(Bill) Yuchen Lin, PhD

• http://yuchenlin.xyz/ • billyuchenlin@gmail.com • Google Scholar

Employment

xAI — Palo Alto, CA. USA

Member of Technical Staff: Core Contributor to Grok 4.5 & 4.6 — *Nov 2024 - **Now***

General Post-Training

- Grok 3 post-training: preference learning via rubric-based RL to improve general chat and reasoning ability; achieved best LMSYS Arena Elo at final release.
- Grok 3 Mini post-training: long-form system instruction following, especially coding system prompts (e.g., Cursor); significantly improved API usability in production.

Agentic coding models:

- Mid-training & SFT: long-context, multi-user-turn data studies; synthesized multi-turn data with a user simulator for long-horizon coding tasks; studied data recipes using RL traces and product traffic.
- RL with non-verifiable tasks: used PRs without unit tests in RL training; ran recipe studies with an agentic grader for multi-rollout, group-wise reward signals; researched data mixture, reward signal types, and sample efficiency.
- Terminal-use agent RL: created terminal-heavy verifiable tasks (especially Docker build) to scale RL data for coding; covered diverse terminal-use cases including SRE tasks.
- Eval infra: automatic evaluation of patch quality for non-testable tasks using SWE expert feedback; sandboxed real coding IDE apps for diverse harness-based evaluation.

Allen Institute for AI (AI2) — Seattle, WA, U.S.

Research Scientist (advised by Prof. Yejin Choi) — *Feb 2023 – Nov 2024*

LLM research. Highlight projects:

- LLM Agents: **SwiftSage** (one of the first multi-agent frameworks to enhance agent efficiency in complex interactive tasks); **LUMOS** (unified training framework and data recipe for tool-based training of open-source LLM agents); **ETO** (self-improvement approach for continuous improvement of agent models via DAgger-style offline RL).
- Post-training and Reasoning: **URIAL** (investigation into the potential of base LLMs through in-context learning and token-level logit differences to inform future research); **MAGPIE** (large-scale data synthesis to enable open-source LLM training); **ZebraLogic** (analysis of LLM reasoning capacity limits on highly compositional tasks).
- Benchmarking: **WildBench** (first work using rubrics for automatic LLM-as-a-judge evaluation on diverse real-world hard tasks, yielding rankings most correlated with LLM-Arena Elo); **WildVision** (a large-scale real human preference dataset for vision-language tasks).

Facebook/Meta AI Research (FAIR) — CA, U.S.

Research Intern (hosted by Scott Yih; Continual learning.) — *May 2021 – Dec 2021*

Google AI — CA, U.S.

Research Intern (hosted by William W. Cohen; Knowledge reasoning) — *May 2020 – Aug 2020*

Research Intern (hosted by Sandeep Tata; Information extraction.) — *May 2019 – Aug 2019*

Education

University of Southern California — Los Angeles, California
Ph.D. in Computer Science — *Aug 2018 – Dec 2022*

Shanghai Jiao Tong University — Shanghai, China
B.Sc. in Computer Science (IEEE Honor Class) — *Spet 2014 – Jun 2018*

Professional Services

- Senior Area Chairs: ACL 2025
- Area Chairs: ICML 2026, ICLR 2025, ACL 2023, EMNLP 2023, EMNLP 2024

Selected Publication[1]

