Rachneet Sachdeva

AI/NLP Researcher, Frankfurt, Germany
rachneet.sachdeva@tu-darmstadt.de | +49 176 47194617
Portfolio | LinkedIn | Github

Ph.D. researcher in NLP and AI with 5+ years of experience building safe, explainable, and production-ready large language model systems. Strong engineering background in scalable ML infrastructure, adversarial robustness, and hallucination mitigation. Proven ability to translate cutting-edge research into real-world, high-impact applications.

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

RESEARCH EXPERIENCE

Ubiquitous Knowledge Processing Lab, TU Darmstadt
Ph.D. Student September 2021-current

- Co-led the development of **UKP-SQuARE**, a scalable QA evaluation platform integrating LLMs; used by **1000+ users** for live deployment of custom models with explainability and adversarial testing features.
- Designed a **contrastive reasoning-based jailbreak attack** against GPT-4, LLaMA3, and others, achieving a **40% increase** in attack success over baselines; proposed an effective defense using chain-of-thought prompting.
- Built a span-level **hallucination detection dataset** (1.8k+ annotations); trained error-detection models and implemented an **LLM feedback loop** to reduce errors in long-form QA.
- Led counterfactual data augmentation experiments using RAG-based LLM pipelines to improve **out-of-domain generalization** by 4% and **model calibration** accuracy by 5% on QA tasks.
- Advised BSc/MSc thesis students and served as TA/instructor for graduate-level NLP courses.

Institute for Networked Systems, RWTH Aachen University
Research Student March-December 2020

- Trained deep learning models to classify wireless signal modulations from real-world noisy SDR data.
- Co-authored a journal publication outlining the model architecture and signal preprocessing pipeline.

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 in the journal of *Frontiers in Big Data (2022)*.

INDUSTRY EXPERIENCE

Convaise
Machine Learning Engineer (Intern) February-June 2021

- Developed an internal platform to **fine-tune and deploy SOTA language models** (e.g., T5, BART) in the AWS cloud with a **single API call**; reduced the deployment effort from 2 days to 10 minutes.
- Built pipelines for training **translation, summarization, and QA models**, integrating evaluation metrics and version control; reduced manual training setup time from several hours to minutes.
- Improved model inference time by optimizing batching and caching strategies in the backend service.

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

- Automated Salesforce UI testing using Selenium, boosting test coverage and reducing manual QA effort.
- Designed CI/CD pipelines using Jenkins to streamline DevOps workflows from code commit to deployment.

TECHNICAL SKILLS

- **Programming:** Python (10+ yrs), Java, C/C++, SQL
- **ML/NLP Frameworks:** PyTorch, HuggingFace Transformers, Scikit-learn, XGBoost, LangChain
- **Developer Tools:** Docker, Kubernetes, GitHub, FastAPI, AWS, Azure, MongoDB
- **Other:** LaTeX, Pandas, NumPy, Matplotlib/Seaborn
- **Natural Languages:** English, German (A2), Hindi, Punjabi, Spanish (A1), Korean (A1)

SELECTED 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 40% higher attack success rates than baselines on six major LLMs, including GPT-4 and LLaMA3.
- Bypassed seven state-of-the-art LLM defense mechanisms, demonstrating POATE's robustness.
- Proposed a chain-of-thought prompting defense that effectively mitigates POATE-style jailbreaks.

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.

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.