HAAU-SING (XIAOCHENG) LI

Tel: (+351) 910-414-857 | Email: hli@ukp.tu-darmstadt.de | LinkedIn: haau-sing-li-152177142 | GitHub: lhaausing

EDUCATION

European Lab for Learning and Intelligent Systems (ELLIS) Darmstadt, Germany & Lisbon, Portugal
Ph.D. in Computer Science, Technische Universität Darmstadt *July 2021 – Now*
ELLIS Exchange at Instituto de Telecomunicações *July. 2023 – June. 2024*
Supervisors: Iryna Gurevych, André F. T. Martins
Research Topics: AI for programming languages, inference of LLMs, reranking, quality estimation.

Center for Data Science, New York University New York, NY
M.S. in Data Science *Sept. 2019 – May 2021*
GPA: 4.0/4.0 **Awards**: Moore Sloan Summer Research Initiative **Teaching**: Machine Learning (Spring 2021)
Selected Courses: Computer Vision, Deep Learning, Inference & Representations, NLP, NLU, Machine Learning.

Renmin University of China Beijing, China
B.A. in English Linguistics, with coursework in Computer Science *Sept. 2015 – June 2019*
GPA: 3.7/4.0; NLP related 3.82/4.0 **Awards**: Graduate with Distinction, ICM Meritorious Award
Selected Courses: Computational Linguistics, Optimization, Data Structures, Discrete Mathematics, Stochastic Process.

PAPERS

Haau-Sing Li, Patrick Fernandes, Iryna Gurevych, André F.T. Martins. *DOCE: Finding the sweet spot for execution-based code generation.* **under review**.
António Farinhas, **Haau-Sing Li**, André F.T. Martins. *Reranking laws for language generation: A communication-theoretic perspective.* **under review**.
Joris Baan*, Nico Daheim*, Evgenia Ilia*, Dennis Ulmer*, **Haau-Sing Li**, Raquel Fernández, Barbara Plank, Rico Sennrich, Chrysoula Zerva, Wilker Aziz. *Uncertainty in natural language generation: From theory to applications.* **under review**.
Haau-Sing Li, Mohsen Mesgar, André F.T. Martins, Iryna Gurevych. *Python code generation by asking clarification questions.* **ACL 2023**.
Yian Zhang*, Alex Warstadt*, **Haau-Sing Li**, Samuel R. Bowman. *When do you need billions of words of pretraining data?* **ACL-IJCNLP 2021**.
Alex Warstadt, Yian Zhang, **Haau-Sing Li**, Haokun Liu, Samuel R. Bowman. *Learning Which Features Matter: RoBERTa Acquires a Preference for Linguistic Generalizations.* **EMNLP 2020**.

SKILLS

Programming: Python, C/C++, SQL **Maching Learning**: Pytorch, Lightning, vLLM, TRL, DeepSpeed, Transformers

RELATED EXPERIENCE

UKP Lab & SARDINE Lab (under ELLIS PhD Program) Darmstadt & Lisbon
Research Assistant, with Prof. Iryna Gurevych and Prof. André F.T. Martins *July 2021 – Now*
Research Topic: Code Generation, Decoding of LLMs, Reranking, Dialogue
- Proposed framework of sampling candidates, reranking, and reranking upper bound improvement on execution-based code generation with state-of-the-art performance. Paper under review.
- Performed experiments to test communication-theory-based reranking laws on execution-based code generation with Minimum Bayes Risk decoding based on execution outputs tested. Paper under review.
- Formulated clarifications on text-to-code generation about missing API calls using control flow graphs of API usage of code. Built pipeline (classifier, ranker, generator) and showcased effectiveness of clarifications on code generation. Paper accepted at ACL 2023.[code] [paper]

NYU Langone Health New York, NY
Research Assistant, with Prof. Narges Razavian *June 2020 – May 2021*
Research Topic: Biomedical NLP
- Ensembled transformers fine-tuned medical texts and n-grams for electronic health records data with state-of-the-art performances on sentence encoders.[code] [report]
- Pretrained and probed medical encoders with state-of-the-art performance among pretrained models. [code] [report]

ML^2 Group, New York University New York, NY
Research Assistant, with Prof. Sam Bowman *Feb. 2020 – May 2021*
Research Topic: Interpretability, Emergence
- Tested emergent behaviors of RoBERTas on supervised and unsupervised tasks. Paper accepted at ACL 2021. [paper]
- Designed synthetic dataset to test linguistic behaviors of encoders. Scaled training with RoBERTa at different pretraining sizes. Paper accepted at EMNLP 2020. [code] [paper]