

Victoria X. Lin

SENIOR RESEARCH SCIENTIST

1 Hacker Way, Menlo Park, CA 94025, USA

victorialin.net | victorialin@meta.com | [@VictoriaLinML](https://twitter.com/VictoriaLinML) | [xivictorialin](https://www.linkedin.com/in/xivictorialin) | [todpole3](https://github.com/todpole3) | [victorialin](https://www.facebook.com/victorialin)

Research Interest

I am passionate about building general intelligent systems that process information at scale and assist humans in various knowledge-intensive tasks. My recent work focuses on large-scale language modeling and neural information retrieval.

Experience

Meta - Foundational AI Research (FAIR)

SENIOR RESEARCH SCIENTIST

- Large-scale language modeling (LLM), retrieval-augmented LLM, LLM alignment and multilinguality

Menlo Park, CA, USA

Jan. 2021 - present

Salesforce Research

SENIOR RESEARCH STAFF

- Natural language to code, question answering and knowledge graph reasoning

Palo Alto, CA, USA

Oct. 2017 - Dec. 2020

Microsoft Research

RESEARCH INTERN

Redmond, WA, USA

Jun. 2015 - Sep. 2015

Allen Institute for Artificial Intelligence

RESEARCH INTERN

Seattle, WA, USA

Jul. 2014 - Sep. 2014

Preprints

* denotes equal contribution

P4. RA-DIT: Retrieval-Augmented Dual Instruction Tuning

[Xi Victoria Lin*](#), Xilun Chen, Mingda Chen, Weijia Shi, Maria Lomeli, Rich James, Pedro Rodriguez, Jacob Kahn, Gergely Szilvasy, Mike Lewis, Luke Zettlemoyer, Scott Wen-tau Yih.

ArXiv 2023

P3. OPT-IML: Scaling language model instruction meta learning through the lens of generalization

Srinivasan Iyer*, [Xi Victoria Lin*](#), Ramakanth Pasunuru*, Todor Mihaylov, Daniel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, Xian Li, Brian O'Horo, Gabriel Pereyra, Jeff Wang, Christopher Dewan, Asli Celikyilmaz, Luke Zettlemoyer, Ves Stoyanov.

ArXiv 2022

P2. OPT: Open pre-trained transformer language models

Susan Zhang*, Stephen Roller*, Naman Goyal*, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, [Xi Victoria Lin](#), Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, Luke Zettlemoyer.

ArXiv 2022

P1. FOLIO: Natural Language Reasoning with First-Order Logic

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Luke Benson, Lucy Sun, Ekaterina Zubova, Yujie Qiao, Matthew Burtell, David Peng, Jonathan Fan, Yixin Liu, Brian Wong, Malcolm Sailor, Ansong Ni, Linyong Nan, Jungo Kasai, Tao Yu, Rui Zhang, Shafiq Joty, Alexander R. Fabbri, Wojciech Kryscinski, [Xi Victoria Lin](#), Caiming Xiong, Dragomir Radev.

ArXiv 2022

Conference Publications

C21. Towards A Unified View of Sparse Feed-Forward Network in Pretraining Large Language Model

Leo Z. Liu, Tim Dettmers, [Xi Victoria Lin*](#), Veselin Stoyanov, Xian Li.

EMNLP 2023

C20. LEVER: Learning to Verify Language-to-Code Generation with Execution.

Ansong Ni, Srinu Iyer, Dragomir Radev, Ves Stoyanov, Scott Wen-tau Yih, Sida I. Wang*, [Xi Victoria Lin*](#).

ICML 2023

C19. Training Trajectories of Language Models Across Scales.

Mengzhou Xia, Mikel Artetxe, Chunting Zhou, [Xi Victoria Lin](#), Ramakanth Pasunuru, Danqi Chen, Luke Zettlemoyer, Ves Stoyanov.

ACL 2023

C18. Reimagining Retrieval Augmented Language Models for Answering Queries.