1. SwiftSage: A Generative Agent with Fast and Slow Thinking for Complex Interactive Tasks
   NeurIPS 2023 Spotlight
   Bill Yuchen Lin, Yicheng Fu, Karina Yang, Faeze Brahman, Shiyu Huang, Chandra Bhagavatula, Prithviraj Ammanabrolu, Yejin Choi, Xiang Ren
2. Agent Lumos: Unified and Modular Training for Open-Source Language Agents | ***ACL 2024***
   Da Yin, Faeze Brahman, Abhilasha Ravichander, Khyathi Chandu, Kai-Wei Chang, Yejin Choi, **Bill Yuchen Lin**[@]
3. The Unlocking Spell on Base LLMs: Rethinking Alignment via In-Context Learning | ***ICLR 2024***
   Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, Yejin Choi
4. LLM-Blender: Ensembling LLMs with Pairwise Comparison and Generative Fusion | ***ACL 2023***
   Dongfu Jiang, Xiang Ren, **Bill Yuchen Lin**[@]
5. Magpie: Alignment Data Synthesis from Scratch by Prompting Aligned LLMs with Nothing | ***arXiv***
   Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, **Bill Yuchen Lin**[@]
6. ZebraLogic: On the Scaling Limits of LLMs for Logical Reasoning | ***ICML 2025***
   Bill Yuchen Lin, Ronan Le Bras, Kyle Richardson, Ashish Sabharwal, Radha Poovendran, Peter Clark, Yejin Choi
7. LoraHub: Efficient Cross-Task Generalization via Dynamic LoRA Composition | ***COLM 2024***
   Chengsong Huang*, Qian Liu*, **Bill Yuchen Lin***, Tianyu Pang, Chao Du, Min Lin
8. Trial and Error: Exploration-Based Trajectory Optimization for LLM Agents | ***ACL 2024***
   Yifan Song, Da Yin, Xiang Yue, Jie Huang, Sujian Li, **Bill Yuchen Lin**[@]
9. WildBench: Benchmarking LLMs with Challenging Tasks from Real Users in the Wild | ***arXiv***
   Bill Yuchen Lin, Yuntian Deng, Khyathi Chandu, Faeze Brahman, Abhilasha Ravichander, Valentina Pyatkin, Nouha Dziri, Ronan Le Bras, Yejin Choi
10. WildVision: Evaluating Vision-Language Models in the Wild with Human Preferences | ***NeurIPS 2024 D&B***
    Yujie Lu, Dongfu Jiang, Wenhu Chen, William Yang Wang, Yejin Choi, **Bill Yuchen Lin**[@]

[1] I publish with the name "Bill [Yuchen/Y.] Lin". [*] = equal contributions. [@] = corresponding author.

All Publication

↘ ***LLM Alignment.***

1. The Unlocking Spell on Base LLMs: Rethinking Alignment via In-Context Learning | ***ICLR 2024***
   Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, Yejin Choi

2. LLM-Blender: Ensembling LLMs with Pairwise Comparison and Generative Fusion | ***ACL 2023***
   Dongfu Jiang, Xiang Ren, **Bill Yuchen Lin**[@]

3. Unsupervised Cross-Task Generalization via Retrieval Augmentation | ***NeurIPS 2022***
   Bill Yuchen Lin, Kangmin Tan, Chris Miller, Beiwen Tian, Xiang Ren

4. TIGERScore: Towards Building Explainable Metric for All Text Generation Tasks | *Transactions on Machine Learning Research* (***TMLR***) *2024*
   Dongfu Jiang, Yishan Li, Ge Zhang, Wenhao Huang, **Bill Yuchen Lin**, Wenhu Chen

5. Inference-Time Policy Adapters: Tailoring Extreme-Scale LMs without Fine-tuning | ***EMNLP 2023***
   Ximing Lu, Faeze Brahman, Peter West, Jaehun Jang, Khyathi Chandu, Abhilasha Ravichander, Lianhui Qin, Prithviraj Ammanabrolu, Liwei Jiang, Sahana Ramnath, Nouha Dziri, Jillian Fisher, **Bill Yuchen Lin**, Skyler Hallinan, Xiang Ren, Sean Welleck, Yejin Choi

6. Faith and Fate: Limits of Transformers on Compositionality | ***NeurIPS 2023 Spotlight***
   Nouha Dziri**, Ximing Lu**, Melanie Sclar**, Xiang Lorraine Li*, Liwei Jiang*, **Bill Yuchen Lin***, Peter West, Chandra Bhagavatula, Ronan Le Bras,Jena Hwang,Soumya Sanyal,Sean Welleck,Xiang Ren, Allyson Ettinger, Zaid Harchaoui, Yejin Choi

↘ **LLM Agents and Vision-Language Models**

1. SwiftSage: A Generative Agent with Fast and Slow Thinking for Complex Interactive Tasks
   NeurIPS 2023 Spotlight
   Bill Yuchen Lin, Yicheng Fu, Karina Yang, Faeze Brahman, Shiyu Huang, Chandra Bhagavatula, Prithviraj Ammanabrolu, Yejin Choi, Xiang Ren

