Tao Yu

CONTACT INFORMATION	https://taoyds.github.io/ Rm 204E, Chow Yei Ching Building, Pokfulam, Hong Kong	tyu@cs.hku.hk +852-5392-4676

RESEARCH INTERESTS

Natural Language Processing. The development and design of the next generation of natural language interfaces with language model agents that transform ("grounding") language instructions into code or actions executable in real-world environments, including databases (data agent), web applications (plugins/web agent), and the physical world (robotic agent) etc,. This involves executable language grounding, code generation, semantic parsing, LLM + tool use, efficient and generalizable LLMs, and dialog or interactive systems.

ACADEMIC POSITION

University of Hong Kong, Hong Kong   08/2021 - *Present*
Assistant Professor in Computer Science
Director, XLANG Lab, co-director, HKU NLP Group

University of Washington, Seattle, WA   09/2021 - 08/2022
Postdoctoral Researcher, UW NLP Group
Working with Noah Smith, Luke Zettlemoyer, and Mari Ostendorf

EDUCATION

Yale University, New Haven, CT   08/2017 - 05/2021
Ph.D. Computer Science
Advisors: Dragomir Radev

Columbia University, New York, NY   09/2015 - 05/2017
M.S. Data Science
Advisors: Owen Rambow and Kathleen McKeown

University of Utah, Salt Lake City, UT   01/2012 - 05/2015
B.S. Mathematics and Economics

HONORS & AWARDS

Google Research Scholar Award, aka *Google Faculty Research Award*	2023
Amazon Research Award, aka *Amazon Faculty Research Award*	2022
Yale PhD Fellowship	2017
Connie and Robert White Honor Roll Scholarship	2014
Mathematics Department Scholarship	2014
International Student and Scholar Services Scholarship	2014
NSF Research Experience for Undergraduates (REU) Fellowship	2014

INVITED TALKS

Tutorial on Language Agents: Foundations, Prospects, and Risks at EMNLP 2024   2024

Advancing Natural Language Interfaces with Language Models as Agents
Columbia NLP seminar, April 2023
Cornell DB seminar, May 2023
Microsoft Research Asia, May 2023
Apple KP Tech Talks, June 2023
Keynote at LLMDB Workshop at VLDB 2023, Sept. 2023
Morgan Stanley ML Speaker Seminar, Dec. 2023
ML4Code Seminar at MILA, Dec. 2023

[TRL Workshop at NeurIPS 2023](#), Dec. 2023
Instacart Distinguished Speaker Series, Jan. 2024

[Tutorial on Complex Reasoning over Natural Language at ACL 2023](#)   2023
"Neuro-Symbolic Approaches: Large Language Models + Tool Use"

Building Natural Language Interfaces with Large Language Models
Amazon AWS, Nov. 2022

Few-shot In-context Learning with Large Language Models
AllState Tech Talks, Jun. 2022

UnifiedSKG: Unifying and Multi-Tasking Structured Knowledge Grounding with Text-to-Text
Language Models
ServiceNow Research (Prev. ElementAI), Feb. 2022

SCoRe: Pre-Training for Context Representation in Conversational Semantic Parsing
Google Research, Apr. 2021

PUBLICATIONS &
PAPERS IN
PREPARATION

1. Yiheng Xu, Hongjin Su, Chen Xing, Boyu Mi, Qian Liu, Weijia Shi, Binyuan Hui, Fan Zhou, Yitao Liu, Tianbao Xie, Zhoujun Cheng, Siheng Zhao, Lingpeng Kong, Bailin Wang, Caiming Xiong, **Tao Yu**
   [Lemur: Harmonizing Natural Language and Code for Language Agents](#)
   Preprint 2023

2. Tianbao Xie*, Fan Zhou*, Zhoujun Cheng*, Peng Shi*, Luoxuan Weng*, Yitao Liu*, Toh Jing Hua, Junning Zhao, Qian Liu, Che Liu, Leo Z. Liu, Yiheng Xu, Hongjin Su, Dongchan Shin, Caiming Xiong, **Tao Yu**
   [OpenAgents: An Open Platform for Language Agents in the Wild](#)
   Preprint 2023