Wang-Chiew Tan, Yuliang Li, Pedro Rodriguez, Richard James, [Xi Victoria Lin](#), Alon Halevy, Scott Wen-tau Yih.

Findings of ACL 2023

C17. Few-shot Learning with Multilingual Language Models.

[Xi Victoria Lin](#)^{*}, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, Xian Li^{*}.

EMNLP 2022

C16. Efficient Large Scale Language Modeling with Mixtures of Experts.

Mikel Artetxe^{*}, Shruti Bhosale^{*}, Naman Goyal^{*}, Todor Mihaylov^{*}, Myle Ott^{*}, Sam Shleifer^{*}, [Xi Victoria Lin](#), Jingfei Du, Srinivasan Iyer, Ramakanth Pasunuru, Giri Anantharaman, Xian Li, Shuohui Chen, Halil Akin, Mandeep Baines, Louis Martin, Xing Zhou, Punit Singh Koura, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Mona Diab, Zornitsa Kozareva, Ves Stoyanov.

EMNLP 2022

C15. Lifting the Curse of Multilinguality by Pre-training Modular Transformers.

Jonas Pfeiffer, Naman Goyal, [Xi Victoria Lin](#), Xian Li, James Cross, Sebastian Riedel, Mikel Artetxe.

NAACL 2022

C14. On Continual Model Refinement in Out-of-Distribution Data Streams.

Bill Yuchen Lin, Sida Wang, [Xi Victoria Lin](#), Robin Jia, Lin Xiao, Xiang Ren, Scott Wen-tau Yih.

ACL 2022

C13. Pretty Princess vs. Successful Leader: Gender Roles in Greeting Card Messages. *Best Paper Honorable Mention*

Jiao Sun, Tongshuang Wu, Yue Jiang, Ronil Awalegaonkar, [Xi Victoria Lin](#), Diyi Yang.

CHI 2022

C12. FeTaQA: Free-form Table Question Answering

Linyong Nan, Chiachun Hsieh, Ziming Mao, [Xi Victoria Lin](#), Neha Verma, Rui Zhang, Wojciech Kryściński, Nick Schoelkopf, Riley Kong, Xiangru Tang, Murori Mutuma, Ben Rosand, Isabel Trindade, Renusree Bandaru, Jacob Cunningham, Caiming Xiong, Dragomir Radev.

TACL 2022

C11. GraPPa: Grammar-Augmented Pre-Training for Table Semantic Parsing

Tao Yu, Chien-Sheng Wu, [Xi Victoria Lin](#), Bailin Wang, Yi Chern Tan, Xinyi Yang, Dragomir Radev, Richard Socher, Caiming Xiong

ICLR 2021

C10. Learning to Synthesize Data for Semantic Parsing.

Bailin Wang, Wenpeng Yin, [Xi Victoria Lin](#) and Caiming Xiong.

NAACL 2021 (short)

C9. DART: Open-Domain Structured Data Record to Text Generation

Linyong Nan, Dragomir Radev, Rui Zhang, Amrit Rau, Abhinand Sivaprasad, Chiachun Hsieh, Xiangru Tang, Aadit Vyas, Neha Verma, Pranav Krishna, Yangxiaokang Liu, Nadia Irwanto, Jessica Pan, Faiaz Rahman, Ahmad Zaidi, Mutethia Mutuma, Yasin Tarabar, Ankit Gupta, Tao Yu, Yi Chern Tan, [Xi Victoria Lin](#), Caiming Xiong, Richard Socher and Nazneen Fatema Rajani.

