Yi (Owen) Yang

owen-yang-18.github.io | owen.yang@duke.edu | 470.943.9490

Education

Duke University, Trinity College of Arts and Science **Durham, NC**
Doctor of Philosophy in Computer Science Sep. 2023 - Present

- *Cumulative GPA:* 4.00/4.00
- *Teaching Assistantship:* Introduction to Databases, Fall 2024; Computational Imaging, Spring 2025
- *Relevant Coursework:* Foundations of Blockchain, Theory and Algorithms of Machine Learning, Database Systems, Circuit Complexity, Graph Matrix Analysis

Emory University, College of Arts and Science **Atlanta, GA**
Bachelor of Science in Computer Science and Mathematics Aug. 2019 - May. 2023

- *Cumulative GPA:* 3.92/4.00; Dean's List: Fall 2021 & Spring 2022 & Fall 2022
- *Teaching Assistantship:* Data Structures and Algorithms, Fall 2021 & Fall 2022
- *Relevant Coursework:* Artificial Intelligence, Partial Differential Equations, Data Mining, Computer Architecture, Convex Optimization, Mathematical Statistics, Advanced Algorithms

Publication

Yang, Yi, Yanqiao Zhu, Hejie Cui, Xuan Kan, Lifang He, Ying Guo, and Carl Yang (2022). "Data-Efficient Brain Connectome Analysis via Multi-Task Meta-Learning". In: *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining.*

Yang, Yi, Hejie Cui, and Carl Yang (2023). "PTGB: Pre-Train Graph Neural Networks for Brain Network Analysis". In: *the Conference on Health, Inference, and Learning.* (**Oral**).

Han, Kaiqiao, **Yang, Yi**, Zijie Huang, Xuan Kan, Yang Yang, Ying Guo, Lifang He, Liang Zhan, Yizhou Sun, Wei Wang, and Carl Yang (2024). "BrainODE: Dynamic Brain Signal Analysis via Graph-Aided Neural Ordinary Differential Equations". In: *Proceedings of the IEEE International Conference on Biomedical and Health Informatics.*

Li, Alexis, **Yang, Yi**, Hejie Cui, and Carl Yang (2024). "Brainsteam: A practical pipeline for connectome-based fmri analysis towards subject classification". In: *Pacific Symposium on Biocomputing.* (**Oral**).

Yang, Yi, Han Xie, Hejie Cui, and Carl Yang (2024). "Fedbrain: Federated training of graph neural networks for connectome-based brain imaging analysis". In: *Pacific Symposium on Biocomputing.*

Yang, Yi, Jian Pei, Jun Yang, and Jichun Xie (2026). "Finding Non-Redundant Simpson's Paradox in Multidimensional Data". In: *Proceedings of the VLDB Endowment.*

Zhang, Minxing, **Yang, Yi**, Roy Xie, Bhuwan Dhingra, Shuyan Zhou, and Jian Pei (Apr. 2026). "Generalizability of Large Language Model-Based Agents: A Comprehensive Survey". In: *ACM Comput. Surv.* 58.10. ISSN: 0360-0300. DOI: 10.1145/3794858. URL: https://doi.org/10.1145/3794858.

Zhang, Minxing†, **Yang, Yi**†, Zhuofan Jia, Xuan Yang, Jian Pei, Yuchen Zang, Xingwang Deng, and Xianglong Chen (2026). "MPCEval: A Benchmark for Multi-Party Conversation Generation". In: *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining.* (**Oral**).

Research Experiences

Duke University, Department of Computer Science **Durham, NC**
Research Assistant, Advisor: Jian Pei, Ph.D. Sep. 2023 - Present

- **MPCEval: A Benchmark for Multi-Party Conversation Generation** *(KDD'26 Dataset & Benchmark Track, PDF)*
  - Introduced MPCEval, the first task-aware evaluation and benchmarking suite for multi-party conversation generation, decomposing generation quality into three orthogonal dimensions: speaker modeling, content quality, and speaker–content consistency, while explicitly distinguishing local next-turn prediction from global full-conversation generation.
  - Developed a suite of novel, quantitative, reference-free, and reproducible metrics that scale across datasets and models, addressing fundamental limitations of existing reference-based metrics (e.g., BLEU, BERTScore) and LLM-as-a-judge approaches in open-ended multi-party settings.
  - Conducted empirical studies across diverse public and real-world datasets, revealing systematic dimension-specific model behaviors and demonstrating that human-authored conversations are not a universal gold standard across all evaluation dimensions.

