Rachneet Sachdeva

NLP Researcher, Frankfurt, Germany
rachneet.sachdeva@tu-darmstadt.de
Portfolio | LinkedIn | Github

Ph.D. student at UKP Lab, TU Darmstadt. Experience in explanability, safety, and rigorous evaluation of large language model (LLM) capabilities at scale.

EDUCATION

Ph.D. Student (Computer Science), UKP Lab, TU Darmstadt -
Advised by Prof.'in Dr. Iryna Gurevych September 2021-current

Master of Science, RWTH Aachen University 1.5/5.0
Electrical engineering with a focus on machine learning and telecommunications September 2017-August 2021

Bachelor of Engineering, Panjab University 1.8/5.0
Electronics and Communications Engineering August 2011-May 2015

PUBLICATIONS

Turning Logic Against Itself: Probing Model Defenses Through Contrastive Questions
__Rachneet Sachdeva__, Rima Hazra, Iryna Gurevych **Preprint 2025**

- Introduced POATE, a jailbreak attack using contrastive reasoning to bypass LLM safety.
- Achieved 44% higher attack success rates across six major LLMs, including GPT-4 and LLaMA3.
- Bypassed seven state-of-the-art LLM defense mechanisms.
- Proposed a chain-of-thought prompting defense to mitigate POATE attacks.

Localizing and Mitigating Errors in Long-form Question Answering
__Rachneet Sachdeva__, Yixiao Song, Mohit Iyyer, Iryna Gurevych **Preprint 2024**

- First hallucination dataset with localized error annotations for human and LLM-generated long-form answers.
- 1.8k span-level error annotations across five error types to analyze shortcomings in long-form answers.
- Trained a feedback model to detect errors and provide justifications.
- Developed an error-informed refinement method to reduce errors using model feedback.

Are Emergent Abilities in Large Language Models just In-Context Learning?
Sheng Lu, Irina Bigoulaeva, __Rachneet Sachdeva__, Harish Tayyar Madabushi, Iryna Gurevych **ACL 2024**

- Challenged the concept of "emergent abilities" in LLMs, attributing them to known underlying competencies.
- Proposed a novel theory explaining emergent abilities as a combination of in-context learning, model memory, and linguistic knowledge.
- Validated this theory with 1000+ experiments, revealing key confounding factors in LLM evaluation.
- Provided practical insights for efficient LLM deployment, preventing inflated capability assessments.

CATfOOD: Counterfactual Augmented Training for Improving Out-of-Domain Performance and Calibration
__Rachneet Sachdeva__, Martin Tutek, Iryna Gurevych **EACL 2024**

- Proposed a methodology to generate diverse counterfactual (CF) training data using LLMs.
- Consistently improved out-of-domain (OOD) performance and calibration of models with CF augmentation.

UKP-SQuARE v2: Explainability and Adversarial Attacks for Trustworthy QA
__Rachneet Sachdeva__, Haritz Puerto, Tim Baumgärtner, Sewin Tariverdian, Hao Zhang, Kexin Wang, Hossain Shaikh Saadi, Leonardo FR Ribeiro, Iryna Gurevych **AACL 2022**

- Designed a framework for explaining model predictions using saliency maps and graph-based explanations.

- Integrated adversarial attack techniques to evaluate and enhance model robustness.

UKP-SQUARE: An Online Platform for Question Answering Research

*Tim Baumgärtner, Kexin Wang, **Rachneet Sachdeva**, Gregor Geigle, Max Eichler, Clifton Poth, Hannah Sterz, Haritz Puerto, Leonardo F. R. Ribeiro, Jonas Pfeiffer, Nils Reimers, Gözde Şahin, Iryna Gurevych* **ACL 2022**

- Co-developed UKP-SQuARE, an extensible QA platform to explore and compare language model capabilities.
- Designed a framework enabling seamless deployment and evaluation of user-trained models.

RESEARCH EXPERIENCE

Ubiquitous Knowledge Processing Lab, TU Darmstadt

Ph.D. Student September 2021-current

- Co-led development of UKP-SQuARE, a QA platform for exploring and comparing language model capabilities.
- Introduced novel LLM-based counterfactual data augmentation, improving out-of-domain performance and calibration of small language models.
- Created a hallucination dataset with expert-annotated span-level errors in long-form answers (human & LLM generated).
- Designed a robust jailbreak attack to expose the vulnerabilities of SOTA LLMs to reasoning-based threats.

Institute for Networked Systems, RWTH Aachen University

Research Student March-December 2020

- Developed deep learning models for classifying over-the-air signal modulations in real-world scenarios.
- Wrote a journal paper with our findings.

Computation Social Sciences and Humanities Institute, RWTH Aachen University

Student Research Assistant May 2018-April 2020

- Developed machine learning algorithms to analyze gender bias effects on online social platforms.
- Published our research paper in the journal of Frontiers in Big Data (2022).

INDUSTRY EXPERIENCE

Convaise

Machine Learning Engineer (Intern) February-June 2021

- Developed a centralized hub to deploy and use state-of-the-art machine learning models.
- Trained state-of-the-art language models for machine translation, text summarization, and QA tasks.

Infosys Limited

Systems Engineer (Full-time) June 2015-August 2017

- Worked in an agile environment to automate workflows from development to testing using DevOps.
- Developed automated test cases using Selenium to find potential flaws in organization workflows.

POSITIONS OF RESPONSIBILITY

- **Reviewer** for ACL Rolling Review (ARR).
- **Supervisor** for bachelor and master thesis students at UKP Lab, TU Darmstadt.
- **Teaching Assistant** for *NLP Ethics* course at the master level (TU Darmstadt).
- **Instructor** for *Data Analysis Software Project for Natural Language* course at master level (TU Darmstadt).
- **Event Manager** at *Teach a Child*. Spearheaded fundraising and educational initiatives for underprivileged children through impactful events.

SKILLS

- **Programming:** Python (10 years), Java, C, C++
- **Frameworks:** PyTorch, Transformers, Scikit-Learn, XGBoost, Pandas, Numpy, Matplotlib/Seaborn
- **Developer Tools:** Docker, AWS, GitHub, LaTeX
- **Natural Languages:** English, German (A2), Hindi, Punjabi, Spanish (A1), Korean (A1)