Rachneet Sachdeva

+49 176 47194617 | rachneet1993@gmail.com | linkedin.com/in/rachneetsachdeva | github.com/Rachneet | rachneet.github.io

Ph.D. researcher in Natural Language Processing (NLP) and AI with 5+ years of applied research and R&D experience, focused on shipping safe, explainable, and production-ready large language model systems. Experienced in scalable NLP infrastructure, adversarial robustness, and low-latency AI services. Proficient in Python and modern ML frameworks, with excellent verbal and written communication skills.

Work Experience

Ph.D. Student — Sep 2021 - Present

Ubiquitous Knowledge Processing Lab, TU Darmstadt | *Darmstadt*

- Co-led the collaborative 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 **RAG-based counterfactual augmentation** experiments, improving language models' **out-of-domain generalization by 4%** and **calibration accuracy by 5%**.
- **DocChat: Multi-Agent RAG System** - Built a document analysis tool using hybrid retrieval (BM25 + vector search) with verification agents to eliminate hallucinations and extract accurate information from complex PDFs.

Machine Learning Engineer (Intern) — Feb 2021 - Jun 2021

Convaise | *Munich*

- 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**.
- Contributed to the research and development of pipelines for training **translation, summarization, and QA models**, integrating evaluation metrics and version control; **reduced manual training setup time by 90%** (from 4 hours to under 20 minutes).
- **Improved model inference time by 50%** through optimized batching and caching strategies in the backend service.

Research Assistant — May 2018 - Apr 2020

CSSH Institute, RWTH Aachen University | *Aachen*

- Processed **80 million+ Amazon reviews** by **21 million users** across **9 million products**, providing a large-scale dataset for gender bias analysis on online review platforms.
- Applied deep learning algorithms to **infer author gender for reviews** lacking explicit name signals, achieving **82% precision** and extending bias detection to previously unlabeled data.

Systems Engineer — Jun 2015 - Aug 2017

Infosys Limited | *Chandigarh*

- Automated Salesforce UI testing using Selenium, boosting test coverage and **reducing manual QA effort by more than 90%**.
- Architected reliable Jenkins-based CI/CD pipelines, **increasing the deployment frequency by 400%** - from weekly to daily releases.

Core Skills

- **Programming Languages:** Python (10+ yrs), Java, C/C++, SQL
- **ML/NLP Frameworks:** PyTorch, TensorFlow, HuggingFace Transformers, Scikit-learn, SpaCy, XGBoost, LangChain, LangGraph, LangSmith, LlamaIndex, MCP (Model Context Protocol), Weights and Biases
- **Developer Tools:** Docker, Kubernetes, GitHub, LaTeX, FastAPI, AWS (Sagemaker, S3), Azure, MongoDB
- **Libraries:** Pandas, NumPy, Pydantic, Matplotlib
- **Natural Languages:** English, German (A2), Hindi, Punjabi, Spanish (A1), Korean (A1)

Education

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

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

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

Selected Publications

Turning Logic Against Itself: Probing Model Defenses Through Contrastive Questions
***Rachneet Sachdeva**, Rima Hazra, Iryna Gurevych* **EMNLP 2025**

- Introduced POATE, a jailbreak attack using contrastive reasoning to bypass LLM safety.
- Achieved 40% higher attack success rates than baselines on 6 major LLMs, including GPT-4 and LLaMA3.
- Bypassed 7 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* **ACL 2025**

- First hallucination dataset with localized error annotations for human and LLM-generated long-form answers.
- 1.8k span-level error annotations across 5 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's and master's thesis students at UKP Lab, TU Darmstadt.
- **Teaching Assistant** for the *NLP Ethics* course; taught 100+ bachelor and master students from diverse academic backgrounds.
- **Instructor** for the *Data Analysis Software Project for Natural Language* course at the master's level (TU Darmstadt).
- **Event Manager** at *Teach a Child*; led fundraising and educational initiatives with a team-first approach to support underprivileged children.
- **Mentored 13 BSc/MSc students** and led collaborative research efforts, demonstrating leadership, teamwork, and interpersonal skills in academic settings.