Qing Lyu

Last updated: Feb 2026

veronica320.github.io
lyuqing@sas.upenn.edu

RESEARCH INTERESTS

Agents, Memory, Interpretability, Computational Linguistics

EDUCATION

University of Pennsylvania, Philadelphia, USA — Aug 2019 – Aug 2024
Ph.D. Computer and Information Science — GPA: 4.00/4.00
Advisors: Chris Callison-Burch and Marianna Apidianaki
Thesis: Towards Faithful and Useful Interpretation of Language Models

Tsinghua University, Beijing, China — Sept 2015 – Jul 2019
B.A. English Language and Literature (Linguistics track) — GPA: 3.88/4.00

INDUSTRY EXPERIENCE

Research Scientist, Post-Training Team — Sept 2025 – now
Databricks Mosaic Research — San Francisco, USA

Research Intern — May 2023 – Aug 2023
Allen Institute for Artificial Intelligence (AI2), AllenNLP — Seattle, USA

Research Intern — May 2022 – Aug 2022
Tencent, AI Lab — Seattle, USA

Algorithm Intern — Sept 2018 – Oct 2018
Tomorrow Advancing Life (TAL) Education Group, AI Lab — Beijing, China

PUBLICATIONS AND MANUSCRIPTS

[Google Scholar]

[19] **Q. Lyu***, K. Sreenivasan, S. Moorjani, A. Polyzotis, S. Havens, M. Carbin, M. Bendersky, M. Zaharia, X. Chen. *MemAlign: Building Better LLM Judges From Human Feedback With Scalable Memory.*

Databricks blog post.

[18] **Q. Lyu***, K. Sreenivasan, M*. Lee, M. Bendersky, A. Polyzotis, X. Meng, O. Khattab, S. Havens, M. Carbin and M. Zaharia. *Agent Learning from Human Feedback (ALHF): A Databricks Knowledge Assistant Case Study.*

Databricks blog post.

[17] G. Dou†, Z. Liu, **Q. Lyu**, K. Ding, E. Wong. *Avoiding copyright infringement via large language model unlearning.*
In **NAACL 2025 Findings**.

[16] **Q. Lyu***, K. Shridhar*, C. Malaviya, L. Zhang, Y. Elazar, N. Tandon, M. Apidianaki, M. Sachan, C. Callison-Burch. *Calibrating Large Language Models with Sample Consistency.*
In **AAAI 2025**.

[15] J. M. Ludan†, **Q. Lyu**, Y. Yang, L. Dugan, M. Yatskar, C. Callison-Burch. *Interpretable-by-Design Text Classification with Iteratively Generated Concept Bottleneck*
ArXiv preprint.

[14] **Q. Lyu**, M. Apidianaki, C. Callison-Burch. *Towards Faithful Model Explanation in NLP: A Survey.*
In **Computational Linguistics 2024**.

[13] **Q. Lyu**, S. Havaldar*, A. Stein*, L. Zhang, D. Rao, E. Wong, M. Apidianaki, C. Callison-Burch. *Faithful Chain-of-Thought Reasoning.*

In **IJCNLP-AACL 2023**. **Area Chair Award** (Interpretability and Analysis of Models for NLP).

[12] **Q. Lyu**, M. Apidianaki, C. Callison-Burch. *Representation of Lexical Stylistic Features in Language Models' Embedding Space.*
In ***SEM 2023**.

[11] J. M. Ludan†, Y. Meng*†, T. Nguyen*†, S. Shah*†, **Q. Lyu**, M. Apidianaki, C. Callison-Burch. *Explanation-based Finetuning Makes Models More Robust to Spurious Cues.*
In **ACL 2023**.

[10] **Q. Lyu**, H. Zheng, D. Li, L. Zhang, M. Apidianaki, C. Callison-Burch. *Is "My Favorite New Movie" My Favorite Movie? Probing the Understanding of Recursive Noun Phrases.*
In **NAACL 2022**.

[9] A. Srivastava, …, L. Zhang, **Q. Lyu**, C. Callison-Burch, … *Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models.*

In **TMLR 2022**.

[8] X. Du, Z. Zhang, S. Li, P. Yu, H. Wang, T. Lai, X. Lin, Z. Wang, I. Liu, B. Zhou, H. Wen, M. Li, D. Hannan, J. Lei, H. Kim, R. Dror, H. Wang, M. Regan, Q. Zeng, **Q. Lyu**, C. Yu, C. Edwards, X. Jin, Y. Jiao, G. Kazeminejad, Z. Wang, C. Callison-Burch, M. Bansal, C. Vondrick, J. Han, D. Roth, S. Chang, M. Palmer, H. Ji. *RESIN-11: Schema-guided Event Prediction for 11 Newsworthy Scenarios.*

In **NAACL 2022** (demo track).

[7] S. Zhou*, L. Zhang*, Y. Yang, **Q. Lyu**, G. Neubig, C. Callison-Burch. *Show Me More Details: Discovering Event Hierarchies from WikiHow*.
In **ACL 2022**.

