

Jiali Cheng

jiali_cheng@uml.edu | [Google Scholar](#) | [Personal Website](#)

SUMMARY

Hands-on exp in LLM Post-Training (70B, RL, OPD), Evolution in Agents, Harness, Trustworthy AI
Contributor to RL-Kernel, a high performance RL training infrastructure
2-time Amazon Applied Scientist Intern on Long-Horizon Web Agent, Agent Memory, Multi-Agent Eval
3-year Data Scientist / Engineer building AI for Science products at a Khosla Venture-backed startup

EDUCATION

UMass Lowell, USA Ph.D. in Computer Science 2023 - present
Northeastern University, Boston, USA M.S. in Computer Engineering 2017 - 2019

EXPERIENCE

Amazon Alexa AI May - Aug, 2026
Applied Science Intern Advised by: Oleg Rokhlenko, Xingzhi Guo, Jialiang Xu

- Proposed a novel setting and benchmark & environment to evaluate long-horizon decentralized multi-agent collaboration at large scale
- Designed graph-aware harness to improve overtime (RSI); Train small models with RL to improve collaboration, outperforming GPT-5.5 and Claude Opus 4.8.

Amazon Alexa AI May - Aug, 2025
Applied Science Intern Advised by: Oleg Rokhlenko, Anjishnu Kumar

- Designed web agents with episodic memory, efficient action-simulation, and environment exploration to support adaptive decision-making and prevent errors. Achieved SOTA results on WebArena-Lite.

Computational Language Understanding Lab, UMass Lowell 2023 - Present
Research Assistant Advised by: Hadi Amiri

- Built benchmarks to evaluate LLM web agents on urban / civil domain websites and tasks
- Built benchmarks to evaluate knowledge conflicts when agent interacts with environment (tool-memory conflict, skill conflict, instruction conflict)
- Develop post-training methods (RL, DPO, GRPO, OPD) to make AI models more trustworthy and controllable (unlearning, editing, robustness, bias and fairness)
- Design mechanistic interpretability methods (SAE, circuits, natural language autoencoder NLA) to 1) edit AI models to produce less harmful responses, 2) control LLM reasoning effort

Foodome Inc., Data Scientist / Data Engineer 2020 - 2022

- Built NLP and KG methods to mine large-scale literature, build domain-specific knowledge graphs and predict new relationships, increasing Acc by 53%+
- Managed reproducible AWS workflows for data acquisition, knowledge extraction, relation prediction, and drug repurposing pipeline (Terraform, GLUE, SageMaker)

Zitnik Lab, Harvard University 2022 - 2023
Research Assistant Advised by: Marinka Zitnik

Barabasi Lab, Network Science Institute 2018 - 2020
Research Associate Advised by: Albert-Laszlo Barabasi

Research Computing Team, NEU, Assistant 2018 - 2020
· Manage High-Performance Computing cluster; Monitor, install hardware, software; Teach hands-on tutorials
· Help users parallelize code in different programming languages to run faster; Max speedup 10.3x

AWARDS

2025 Outstanding Graduate Research Award for Ph.D. Students, UMass Lowell School of CS

PUBLICATIONS

Workshop & Preprint

- [1] WebATLAS: Agentic Task-completion with Look-ahead Action Simulation
Jiali Cheng, Anjishnu Kumar, et al. [LAW @ NeurIPS 2025](#) Link
- [2] Investigating Tool-Memory Conflict in Tool-Augmented LLMs
Jiali Cheng, Rui Pan, Hadi Amiri [R2F @ ICML 2025](#) Link
- [3] Do Students Debias Like Teachers? On the Distillability of Bias Mitigation Methods
Jiali Cheng, Chirag Agarwal, Hadi Amiri [R2F @ ICML 2025](#) Link
- [4] Linguistic Blind Spots of Large Language Models
Jiali Cheng, Hadi Amiri [CMCL Workshop @ NAACL 2025](#) Link
- [5] MU-Bench: A Multi-task Multi-modal Benchmark for Machine Unlearning
Jiali Cheng, Hadi Amiri [SafeGenAI Workshop @ NeurIPS 2024](#) Link

Conference

- [6] TRACER: Token ReAssignment for Concept ERasure in Generative Recommendation
Ziheng Chen, **Jiali Cheng**, Zezhong Fan, Hadi Amiri, Diyuan Wu, Gabriele Tolomei, Yang Zhang
[NeurIPS 2026](#) Link
- [7] CURE: Circuit-Aware Unlearning for LLM-based Recommendation
Ziheng Chen, Zezhong Fan, **Jiali Cheng**, Hadi Amiri, Yunzhi Yao, Yang Zhang, Xiangguo Sun
[CIKM 2026](#) Link
- [8] Understanding Machine Unlearning Through The Lens of Mode Connectivity
Jiali Cheng, Hadi Amiri [COLM 2026](#) Link
- [9] A Mechanistic Perspective and Circuit-Guided Difficulty Metric for Unlearning
Jiali Cheng, Ziheng Chen, Chirag Agarwal, Hadi Amiri [ACL 2026 Findings](#) Link
- [10] Fair Cognitive Impairment Detection Through Unlearning
William Nguyen, **Jiali Cheng**, Hadi Amiri [Interspeech 2026](#) Link

