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

Rhea Sukthanker

Last updated: February 4th

Research Interests

My research focuses on automating and optimizing foundation model inference—particularly for Large Language Models (LLMs) and vision models—to enable efficient, scalable, and cost-effective deployment in real-world applications. While training these models incurs significant one-time compute, their inference costs dominate long-term energy consumption, latency, and memory usage across deployment contexts. To address this challenge, my work develops novel and automated techniques for model compression, including structured and unstructured pruning, low-bit quantization, and knowledge distillation, with an emphasis on reducing manual hyperparameter tuning and expert intervention. By integrating hardware-aware optimization, automated algorithm selection, and scalable search strategies, I aim to make foundation models more accessible, sustainable, and practical for diverse domains ranging from edge computing to large-scale cloud services.

As part of this goal, I actively contribute to open-source tools and research infrastructure. I co-develop and maintain the library whittle, a framework for automated model compression of large language models, and contribute to the neural architecture search library NASLib, supporting research on reproducible architecture search methods.

My recent research interests include:

- Automated Foundation Model Compression: Developing scalable, hardware-aware algorithms for pruning, low-bit quantization and knowledge distillation.
- Efficient and Multi-Objective Neural Architecture Search: Designing gradient-based, weight-sharing, and multi-objective optimization techniques for scalable neural architecture search under compute and memory constraints.

Education

Department of Computer Science, University of Freiburg Freiburg, Germany
Ph.D. in Computer Science Feb 2022 - May 2026 *(expected)*

- Advisor: Prof. Frank Hutter
- Research focus: Efficient and Scalable Multi-Objective Optimization for Foundation Models

Department of Computer Science, ETH Zurich Zurich, Switzerland
Masters in Data Science Sept 2018 - July 2021

- GPA: 5.39/6
- Awarded the ESOP Scholarship.

Department of Information Technology, VIT University Vellore, India
Bachelor's in Information Technology June 2014 - June 2018

- GPA: 9.75/10; Rank: 2nd
- Awarded merit scholarship for academic excellence.

Research and Internship Experience

Microsoft Research Cambridge, UK
Research Intern (Applied Sciences Group) May 2025 - July 2025

- Supervisors: Dr. Pashmina Cameron and Dr. James Hensman
- Project: **Automated Quantization of LLMs** (**AQUA**)
- Research Contributions:
  - Contributed to AQUA, an internal Microsoft Research framework for automated end-to-end post-training quantization of large language models (LLMs); designed and implemented a LoRA + knowledge distillation fine-tuning pipeline enabling stable, parameter-efficient adaptation of highly compressed (2-bit) models without reverting to higher-precision weights.
  - Advanced Microsoft's internal 2-bit vector quantization framework by integrating LoRA A/B matrices directly into the quantized setting, investigating dataset mixing strategies for next-generation Phi models, and conducting extensive empirical studies on fine-tuning and distillation to derive best practices for training ultra-compressed LLMs under extreme memory and hardware constraints.
  - Demonstrated performance surpassing Quantization-Aware Training (QAT) methods at a fraction of the computational and training cost, significantly improving the practicality of 2-bit model deployment; contributing inventor on the patent "Vector Quantization using a Learnable Codebook."

Computer Vision Lab, ETH Zurich Zurich, Switzerland
Student Researcher March 2021 - April 2022

- Advisors: Dr. Zhiwu Huang and Dr. Suryansh Kumar
- Research Contributions:

  Neural Architecture Search of SPD Manifold Networks (**Master's Project , ETH Zurich**)

  - Formulated a new neural architecture search (NAS) problem for Symmetric Positive Definite (SPD) manifold networks and defined a geometry-aware search space tailored to non-Euclidean data representations.
  - Developed a one-shot NAS method using a differentiable supernet to efficiently explore SPD architecture candidates while preserving manifold structure.
  - Demonstrated improved performance over state-of-the-art handcrafted SPD networks and traditional NAS algorithms on multiple vision benchmarks with significantly lighter architectures.

  Generative Flows with Invertible Attentions (**Master's Thesis, ETH Zurich**)

  - Proposed novel normalizing flow architectures with invertible attention mechanisms to improve expressivity while preserving exact likelihood computation.
  - Conducted large-scale experiments showing improved generative modeling performance over baseline flow models.
  - Resulted in publication at *Computer Vision and Pattern Recognition (CVPR) 2022*.

Computational Intelligence Laboratory, NTU Singapore
Research assistant May 2017-July 2017 and Jan 2018 – May 2018

- Advisor: Dr. Erik Cambria
- Research Contributions:

  Anaphora and Coreference Resolution: A Review

  - Authored a comprehensive survey of coreference and anaphora resolution methods, analyzing rule-based, feature-based, and neural approaches.
  - Synthesized evaluation metrics, benchmark datasets, and emerging research trends to highlight challenges and future directions in document-level NLP.

Preprints

1. Arjun Krishnakumar*, **Rhea Sanjay Sukthanker***, Hannan Javed Mahadik*, Gabriela Kadlecová, Vladyslav Moroshan, Timur Carstensen, Frank Hutter, Aaron Klein. Where to Begin: Efficient Pretraining via Subnetwork Selection and Distillation. *(under review)*.

Journal Publications

1. **Rhea Sukthanker**, Soujanya Poria, Erik Cambria, Ramkumar Thirunavukarasu. Anaphora and coreference resolution: A review. *Information Fusion (IF:15.5)*.

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*.

Conference Publications

*: equal contribution

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*.

Patents

1. A. Zela, B. S. Staffler, F. Hutter, M. Safari, and **R. S. Sukthanker** (Feb. 19, 2025b). “Method and/or apparatus for architecture search”. U.S. pat. req. US20250272576A1. Robert Bosch GmbH. URL: https://patents.google.com/patent/US20250272576A1/. Published.
2. A. Zela, B. S. Staffler, F. Hutter, M. Rapp, and **R. S. Sukthanker** (May 19, 2025a). “Method and device for reducing a network dimension of a base model”. U.S. pat. req. US20250378332A1. Robert Bosch GmbH. URL: https://patents.google.com/patent/US20250378332A1/. Published.
3. R. Grazzi, **R. S. Sukthanker**; E. Portugaly, H. R. Jackson-FLux, P. J. Cameron, and J. J. Hensman.(Feb 2, 2026). “Vector Quantization using a Learnable Codebook”. U.S. pat. (filed)

Academic Services

Reviewer

- NeurIPS: 2023, 2024
- ICML: 2024, 2025, 2026
- ICLR: 2024, 2025, 2026
- AutoML: 2024

Diversity and Inclusion Chair

- AutoML 2024

Teaching

- Deep Learning Lab (Semester Course: 2022)
- Foundations of Deep Learning (Semester Course: 2023, 2024)
- Pruning and Efficiency in Large Language Models (Seminar Course: 2024)

Awards and Honors

- Awarded Goa Scholars 2018-19
- Awarded ETH Zurich Excellence Scholarship
- 1st place in AutoML Cup organized at AutoML 2023
- 3rd place in AutoML Decathlon organized at NeurIPS 2022
- Oral Presentation at NeurIPS 2023

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: Prior Labs, ELLIS Institute Tübingen and University of Freiburg
Email: fh@cs.uni-freiburg.de

Dr. Aaron Klein: ELLIS Institute Tübingen
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

Dr. Pashmina Cameron: Microsoft, Redmond, Seattle, USA
Email: pcameron@microsoft.com

Dr. James Hensman: Microsoft Research, Cambridge, UK
Email: jameshensman@microsoft.com