# Victoria X. **Lin**

Research Scientist

*1 Hacker Way, Menlo Park, CA 94025, USA*

victorialin.org | victoriaxlin.ai@gmail.com | @VictoriaLinML | VictoriaLinML | todpole3 | 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 efficient multi-modal LLM pre-training, retrieval-augmentation and neural information retrieval.

## Experience

**Meta, AGI Foundation** — *Menlo Park, CA, USA*

Research Scientist — *Jan. 2021 - present*

- ***Llama4 multimodal pretraining:*** scaling laws, architecture ablation, and data curriculum
- *Efficient and sparse architecture design for early-fusion multimodal LLMs:* Mixture-of-transformers, MoMa, Chameleon
- *RAG and neural information retrieval*: RA-DIT, ReasonIR, DRAMA
- *LLM fine-tuning and few-shot learning*: OPT-IML, XGLM, OPT

**Salesforce Research** — *Palo Alto, CA, USA*

Research Scientist — *Oct. 2017 - Dec. 2020*

- Code generation, reasoning over structured data, question answering

## Education

**University of Washington** — *Seattle, WA, USA*

Ph.D. in Computer Science — *Advisor: Prof. Luke Zettlemoyer*

**University of Pennsylvania** — *Philadelphia, PA, USA*

M.Sc. in Computer Science (Ph.D. transfer)

**University of Oxford** — *Oxford, UK*

M.Sc. in Computer Science

**The Hong Kong Polytechnic University** — *Kowloon, HK*

B.Eng. in Electronic and Information Engineering

## Prepirnts

* denotes equal contribution  # research interns hosted by me

P2. **MoMa: Efficient Early-Fusion Pre-training with Mixture of Modality-Aware Experts**
Xi Victoria Lin*, Akshat Shrivastava*, Liang Luo, Srinivasan Iyer, Mike Lewis, Gargi Ghosh, Luke Zettlemoyer, Armen Aghajanyan*. — ArXiv 2024

P1. **Chameleon: Mixed-Modal Early-Fusion Foundation Models**
Chameleon Team. — ArXiv 2024

## Conference and Journal Publications

J1. **Mixture-of-Transformers: A Sparse and Scalable Architecture for Multi-Modal Foundation Models**
Weixin Liang (#), Lili Yu, Liang Luo, Srinivasan Iyer, Ning Dong, Chunting Zhou, Gargi Ghosh, Mike Lewis, Wen-tau Yih, Luke Zettlemoyer, Xi Victoria Lin. — TMLR 2025

C32. **LMFusion: Adapting Pretrained Language Models for Multimodal Generation**
Weijia Shi, Xiaochuang Han, Chunting Zhou, Weixin Liang, Xi Victoria Lin, Luke Zettlemoyer, Lili Yu. — NeurIPS 2025

C31. **ReasonIR: Training Retrievers for Reasoning Tasks**
Rulin Shao, Rui Qiao, Varsha Kishore, Niklas Muennighoff, Xi Victoria Lin, Daniela Rus, Bryan Kian Hsiang Low, Sewon Min, Wen-tau Yih, Pang Wei Koh, Luke Zettlemoyer. — COLM 2025

C30. **DRAMA: Diverse Augmentation from Large Language Models to Smaller Dense Retrievers**
Xueguang Ma, Xi Victoria Lin, Barlas Oguz, Jimmy Lin, Wen-tau Yih, Xilun Chen. ACL 2025

C29. **SelfCite: Self-Supervised Alignment for Context Attribution in Large Language Models**
Yung-Sung Chuang, Benjamin Cohen-Wang, Shannon Zejiang Shen, Zhaofeng Wu, Hu Xu, Xi Victoria Lin, James Glass, Shang-Wen Li, Wen-tau Yih. ICML 2025

C28. **Nearest Neighbor Speculative Decoding for LLM Generation and Attribution**
Minghan Li (#), Xilun Chen, Ari Holtzman, Beidi Chen, Jimmy Lin, Wen-tau Yih, Xi Victoria Lin. NeurIPS 2024