2. Agent Lumos: Unified and Modular Training for Open-Source Language Agents | ***ACL 2024***
   Da Yin, Faeze Brahman, Abhilasha Ravichander, Khyathi Chandu, Kai-Wei Chang, Yejin Choi, **Bill Yuchen Lin**[@]

3. Trial and Error: Exploration-Based Trajectory Optimization for LLM Agents | ***ACL 2024***
   Yifan Song, Da Yin, Xiang Yue, Jie Huang, Sujian Li, **Bill Yuchen Lin**[@]

4. Suspicion-Agent: Playing Imperfect Information Games with Theory of Mind Aware GPT4 | ***COLM 2024***
   Jiaxian Guo, Bo Yang, Paul Yoo, **Bill Yuchen Lin**, Yusuke Iwasawa, Yutaka Matsuo

5. VisualWebBench: How Far Have Multimodal LLMs Evolved in Web Page Understanding and Grounding? | ***COLM 2024***
   Junpeng Liu, Yifan Song, **Bill Yuchen Lin**, Wai Lam, Graham Neubig, Yuanzhi Li, Xiang Yue

6. VideoScore: Building Automatic Metrics to Simulate Fine-grained Human Feedback for Video Generation | ***EMNLP 2024***
   Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, ⋯, **Bill Yuchen Lin**, Wenhu Chen

7. OpenCodeInterpreter: Integrating Code Generation with Execution and Refinement | ***ACL 2024 Findings***
   Tianyu Zheng, Ge Zhang, Tianhao Shen, Xueling Liu, **Bill Yuchen Lin**, Jie Fu, Wenhu Chen, Xiang Yue

8. The BiGGen Bench: A Principled Benchmark for Fine-grained Evaluation of Language Models with Language Models | ***NAACL-HLT 2025*** (***Best paper awared***)

↘ ***Safety and Robustness.***

1. SafeDecoding: Defending against Jailbreak Attacks via Safety-Aware Decoding. | ***ACL 2024***
   Zhangchen Xu, Fengqing Jiang, Luyao Niu, Jinyuan Jia, **Bill Yuchen Lin**, Radha Poovendran

2. RockNER: A Simple Method to Create Adversarial Examples for Evaluating the Robustness of NER Models. | ***EMNLP 2021***
   Bill Yuchen Lin, Wenyang Gao, Jun Yan, Ryan Moreno, Xiang Ren

3. On the Robustness of Reading Comprehension Models to Entity Renaming. | ***NAACL 2022***
   Jun Yan, Yang Xiao, Sagnik Mukherjee, **Bill Yuchen Lin**, Robin Jia, Xiang Ren

↘ **Machine Learning for NLP**

1. On Continual Model Refinement in Out-of-Distribution Data Streams | ***ACL 2022***
   Bill Yuchen Lin, Sida Wang, Xi Victoria Lin, Robin Jia, Lin Xiao, Xiang Ren, Scott Yih

2. LoraHub: Efficient Cross-Task Generalization via Dynamic LoRA Composition | ***COLM 2024***
   Chengsong Huang*, Qian Liu*, **Bill Yuchen Lin***, Tianyu Pang, Chao Du, Min Lin

3. FedNLP: Benchmarking Federated Learning Methods for Natural Language Processing Tasks | ***NAACL 2022*** *Findings*
   Bill Yuchen Lin*, Chaoyang He*, Zihang Zeng, Hulin Wang, Yufen Huang, Christophe Dupuy, Rahul Gupta, Mahdi Soltanolkotabi, Xiang Ren, Salman Avestimehr

4. CrossFit: A Few-shot Learning Challenge for Cross-Task Generalization in NLP. | ***EMNLP 2021***
   Qinyuan Ye, **Bill Yuchen Lin**, Xiang Ren

5. Learn Continually, Generalize Rapidly: Lifelong Knowledge Accumulation for Few-shot Learning. | ***EMNLP 2021*** *Findings*
   Xisen Jin, **Bill Yuchen Lin**, Mohammad Rostami, Xiang Ren

