


# Sree Harsha Nelaturu

 [nelaturuharsha.github.io](https://github.com/nelaturuharsha)  
 [nelaturu.harsha\(at\)gmail\(dot\)com](mailto:nelaturu.harsha(at)gmail(dot)com)

 [/sree-harsha-nelaturu](https://www.linkedin.com/in/sree-harsha-nelaturu)  
 [/nelaturuharsha](https://github.com/nelaturuharsha)

## Education

**Universität des Saarlandes** || *MSc Visual Computing (GPA: 1.7)\*\** || Saarbrücken, DE Oct 2021 - Present

**Massachusetts Institute of Technology** || *Special Student in EECS (GPA: 5.0/5.0)* || Cambridge, MA, USA Sept - Dec 2018

**SRM Institute of Science and Technology** || *B.Tech ECE (86.18%)* || Chennai, TN, India July 2016 - May 2020

[\*\* = In the german system, 1.0 is the highest achievable grade]

## Publications and pre-prints

- **INCLUDE: Evaluating Multilingual Language Understanding with Regional Knowledge:** (Angelika Romanou, Negar Foroutan, Anna Sotnikova, Zeming Chen, **Sree Harsha Nelaturu**, Shivalika Singh, and other authors) – Core contributor, **Spotlight @ ICLR 2025**. Advised by Marzieh Fadaee, Sara Hooker, Antoine Bosselut.
- **On the Fairness Impacts of Hardware Selection in Machine Learning** (**Sree Harsha Nelaturu\***, Nishaanth Kanna Ravichandran\*, Cuong Tran, Sara Hooker, Ferdinando Fioretto). Accepted **Poster @ ICML 2024** [\* = equal contribution]
- **End to End learnable masks with differentiable indexing.** (Dibyanshu Shekhar\*, **Sree Harsha Nelaturu\***, Ashwath Shetty\*, Ilia Sucholutsky). Accepted for archival at **Tiny Papers @ ICLR2023** [\* = equal contribution]
- **Accelerated CNN Training through Gradient Approximation.** (Ziheng Wang, **Sree Harsha Nelaturu**, Saman Amarsinghe). Published at *EMC<sup>2</sup> Workshop* at the International Symposium on Computer Architecture (**ISCA 2019**).

## Experience

**Amazon Web Services** || *Applied Scientist Intern* || Tübingen, Germany November 2024 - May 2025

- **(November 2024 - May 2025) Manager: Dr. Jonas Kübler.** Worked on evaluating the impact of quantization and worked on disaggregated inference. Internship report submitted to internal conference (AMLC)

**Max Planck Institut für Informatik** || *Research Assistant (HiWi)* || Saarbrücken, Germany August – October 2024

- **(August - October 2024) Advisor: Dr. Jonas Fischer.** Working on the Mechanistic Interpretability of f-MRI + Image reconstruction models.

**CISPA Helmholtz Institute for Information Security** || *Research Assistant (HiWi)* || Saarbrücken, Germany July 2022 – July 2024

- **(August 2023 - July 2024) Advisor: Dr. Rebekka Burkholz.** Developed techniques for perturbation aware and accelerated methods for sparse optimization. Open sourced [TurboPrune](#) - 21x faster ground up rewrite of group's codebase.
- **(July 2022 - July 2023) Advisor: Dr. Sebastian Stich.** Worked on communication and compute efficient algorithms for federated/distributed optimization using knowledge distillation and sparsity.

**Rediscovery.io** || *Jr. Deep Learning Research Scientist* || Remote - London, UK July. 2020 – May 2021

- Contributed to the development of the remo.ai - a dataset management and visualization tool SDK and integrated supervised/self-supervised learning methods for [classification, segmentation, object detection] in the open source SDK.

**Myelin Foundry** || *Deep Learning Intern* || Bengaluru, IN

- **(March - June 2020)** Designed an end-to-end pipeline for media restoration, upscaling and enhancement for old movies/TV-shows. Involved market research and development of on-device super-resolution for 540p -> 4K upscaling.
- **(June 2019)** Developed an optimized pipeline for training and edge deployment of ASR (Automatic Speech Recognition) for low-resource languages.

**RunwayML** || *ML Researcher (Consultant)* || Remote - Brooklyn, USA Sept. 2019 – Jan. 2020

- Added 22+ optimized CV, NLP models to the Runway model zoo – including generative, processing and task oriented models via an intuitive interface in the SDK easily accessible by creatives/artists. Details [here](#).

**Response Environments, MIT Media Lab** || *Undergraduate Researcher* || Cambridge, MA, USA Sept., - Dec., 2018

- Developed an information delivery pipeline using DNNs to classify and subsequently modifying a user's audio-stream. Achieved highest possible "A" grade as part of course 6.100 - EECS Project.

## Communities and Volunteering

**Cohere Labs (Formerly C4AI)** || *Community Lead and Researcher* || *Remote*

2022 - Present

- Founded and co-led the ML Theory group and currently co-lead the ML efficiency group. I present research papers, organize guest lectures and workshops in the community. Top 1% active community members.

## Awards and Conferences

- **Expedition Aya: Most Promising Award (May 2025):** explored multilingual speculative decoding to make LLMs more efficient
- **Federated Learning Practical, Deep Learning Indaba (Sept 2024):** In collaboration with Andrej Jovanović and Luca Powell
- **Best use of OpenAI API (Feb 2021):** Stanford TreeHacks
- **Eastern European Machine Learning School (EEML) (2021, 2022):** Accepted based on original research proposal.
- **Silver Medal (Feb 2019):** SRM Research Day
- **First Place Winner (Dec 2017):** Microsoft GAINS AI Hackathon
- **First Place Winner, (Dec 2017):** ImagingHub Smart Home Competition
- **Innovation Award, March 2017:** Smart India Hackathon (Ministry of Electronics and IT)

## References

- **(Thesis Advisor) Dr. Rebekka Burkholz, CISPA Helmholtz Center for Information Security:** burkholz@cispa.de
- **(Research Advisor) Dr. Sara Hooker, Cohere For AI:** sarahooker@cohere.com
- **(Research Advisor) Dr. Ferdinando Fioretto, University of Virginia:** fioretto@virginia.edu

## Skills and Interests

- **Tools and frameworks:** PyTorch, TensorRT, JAX, OpenVINO, CUDA, DeepSpeed, Transformers, HuggingFace, TVM, vLLM,
- **Interests:** Efficient training/optimization methods [distributed, federated] and inference, transformers, large language models, Sparsity, Pruning, Quantization, multilingual, multimodal.

## Links

- **Website:** <https://nelaturuharsha.github.io/>
- **INCLUDE: Evaluating Multilingual Language Understanding with Regional Knowledge :** <https://arxiv.org/abs/2411.19799>
- **On The Fairness Impacts of Hardware Selection in Machine Learning :** - <https://arxiv.org/abs/2312.03886>
- **Accelerated CNN Training Through Gradient Approximation:** <https://www.emc2-ai.org/assets/docs/isca-19/emc2-isca19-paper3.pdf>
- **TurboPrune:** <https://github.com/nelaturuharsha/TurboPrune>