- **Learning Database Semantics through Usage Patterns** *(On-going)*
  - Investigating the recovery of missing or undocumented database metadata – including table/column semantics, column roles, and join relationships – from historical query-SQL usage patterns, addressing a real-world scenario where schema documentation is absent or incomplete.
  - Designing an agentic reasoning framework that combines LLM-based semantic inference with structured SQL pattern analysis (e.g., JOIN structures, WHERE clauses, aggregation patterns), and developing active learning strategies for example selection (coverage-based, semantic diversity, confusion-driven) to maximize semantic recovery under limited query budgets.
  - Demonstrating that recovered semantics directly improve downstream natural language interfaces to databases, achieving up to 80% column description accuracy, 90%+ join graph reconstruction F1, and significant improvements in end-to-end text-to-SQL execution accuracy.

- **General Pattern Co-clustering of Sparse Tabular Data** *(In submission, PDF)*
  - Developing an unsupervised framework for co-cluster discovery in sparse tabular data by learning latent representations that model similarity structures between row and column objects.
  - Performing empirical studies on diverse real and synthetic datasets, providing interpretations to discovered co-clusters to support downstream applications such as recommendation systems.

- **Generalizability of Large Language Model-Based Agents** *(ACM Computing Surveys, PDF)*
  - Conducted comprehensive literature review of 148+ papers on LLM-based agents, establishing the first formal definition of LLM-based agent generalizability and developing a structured evaluation framework for measuring generalizability across four dimensions: instruction-level, task-level, environment-level, and domain-level.
  - Developed systematic taxonomy of strategies for enhancing agent generalizability, organizing 50+ methodological approaches into three categories: (1) backbone LLM optimization through diverse training data and multi-objective loss functions, (2) agent specialized component enhancement including perception filtering, adaptive memory retrieval, and tool diversity, and (3) inter-component interaction mechanisms for effective information transfer and coordination.
  - Identified critical open challenges including development of standardized domain-task ontologies for fair comparison of generalizability claims, variance-based and cost-based evaluation metrics for measuring generalizability beyond aggregate success rates, limitations in existing training

and architectural approaches, and novel frameworks for improving cross-component integration.

- **Paradox Learning and Data Storytelling** *(VLDB'26, PDF)*
  - Discovering that multiple instances of association reversal (*e.g.,* Simpson's paradox) in a multi-attribute dataset can describe the same data subspace – a phenomenon termed redundancy – and conjecturing that redundancy among association reversals establishes an equivalence relation.
  - Designing computational methods for efficient discovery of paradox redundancy by employing pruning strategies to eliminate exhaustive and repetitive data subspace searches, and implementing parallelism and simplified data structures to improve execution and memory efficiency.
  - Constructing a synthetic data generator to uncover relevant conditions and properties of data distributions that precipitate formation of redundant instances of statistical paradoxes.

Emory University, Department of Computer Science **Atlanta, GA**
Undergraduate Research Assistant, Advisor: Carl Yang, Ph.D. Aug. 2021 - Apr. 2023

- **Meta-learning for Brain Network Analysis** *(KDD'22, PDF)*
  - Developed a meta-learning framework for graph neural networks, improving knowledge generalization across brain network systems and adaptation to local datasets with limited samples.
  - Introduced adaptive and task-specific learning trajectories to reconcile heterogeneous cross-tasks model performance due to varying task difficulties, resulting in faster model convergence and a 21% relative improvement in downstream performance over the strongest baseline.

- **GNN Pre-training for Brain Network Analysis** *(CHIL'23, PDF)*
  - Proposed a self-supervised GNN pre-training for brain network data by optimizing a novel contrastive objective that leverages functional and spatial dependencies among cerebral regions, achieving a 13% relative increase in disease prediction accuracy over supervised baselines.
  - Designed a data-driven pre-processing pipeline to address the format misalignment issues in cross-dataset learning with a focus on the preservation of original brain parcellation structures.

- **Federated Learning for Brain Network Analysis** *(PSB'24, PDF)*
  - Implemented a federated learning framework for brain network representation learning, enabling collaborative and distributed learning among multiple institutions with privately enclosed data.
  - Developed a hierarchical clustering of FL participants using clinical prior knowledge and neural circuitry metadata to enhance client personalization, leading to a 14% relative improvement in biomarker and disease prediction accuracy over state-of-the-art FL architectures.

Professional Experiences

Samsung Research America **Mountain View, CA**
Research Engineer Intern @ Mobile Platform and Solutions Jun. 2025 - Aug. 2025
- Developed a heterogeneous GNN framework to model low-level behavioral activities of Android applications, achieving 97.44% multi-class malware classification accuracy across 11,598 APK samples.
- Built an Android client deploying the malware detection pipeline on mobile CPU, performing resource efficient GNN inference on sampled APKs with minimal memory footprint.

Amazon.com **Seattle, WA**
Applied Scientist Intern May. 2026 - Aug. 2026
- Developed an evidence-based explanation framework for LLM-based household attribute prediction from large-scale e-commerce behavior data, connecting purchase/search histories and product metadata to model predictions to improve interpretability and guide more reliable fine-tuning.