Rishabh Bhardwaj

rishabhbhardwaj15@gmail.com
Webpage
LinkedIn
GitHub
(+65)85488773

Research Interest

- My research primarily centers on making NLP models more robust and safe for use. I am currently exploring ways to make LLMs more harmless (safe) while staying helpful (performant). As a part of my safety research, we released safety benchmarks such as Red-Eval and WalledEval, and an alignment approach Red-Instruct. During my phase as a researcher, I have published over 20 research papers. Recent papers at top places (ACL, EMNLP, ICASSP) involve parameter-efficient (PE) learning in language models including dynamic prompts, hypernetworks, and adapters for speech. In the line of PE, I have also explored ways to prune adapters using hypersurface geometry. At the beginning of my PhD, I worked on gender-debiasing algorithms in language models where we removed information from directions in the embedding space that encodes gender information. Another work studied the interpretability of Transformers where we theoretically analyze if the self-attention is identifiable and propose identifiable transformers (accepted at ACL 2021). I was fortunate to work at Salesforce as an NLP research intern in making prompt-tuning dynamic (EMNLP 2022), and a research intern at AWS AI at Amazon, California, in converting LMs semi-parametric (ACL 2023).

Education

2020 – **Ph.D. in Natural Language Processing**, Singapore University of Technology and Design, Singapore.
CGPA: ***5.0/5.0***

2014 – 2018 **B.E. (Hons) Electrical and Electronics**, Birla Institute of Technology and Science Pilani, India.
CGPA: ***8.95/10***

2012 – 2013 **CBSE Higher Secondary Education**, SSAV, Agra, India.
Percentage: ***85.8***

2010 – 2011 **CBSE Secondary Education**, KV Bharatpur, India.
CGPA: ***10/10***

Employment History

June 2022 – Sept 2022 **Applied Scientist Intern** Natural Language Processing at *Amazon, California, USA* (work accepted at ACL 2023).

Jan 2022 – April 2022 **Research Intern** Natural Language Processing at *Salesforce Research Singapore* (work accepted at EMNLP 2022).

April 2019 – Jan 2020 **Research Assistant** under Dr. Prateek Saxena, Associate Professor, Department of Computer Science, *National University of Singapore*.

Jan 2018 – Mar 2019 **Visiting Researcher** under Dr. Lu Wei, Associate Professor, ISTD Pillar, *Singapore University of Technology and Design*.

May 2017 – Aug 2017 **Research Intern** under Dr. Manisha Gupta, Assistant Professor, Electrical and Computer Engineering, *University of Alberta, Canada*.

Research Publications

Conference Proceedings

1. Gupta, P., ... (Lead Contributor) Bhardwaj, R. & Poria, S. (2024). Walledeval: A comprehensive safety evaluation toolkit for large language models
EMNLP 2024 Demo Track (under review)
[https://github.com/walledai/walledeval].

2. Bhardwaj, R., Do, D. A. & Poria, S. (2024). Language models are Homer simpson! safety re-alignment of fine-tuned language models through task arithmetic
ACL 2024
[https://aclanthology.org/2024.acl-long.762].

3. Pala, T. D., Toh, V. Y., Bhardwaj, R. & Poria, S. (2024). Ferret: Faster and effective automated red teaming with reward-based scoring technique
AAAI 2024 (under review)
[https://arxiv.org/abs/2408.10701].

4. Han, V. T. Y., Bhardwaj, R. & Poria, S. (2024). Ruby teaming: Improving quality diversity search with memory for automated red teaming
AAAI 2024 (under review)
[https://arxiv.org/abs/2406.11654].

5. Li*, Y., Bhardwaj*, R., Ambuj, M. & Poria, S. (2024). Hypertts: Parameter efficient adaptation in text to speech using hypernetworks.
COLING 2024
[https://github.com/declare-lab/HyperTTS].

6. Bhardwaj, R. & Poria, S. (2023a). Language model unalignment: Parametric red-teaming to expose hidden harms and biases.
arXiv 2023
[https://github.com/declare-lab/red-instruct].

7. Bhardwaj, R. & Poria, S. (2023b). Red-teaming large language models using chain of utterances for safety-alignment
[https://github.com/declare-lab/red-instruct].

8. Bhardwaj, R., Vaidya, T. & Poria, S. (2023). Adapter pruning using tropical characterization.
EMNLP 2023
[https://aclanthology.org/2023.findings-emnlp.116/].

9. Bhardwaj, R., Li, Y., Majumder, N. & Poria, S. (2023). Knn-cm: A non-parametric inference-phase adaptation of parametric text classifiers.
EMNLP 2023.

