

# Guangya (Wayne) Wan

857-207-5945 | <https://github.com/wan19990901> | <https://www.linkedin.com/in/guangya-wan-90b289189/>  
wxr9et@virginia.edu | <https://gwan.netlify.app>

## Professional Summary

---

Ph.D. researcher in Data Science advancing the frontiers of Large Language Model (LLM) capabilities and making it trustworthy through novel approaches in reasoning mechanisms, uncertainty estimation, and human-aligned behaviors. Specialized in developing efficient and reliable LLM frameworks with demonstrated applications in healthcare and social science. Seeking summer 2025 internship opportunities to drive innovation in responsible LLM development and deployment in industry.

## Education

---

**University of Virginia** 08/2023 - 05/2027 (Expected)  
**Doctor of Philosophy in Data Science** (Research Interest: **LLM Responsible Reasoning & Agents & Social Intelligence**)  
Advisor: Sheng Li & Tom Hartvigsen

**Harvard University** 08/2021- 05/2023  
**Master of Science in Biostatistics** (With a Concentration in Data Science) **GPA: 3.89**  
Relevant courses: Machine Learning in Healthcare, Natural Language Processing, Parallel Computing, Computer Vision

**University of Illinois-Urbana, Champaign** 09/2017- 12/2020  
Major: Statistics (Summa Cum Laude & James Scholar) **Minor:** Computer Science & Mathematics **GPA: 3.94**  
Relevant courses: Data Mining, Data Structures & Algorithm, Database System, Information Retrieval, Machine Learning

## Skills

---

**Machine Learning & AI:** PyTorch, Hugging Face, LangChain/LangGraph, Transformers, Unsloth, LightGBM

**Big Data & Cloud:** PySpark, AWS (EC2/SageMaker), Snowflake, Vertica, Dask

**Programming:** Python, C/C++, R, SQL, NoSQL, Bash

**Development Tools:** Docker, Git, Unix, VSCode, Cursor

## Professional Experience

---

**Chewy** **Boston, MA**  
**Data Scientist Intern** 06/2022-08/2022

- **People Analytics:** Extracted and analyzed product attributes using Snowflake SQL/PySpark tools for consumer insights.
- **Forecast Modelling:** Enhanced demand forecasting accuracy by 20% by replacing KNN with advanced time-series clustering (DTW/tslearn) and implementing a LightGBM pipeline on AWS SageMaker for new product predictions

**Boston Children Hospital** **Boston, MA**  
**Machine Learning Research Intern (Computational Health Informatics Program)** 01/2022-04/2023

- **Fairness & Text analytics:** Assessed text based ML model fairness in ED admissions using differential demography and equalized odds metrics, experimenting with various text encoding methods including word2vec and Sentence Transformer.
- **Package Development:** Developed a Python package for fairness-aware ML algorithms for Electronics Health Records, including data processing, model fitting (scikit-learn/XGB), and full documentation with version control in git.

**National Center for Super Computing Application** **Champaign, IL**  
**Machine Learning Research Intern (Student Pushing Innovation Program)** 06/2019-12/20

- **Big Data Processing:** Architected end-to-end ML pipelines on HPC infrastructure to analyze 10GB+ virus outbreak data, implementing containerized workflows with Docker for reproducibility.
- **Inference Acceleration:** Optimized data processing through CUDA-accelerated ML models

## Relevant Research Projects

---

**RASC: A Novel Framework for Efficient and Faithful Large Language Model Reasoning**

- Introduced RASC framework that systematically reduces LLM Self-Consistency sampling by ~70% while maintaining accuracy by evaluating both reasoning paths and final answers.
- Demonstrated effectiveness across various reasoning tasks and models, with improved reasoning quality and rationale quality and reduced computational costs compared to existing methods.

**CoT Rerailer: Fixing LLM Reasoning Errors Through Step-level Mistakes Detection and Correction**

- Novel two-stage system that first identifies flawed reasoning (Derailment) then fixes errors (Rerailment) in LLM outputs using consistency checks and multi-agent debate.
- Achieved better accuracy with fewer API calls compared to existing methods across multiple datasets.

**Comparing Direct vs. Step-wise Chain of Thought Improvements in LLMs: A Causal Analysis**

- Study compares two Chain of Thought improvement methods (Direct vs. step-by-step) to understand their causal impact on LLM accuracy and implications on future LLM reasoning improvement directions.