NAACL 2021

C8. Bridging Textual and Tabular Data for Cross-Domain Text-to-SQL Semantic Parsing

[Xi Victoria Lin](#), Richard Socher, Caiming Xiong

Findings of EMNLP 2020

C7. Double-Hard Debias: Tailoring Word Embeddings for Gender Bias Mitigation

Tianlu Wang, [Xi Victoria Lin](#), Nazeen Fatema Rajani, Bryan McCann, Vicente Ordonez and Caiming Xiong

ACL 2020

C6. CoSQL: A Conversational Text-to-SQL Challenge Towards Cross-Domain Natural Language Interfaces to Databases

Tao Yu, Rui Zhang, Heyang Er, Suyi Li, Eric Xue, Bo Pang, [Xi Victoria Lin](#), Yi Chern Tan, Tianze Shi, Zihan Li, Youxuan Jiang, Michihiro Yasunaga, Sungrok Shim, Tao Chen, Alexander Fabbri, Zifan Li, Luyao Chen, Yuwen Zhang, Shreya Dixit, Vincent Zhang, Caiming Xiong, Richard Socher, Walter Lasecki and Dragomir Radev

EMNLP 2019

C5. Editing-based SQL Query Generation for Cross-Domain Context-Dependent Questions

Rui Zhang, Tao Yu, Heyang Er, Sungrok Shim, Eric Xue, [Xi Victoria Lin](#), Tianze Shi, Caiming Xiong, Richard Socher and Dragomir Radev

EMNLP 2019

C4. SPaRC: Cross-Domain Semantic Parsing in Context

Tao Yu, Rui Zhang, Michihiro Yasunaga, Yi Chern Tan, [Xi Victoria Lin](#), Suyi Li, Heyang Er, Irene Li, Bo Pang, Tao Chen, Emily Ji, Shreya Dixit, David Proctor, Sungrok Shim, Jonathan Kraft, Vincent Zhang, Caiming Xiong, Richard Socher, Dragomir Radev

ACL 2019

C3. Multi-Hop Knowledge Graph Reasoning with Reward Shaping

[Xi Victoria Lin](#), Richard Socher and Caiming Xiong

EMNLP 2018

C2. NL2Bash: A Corpus and Semantic Parser for Natural Language Interface to the Linux Operating System

[Xi Victoria Lin](#), Chenglong Wang, Luke Zettlemoyer and Michael D. Ernst

LREC 2018

C1. Compositional Learning of Embeddings for Relation Paths in Knowledge Bases and Text

Kristina Toutanova, [Xi Victoria Lin](#), Scott Wen-tau Yih, Hoifung Poon and Chris Quirk

ACL 2016

Other Publications

O7. Testing Cross-Database Semantic Parsers Using Canonical Utterances. *Best Paper Award*

Heather Lent, Semih Yavuz, Tao Yu, Tong Niu, Yingbo Zhou, Dragomir Radev, [Xi Victoria Lin](#).

Eval4NLP @EMNLP 2021

O6. NeurIPS 2020 NLC2CMD Competition: Translating Natural Language to Bash Commands.

Mayank Agarwal, Tathagata Chakraborti, Quchen Fu, David Gros, [Xi Victoria Lin](#), Jaron Maene, Kartik Talamadupula, Zhongwei Teng, Jules White.

NeurIPS 2020
Competition Track

O5. ColloQL: Robust Text-to-SQL Over Search Queries

Karthik Radhakrishnan, Arvind Srikantan, [Xi Victoria Lin](#)

Intex-Sempar @EMNLP 2020

O4. Photon: A Robust Cross-Domain Text-to-SQL System

Jichuan Zeng*, [Xi Victoria Lin](#)*, Caiming Xiong, Richard Socher, Michael R. Lyu, Irwin King, Steven C.H. Hoi

ACL 2020 Demonstration Track

O3. Program Synthesis from Natural Language Using Recurrent Neural Networks

[Xi Victoria Lin](#), Chenglong Wang, Deric Pang, Kevin Vu, Luke Zettlemoyer, Michael D. Ernst

UWCSE-TR 2017

O2. Multi-label Learning with Posterior Regularization

[Xi Victoria Lin](#), Sameer Singh, Luheng He, Ben Taskar, and Luke Zettlemoyer

MLNLP @NeurIPS 2014

O1. Fine-grained Named Entity Classification in Machine Reading

[Xi Victoria Lin](#)

