Jialiang Xu

Email: xjl@stanford.edu | Website: liamjxu.github.io | LinkedIn: www.linkedin.com/in/xjl

EDUCATION

Stanford University

Master of Science, Computer Science — *2023-2025*

University of Illinois at Urbana Champaign

Bachelor of Science, Computer Engineering — *2018-2022, GPA 3.98 / 4.0*

Minor, Computer Science — *2019-2021, GPA 4.0 / 4.0*

RESEARCH EXPERIENCES

Alexa AI, Amazon Science — *Jun 2023 – Sep 2023 (Expected)*
Applied Scientist Intern, *Alexa Proactive Experience* — *Seattle, WA*

- Working on domain-specific applications of the latest developments in NLP.

Bosch Research — *Mar 2023 – Jun 2023*
Research Intern, *Natural Language Processing* — *Sunnyvale, CA*

- Adapted open-sourced Large Language Models (LLM) to domain-specific chat applications. The motivation behind is to create open-sourced alternatives for popular close-sourced chat LLMs, therefore alleviating concerns for private data security observed in the common API-based approach.
- Finetuned a series of Flan-T5 models which outperform the vanilla version and popular API-based chat systems on corporate-owned data. Applied techniques of domain adaptation and prompt tuning to inject domain knowledge to LLM while avoiding catastrophic forgetting of the general language capabilities.

Microsoft Research — *Jul 2021 – Jul 2022*
Research Intern, *Data Knowledge Intelligence Group* — *Beijing, China*

- Research: Focused on understanding semi-structured data (e.g., tables, forms, logs) with Natural Language Processing techniques. Produced 3 research papers in the review process for top-tier conferences including ACL, KDD, and EMNLP. The pdf, slides, posters, and code repo can be found at https://liamjxu.github.io/publications/.
- Paper 1 (ACL 2023 findings): Built a model that extracts table field metadata such as the field property, field roles, semantic field type, and default aggregation. Collected a large-scale corpus and proposed a strong baseline for the metadata tasks.
- Paper 2: Proposed a series of techniques to improve tabular models' capability to understand numeracy, including a novel method to tokenize numbers, a novel embedding approach to represent numbers, and a novel pre-training loss that encourages numeracy. Improved results of existing models such as BERT, TAPAS, and RoBERTa on a series of tabular-related datasets such as TabFact, TATQA, and WTQ.
- Paper 3 (EMNLP 2022): *Towards Robust Numerical Question Answering: Diagnosing Numerical Capabilities of NLP Systems*. In this paper, we propose to conduct systematic perturbations to Numerical QA datasets as a probe into the weakness in Language Models' numerical capabilities.
- Tech-Transfer: Cooperated with product teams from Bing, Azure, and Excel on transferring research output into features for Microsoft products including Edge, Synapse Notebook, and Excel. Provided fundamental tools that allow 1) Bing to identify table fields for best visualization, 2) Azure to automate pivot table generation, and 3) Excel to intelligently assist users to generate analysis and visualizations for their spreadsheets.

Blender Lab — *Jul 2022 – Dec 2022*
Undergraduate Research Assistant, *Supervised by Prof. Heng Ji* — *Urbana-Champaign Area, IL*

- Working on the DARPA MIPs Project, conducting research on the linguistic phenomenon of the framing effect.
- Collected a framing dataset of 195,579 news articles from a variety of news agencies annotated with political bias and factuality information for contrastive learning (20 times larger than previous datasets). The articles are clustered into 378 different topics. Multiple baseline models were run, and the empirical results showed that the dataset is challenging to existing models.
- Currently working on a generative contrastive learning framework that improves pre-trained language models' performance on political stance identification in an interpretable manner. Targeting EMNLP 2023 for paper submission, the manuscript is currently in the preparation process.

