Yifan (Holly) Cui

Durham, NC | (805)-280-8850 | yifan.cui@duke.edu | https://hollyyfc.github.io/

EDUCATION

Duke University — Durham, NC

***M.S., Statistics**; GPA: 3.97/4.0* — Aug 2023 – May 2025

***B.S., Statistical Science** (Data Science Concentration), Minor in Computer Science; GPA: 3.958/4.0* — Aug 2021 – May 2023

- **Honors**: Posit PBC Opportunity Scholar, The R Foundation for Statistical Computing Invited Developer, American Statistical Association Member, Harvard Datathon@LISH Top10 Winner, Duke Graduation Magna Cum Laude, Duke Dean's List with Distinction, Duke Chief Student Marshal (2022), Duke Stats Teaching Assistant of the Year Nomination (2022-2024)
- **Coursework**: Predictive Modeling, Multilevel & Hierarchical Models, Bayesian Analysis, Causal Inference, Advanced Data Visualization, Inference Theory, High-Dimensional Statistics, Statistical Computing, Statistical Consulting, Study Design, Machine Learning (ML) & Data Mining, Stochastic Modeling & Probabilistic ML, Natural Language Processing (NLP), Cloud Computing, Data Structures & Algorithms, Computer Systems, Software Quality Management

EXPERIENCE

Eli Lilly and Company — Indianapolis, IN

Master's Statistician Intern *Diabetes Common Safety Tables, Figures, Lists (TFLs) Automation* — May 2024 – Aug 2024

- Developed and launched a Shiny app to automate the creation, execution, and review of common safety TFLs, integrating R and SAS code with output formatting, progress tracking, and error reporting through front-end UI design and back-end cloud system engineering; consolidated 30+ common safety TFLs from 300+ listings across 5+ Diabetes study by building a flexible internal TAFFY template project; reimagined the clinical reporting pipeline with enhanced efficiency and consistency
- Orchestrated regular meetings with senior leadership; pitched the app to 600+ global employees; achieved successful implementations in Diabetes, with ongoing rollouts to Neuroscience and other therapeutic areas

Duke University — Durham, NC

Student Research Affiliate *Machine Learning in Healthcare: Lab Test Harmonization* — May 2022 – Dec 2022

- Optimized lab test deduplication of grouper labels by fine-tuning Bio-BERT, an NLP model pre-trained on biomedical corpora; established a new method of cross-comparison similarity evaluation based on ground-truth text embeddings; uncovered a 95% performance boost in the application to Duke Hospital's lab database
- Demonstrated academic distinction by contributing to the Duke AI Health 2022 cohort as the sole undergraduate participant; effectively communicated research outcomes through a well-received presentation at the Duke AI Health Poster Showcase 2022

Hiya Inc. — Seattle, WA

Data Science Intern *Hiya Shield Project: Robocall Identification & Screening* — May 2022 – Aug 2022

- Spearheaded an NLP-based robocall detection system based on internal audio databases, leveraging SBERT, unsupervised learning, statistical analysis, and AWS Cloud on text- and audio-space manipulation
- Enhanced classification efficiency by discovering optimal audio truncation length and similarity thresholds, driving a 67% faster user experience with a customizable accuracy screening feature for Hiya mobile app

Tsinghua University — Beijing, China

Lead Author & Research Assistant *Cross-Media Retrieval Based on Big Data Technology* — Jun 2020 – Mar 2021

- Refined traditional permutation invariant training with mean squared error loss through BLSTM/LSTM and CNN in a key media separation technique; innovated two new separation methods – the FIX strategy and the masking-based data augmentation strategy, demonstrating notable performance gains
- Publication: *Audio-Visual Single-Channel Signal Separation based on Big Data Augmentation* in IEEE (IICSPI 2020)

PROJECTS

- **TidyTuesday for Python**: Scraped and processed TidyTuesday metadata into Python-friendly formats (CSV/JSON), publishing on Hugging Face with 70+ stars to foster cross-community collaboration in data science and machine learning
- **Operationalizing LLM with Rust & AWS**: Deployed a text summarization service based on BART-large-XSum model using Rust, Docker, and Kubernetes on AWS EKS, incorporating CI/CD pipelines, service monitoring, and load testing
- **Colorblindness Awareness App:** An R Shiny app with generative arts and color manipulation to test and simulate colorblindness
- **Causal Review on Principal Stratification with Noncompliance**: Analyzed causal inference models addressing noncompliance and missing outcome data using mixture models, comparing Bayesian and likelihood-based approaches
- **Health Insurance Cost Prediction**: Utilized regression models and GAM to forecast medical costs for insurance pricing

LEADERSHIP ACTIVITIES

- **Duke Statistical Science Majors Union** | *Chief Technology Officer* (May 2022 - May 2023): Organized 15+ departmental events for 280+ members; oversaw Duke Datathon, alumni networks, club recruitment and advertisement
- **Duke Impact Investing Group** | *Project Manager & Data Analyst* (Aug 2021 - Dec 2022): Managed data analysis for start-up partners with strategic business insights; coordinated student analysts on project outlining, meetings, and client communication

SKILLS & INTERESTS

Technical: R, Python, SQL, SAS, Java, Rust, C || Shiny, Quarto, Git, LaTeX, AWS Cloud, Microsoft Office Suite

Languages: English, Mandarin; **Interests**: Boxing, Cycling, Piano (Level 9), Music Producing, Baking