Rishabh Maheshwary

rf.rishabh@gmail.com | +91 – 8427119320 | GitHub | LinkedIn | Scholar | Website

Summary

My research interests broadly encompass training, alignment, and benchmarking of language models. With the growing integration of AI into our day to day lives, I am passionate about developing AI systems that are not only safe but also robust and reliable.

Work Experience

- **ServiceNow** - *Applied Scientist* — **Hyderabad, India**
  My research focuses on enhancing the overall capabilities of language models and aligning them with curated human feedback to control their behaviors in real-world applications. *Aug 2023 – Present*

- **Facebook AI Research** - *AI Resident* — **California, U.S.**
  My research was focused on designing intelligent and reliable systems having joint understanding of vision and language modalities. *Nov 2021 – Dec 2022*

- **Verisk AI** – *Research Intern* — **Hyderabad, India**
  I worked on joint language and vision understanding of multimodal content and semantic understanding of natural language documents. *May 2021 – Oct 2021*

- **Google Summer of Code** – *Software Developer Intern* — **Remote**
  I developed an application enabling users to report incidents thus facilitating nearby assistance and communication. *Apr 2018 – Sept 2018*

Publications

1. Teaching Language models what not to do — Aligning Language Models with Incremental Pairwise Preferences. ***Under Submission in ACL 2024***.

2. Corentin Dancette, Spencer Whitehead, **Rishabh Maheshwary**, Ramakrishna Vedantam, Stefan Scherer, Xinlei Chen, Matthieu Cord, Marcus Rohrbach. Improving Selective Visual Question Answering by Learning from Your Peers. In the *Proceedings of the* ***IEEE/CVF Conference on Computer Vision and Pattern Recognition*** (***CVPR, 2023***) Vancouver, Canada.

3. Vivek Kumar, **Rishabh Maheshwary**, Vikram Pudi. Practice Makes a Solver Perfect: Data Augmentation methods for Math Word Problem Sovers. In the *Proceedings of* ***North American Chapter of the Association for Computational Linguistics*** (***NAACL 2022***), Seattle, Washington.

4. **Rishabh Maheshwary***, Saket Maheshwary*, Vikram Pudi. A Strong Baseline for Query Efficient Attacks in a Black Box Setting. In the *Proceedings of* ***Empirical Methods in Natural Language Processing*** (***EMNLP***) ***2021***, Punta Cana, Dominican Republic.

5. **Rishabh Maheshwary***, Vivek Kumar*, Vikram Pudi. Adversarial Examples for Evaluating Math Word Problem Solvers. In the *Findings of ACL*: ***Empirical Methods in Natural Language Processing*** (***EMNLP***) ***2021***, Punta Cana, Dominican Republic.

6. **Rishabh Maheshwary**, Saket Maheshwary, Vikram Pudi. Generating Natural Language Attacks in a Hard Label Black Box Setting. In the *Proceedings of* ***Association for the Advancement of Artificial Intelligence*** (***AAAI***) ***2021***, Vancouver, Canada.

7. **Rishabh Maheshwary**, Saket Maheshwary, Vikram Pudi. A Context Aware Approach for Generating Natural Language Attacks. In the *Proceedings of* ***Association for the Advancement of Artificial Intelligence*** (***AAAI***) ***2021***, Vancouver, Canada.

---

* Equal Contribution

Education

- **International Institute of Information Technology, Hyderabad** **Hyderabad, India**
  MS by Research in Computer Science and Engineering | CGPA: 8.7/10 *2019 – 2021*

- **University Institute of Engineering and Technology, Panjab University** **Chandigarh, India**
  BTech in Computer Science and Engineering | CGPA: 8.4/10 *2015 – 2019*

Major Projects

- **Information Extraction from Form like Documents** *2021*
  The aim is to design an intelligent reading system that is expected to respond to ad-hoc requests for information, expressed in natural language questions by human users.

- **Generating Adversarial Attacks on Natural Language Processing Models** *2019*
  The aim is to evlauate the robustness and generalization of text classification, entailment, question answering and language modelling systems.

- **MultiHop Question Answering** *2019*
  The aim is to answer questions which require reasoning over multiple supporting documents.

- **Deep Learning for detecting Hate Speech Tweets** *2018*
  The aim is to identify abusive language, flag offensive content using natural language processing.

Miscellaneous

- Actively reviewing for ECCV, TMLR, CoNLL and NeurIPS. *2021 - Present*
- **Google Summer of Code** and **Google CodeIn** mentor. *2018*
- Ranked $1st$ out of 100+ teams in **CODETRIX** (National level coding contest). *2017*
- Ranked $3rd$ out of 100+ teams in **CODE-IT** (National level coding contest). *2016*

Programming Languages and Technologies

- Python, PyTorch, C++, C, Shell, Git
- Machine Learning, Deep learning, Reinforcement learning
- NLP, Multimodal vision & language