Last updated: August 15th

# Rhea Sukthanker

EMAIL: sukthank@cs.uni-freiburg.de
LINKS: HOMEPAGE, GOOGLE SCHOLAR, TWITTER

## Research Interests

My research focuses on automating and optimizing foundation model inference—particularly for large language models (LLMs) and vision models—to facilitate inference efficiency in their real-world applications. To address this, my work develops novel techniques for pruning, quantization, and knowledge distillation, reducing the manual effort in tuning these methods. Ultimately, my goal is to make foundation models more accessible and sustainable across diverse domains. As an open-source effort towards this goal, I develop and maintain the library whittle, with several others.

- Automated Foundation Model Compression: Pruning, Quantization, Distillation
- Efficient Neural Architecture Search: Gradient-based search, Weight-Sharing

## Education

**Department of Computer Science, University of Freiburg** — Freiburg, Germany
*Ph.D. in Computer Science* — 2022 - 2026 *(expected)*
- Advisor: Prof. Frank Hutter
- Research area: Architecture Inference Optimization

**Department of Computer Science, ETH Zurich** — Zurich, Switzerland
*Masters in Data Science* — 2018 - 2021
- GPA: 5.39/6

**Department of Information Technology, VIT University** — Vellore, India
*Bachelor's in Information Technology* — 2014 - 2018
- GPA: 9.75/10

## Research and Internship Experience

**Microsoft Research** — Cambridge
*Applied Science Intern* — May 2025 - July 2025
- Supervisors: Dr. Pashmina Cameron and Dr. James Hensman
- Project: **Automated Quantization of LLMs (AQUA)**
- Research focus: Knowledge Distillation on 2-bit (Vector) Quantized LLMs. I worked on applying knowledge distillation based recovery finetuning for 2-bit quantized LLMs achieving *state-of-the-art*, 2-bit language models, outperforming Quantization Aware Training methods at a fraction of their cost.

**Computer Vision Lab, ETH Zurich** — Zurich, Switzerland
*Student Researcher* — March 2021 - April 2022
- Advisors: Dr. Zhiwu Huang and Dr. Suryansh Kumar
- Research area: Neural Architecture Search and Generative Models

**Computational Intelligence Laboratory, NTU** — Singapore
*Research assistant* — May 2017-July 2017 and Jan 2018 – May 2018
- Advisor: Dr. Erik Cambria
- Research area: Anaphora and Coreference Resolution

## Journal Publications

1. **Rhea Sukthanker**, Soujanya Poria, Erik Cambria, Ramkumar Thirunavukarasu. Anaphora and coreference resolution: A review. *Information Fusion (IF:14.7)*.

## Workshop Publications

1. **Rhea Sukthanker**, Benedikt Staffler, Frank Hutter, Aaron Klein. Large Language Model Compression with Neural Architecture Search. *NeurIPS 2024 Compression Workshop*.
2. **Rhea Sukthanker***, Arber Zela*, Benedikt Staffler, Samuel Dooley, Josif Grabocka, Frank Hutter. Multi-Objective Differentiable Architecture Search. *ICML 2024 WANT Workshop*.
3. Yan Wu, Zhiwu Huang, Suryansh Kumar, **Rhea Sanjay Sukthanker**, Radu Timofte, Luc Van Gool. Trilevel Neural Architecture Search for Efficient Single Image Super-Resolution. *CVPR 2022 NAS Workshop*.

*: equal contribution

## Conference Publications

1. **Rhea Sukthanker***, Arber Zela*, Benedikt Staffler, Samuel Dooley, Josif Grabocka, Frank Hutter. Multi-Objective Differentiable Architecture Search. *International Conference on Learning Representations (ICLR 2025) , Singapore*.
2. **Rhea Sukthanker**, Arber Zela, Benedikt Staffler, Aaron Klein, Lennart Purucker, Jörg K. H. Franke, Frank Hutter. HW-GPT-Bench: Hardware-Aware Architecture Benchmark for Language Models. *Neural Information Processing Systems DBT Track (NeurIPS 2024), Vancouver, Canada*.
3. **Rhea Sukthanker**, Arjun Krishnakumar, Mahmoud Safari, Frank Hutter. Weight-Entanglement Meets Gradient-Based Neural Architecture Search. *International Conference on Automated Machine Learning (AutoML 2024), Paris, France*.
4. Samuel Dooley*, **Rhea Sukthanker***, John P. Dickerson, Colin White, Frank Hutter, Micah Goldblum. Rethinking bias mitigation: Fairer architectures make for fairer face recognition **oral**. *Neural Information Processing Systems (NeurIPS 2023), New Orleans, USA*.
5. Simon Schrodi, Danny Stoll, Binxin Ru, **Rhea Sukthanker**, Thomas Brox, Frank Hutter. Construction of Hierarchical Neural Architecture Search Spaces based on Context-free Grammar. *Neural Information Processing Systems (NeurIPS 2023), New Orleans, USA*.
6. **Rhea Sukthanker**, Zhiwu Huang, Suryansh Kumar, Radu Timofte, Luc Van Gool. Generative flows with invertible attentions. *Computer Vision and Pattern Recognition (CVPR 2022), New Orleans, USA*.
7. **Rhea Sukthanker**, Zhiwu Huang, Suryansh Kumar, Radu Timofte, Luc Van Gool. Neural Architecture Search of SPD Manifold Networks. *International Joint Conferences on Artificial Intelligence (IJCAI 2021), Montreal, Canada*.

## Academic Services

**Reviewer**

- NeurIPS: 2023, 2024
- ICML: 2023, 2024
- ICLR: 2023, 2024
- AutoML: 2023, 2024
- AISTATS: 2024

**Diversity and Inclusion Chair**

- AutoML 2024

**Teaching**

- Foundations of Deep Learning (Semester Course: 2023, 2024)
- Pruning and Efficiency in Large Language Models (Seminar 2024)

## Awards and Honors

- Awarded Goa Scholars 2018-19
- Awarded ETH Zurich Excellence Scholarship

## Invited Talks

- NeurIPS 2023 Oral Talk : "Rethinking bias mitigation: Fairer architectures make for fairer face recognition"
- AutoML Seminar 2024 : "Rethinking bias mitigation: Fairer architectures make for fairer face recognition"

## References

Prof. Dr. Frank Hutter: ELLIS Institute Tübingen and University of Freiburg, Germany
Email: fh@cs.uni-freiburg.de

Dr. Aaron Klein: ScaDS.AI, Leipzig, Germany
Email: kleiaaro@gmail.com

Dr. Zhiwu Huang: University of Southampton, UK
Email: Zhiwu.Huang@soton.ac.uk

Dr. Suryansh Kumar: Texas A&M University College Station, USA
Email: suryanshkumar@tamu.edu