**Proactive LLM Doctor for Mental Health Differential Diagnosis**

- Developed a novel framework integrating structured domain knowledge with LLMs to create proactive conversational agents for Mental Health Differential Diagnosis.

### Mind AI's Mind: Towards Transparent Automated Depression Diagnosis

- Developed a novel medical AI explanation pipeline combining XAI techniques and LLM-based reasoning for transparent depression diagnosis, achieving significant AI-Expert agreement scores.
- Implemented a RAG-enhanced LLM reasoning system to translate statistical model explanations into natural language diagnostic reports, validated by psychiatric experts across multiple depression diagnostic datasets.

### Capability Drives Confidence: Measuring Model Uncertainty through Persuasion Resistance.

- Introduce a novelty perspective to under LLMs' overconfidence by linking it to social-psychological theory.
- Provide a new method to efficiently detect the issues of LLMs' overconfidence measurement.

### Evaluating Linguistic Bias in LLMs Through Race Role-Playing Scenarios

- Developed systematic testing protocols to compare LLM responses between standardized English and African American English Preference tuned LLM, identifying bias patterns across social scenarios and proposing targeted interventions for fairness improvement.

## Honors and Awards

---

- Computational Health Informatics Programs Fellowship, Harvard University (2022)
- Honorable Mention Paper Award, Joint Mathematics Meetings (JMM) (2021)
- Summa Cum Laude (Graduation with highest honors), University of Illinois-Urbana Champaign (2020)
- Student Pushing Innovation (SPIN) Fellowship (2020)
- Canadian Undergraduate Mathematical Conference, Best Presentation Nominee (2020)
- Edmund J. James Scholar, University of Illinois-Urbana Champaign (2018-2020)
- Excellence in Undergraduate Research, Department of Mathematics, University of Illinois-Urbana Champaign (2019)
- Dean's List, University of Illinois-Urbana Champaign (2017-2020)
- Regional Champion, Canadian Senior Mathematics Competition, University of Waterloo (2016)

## Teaching Experience

---

### University of Virginia

- CS 5012: Foundations of Computer Science (Summer 2024)
- DS 6310: Statistics Inference Theory II: Inference & Prediction (Spring 2024)
- DS 6030: Statistical Learning (Fall 2023)

### Harvard University

- CS 109B: Data Science 2: Advanced Topics in Data Science (Spring 2023)

## Service

---

### Conference Reviewer

- Neural Information Processing Systems (NeurIPS) 2024
- International Conference on Learning Representations (ICLR) 2025
- International Conference on Artificial Intelligence and Statistics (AISTATS) 2025

### Journal Reviewer

- IEEE Computational Intelligence Magazine (CIM)

## Selected Publications

---

1. **Wan G**, Wu Y, Hu M, Chu Z, Li S. "Bridging causal discovery and large language models: A comprehensive survey of integrative approaches and future directions." *arXiv preprint* arXiv:2402.11068 (2024).
2. **Wan G**, Wu Y, Chen J, Li S. "Dynamic Self-Consistency: Leveraging Reasoning Paths for Efficient LLM Sampling." *arXiv preprint* arXiv:2408.17017 (2024).
3. **Wan G**, Wu Y, Chen J, Li S. "CoT Rerailer: Enhancing the Reliability of Large Language Models in Complex Reasoning Tasks through Error Detection and Correction." *arXiv preprint* arXiv:2408.13940 (2024).
4. McCoy JA, Levine LD, **Wan G**, et al. "Intrapartum electronic fetal heart rate monitoring to predict acidemia at birth with the use of deep learning." *American Journal of Obstetrics and Gynecology* (2024).
5. **Wan G**, Allen J, Ge W, et al. "Two-step light gradient boosted model to identify human west nile virus infection risk factor in Chicago." *PLOS ONE* 19(1): e0296283 (2024).
6. La Cava WG, Lett E, **Wan G**. "Fair admission risk prediction with proportional multicalibration." *Conference on Health, Inference, and Learning*, 350-378 (2023).
7. **Wan G**, Yang L, Bachina P, Qiu RN. "Testing Mathematical Models of Diabetes Against Blood Glucose Data." Canadian Undergraduate Mathematical Conference (2020).