- [11] FUTURE: Flexible Unlearning for Tree Ensemble
Ziheng Chen, Jin Huang, **Jiali Cheng**, Yuchan Guo, Mengjie Wang, Lalitesh Morishetti, Kaushiki Nag, Hadi Amiri [CIKM 2025](#) Link
- [12] FROG: Fair Removal on Graphs
Ziheng Chen, **Jiali Cheng**, Gabriele Tolomei, Sijia Liu, Hadi Amiri, Yu Wang, Kaushiki Nag, Lu Lin
[CIKM 2025](#) Link
- [13] Speech Unlearning
Jiali Cheng, Hadi Amiri [Interspeech 2025](#), [Blue Sky Track](#) Link
- [14] Tool Unlearning for Tool Augmented LLMs
Jiali Cheng, Hadi Amiri [ICML 2025](#) Link
- [15] EqualizeIR: Mitigating Linguistic Biases in Retrieval Models
Jiali Cheng, Hadi Amiri [NAACL 2025 \(Short\)](#) Link
- [16] FairFlow: Mitigating Dataset Biases through Undecided Learning for NLU
Jiali Cheng, Hadi Amiri [EMNLP 2024 \(Long\)](#) Link
- [17] MultiDelete for Multimodal Machine Unlearning
Jiali Cheng, Hadi Amiri [ECCV 2024](#) Link
- [18] MedDec: A Dataset for Extracting Medical Decisions from Discharge Summaries
Mohamed Elgaar, **Jiali Cheng**, Nidhi Vakil, Hadi Amiri, Leo Anthony Celi
[Findings of ACL 2024 \(Long\)](#) Link
- [19] CogniVoice: Multimodal and Multilingual Fusion Networks for Mild Cognitive Impairment Assessment from Spontaneous Speech
Jiali Cheng, Mohamed Elgaar, Nidhi Vakil, Hadi Amiri [Interspeech 2024](#) Link
- [20] Exploring the Impact of Model Scaling on Parameter-Efficient Tuning
Yusheng Su, Chi-Min Chan, **Jiali Cheng**, Yujia Qin, Yankai Lin, Shengding Hu, Zonghan Yang, Ning Ding, Xingzhi Sun, Guotong Xie, Zhiyuan Liu, Maosong Sun
[EMNLP 2023 \(Long\)](#) Link
- [21] GNNDelete: A General Strategy for Unlearning in Graph Neural Networks
Jiali Cheng, George Dasoulas, Huan He, Chirag Agarwal, Marinka Zitnik [ICLR 2023](#) Link

Journal

- [22] Language-Specific Representation of Emotion-Concept Knowledge Causally Supports Emotion Inference
Ming Li, Yusheng Su, Hsiu-Yuan Huang, **Jiali Cheng**, Xin Hu, Xinmiao Zhang, Huadong Wang, Yujia Qin, Xiaozhi Wang, Zhiyuan Liu, Dan Zhang
[iScience Volume 27, Issue 12, 111401, Cell Press](#) Link

TEACHING & TALKS

Machine Learning

TA, UMass Lowell, Fall 2025

Graph Machine Unlearning Guest Lecture, Advanced Social Computing Course @ UMass Lowell, Spring 2023

Computing I & II Lab (C Programming)

TA, UMass Lowell, 2023, 2024

Introduction to HPC Cluster with SLURM

Lecturer, Northeastern University, 2018 - 2019

Introduction to Python

Lecturer, Northeastern University, 2018 - 2019

SERVICE

- Committee for SemEval Task Proposals 2027

Reviewer:

- ICLR since 2024, COLM since 2026, ARR since 2024
- KDD since 2025, Interspeech since 2026
- IEEE TPAMI, WACV 2023, EMNLP 2022, NAACL 2022, IEEE Vis 2021, 2022

Mentoring:

- William Nguyen, BS @ UMass Lowell
- Carlo Reyes, BS @ UMass Lowell
- Anugrah Vaishnav, MS @ UMass Lowell
- HaEun Kim, MS @ UMass Lowell
- Tolotra Samuel Randriakotonjanahary, MS @ UMass Lowell
- Camila Camacho, BS @ UMass Lowell