[6] Y. Yang, A. Panagopoulou, **Q. Lyu**, L. Zhang, M. Yatskar, C. Callison-Burch. *Visual Goal-Step Inference using wikiHow*.
In **EMNLP 2021**.

[5] **Q. Lyu**, H. Zhang, E. Sulem, D. Roth. *Zero-shot Event Extraction via Transfer Learning: Challenges and Insights.*
In **ACL 2021.**

[4] **Q. Lyu***, L. Zhang*, C. Callison-Burch. *Goal-Oriented Script Construction*.
In **INLG 2021**.

[3] H. Wen, Y. Lin, T. Lai, X. Pan, S. Li, X. Lin, B. Zhou, M. Li, H. Wang, H. Zhang, X. Yu, A. Dong, Z. Wang, Y. Fung, P. Mishra, **Q. Lyu**, D. Surís, B. Chen, Susan W. Brown, M. Palmer, C. Callison-Burch, C. Vondrick, J. Han, D. Roth, S-F. Chang, H. Ji. *RESIN: A Dockerized Schema-Guided Cross-document Cross-lingual Cross-media Information Extraction and Event Tracking System*.
In **NAACL 2021** (demo track).

[2] L. Zhang, **Q. Lyu**, C. Callison-Burch. *Intent Detection with WikiHow*.
In **AACL-IJCNLP 2020**.

[1] L. Zhang*, **Q. Lyu***, C. Callison-Burch. *Reasoning about Goals, Steps, and Temporal Ordering with WikiHow*.
In **EMNLP 2020**; Spotlight presentation at the Workshop on Enormous Language Models at ICLR 2021.

(*: equal contribution. †: undergraduate/master's mentee.)

SERVICES AND ACTIVITIES

- **Action Editor / Area Chair** for ARR 2024-2025
- **Co-organizer** of Tutorial: Explanations in the Era of Large Language Models (to appear in NAACL'24) 2024
- **Program Committee member** of the 9th Mid-Atlantic Student Colloquium on Speech, Language and Learning (MASC-SLL) 2022
- **Panelist** at WiCS x FemmeHacks CIS PhD Panel 2022

- **Reviewer** for the Beyond the Imitation Game Benchmark (BIG-BENCH) initiated by Google Research 2021
- **Reviewer** for ACL, EMNLP, NAACL, ACL Rolling Review 2021 – now
- **Co-organizer** of CLUNCH, Penn's NLP seminar series 2020

TEACHING EXPERIENCE

Teaching Assistant

- CIS 530 (Computational Linguistics) - Fall 2021, University of Pennsylvania
- CIS 419/519 (Applied Machine Learning) - Fall 2019, University of Pennsylvania
- Computational Linguistics - Fall 2018, Tsinghua University

INVITED TALKS

- "Towards Faithful Model Explanation in NLP" – Guest Lecture in LING 401 (Introduction to Computational Linguistics), UNC-Chapel Hill, Nov 2025
- "Towards Faithful Model Explanation in NLP" – Guest Lecture in NLP 244 (Advanced Machine Learning for NLP), University of California, Santa Cruz, Mar 2023
- "Towards Faithful Model Explanation in NLP" – Guest Lecture in CIS 530 (Computational Linguistics), University of Pennsylvania, Dec 2023
- "Faithful Chain-of-Thought Reasoning" – Talk at University of Colorado at Boulder NLP lab seminar, Mar 2024
- "Towards Faithful Model Explanation in NLP" – Talk at Microsoft Research Montreal NLP lab seminar, 2024 (upcoming)

SIDE PROJECTS

A Societal Model Built from Scratch [demo] Aug 2023

Project at Allen Institute for Artificial Intelligence (AI2)'s Hackathon

- In 3 days, we built a 3D simulation of a neighborhood in Green Lake, Seattle, with the Unity engine, leveraging realistic data from OpenStreetMap and satellite imagery and generating building interiors with Procthor.
- I led the creation of 8 generative agents powered by LLMs, each with their unique personality and memory. I ran a mini-social-experiment, a group speed dating event, matching the agents based on their interaction and conversation with each other.
- Our project won the "I Can't Believe It Worked!" Award.

HONORS

Area Chair Award (Interpretability and Analysis of Models for NLP) at AACL-ICJNLP'23 2023

Excellent Graduation Thesis Award, Tsinghua University 2019

National Scholarship, Chinese Ministries of Education and Finance 2018

3rd Place at "Sentiment analysis of Chinese Metaphor", Shared Task at the 17th China National Conference on Computational Linguistics (CCL 2018) 2018

Jiang Nanxiang Scholarship, Tsinghua University 2017

Merit-based Scholarship of all school years, Tsinghua University 2015 – 2019

First Prize (Individual Contest), National Linguistics Olympiad (NOL) 2014

SKILLS

Programming Skills

Python, C/C++, SQL, MATLAB, HTML, JavaScript

Language Skills

Chinese (native), English (proficient), French (conversational)