6. IsoBN: Fine-Tuning BERT with Isotropic Batch Normalization. | ***AAAI 2021***
   Wenxuan Zhou, **Bill Yuchen Lin**, Xiang Ren

7. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. | *Transactions on Machine Learning Research* (***TMLR***) *2023*
   442 authors including **Bill Yuchen Lin**

↘ **Commonsense Reasoning**

1. KagNet: Knowledge-Aware Graph Networks for Commonsense Reasoning. | ***EMNLP 2019***
   Bill Yuchen Lin, Xinyue Chen, Jamin Chen, Xiang Ren

2. CommonGen: A Constrained Text Generation Challenge Towards Generative Commonsense Reasoning. | ***EMNLP 2020*** *Findings*
   Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, Xiang Ren

3. Common Sense Beyond English: Evaluating and Improving Multilingual Language Models for Commonsense Reasoning. | ***ACL 2021***
Bill Yuchen Lin, Seyeon Lee, Xiaoyang Qiao, Xiang Ren

4. Differentiable Open-Ended Commonsense Reasoning. | ***NAACL 2021***
Bill Yuchen Lin, Haitian Sun, Bhuwan Dhingra, Manzil Zaheer, Xiang Ren and William W. Cohen

5. Birds have four legs?! NumerSense: Probing Numerical Commonsense Knowledge of Pre-trained Language Models. | ***EMNLP 2020*** (*short*)
Bill Yuchen Lin, Seyeon Lee, Rahul Khanna and Xiang Ren

6. RiddleSense: Reasoning about Riddle Questions Featuring Linguistic Creativity and Commonsense Knowledge. | ***ACL 2021*** *Findings*
Bill Yuchen Lin, Ziyi Wu, Yichi Yang, Dong-Ho Lee, Xiang Ren

7. On Grounded Planning for Embodied Tasks with Language Models | ***AAAI 2023***
Bill Yuchen Lin*, Chengsong Huang*, Qian Liu, Wenda Gu, Sam Sommerer, Xiang Ren

8. Automatic Extraction of Commonsense LocatedNear Knowledge. | ***EMNLP 2018***
Frank F. Xu*, **Bill Yuchen Lin***, Kenny Q. Zhu

9. Reflect, Not Reflex: Inference-Based Common Ground Improves Dialogue Response Quality | ***EMNLP 2022***
Pei Zhou, Hyundong J. Cho, Pegah Jandaghi, Dong-Ho Lee, **Bill Yuchen Lin**, Jay Pujara, Xiang Ren

10. Pre-training Text-to-Text Transformers for Concept-Centric Common Sense | ***ICLR 2021***
Wangchunshu Zhou, Dong-Ho Lee, Ravi Kiran Selvam, Seyeon Lee, **Bill Yuchen Lin**, Xiang Ren

11. RICA: Evaluating Robust Inference Capabilities Based on Commonsense Axioms | ***EMNLP 2021***
Pei Zhou, Rahul Khanna, Seyeon Lee, **Bill Yuchen Lin**, Daniel Ho, Jay Pujara, Xiang Ren

12. Probing Commonsense Explanation in Dialogue Response Generation | ***EMNLP 2021*** *Findings*
Pei Zhou, Pegah Jandaghi, Hyundong Cho, **Bill Yuchen Lin**, Jay Pujara, Xiang Ren

↘ **Information Extraction**

1. Neural Adaptation Layers for Cross-domain Named Entity Recognition. | ***EMNLP 2018***
Bill Yuchen Lin, Wei Lu

2. FreeDOM: A Transferable Neural Architecture for Structured Information Extraction on Web Documents. | ***KDD 2020*** (***Research Track***)
Bill Yuchen Lin, Ying Sheng, Nguyen Vo, Sandeep Tata

3. TriggerNER: Learning with Entity Triggers as Explanation for Named Entity Recognition. | ***ACL 2020***
Bill Yuchen Lin*, Dongho Lee*, Ming Shen, Ryan Moreno, Xiao Huang, Prashant Shiralkar, Xiang Ren

4. Mining Cross-Cultural Differences and Similarities in Social Media.
ACL 2018
Bill Yuchen Lin*, Frank F. Xu*, Kenny Q. Zhu, Seung-won Hwang

