Hassan Soliman

hassansoliman.com/ | hassanmahmoudsd@gmail.com | **Location**: Mannheim, Germany
linkedin.com/in/hassansolimann/ | github.com/hassanmahmoudd | **Mobile**: (+49) 1782554815

Experience

German Research Center for AI (DFKI) — Berlin, Germany
AI Researcher — *Jan 2023 - Present*

- Led two projects in the Educational Technology lab, managing technical implementation and supervising two students.
- Developed a chatbot for a graduate-level course that answered student queries with **87%** accuracy. One of the two papers published on the project was nominated for the **Best Demo Award** at **ECTEL 2024**.
- Applied advanced Retrieval-Augmented Generation **(RAG)** techniques, including Hybrid Ensemble Search and Reranking Mechanism, to enhance chatbot interactions and improve the retrieval of course materials.
- Supported mentoring-style conversations by leveraging flexible **agentic workflows** with LangGraph, utilizing multiple small **open-source** models hosted on **Azure**, and using databases for user usage tracking and monitoring.
- Implemented a sub-module for adaptive dialogue systems, customizing responses based on user emotional state and demographics, and benchmarking performance using **OpenAI** LLMs and **open-source** models.

Bosch Center for AI (BCAI) — Renningen, Germany
Applied Scientist Intern — *May 2022 - Aug 2022*

- Contributed to the NLP & Semantic Reasoning group, applying findings from my master's thesis on Neural Entity Linking to a high-impact industrial project using real data at Bosch.
- Refactored, tested, and documented **production-level** code for machine learning models, ensuring scalability and efficiency for real-world deployment, leveraging the in-house **GPU cluster** for model fine-tuning.
- Trained and evaluated machine learning models on a **large-scale** domain-specific dataset, achieving **77%** end-to-end recall for top-3 entity predictions, outperforming existing models.

Bosch Center for AI (BCAI) — Renningen, Germany
Master's Thesis Student — *Jun 2021 - Jan 2022*

- Joined the NLP & Semantic Reasoning group and worked on a unified system for linking named entities to general-domain (Wikipedia) and domain-specific knowledge bases (KBs), using context-aware embeddings (BERT) to learn a joint vector space. A pre-print of the thesis is available on arXiv. https://arxiv.org/abs/2210.15616.
- Optimized a state-of-the-art model for cross-domain applications, supporting domain extension and identifying optimal data sources for **fine-tuning**, and improved **GPU memory** utilization for efficient embedding calculations.
- Achieved a **9%** increase in Average Precision for the top-1 entity and a **20%** gain in Mean Average Precision (MAP) for top-10 entity linking across four domain-specific KBs, resulting in a workshop publication at **ACL 2022**.

Max Planck Institute for Informatics (MPII) — Saarbrücken, Germany
Research Assistant — *Nov 2020 - May 2021*

- Developed a model prototype within the Database & Information Systems group to identify diverse peer groups for entities, contributing to advanced **set expansion** techniques.
- Implemented a baseline model for entity set expansion, leveraging **Wikipedia lists** as a knowledge source to enhance the model's accuracy and comprehensiveness in the expanded sets.
- Optimized the algorithm's performance by achieving a **3x** faster runtime using efficient sparse matrix multiplication techniques, significantly improving computational efficiency.

Amazon — Luxembourg, Luxembourg
Software Development Engineer Intern — *Aug 2019 - Feb 2020*

- Maintained a web-based simulation tool for the Fulfillment Acceleration team using the **AWS** cloud platform, working as a full-stack software engineer.
- Enhanced **delivery speed simulations** for prime customers, contributing to a successful report on fulfillment operations and improving delivery efficiency.
- Collaborated as a system administrator in an Agile environment, managing **server infrastructure** and providing technical support for team tools.

Education

Technische Universität Darmstadt, Ubiquitous Knowledge Processing Lab — Darmstadt, Germany
Doctor of Philosophy (Ph.D.) in Computer Science; Supervised by: Prof. Dr. Iryna Gurevych — *Jan 2025 - Dec 2028*

Universität des Saarlandes, Saarland Informatics Campus — Saarbrücken, Germany
Master of Science (M.Sc.) in Computer Science; GPA: 1.40 / 1.00 — *Oct 2018 - Sep 2022*

Alexandria University, Faculty of Engineering — Alexandria, Egypt
Bachelor of Science (B.Sc.) in Computer and Communication Engineering; GPA: 3.96 / 4.00 — *Sep 2013 - Sep 2018*

Skills

Fields: Machine Learning, Deep Learning, Natural Language Processing (NLP), Natural Language Understanding, Large-Scale Language Modeling, Generative AI, Conversational AI, Dialogue Systems, Neural Machine Translation, Information Extraction, Data Analysis, Data Visualization, Distributed Computing.

