Yi Yang

(+1) 8802208528   (+86) 17201577950
yiy122@ucsd.edu   yanggnay@mail.ustc.edu.cn
Website: https://yiyang-github.github.io/

Education

University of California San Diego (UCSD) — San Diego, United States
Ph.D. student at Halıcıoglu Data Science Institute (HDSI) — Sep. 2025 - Now

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

Research Interest

Applying statistical principles from causality, bayesian, and information theory to better understand and improve complex systems, i.e., generative AI, cognition, evolution.
- Utilizing ML to build useful tools and act as test subjects in different areas
- Building efficient, reliable, infant-like ML algorithms.

Research Experience

Identifying Causal Structures w/o Invertibility — July. 2025 - Now
University of California San Diego, Advisor: Prof. Biwei Huang
Keywords: causal discovery, identification, learning theory
- Proposed identification theory about causal structures in the presence of latent confounders w/o the long-used but inapplicable invertiable functional assumptions.

Bayesian Phylogenetics for Sino-Tibetan Language Evolution — Jul. 2024 - Sep. 2024
University of Pennsylvania, Advisor: Prof. Mark Liberman
Keywords: language evolution, bayesian phylogenetics, Sino-Tibetan languages
- Generated cognate datasets and proposed quantitative metrics to evaluate Bayesian Phylogenetics with ground truth, revisited the mathematics behind.

Testing & Developing Speech ML in Stimulus Poverty — Oct. 2023 - Oct. 2024
University of Science and Technology of China, Advisor: Prof. Jiahong Yuan
Keywords: language acquisition, wav2vec2.0; tones, low-resource nlp, documentation
- 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.
- Proposed the first automated tone transcription and clustering methods for documentation, pitch-based similarity representations `Tone2Vec` for 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.

Machine Learning for Relational Databases — May. 2023 - Oct. 2023
University of Illinois Urbana-Champaign, Advisor: Jiaxuan You
- 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

Unlearning for Graph Neural Networks — May. 2022 - May. 2023
University of Science and Technology of China, Advisor: Prof. Xiangnan He
Keywords: machine unlearning, interpretability, graph neural networks, statistics
- 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

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*