10. Bhardwaj, R., Polovets, G. & Sunkara, M. (2023). Adaptation approaches for nearest neighbor language models.
ACL 2023
(work done at Amazon AWS AI, California.)
[https://aclanthology.org/2023.findings-acl.73.pdf].

11. Hong*, P., Bhardwaj*, R., Majumder, N. & Poria, S. (2023). A robust information-masking approach for generating domain counterfactuals.
ACL 2023
[https://aclanthology.org/2023.findings-acl.231].

12. Li, Y., Mehrish, A., Bhardwaj, R., Majumder, N., Cheng, B., Zhao, S., … Poria, S. (2023). Evaluating parameter-efficient transfer learning approaches on sure benchmark for speech

understanding.
ICASSP 2023
[https://ieeexplore.ieee.org/document/10095656].

13 Bhardwaj, R., Saha, A. & Hoi, S. C. (2022). Vector-quantized input-contextualized soft prompts for natural language understanding.
EMNLP 2022
(work done at Salesforce research, Singapore.)
[https://aclanthology.org/2022.emnlp-main.455/].

14 Bhardwaj, R., Vaidya, T. & Poria, S. (2022a). Federated distillation of natural language understanding with confident sinkhorns.
COLING 2022 (presentation)
[https://arxiv.org/abs/2110.02432].

15 Bhardwaj, R., Majumder, N., Poria, S. & Hovy, E. (2021). More identifiable yet equally performant transformers for text classification.
ACL 2021
[https://github.com/declare-lab/identifiable-transformers].

16 Pan*, J., Bhardwaj*, R., Wei, L., Chieu, H. L., Pan, X. & Puay, N. y. (2019). Twitter homophily: Network based prediction of user's occupation.
ACL 2019
[https://www.aclweb.org/anthology/P19-1252/].

Journals

1 Mehrish, A., Majumder, N., Bharadwaj, R., Mihalcea, R. & Poria, S. (n.d.). A review of deep learning techniques for speech processing
Information Fusion, 2023
[https://www.sciencedirect.com/science/article/abs/pii/S1566253523001859].

2 Bhardwaj, R., Vaidya, T. & Poria, S. (2022b). Towards solving nlp tasks with optimal transport loss.
Journal of King Saud University-Computer and Information Sciences, 2022 (impact factor 8.839)
[https://www.sciencedirect.com/science/article/pii/S1319157822003986].

3 Bhardwaj, R., Majumder, N. & Poria, S. (n.d.). Investigating gender bias in bert
Cognitive Computation, 2021
[https://arxiv.org/abs/2009.05021].

4 Poria, S., Majumder, N., Hazarika, D., Ghosal, D., Bhardwaj, R., Jian, S. Y. B., …, Chhaya, N. et al. (n.d.). Recognizing emotion cause in conversations
Cognitive Computation, 2021
[https://arxiv.org/abs/2012.11820].

5 Majumder*, N., Bhardwaj*, R., Poria, S., Zadeh, A., Gelbukh, A., Hussain, A. & Morency, L.-P. (n.d.). Improving aspect-level sentiment analysis with aspect extraction
Neural Computing and Applications, 2020
[https://arxiv.org/abs/2005.06607].

(In total, I have more than 25 published research papers with over 15 as first author. Please find more about my past works at: google scholar)

Reviewer & Supervision

- **Reviewer** of top-tier NLP conferences such as NeurIPS 2024, AAAI 2024, ACL 2024, EMNLP 2024, AAAI 2023, EMNLP 2023, ACL 2023, EMNLP 2022, ACL 2022, ACL 2021, NAACL 2021, ACL 2020 and Journals such as Information Fusion (impact factor 13.7) and Neural Computing and Applications (impact factor 4.774). In total, I have served as a reviewer for more than 9 scientific communities.
- **Teaching Assistant** I have been a teaching assistant of two undergraduate level courses—Theory & Practice of Deep Learning and Information Retrieval (IR).
- **Supervisor** I have been an official supervisor of two outstanding undergrad students Mridula Ratheesh Thekkedath and Phang Teng Fone working on Complex Question Answering.

Skills

Coding	Python, Matlab, C, Java, LaTeX.
Machine Learning	PyTorch, Tensorflow, Keras, Pandas, Numpy, SciPy, NLTK.

Awards

- **Kaggle (Sentiment Analysis)** 2^{nd} rank in Shopee Sentiment Analysis Challenge.
  Link: `https://www.kaggle.com/c/shopee-sentiment-analysis/leaderboard`
- **Kaggle (Product Detection)** 2^{nd} rank in Shopee Product Detection Challenge.
  Link: `https://www.kaggle.com/c/shopee-product-detection-open/leaderboard`
  (Both the Kaggle competitions were hosted amongst 6 countries)
- **MITACS Globalink** research internship award for research at University of Alberta, Edmonton, Canada (2017).