3. Tianbao Xie*, Siheng Zhao*, Chen Henry Wu, Yitao Liu, Qian Luo, Victor Zhong, Yanchao Yang, **Tao Yu**
   [Text2Reward: Automated Dense Reward Function Generation for Reinforcement Learning](#)
   Preprint 2023

4. Hongjin Su*, Weijia Shi*, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A Smith, Luke Zettlemoyer,**Tao Yu**
   [One Embedder, Any Task: Instruction-Finetuned Text Embeddings](#)
   ACL Findings 2023

5. Yuhang Lai*, Chengxi Li*, Yiming Wang*, Tianyi Zhang*, Ruiqi Zhong*, Luke Zettlemoyer, Scott Wen-tau Yih, Daniel Fried, Sida Wang, **Tao Yu**
   [DS-1000: A Natural and Reliable Benchmark for Data Science Code Generation](#)
   ICML 2023

6. Tianyi Zhang, **Tao Yu**, Tatsunori B Hashimoto, Mike Lewis, Wen-tau Yih, Daniel Fried, Sida I Wang
   [Coder Reviewer Reranking for Code Generation](#)
   ICML 2023

7. Jiacheng Ye, Zhiyong Wu, Jiangtao Feng, **Tao Yu**, Lingpeng Kong
   [Compositional Exemplars for In-context Learning](#)
   ICML 2023

8. Zhoujun Cheng*, Tianbao Xie*, Peng Shi, Chengzu Li, Rahul Nadkarni, Yushi Hu, Caiming Xiong, Dragomir Radev, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, **Tao Yu**
   [Binding Language Models in Symbolic Languages](#)
   ICLR 2023

9. Hongjin Su, Jungo Kasai, Chen Henry Wu, Weijia Shi, Tianlu Wang, Jiayi Xin, Rui Zhang, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, **Tao Yu**
Selective Annotation Makes Language Models Better Few-Shot Learners
ICLR 2023

10. Tianbao Xie*, Chen Henry Wu*, Peng Shi, Ruiqi Zhong, Torsten Scholak, Michihiro Yasunaga, Chien-Sheng Wu, Ming Zhong, Pengcheng Yin, Sida I. Wang, Victor Zhong, Bailin Wang, Chengzu Li, Connor Boyle, Ansong Ni, Ziyu Yao, Dragomir Radev, Caiming Xiong, Lingpeng Kong, Rui Zhang, Noah A. Smith, Luke Zettlemoyer, **Tao Yu**
UnifiedSKG: Unifying and Multi-Tasking Structured Knowledge Grounding with Text-to-Text Language Models
EMNLP 2022

11. Jiacheng Ye*, Jiahui Gao*, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, **Tao Yu**, Lingpeng Kong
ZeroGen: Efficient Zero-shot Learning via Dataset Generation
EMNLP 2022

12. Yushi Hu, Chia-Hsuan Lee, Tianbao Xie, **Tao Yu**, Noah A. Smith, Mari Ostendorf
In-Context Learning for Few-Shot Dialogue State Tracking
EMNLP Findings 2022

13. Jiacheng Ye, Jiahui Gao, Zhiyong Wu, Jiangtao Feng, **Tao Yu**, Lingpeng Kong
ProGen: Progressive Zero-shot Dataset Generation via In-context Feedback
EMNLP Findings 2022

14. Qi Liu, Zihuiwen Ye, **Tao Yu**, Phil Blunsom, Linfeng Song
Augmenting Multi-Turn Text-to-SQL Datasets with Self-Play
EMNLP Findings 2022

15. Yiru Chen, Ryan Li, Austin Mac, Tianbao Xie, **Tao Yu**, Eugene Wu
NL2INTERFACE: Interactive Visualization Interface Generation from Natural Language Queries
IEEE Visualization Conference NLVIZ Workshop, 2022

16. Ziming Mao*, Chen Henry Wu*, Ansong Ni, Yusen Zhang, Rui Zhang, **Tao Yu**, Budhaditya Deb, Chenguang Zhu, Ahmed H Awadallah, Dragomir Radev
DYLE: Dynamic Latent Extraction for Abstractive Long-Input Summarization
ACL 2022

