(+86) 172-0157-7950
yanggnay667@gmail.com
https://yiyang-github.github.io/

Yi Yang

Education

University of California San Diego (UCSD) San Diego, United States
Ph.D. student at Halıcıoglu Data Science Institute (HDSI)

- Advisor: Prof. Biwei Huang
- Interests: *Causal Discovery*, *Causality-empowered Foundation Models*

University of Science and Technology of China (USTC) Hefei, China
B.Sc. in Statistics *School of the Gifted Young (special program)* Sep. 2021 - Jul. 2025

- Enrolled one year younger than typical students

Publication

1. **Yi Yang***, Yiming Wang*, Jiahong Yuan. Transformer-based Speech Model Learns Well as Infants and Encodes Abstractions through Exemplars in the Poverty of the Stimulus Environment. *International Conference on Computational Linguistics (COLING 2025, **Oral**)* [PDF]
2. **Yi Yang**, Yiming Wang, ZhiQiang Tang, Jiahong Yuan. Automated Tone Transcription and Clustering with Tone2Vec. *Findings of the Association for Computational Linguistics: EMNLP*, 2024. [PDF]
3. Jiancan Wu*, **Yi Yang***, Yuchun Qian, Yongduo Sui, Xiang Wang, Xiangnan He. GIF: A General Graph Unlearning Strategy via Influence Function. *Proceedings of the ACM Web Conference (WWW)*, 2023. [PDF]
4. Zizhao Zhang*, **Yi Yang***, Lutong Zou*, He Wen*, Tao Feng, Jiaxuan You. RDBench: ML Benchmark for Relational Databases. *arXiv:2310.16837*. [PDF]
5. Yiming Wang, **Yi Yang**, Jiahong Yuan. Normalization through Fine-tuning: Understanding Wav2vec 2.0 Embeddings for Phonetic Analysis. *arXiv:2503.04814*. [PDF]

* indicates co-first authorship

Peer-reviewed Presentations

1. **Yi Yang**, Yiming Wang, Jiahong Yuan. Saving Voices: How AI Can Rescue Endangered Languages in the Digital World? *74th Annual International Communication Association Conference, Beijing Regional Hub (ICA Beijing)*, 2024.
2. **Yi Yang**, Yiming Wang, ZhiQiang Tang, Jiahong Yuan. Automatic Transcription and Representations for Lexical Tones in Sino-Tibetan Languages. *10th International Conference on Computational Social Science (IC2S2)*, 2024.
3. **Yi Yang**, Yiming Wang, Jiawei Yang, Mingjie Zhang, Jiahong Yuan. Quantifying Language Evolution with Transcriptions Only. *The 15th International Conference in Evolutionary Linguistics (CIEL)*, 2024.

Skills

Languages: Chinese (Standard Mandarin, Jianghuai Mandarin), English

Programming: `Python`, `C++`, `R`, machine learning frameworks like `Pytorch`, Bayesian phylogenetic analysis software `BEAST`, familiar with `Linux`

Awards and Honors

• **Guo Moruo Scholarship** (Highest honor for USTC undergrad students)	2025
• **Baosteel Scholarship**, USTC	2024
• **Class 87 Scholarship**, USTC	2024
• **First Prize of the Chinese Mathematics Competitions** (top 8%), Anhui division	2022

Reviewers for

Conferences: *ICLR 2025*, *COLING 2025*, *ACL ARR 2025 February*, *IC^2S^2 2025*

Journals: *IEEE Transactions on Artificial Intelligence (TAI)*, *Telematics and Informatics R*

Workshops: *ICLR 2024 AGI Workshop*

Research Experience

University of Pennsylvania, Advisor: Prof. Mark Liberman Jul. 2024 - Sep. 2024

Is Bayesian Phylogenetics Really Reliable for Language Evolutions?

Keywords: language evolution, bayesian phylogenetics, cognate sets

- Generated cognate datasets and proposed quantitative metrics to evaluate Bayesian Phylogenetics with ground truth, revisited the mathematics behind. (Manuscript)

USTC, Advisor: Prof. Jiahong Yuan Oct. 2023 - Oct. 2024

Can Machines Perceive Speech in Human-like Poverty of the Stimulus Environments?

Keywords: language acquisition, wav2vec2.0, speech recognition

- Designed sparsity and noise scenarios on the phoneme and tone recognition to simulate the Poverty of the Stimulus environments along with three metrics for abstraction: label correction, categorical patterns, and clustering effects
- *wav2vec2.0* can learn, correct, and re-represent exemplars—speech and labels—into abstractions as parameters, moving beyond simple memorization.

Automated Tone Toolkit for Low-resource Indigeneous Sino-Tibetan Languages

- Proposed the first automated tone transcription and clustering methods for documentation, pitch-based similarity representations `Tone2Vec` for analysis
- Released `ToneLab`, to facilitate automated fieldwork and cross-regional analysis.
- Experiments demonstrate that these algorithms are especially beneficial for low resources indigeneous languages, which perform well in transcription and clustering with a small amount of data.

UIUC, Advisor: Prof. Jiaxuan You May. 2023 - Oct. 2023

Machine Learning for Relational Databases

- Defined machine learning tasks as the column value prediction for relational databases and transformed databases into graphs
- Collected hierarchical datasets along and designed multiple tasks to enable meaningful comparisons between ML methods from diverse domains

USTC, Advisor: Prof. Xiangnan He May. 2022 - May. 2023

Machine Unlearning for Structural Data Privacy

Keywords: responsible AI, machine unlearning, interpretability, graph neural networks

- Presented a unified problem formulation of diverse graph unlearning tasks w.r.t. node, edge, and feature by different privacy and security requests.
- Proposed `GIF`, a model-agnostic unlearning method for graphs, which considered the inter-dependency between connected neighbors.
- Deduced the closed-form solution of parameter changes on one-layer graph convolution networks to provide a better understanding of the unlearning mechanism