Xiaocheng (Haau-Sing) 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 *July 2021 – Dec. 2025 (Expected)*

- Supervisors: Iryna Gurevych, André F. T. Martins.
- Research Topics: LLM Inference, Code Generation, Agent, 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.

Publications

- Haau-Sing Li, Patrick Fernandes, Iryna Gurevych, André F.T. Martins. *DOCE: Finding the Sweet Spot for Execution-Based Code Generation*. **In Submission.**
- António Farinhas, Haau-Sing Li, André F.T. Martins. *Reranking Laws for Language Generation: A Communication-Theoretic Perspective* **NeurIPS 2024 (Spotlight)**.
- 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.**

Research Experience

UKP Lab & SARDINE Lab Darmstadt & Lisbon
Ph.D. student *July 2021 – Present*

- Presented a unified inference framework for code generation, with state-of-the-art results and highlights on filtering, execution-based reranking, and multiple-candidate refinement. **In submission.**
- Proposed a framework linking inference scaling law and communication theory. (**NeurIPS 2024 Spotlight**).
- Proposed a novel approach to build a synthetic dataset that clarifies missing API calls with API control flows. Pipeline showcased the 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.

- Comprehensively studied emergent behaviors of masked language models (MLMs) (**ACL 2021**).
- Created hallucination dataset and tested MLMs. Scaled training of RoBERTas 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.

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