SHAILZA JOLLY

Berlin, Germany

shailzajolly@gmail.com in shailzajolly Scholar +49 176 67279750

PROFILE

Senior ML Scientist/Engineer with a Ph.D. in CS and 6+ years spanning academic research and industry. Specializes in LLMs, Generative AI, and scalable data pipelines. Brings hands-on depth in modern LLM architectures, efficiency techniques (KV-cache, speculative decoding, FlashAttention), and agentic system design, alongside a track record of top-tier publications (AAAI, NAACL, EMNLP) and production delivery.

SKILLS

- **ML/GenAI:** LLMs, NLU/NLG, multimodal learning, synthetic data, model evaluation.
- **LLM Systems:** Agentic workflows (MCP, LangGraph/LangSmith), RAG/grounding concepts, experiment design & offline evaluation.
- **Frameworks & Libraries:** Python, PyTorch, Hugging Face (Transformers, PEFT, Datasets), PySpark; Weights & Biases (W&B).
- **Search & Retrieval:** Embedding models (e.g., Sentence Transformers), reranking, vector search; Pinecone, Chroma.
- **Infrastructure & Deployment:** AWS, Docker, FastAPI, Git.

WORK EXPERIENCE

Career Break / Independent Study July 2025 - Present
Generative AI & LLM Systems *Berlin, Germany*

- Deepened expertise in modern LLM architectures and efficiency: FlashAttention, RoPE, scaling laws, KV-cache, speculative decoding, and fast inference techniques.
- Studied agentic system design (tool use, MCP) and evaluation strategies for LLM/agentic systems; built small prototypes to test concepts.

Flinn.ai January 2025 - June 2025
Senior AI Engineer *Berlin, Germany*

- Led development of AI-driven product features including data extraction from medical research papers and multilingual complaint monitoring.
- Partnered with product and backend teams to scope problems, define success metrics, and deliver end-to-end features within an agile product lifecycle.
- *Role eliminated due to company-wide strategic shift.*

Parental Leave (Maternity) December 2023 - December 2024

- Parental leave; stayed engaged with research literature and maintained technical skills.

Amazon AI July 2022 - November 2023
Research Scientist *Berlin, Germany*

- Led development of a scalable noise-removal/data-quality pipeline processing billions of tokens to improve training data for LLMs.
- Wrote research plans and technical documentation; presented results and recommendations to senior science/engineering leadership.
- Mentored Master's/Ph.D. interns and collaborated on research deliverables.

German Research Center for Artificial Intelligence (DFKI) March 2019 - June 2022
Machine Learning Scientist *Berlin, Germany*

- Delivered ML research and engineering (coding, experimentation, publications) in BMBF-funded projects: *XAINES* (Explainable AI) and *DeFuseNN* (vision-language systems).
- Supervised interns and BSc/MSc students; provided technical mentorship and project guidance.

NVIDIA Research May - August 2021
Machine Learning Scientist Intern *Santa Clara, United States*

· Built a document understanding pipeline combining table/cell detection, tabular structure retrieval, and OCR for financial documents.

Amazon Alexa August 2019 - December 2019
Applied Scientist Intern *Aachen, Germany*

· Developed a method for generating diverse synthetic training data to improve intent classification and slot labeling for task-oriented NLU.

SAP AI Research May 2018 - February 2019
Research Intern / Master's Thesis *Berlin, Germany*

· Designed and implemented an evaluation metric for Visual Question Answering (VQA) models.

EDUCATION

TU Kaiserslautern, Germany March 2019 - June 2022
Ph.D. in Computer Science
Building Natural Language Generation and Understanding Systems in Data-Constrained Settings; work on VQA, interpretability, and conversational AI.
Overall grade: **"Sehr Gut" 1.0** (1.0 is highest on 1.0–5.0 scale)

University of Copenhagen, Denmark January 2021 - March 2021
Visiting Researcher
Developed an unsupervised post-editing algorithm to generate fluent fact-checking explanations.

TU Kaiserslautern, Germany October 2016 - November 2018
M.Sc. in Computer Science — Minor in Economics
Overall grade: **"Sehr Gut" 1.5**

Kyushu University, Japan November 2017 - February 2018
Semester Abroad
Explainable AI to analyze behavior of deep CNN architectures for image recognition.

Guru Nanak Dev Engineering College, India August 2012 - July 2016
B.Tech in Computer Science & Engineering
Overall grade: **First Division with Distinction**

SELECTED PUBLICATIONS

- **Jolly, S.**, Zhang, Z. X., Dengel, A., & Mou, L. (2022). Search and learn: Improving semantic coverage for data-to-text generation. *AAAI.*
- **Jolly, S.**, Pezzelle, S., & Nabi, M. (2021). EaSe: A diagnostic tool for VQA based on answer diversity. *NAACL-HLT.*
- **Jolly, S.**, & Kapoor, S. (2020). Can pre-training help VQA with lexical variations? *Findings of EMNLP.*
- **Jolly, S.**, Falke, T., Tirkaz, C., & Sorokin, D. (2020). Data-efficient paraphrase generation to bootstrap intent classification and slot labeling. *COLING (Industry Track).*
- **Jolly, S.**, Iwana, B. K., Kuroki, R., & Uchida, S. (2018). How do convolutional neural networks learn design? *ICPR.* (*Best Student Paper*)

AWARDS AND HONORS

- Received **AAAI-22 Scholarship** by Hitachi to attend AAAI 2022 (2022).
- Awarded **AI Newcomer 2021 Award** (German Informatics Society and BMBF, Germany) (2021).
- Received **EU-Cost STSM Grant** to work on Multi3generation project at CopeNLU (2020).
- **Best Student Paper Award** at ICPR (2018).