Cyber Physical Computing Lab *Jan 2021 – May 2021*
Undergraduate Research Assistant, *Supervised by Prof. Tarek Abdelzaher* *Urbana-Champaign Area, IL*

- Worked on the DARPA SocialSim Project, trained a model that ranked the first place in the final evaluation.
- Built pipelines aiming to understand public user behavior on social platforms such as Twitter and YouTube. Trained Linear Regression and LSTM with Scikit-learn and PyTorch to predict statistics such as new user count or new post count based on past social news. Both the input and the output were in the form of time series.

PROFESSIONAL EXPERIENCE

Discovery Partners Institute *Aug 2022 – Dec 2022*
Machine Learning Engineer *Chicago, IL*

- Added Semantic Search functionalities to a multi-source biomedical searching platform 1-Search.
- Finetuned a Biomedical-domain-specific pre-trained language model on a Learning-To-Rank dataset and evaluated the model on both public datasets and an internal dataset collected from 1-Search. Implemented pipelines for two downstream functionalities. Improved the inference latency from over 3 seconds to less than 2 seconds by utilizing dynamic model quantization. Served the model on an Azure virtual machine with TorchServe.

Ansys Inc. *May 2020 – Aug 2020*
Software Development Intern, *Urbana-Champaign Area, IL*

- Added a series of new features to the main EM simulation software, Ansys Electromagnetic Desktop (AEDT). Built and Maintained Python modules facilitating the automation of user project transplantation between the classic EMIT toolkit and the newest AEDT desktop. Features include parameterized component importing, port connectivity, orientation match, RF system configuration combination forming, relative schematic positioning, and others. Assisted FTEs with API modification and unit & regression testing for multiple iterations.
- Modules deployed in the latest Ansys 2022R1, which is used by the whole Ansys customer community, enabling customers to shift to the new platform seamlessly.

PUBLICATIONS

Peer-reviewed Conference and Journal Publications

[P4] Towards Robust Numerical Question Answering: Diagnosing Numerical Capabilities of NLP Systems
Jialiang Xu, Mengyu Zhou, Xinyi He, Shi Han, Dongmei Zhang
EMNLP 2022

[P3] Inferring Tabular Analysis Metadata by Infusing Distribution and Knowledge Information
Xinyi He, Mengyu Zhou, **Jialiang Xu**, Xiao Lv, Tianle Li, Yijia Shao, Shi Han, Zejian Yuan, Dongmei Zhang
ACL 2023 findings

Manuscripts and Pre-prints

[P2] LUNA: Language Understanding with Number Augmentations on Transformers via Number Plugins and Pre-training
Hongwei Han*, **Jialiang Xu***, Mengyu Zhou, Yijia Shao, Shi Han, Dongmei Zhang
ArXiv, submitted to ACL 2023,

[P1] LM-Switch: Lightweight Language Model Conditioning in Word Embedding Space
Chi Han, **Jialiang Xu**, Manling Li, Yi Fung, Chenkai Sun, Nan Jiang, Tarek Abdelzaher, Heng Ji
ArXiv, submitted to ACL 2023, "*" denotes equal contribution.

"*" denotes equal contribution.

COMMUNITY SERVICES

EMNLP: program committee paper reviewer (2022, 2023), conference volunteer (2022).

ACL: program committee paper reviewer (2023).

SKILLS

Languages and frameworks: Python, PyTorch, PyTorch Lightning | **visualization**: seaborn, streamlit
Scientific writing: LaTeX | **collaborating and tracking**: Git, WandB | **web parsing**: BeautifulSoup, Selenium

HONORS AND AWARDS

Microsoft Stars of Tomorrow Award	2022
Horace and Kate Wu Scholarship	2022
Daniel W. and Carol A. Dobberpuhl Scholarship	2022
ECE Visionary Scholarship	2022
Yunni and Maxine Pao Memorial Scholarship	2021
ECE Alumni Association Scholarship	2021
First Place, DARPA SocialSim Final Evaluation	2021
First Place, UIUC EOH Original Undergraduate Research Award	2021
Omron Scholarship	2020
UIUC, Dean's list	2018
UIUC, Edmund J. James Scholarship	2018