Technologies and Libraries: LangChain, LangGraph, LangSmith, LlamaIndex, HuggingFace, Transformers, Scikit-Learn, Keras, PyTorch, TensorFlow, Pandas, NumPy, Matplotlib, Leaflet, Docker, Kubernetes, Airflow, Git, Jira.

Programming and Databases: Python, R, Java, C++/C, Java Spring, Angular.js, SQL, MongoDB, Redshift, DynamoDB.

Projects

SmolLM: Implementing, Fine-Tuning, and Aligning a LLM for Grammatical Error Correction *Code*

- Implemented the **SmolLM-135M** (by **HuggingFace**) language model architecture, including components like Rotary Positional Embeddings, KV Cache, and Grouped-Query Attention, RMS Normalization, and SwiGLU Activation.
- Fine-tuned the model on the **Grammatical Error Correction (GEC)** task using the **Grammarly CoEdIT** dataset.
- Applied **RLAIF** through Direct Preference Optimization **(DPO)** to align model outputs with desired corrections.
- Created a **Colab notebook** to guide users through implementation, fine-tuning, and evaluation processes.
- Achieved significant improvements in grammatical error correction accuracy, scoring an expected **BLUE** score of ~ 0.48.
- Leveraged Python libraries such as PyTorch, Transformers, Datasets, and TRL to build and train the model effectively.

LinguaLexMatch: Enhanced Document Language Detection *Code*

- Developed and evaluated three **language detection** models, including an **Embedding-Based** approach, a **TF-IDF-based Multinomial Naive Bayes** model, and a fine-tuned **Transformer-Based** methodology.
- Implemented an embedding-based approach using the **intfloat/multilingual-e5-large-instruct** model by generating a representative embedding for each language and classifying documents based on cosine similarity.
- Benchmarked models on the **papluca/language-identification** dataset, achieving **99.81% accuracy** with the embedding-based model.
- Analyzed performance metrics such as **Accuracy**, **F1 Scores**, and **Confusion Matrices** across 20 different languages.
- Developed a **Colab notebook** for replicable implementation and evaluation of different language detection models.
- Utilized Python libraries including Datasets, Transformers, and Scikit-learn for model development and evaluation.

Publications

Scalable Mentoring Support with a Large Language Model Chatbot — *ECTEL, 2024*

- **Hassan Soliman**, Miloš Kravčík, Alexander T. Neumann, Yue Yin, Norbert Pengel, Maike Haag. *Technology Enhanced Learning for Inclusive and Equitable Quality Education.* https://doi.org/10.1007/978-3-031-72312-4_37

Generative KI zur Lernenbegleitung in den Bildungswissenschaften — *DELFI, 2024*

- **Hassan Soliman**, Miloš Kravčík, Alexander T. Neumann, Yue Yin, Norbert Pengel, Maike Haag, Heinz-Werner Wollersheim. *22. Fachtagung Bildungstechnologien.* https://doi.org/10.18420/delfi2024_15

Using Large Language Models for Adaptive Dialogue Management in Digital Telephone Assistants — *ACM UMAP, 2024*

- **Hassan Soliman**, Miloš Kravčík, Nagasandeepa Basvoju, Patrick Jähnichen. *32nd ACM Conference on User Modeling, Adaptation and Personalization (UMAP Adjunct '24).* https://doi.org/10.1145/3631700.3664902

A Study on Entity Linking Across Domains: Which Data is Best for Fine-Tuning? — *ACL RepL4NLP, 2022*

- **Hassan Soliman**, Heike Adel, Mohamed H. Gad-Elrab, Dragan Milchevski, Jannik Strötgen. *7th Workshop on Representation Learning for NLP, ACL.* https://doi.org/10.18653/v1/2022.repl4nlp-1.19

Awards

Second Best Demo Paper — *Presented a demo paper on an LLM-based Chatbot* *ECTEL, 2024*
First Class Honor Degree — *Outstanding Academic Performance* *Alexandria University, 2018*

References

Dr. Miloš Kravčík — *Senior Researcher, DFKI GmbH* *milos.kravcik@dfki.de*
Prof. Dr. Heike Adel — *University of Applied Sciences, Stuttgart* *adel-vu@hdm-stuttgart.de*
Dr. Mohamed H. Gad-Elrab — *Research Engineer, BCAI* *mohamed.gad-elrab@de.bosch.com*