17. Yusen Zhang, Ansong Ni, **Tao Yu**, Rui Zhang, Chenguang Zhu, Budhaditya Deb, Asli Celikyilmaz, Ahmed Hassan Awadallah, Dragomir Radev
An Exploratory Study on Long Dialogue Summarization: What Works and What's Next
EMNLP Findings 2021

18. Ansong Ni, Zhangir Azerbayev, Mutethia Mutuma, Troy Feng, Yusen Zhang, **Tao Yu**, Ahmed Hassan Awadallah, Dragomir Radev
SummerTime: Text Summarization Toolkit for Non-experts
EMNLP 2021

19. Heather Lent, Semih Yavuz, **Tao Yu**, Tong Niu, Yingbo Zhou, Dragomir Radev, Xi Victoria Lin
Testing Cross-Database Semantic Parsers Using Canonical Utterances
Evaluation & Comparison of NLP Systems Workshop at EMNLP 2021, **Best Paper Award**

20. Chang Shu, Yusen Zhang, Xiangyu Dong, Peng Shi, **Tao Yu**, Rui Zhang
Logic-Consistency Text Generation from Semantic Parses
ACL Findings 2021

21. **Tao Yu**, Rui Zhang, Oleksandr Polozov, Chris Meek, Ahmed Hassan Awadallah
SCoRe: Pre-Training for Context Representation in Conversational Semantic Parsing
ICLR 2021

22. **Tao Yu**, Chien-Sheng Wu, Xi Victoria Lin, Bailin Wang, Yi Chern Tan, Xinyi Yang, Dragomir Radev, Richard Socher, Caiming Xiong
GraPPa: Grammar-Augmented Pre-Training for Table Semantic Parsing
ICLR 2021

23. Ming Zhong, Da Yin, **Tao Yu**, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, Dragomir Radev
QMSum: A New Benchmark for Query-based Multi-domain Meeting Summarization
NAACL 2021

24. Linyong Nan, Dragomir Radev, Rui Zhang, Amrit Rau, Abhinand Sivaprasad, Chiachun Hsieh, Nazneen Fatema Rajani, Xiangru Tang, Aadit Vyas, Neha Verma, Pranav Krishna, Yangxiaokang Liu, Nadia Irwanto, Jessica Pan, Faiaz Rahman, Ahmad Zaidi, Murori Mutuma, Yasin Tarabar, Ankit Gupta, **Tao Yu**, Yi Chern Tan, Xi Victoria Lin, Caiming Xiong, Richard Socher, Nazneen Fatema Rajani
DART: Open-Domain Structured Data Record to Text Generation
NAACL 2021

25. Ruiqi Zhong, **Tao Yu**, Dan Klein
Semantic Evaluation for Text-to-SQL with Distilled Test Suites
EMNLP 2020

26. Yusen Zhang, Xiangyu Dong, Shuaichen Chang, **Tao Yu**, Peng Shi, Rui Zhang
Did You Ask a Good Question? A Cross-Domain Question Intention Classification Benchmark for Text-to-SQL
Interactive and Executable Semantic Parsing Workshop at EMNLP 2020

27. **Tao Yu**, Rui Zhang, He Yang 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, Dragomir Radev
CoSQL: A Conversational Text-to-SQL Challenge Towards Cross-Domain Natural Language Interfaces to Databases
EMNLP 2019

28. Rui Zhang, **Tao Yu**, He Yang Er, Sungrok Shim, Eric Xue, Xi Victoria Lin, Tianze Shi, Caiming Xiong, Richard Socher, Dragomir Radev
Editing-based SQL Query Generation for Cross-Domain Context-Dependent Questions
EMNLP 2019

29. **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
SParC: Cross-Domain Semantic Parsing in Context
ACL 2019

30. Richard A Plunz, Yijia Zhou, Maria Isabel Carrasco Vintimilla, Kathleen Mckeown, **Tao Yu**, Laura Uguccioni, Maria Paola Sutto
Twitter Sentiment in New York City Parks as Measure of Well-being
Landscape and Urban Planning 2019

31. **Tao Yu**, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, Zilin Zhang, Dragomir Radev
Spider: A Large-Scale Human-Labeled Dataset for Complex and Cross-Domain Semantic Parsing and Text-to-SQL Task
EMNLP 2018