5. AlpacaTag: An Active Learning-based Crowd Annotation Framework for Sequence Tagging. | ***ACL 2019*** (*demo track*)
Bill Yuchen Lin*, Dongho Lee*, Frank F. Xu, Ouyu Lan, Xiang Ren

6. AutoTriggER: Named Entity Recognition with Auxiliary Trigger Extraction. | ***EACL 2023***
Dong-Ho Lee, Ravi Kiran Selvam, Sheikh Muhammad Sarwar, **Bill Yuchen Lin**, Mahak Agarwal, Fred Morstatter, Jay Pujara, Elizabeth Boschee, James Allan, Xiang Ren

7. Learning to Contextually Aggregate Multi-Source Supervision for Sequence Labeling. | ***ACL 2020***
Ouyu Lan, Xiao Huang, **Bill Yuchen Lin**, He Jiang, Liyuan Liu, Xiang Ren

8. NERO: A Neural Rule Grounding Framework for Label-Efficient Relation Extraction. | ***WWW 2020***
Wenxuan Zhou, Hongtao Lin, **Bill Yuchen Lin**, Ziqi Wang, Junyi Du, Leonardo Neves, Xiang Ren

9. ExtRA: Extracting Prominent Review Aspects from Customer Feedback. | ***EMNLP 2018***
Zhiyi Luo, Shanshan Huang, Frank F. Xu, **Bill Yuchen Lin**, Hanyuan Shi, Kenny Q. Zhu

10. LEAN-LIFE: A Label-Efficient Annotation Framework Towards Learning from Explanation. | ***ACL 2020*** (*demo track*)
Ouyu Lan, Xiao Huang, **Bill Yuchen Lin**, He Jiang, Liyuan Liu, Xiang Ren

---

↘ ***Book.***

1. Knowledge-Augmented Methods for Natural Language Processing
Meng Jiang, **Bill Yuchen Lin**, Shuohang Wang, Yichong Xu, Wenhao Yu, Chenguang Zhu
2024 | ***Springer*** | DOI: 10.1007/978-981-97-0747-8

Invited Talks & Tutorials

- Complex Reasoning in Natural Language
  - ACL 2023 Tutorial
- Language-based Reasoning for Complex Interactive Tasks in Real-World Situations
  - Microsoft Research, Montreal (Host: Xingdi Yuan) & NCSU (Host: Prof. Dongkuan Xu)
  - 2023, June & July
- Advancing Large Language Models with Input Augmentation and Output Refinement
  - Microsoft Research, Redmond (Host: Yang Liu & Chenguang Zhu)
  - 2023, April
- Improving and Evaluating the Commonsense Reasoning Ability of Language Models
  - AI2 Mosaic (Host: Yejin Choi)
  - 2022, Spet
- Knowledge-Augmented Methods for Natural Language Processing
  - ACL 2022 Tutorial + WSDM 2023 Tutorial
- Teaching Machines to Think and Talk Like Humans with Common Sense
  - USC-ACM, Online, 11/2020
- Neuro-Symbolic Commonsense Reasoning for Natural Language Processing
  - Fudan University, Online, 06/2020, Host: Prof. Zhongyu Wei
- Commonsense Reasoning for Natural Language Understanding and Generation
  - Microsoft Research Asia, 12/18/2019, Host: Dr. Nan Duan
  - Tsinghua University, China, 12/17/2019, Host: Prof. Zhiyuan Liu
  - Westlake University, China, 11/13/2019, Host: Prof. Yue Zhang

---

Selected Honors and Awards

- Best Paper Award, NAACL 2025.
- Best Paper Award, TrustNLP 2021.
- AI Rising Stars in Chinese Students Selected by Baidu Research (2021)
- Best Paper Runner-Up, WWW 2020
- Travel Support for ACL2018, ACL Walker Fund, 2018
- Best Bachelor Thesis Award (Top 1% at SJTU & Top 1 in the CSE Department), 2018
- Yuanqing Yang's Scholarship, SJTU, 2018
- Annenberg Fellowship Award, University of Southern California, 2018
- Qide Scholarship, Shanghai Jiao Tong University, 2017-2018
- Shanghai Scholarship, Shanghai Jiao Tong University, 2016-2017 (rank 2/69)