HAAU-SING (XIAOCHENG) LI

Personal Page | Twitter | GitHub | LinkedIn | Email | (+351) 910-414-857

EDUCATION

European Lab for Learning and Intelligent Systems (ELLIS) Ph.D. Program — Darmstadt, DE & Lisbon, PT

Ph.D. in Computer Science, TU Darmstadt; Exchange at Instituto de Telecomunicações — *2021 – Now*

- Supervisors: Iryna Gurevych, André F. T. Martins.
- Research Topics: Test-Time Improvement, Code Generation, Reasoning.
- Scholarships: TUDa Ph.D. Scholarship, UTTER (Horizon Europe), ELISE Mobility Program, LxMLS Moderator (2022).

Center for Data Science, New York University — New York, NY

M.S. in Data Science (GPA: 4.0/4.0) — *2019 – 2021*

- Scholarships: Moore Sloan Summer Research Initiative (2020).
- Selected Courses: Deep Learning, Inference & Representations, NLP, NLU, Machine Learning, Computer Vision.

Renmin University of China — Beijing, CN

B.A. in English Linguistics; coursework in Computer Science (GPA: 3.7/4.0; ML-related 3.82/4.0) — *2015 – 2019*

- Awards: Distinctive Graduate (5%), MCM Meritorious Award (10%, 2018), Academic Excellence (10%, 2016-2018).
- Selected Courses: Computational Linguistics, Optimization, Data Structures, Discrete Mathematics, Stochastic Processes.

PAPERS

- Haau-Sing Li, Patrick Fernandes, Iryna Gurevych, André F.T. Martins. *DOCE: Finding the Sweet Spot for Execution-Based Code Generation*. **Preprint**.
- António Farinhas, Haau-Sing Li, André F.T. Martins. *Reranking Laws for Language Generation: A Communication-Theoretic Perspective*. **In Submission**.
- Joris Baan*, Nico Daheim*, Evgenia Ilia*, Dennis Ulmer*, Haau-Sing Li, Raquel Fernández, Babara Plank, Rico Sennrich, Chrysoula Zerva, Wilker Aziz. *Uncertainty in Natural Language Generation: From Theory to Applications*. **In Submission**.
- 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 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**.

WORKING EXPERIENCE

UKP Lab & SARDINE Lab — Darmstadt & Lisbon

Ph.D. student — *July 2021 – Now*

- Proposed framework of test-time compute for code generation, with **state-of-the-art results** and crucial findings on filtering, sampling temperature, and self-debugging on multiple candidates. **Preprint released**.
- Conceptualized reranking laws with communication theory and tested on code generation. **Camera-ready soon**.
- Built synthetic datasets on missing API calls with control flow graph. Built pipeline (classifier, ranker, generator) and showcased effectiveness of clarifications on code generation (**ACL 2023**).

ML^2 Group, New York University — New York, NY

Research Assistant, with Dr. Alex Warstadt and Prof. Samuel R. Bowman — *Feb. 2020 – May 2021*

Research Topics: Interpretability, Emergence.

- Tested emergent behaviors of RoBERTas on supervised and unsupervised tasks (**ACL 2021**).
- Built linguistic dataset to test model behaviors. Scaled pertaining of RoBERTa of different sizes (**EMNLP 2020**).

NYU Langone Health — New York, NY

Research Assistant, with Prof. Narges Razavian — *June 2020 – May 2021*

Research Topic: Biomedical NLP.

- Ensembled fine-tuned medical transformers for electronic health records [code]. Pretrained and probed Longformer with state-of-the-art performance among transformers [code].

SKILLS AND SERVICES

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

Reviewing: ACL (2024), EMNLP (2024), LREC-Coling (2024).

Teaching: Lisbon Machine Learning Summer School (LxMLS) (2022, 2023), Machine Learning (NYU Spring 2021).