32. **<u>Tao Yu</u>**, Michihiro Yasunaga, Kai Yang, Rui Zhang, Dongxu Wang, Zifan Li, Dragomir Radev
SyntaxSQLNet: Syntax Tree Networks for Complex and Cross Domain Text-to-SQL Task
EMNLP 2018

33. **<u>Tao Yu</u>**, Zifan Li, Zilin Zhang, Rui Zhang, Dragomir Radev
TypeSQL: Knowledge-based Type-Aware Neural Text-to-SQL Generation
NAACL 2018

34. Mohammad Sadegh Rasooli, Noura Farra, Axinia Radeva, **<u>Tao Yu</u>**, and Kathleen McKeown
Cross-Lingual Sentiment Transfer with Limited Resources
Machine Translation 2017

35. Owen Rambow, **<u>Tao Yu</u>**, Axinia Radeva, Sardar Hamidian, Alexander R. Fabbri, Debanjan Ghosh, Christopher Hidey, Tianrui Peng, Mona Diab, Kathleen McKeown, Smaranda Muresan
The Columbia-GWU System at the 2016 TAC KBP BeSt Evaluation
NIST TAC KBP Workshop 2016

STUDENTS

Tianbao Xie (HKU PhD, 2022)

Hongjin Su (HKU PhD, 2022)

Yiheng Xu (HKU PhD, 2022, co-advised with Lingpeng Kong)

Jiacheng Ye (HKU PhD, 2022, co-advised with Lingpeng Kong)

Fan Zhou (Intern 2023, SJTU BS/MS)

Leo Zeyu Liu (Intern 2023, UW BS/MS → UT Austin PhD)

Zhoujun Cheng (Intern 2022, SJTU BS/MS)

Chen Henry Wu (Intern 2022, Tsinghua BS → CMU PhD)

Ryan Li (Intern 2022, UW BS → Stanford MS)

Chengzu Li (Intern 2022, Xi'an Jiaotong BS → Cambridge PhD)

Shuyang Jiang (Intern 2023, SJTU BS → Fudan PhD)

Yiming Wang (Intern 2022, PKU BS → Harvard MS)

Yuhang Lai (Intern 2022, BIT BS → Fudan MS)

Chengxi Li (Intern 2022, HIT BS → CUHK PhD)

Ming Zhong (Intern 2020, Fudan MS → UIUC PhD)

Da Yin (Intern 2020, PKU BS → UCLA PhD)

Yusen Zhang (Intern 2020, Emory MS → PSU PhD)

Naihao Deng (Intern 2020, UMich BS → UMich PhD)

Michihiro Yasunaga (Project Student, 2018-19, Yale BS → Stanford PhD)

Suyi Li (Project Student, 2019, Yale MS → HKUST PhD)

Kai Yang (Project Student, 2018, Yale MS → Google SE)

James Ma (Project Student, 2018, Yale BS → LinkedIn DS)

Grants		
	Early Career Scheme, Research Grants Council of Hong Kong (HKD$1,078,487)	2023
	Google Research Scholar Award (USD$35,000)	2023
	Amazon Research Award (USD$40,000 + $80,000 AWS Credits)	2022

Research &
Industry
Experience

Microsoft Research, Redmond, WA   06/2020 - 08/2020

NLP Research Intern
Mentors: Ahmed Hassan Awadallah, Oleksandr Polozov, and Chris Meek

Salesforce Research, Palo Alto, CA   06/2019 - 08/2019

NLP Research Intern
Mentors: Victoria Lin, Caiming Xiong, and Richard Socher

Samsung Research America, Mountain View, CA   06/2018 - 08/2018
NLP Research Intern

Columbia CCLS & NLP Group, New York, NY   05/2016 - 10/2016

Graduate Research Assistant
Mentors: Owen Rambow and Kathleen McKeown

Teaching
Experience

Instructor: Advanced NLP (DATA 8005), University of Hong Kong	2023
Teaching Assistant: Natural Language Processing (CPSC 577), Yale University	2019
Teaching Assistant: Database Systems (CPSC 537), Yale University	2019
Teaching Assistant: Introduction to NLP (CS 4705), Columbia University	2017
Teaching Assistant: Natural Language Processing (Coursera)	2016