C27. **Sirius: Contextual Sparsity with Correction for Efficient LLMs**
Yang Zhou, Zhuoming Chen, Zhaozhuo Xu, Xi Victoria Lin, Beidi Chen. NeurIPS 2024

C26. **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. EMNLP 2024

C25. **Branch-Train-MiX: Mixing Expert LLMs into a Mixture-of-Experts LLM**
Sainbayar Sukhbaatar, Olga Golovneva, Vasu Sharma, Hu Xu, Xi Victoria Lin, Baptiste Rozière, Jacob Kahn, Daniel Li, Wen-tau Yih, Jason Weston, Xian Li COLM 2024

C24. **Instruction-tuned Language Models are Better Knowledge Learners**
Zhengbao Jiang, Zhiqing Sun, Weijia Shi, Pedro Rodriguez, Chunting Zhou, Graham Neubig, Xi Victoria Lin, Wen-tau Yih, Srinivasan Iyer ACL 2024

C23. **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, Wen-tau Yih. ICLR 2024

C22. **In-Context Pretraining: Language Modeling Beyond Document Boundaries**
Weijia Shi, Sewon Min, Maria Lomeli, Chunting Zhou, Margaret Li, Rich James, Xi Victoria Lin, Noah A. Smith, Luke Zettlemoyer, Wen-tau Yih, Mike Lewis. ICLR 2024

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 (#), Srini Iyer, Dragomir Radev, Ves Stoyanov, 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, Wen-tau Yih. Findings of ACL 2023

T2. **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

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, Wen-tau Yih. ACL 2022

T1. **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

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, Wen-tau Yih, Hoifung Poon and Chris Quirk — ACL 2016

## Other Publications

O8. **Towards LLMs for Everyone: Instruction Following, Knowledge Retrieval and Multilingualism**
Xi Victoria Lin — Ph.D. Thesis, University of Washington

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, 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

## Honors & Awards

2022 **Best Paper Honarable Mention**, The ACM CHI Conference on Human Factors in Computing Systems — *CHI 2022*
2021 **Best Paper Award**, The 2nd Workshop on Evaluation & Comparison of NLP Systems — *Eval4NLP @EMNLP 2021*

## Service

### Senior Area Chair & Area Chair

Senior Area Chair — Generation Track, AACL-IJCNLP 2022
Area Editor/Chair — ACL Rolling Review (ARR), 2023-present

### Organizing Committee

Demonstration Chair — NAACL 2021

### Workshops Organized

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 |

## Talks

T9. **Large Language Models for Knowledge Intensive Problem Solving** (invited talk) — OxML 2024

T8. **Retrieval-Augmented Dual Instruction Tuning** (invited talk)
Google NLP Reading Group 2024
Cohere for AI Interactive Reading Group 2023
LlamaIndex Webinar 2023

T7. **Aligning Semi-Parametric Language Models** (guest lecture) — NYU DS-GA.1011 NLP 2023

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)
CMU 18-789: Deep Generative Modeling 2024
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** (co-speaker) — Dreamforce 2019

## Panels

P3. **Building Inclusive Communities at ICML** — Social Event @ICML 2025

P2. **Reasoning Capabilities of LLMs** — KLR @ICML 2023

P1. **Where and how can KRR benefit ML, and what should be explored?** — KR2ML @NeurIPS 2020

## Technical Writings

**Talk to Your Data: One Model, Any Relational Database.** — Salesforce Research Blog 2020

## Internships

**Microsoft Research** — *Redmond, WA, USA*
Research Intern — *Jun. 2015 - Sep. 2015*

**Allen Institute for Artificial Intelligence** — *Seattle, WA, USA*
Research Intern — *Jul. 2014 - Sep. 2014*

## 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 v1.0:** http://tellina.rocks/ — University of Washington, 2017
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.

## Teaching

### CIS 520: Machine Learning

*University of Pennsylvania*

Teaching Assistant

*Sep. 2012 - Dec. 2012*

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