M.Sc. Thesis 2011
University of Oxford

Patents

Multi-hop knowledge graph reasoning with reward shaping

[Xi Victoria Lin](#), Richard Socher, Caiming Xiong

US Patent App. 16/051,309

Talks

T6. Knowledge and Skill Acquisition through LLM Pre-training and Instruction-tuning (invited talk)

KLR @ICML 2023

T5. LLMs as Instructable Task Solvers: Lessons Learned and Future Possibilities (invited talk)

Stanford NLP Seminar
Spring 2023

T4. Bridging Textual and Tabular Data: Is Attention All We Need? (invited talk)

KR2ML @NeurIPS 2020

T3. Natural Language Interfaces to Databases (guest lecture)

NYU CS2590 NLP 2020

T2. Reinforcement Learning for Knowledge Graph Reasoning (invited talk)

Knowledge ConneXions 2020

T1. Creating The Future Of AI: How Salesforce Research Advances AI For CRM (session speaker)

Dreamforce 2019

Technical Writings

Talk to Your Data: One Model, Any Relational Database.

Salesforce Research Blog 2020

Software

Photon v1.1: <https://naturalsql.com/>

Salesforce, 2020

Photon is a deep learning based cross-domain natural language interface to databases that focuses on factual look-up questions. It allows end users to query a number of relational DBs in natural language, including DBs it has never been trained on.

Tellina is an end-user scripting assistant that can be queried via natural language. It translates a natural language sentence typed by the user into a piece of short, executable script.

Education

University of Washington

PH.D. CANDIDATE IN COMPUTER SCIENCE

Seattle, WA, USA

Jun. 2013-Aug. 2017, Jan. 2021-present

University of Pennsylvania

M.SC. IN COMPUTER SCIENCE (PH.D. TRANSFER)

Philadelphia, PA, USA

May 2013

University of Oxford

M.SC. IN COMPUTER SCIENCE

Oxford, UK

Sep. 2011

The Hong Kong Polytechnic University

B.ENG. IN ELECTRONIC AND INFORMATION ENGINEERING

Kowloon, HK

Aug. 2010

Xi'an Jiaotong University

SPECIAL CLASS FOR THE GIFTED YOUNG

Xi'an, Shaanxi, CN

Jan. 2008

Honors & Awards

2022	Best Paper Honorable Mention , The ACM CHI Conference on Human Factors in Computing Systems	
2021	Best Paper Award , The 2nd Workshop on Evaluation & Comparison of NLP Systems	Eval4NLP@EMNLP
2014	Student Travel Award , NeurIPS Women in Machine Learning Workshop (WiML)	Montreal, Québec, CA
2011	Doctoral Research Fellowship , University of Pennsylvania	Philadelphia, PA, USA
2010	Best Academic Performance Award , Department of Electronic and Information Engineering, HKPU	Kowloon, HK
2010	Hong Kong SAR Government Scholarship , H.K. SAR Government	Kowloon, HK

Service

CO-CHAIR

System Demonstration Track

NAACL 2021

ORGANIZING COMMITTEE

1st Workshop on Interactive and Executable Semantic Parsing (Intex-Sempar)

EMNLP 2020

Competition for Automatic Translation of English to Bash (NLC2CMD)

NeurIPS 2020

PROGRAM COMMITTEE

2022	ARR, AACL, ICLR-DL4C
2021	ARR, ACL-NLP4Prog
2020	ACL, EMNLP, AACL, ACL-NLI
2019	ICML, ACL, NAACL
2018	ACL, EMNLP, COLING, CoNLL
2017	ACL, EMNLP
2016	EMNLP
2015	EMNLP

Teaching

CIS 520: Machine Learning

TEACHING ASSISTANT

University of Pennsylvania

Sep. 2012 - Dec. 2012

- Making exam problems; answering Piazza questions; holding office hours; grading

Programming Skills

Languages	Python, Java, C++, Matlab, R, Lua, JavaScript, HTML, CSS
Deep Learning	Pytorch, Tensorflow
Web	Django