

Sunnyvale, CA  
yash-s20.github.io

# YASH SHARMA

(845) 290-4694  
yash.sharma200999@gmail.com  
github.com/yash-s20

## EDUCATION

|   |                      |  |
|---|----------------------|--|
| <b>Cornell University</b><br>MS in <b>Computer Science</b>   Minor in <b>Cognitive Science</b><br>Computational Sustainability, Advanced Language Technologies, Advanced Programming Languages  | <i>Ithaca, NY</i>    | <b>Aug 2022 – May 2024</b><br>GPA: 3.91 / 4  |
| <b>Indian Institute of Technology Bombay</b><br>B.Tech in <b>Computer Science &amp; Engineering</b> (Honors)   Minor in <b>Artificial Intelligence</b><br>Deep Learning for NLP, Advanced Machine Learning, Analysis of Concurrent Programs | <i>Mumbai, India</i> | <b>Aug 2017 – May 2021</b><br>GPA: 9.68 / 10 |

## SOFTWARE SKILLS

**Systems & Programming** | python, C/C++, bash, Rust, Haskell, Java, Javascript, SQL, AVX, Git, Perforce, Docker, KVM  
**Machine Learning** | PyTorch, TensorFlow, TensorRT & onnxruntime

## WORK EXPERIENCE

|   |                              |                             |
|---|------------------------------|-----------------------------|
| <b>ML Research Engineer, Matic Robots</b><br>• Part of the <b>Neural Networks</b> team building robust, secure and autonomous <b>perception and understanding</b><br>• Building, training and evaluating <b>transformer-based</b> 3D reconstruction networks that run <b>real-time</b> on edge devices  | <i>Mountain View, CA</i>     | <b>Jun 2024 – present</b>   |
| <b>Software Engineer, Samsung Electronics</b><br>• Developed high-performance <b>5G-NR</b> virtual L1 layer as part of <b>Physical Uplink Shared Channel</b> team<br>• Utilized Intel®Intrinsics ( <b>AVX-512</b> ) for efficient parallel processing of data, focusing on cache bottleneck optimization<br>• Reduced bottlenecks in uplink signal processing pipeline to achieve upto <b>20% speedup</b> | <i>Suwon, South Korea</i>    | <b>Sep 2021 – Aug 2022</b>  |
| <b>Network Engineer Intern, Samsung Electronics</b><br>• Built an automated network load testing framework to evaluate performance of in-production load balancing services   | <i>remote</i>                | <b>Jun 2020 – July 2020</b> |
| <b>Summer Research Intern, TU Braunschweig</b><br>• Designed and built <b>WeLineation</b> , a full-stack app using <b>Expectation Maximization</b> for medical image segmentation   | <i>Braunschweig, Germany</i> | <b>May 2019 – July 2019</b> |

## RESEARCH EXPERIENCE

|   |                                   |                            |
|---|-----------------------------------|----------------------------|
| <b>Master's Thesis - Prof. Sanjiban Choudhury</b><br>- Built a learning system using <b>Vision-Language transformers</b> to allow the transfer of human skills to household robots<br>- Collaborated on a <b>speech-interactive task planner</b> for human-robot collaborative cooking, and a web-based evaluator | <i>Cornell University</i>         | <b>Feb 2023 – Apr 2024</b> |
| <b>Undergraduate Research - Prof. Preethi Jyothi</b><br><b>Improving code-switched Automatic Speech Recognition (ASR) using Transformers</b><br>- Built a new bilingual <b>speech recognition</b> model conditioned on language using CUDA accelerated dynamic programming  | <i>IIT Bombay &amp; Microsoft</i> | <b>Aug 2020 – Jun 2021</b> |
| <b>Improving Low Resource Code-switched ASR using Augmented Code-switched TTS</b><br>- Used end-to-end ASR models trained on Hindi and English monolingual corpi and code-switched synthetic data to improve performance in low-resource settings   |                                   | <b>Dec 2019 – Jun 2020</b> |

## PUBLICATIONS

- **Demo2Code**: From Summarizing Demonstrations to Synthesizing Code via Extended Chain-of-Thought [*NeurIPS 2023*]
- Improving **low resource code-switched ASR** using augmented code-switched TTS [*INTERSPEECH 2020*]
- **WeLineation**: crowdsourcing delineations for reliable ground truth estimation [*SPIE Medical Imaging 2020*]

## TEACHING ASSISTANTSHIPS

|  |  |
|--|--|
| <b>Cornell University</b><br><b>Intro. to Machine Learning</b> <i>Spring 2024</i><br><b>Intro. to Analysis of Algorithm</b> <i>Summer 2023</i> | <b>Computer System Organization &amp; Programming</b> <i>Fall 2022, 2023</i><br><b>Computational Sustainability</b> <i>Spring 2023</i> |
| <b>IIT Bombay</b><br><b>Software Systems Lab</b> <i>Fall 2019, 2020</i>  | <b>Calculus</b> <i>Fall 2018</i>   |

## KEY PROJECTS

|  |                                    |                |                    |
|--|------------------------------------|----------------|--------------------|
| <b>Psychological analysis of ChatGPT</b><br>Research course exploring decision making of LLMs in risky and ethically ambiguous situations          | <i>Prof. Valerie Reyna</i>         | <b>Cornell</b> | <b>Fall 2023</b>   |
| <b>Modeling misinformation in organizations</b><br>Formalize the effect of corruption in hierarchical organizations using information networks     | <i>Prof. Jon Kleinberg</i>         | <b>Cornell</b> | <b>Spring 2023</b> |
| <b>Few-shot action recognition on egocentric data</b><br>Building a two-head action recognition system for EPIC-Kitchens tackling long-tail labels | <i>Prof. Kilian Weinberger</i>     | <b>Cornell</b> | <b>Fall 2022</b>   |
| <b>Morphological Inflection through Deep Learning</b>  | <i>Prof. Pushpak Bhattacharyya</i> | <b>IITB</b>    | <b>2021</b>        |
| <b>Maze Solving with Evolutionary RL</b>   | <i>Prof. S. Kalyanakrishnan</i>    | <b>IITB</b>    | <b>2020</b>        |