removal pipeline via heterogeneous knowledge distillation with a novel feature affinity module, yielding a **15%** PSNR gain and **20×** model compression.

PUBLICATIONS

- *Intelligent Control of Robotic X-ray Devices using a Language-Promptable Digital Twin*, **IPCAI, 2025**. [arXiv]
- *Rich feature distillation with feature affinity module for efficient image dehazing*, **Optik Journal, 2022**. [arXiv]

AWARDS

- **The Siemens Healthineers Best Paper Award**, IPCAI 2025
- **Best Project Award**, NLP: Self-Supervised (Spring 2025), Johns Hopkins University