Jiacheng Zhu

Master Student in Data Science @ University of Pennsylvania

***Research Interests**: Natural Language Processing, Data-Centric AI, Synthetic Data, LLM*

(+1) 619-548-8389 | jiachzhu@seas.upenn.edu | jiachengzhu.com | jjz5463 | jjz5463

Education

University of Pennsylvania, School of Engineering and Applied Science — *Philadelphia, PA*

Master of Science in Engineering — *May 2025*

Data Science — *GPA: 3.88/4.0*

- *Relevant coursework*: Machine Learning, Big Data Analytics, Natural Language Processing, Large Language Model, Databases and Information Systems, Analysis of Algorithms, Computer Vision, Linear Algebra, Natural Language Processing Research Practicum

The Pennsylvania State University, Eberly College of Science — *University Park, PA*

Bachelor of Science — *Dec 2022*

Data Science — *GPA: 3.92/4.0*

- *Honor*: Magna Cum Laude
- *Relevant coursework*: Probabilities, Mathematical Statistics, Statistical Computing, Statistical Inference, Object-Oriented Programming

Papers

Ajay Patel, **Jiacheng Zhu**, Justin Qiu, Zachary Horvitz, Marianna Apidianaki, Kathleen McKeown, and Chris Callison-Burch (2024). *StyleDistance: Stronger Content-Independent Style Embeddings with Synthetic Parallel Examples*. (Preprint, under review for NAACL). arXiv: `2410.12757 [cs.CL]`. URL: `https://arxiv.org/abs/2410.12757`

Academic and Research Experience

University of Pennsylvania, Penn NLP — *Philadelphia, PA, USA*

NLP Teaching Assistant — *Aug 2024 - Present*

- Assisted Professor Mark Yatskar in teaching CIS 530: Natural Language Processing by conducting office hours, mentoring students, and actively participating in discussion boards.
- Managed grading processes and organized student inquiries to ensure timely and effective feedback.
- Addressed technical and conceptual questions from students, applying debugging skills and expertise in Python, PyTorch, and Hugging Face for NLP models.

NLP Research Assistant — *Mar 2024 - Present*

- Collaborated with Ph.D. student Ajay Patel on the HIATUS (Human Interpretable Attribution of Text using Underlying Structure) project at Penn NLP, under the guidance of Professors Chris Callison-Burch and Marianna Apidianaki.
- Contributed to a paper submitted to NAACL 2025 titled “StyleDistance,” focusing on content-independent style embeddings. Developed a synthetic dataset of near-exact paraphrases with controlled style variations, encompassing positive and negative examples across 40 distinct style features for precise contrastive learning.
- Enhanced skills in debugging NLP models, analyzing results, and adapting to evolving research challenges, thereby establishing a solid foundation for future research in natural language processing.

The Pennsylvania State University, Statistics Department — *University Park, PA, USA*

Research Assistant in Statistics — *May 2022 - Dec 2022*

- Contributed to the PHIA (Population-based HIV Impact Assessment) project under Professor Le Bao, analyzing large-scale global health datasets to uncover critical trends in the global HIV epidemic.
- Applied advanced data analysis techniques, including statistical modeling and computational methods, to derive insights from multi-source health data, providing actionable recommendations for health policy.
- Developed and implemented data pre-processing pipelines, including feature engineering, to optimize datasets for statistical and machine learning models, ensuring data quality and accuracy.
- Collaborated with cross-disciplinary teams, leveraging computational approaches to analyze high-dimensional data, which enhanced my ability to work on large-scale data analysis in applied contexts.

Professional Experience

UISEE, Autonomous Driving Startup — *Shanghai, China*

Data Engineer — *Feb 2023 - Jun 2023*

- Collaborated with a team to support the development of a customized S3-like object storage system and private cloud infrastructure, enabling secure and scalable data management for autonomous driving research.
- Contributed to optimizing data pipelines, focusing on reducing latency and improving data accessibility for machine learning workflows, which enhanced system performance.
- Assisted in developing and debugging APIs for internal storage services, helping to streamline data access for researchers and engineers working on machine learning models.
- Worked alongside cross-functional teams to integrate cloud-based solutions with real-time data streams, facilitating efficient access to large-scale datasets in autonomous vehicle research environments.

Projects

Large Language Model-Generated Text Detection — *May 2024*

Natural Language Processing — *GitHub · Report*

- Developed a binary classification model to distinguish GPT-generated text using author attribution techniques and sentence embeddings, addressing challenges in detecting AI-generated content.
- Created a robust dataset covering single-genre, cross-genre, and style-shifting tasks, achieving 90% accuracy on cross-genre examples and 74% accuracy on style-shifting tasks, highlighting the model's ability to perform well in difficult classification scenarios.
- Conducted a comprehensive failure case analysis, identifying the impact of prompt engineering on classification performance and implementing dimensionality reduction techniques to enhance detection accuracy.
- Proposed improvements for detecting style-shifted GPT-generated text, demonstrating the potential for advancing author attribution and text generation detection techniques in natural language processing research.

Simple Yelp Website — *May 2024*

Database + Web Programming (React, Node.JS) — *GitHub · Report*

- Developed a full-stack web application, "Simple Yelp," enabling users to evaluate businesses through comprehensive reviews. Leveraged React for the frontend, NodeJS and Express for server-side logic, and MySQL hosted on AWS RDS for data storage.
- Utilized open-source data from Yelp, managing large datasets of over 7 million records, decomposing data into 3NF (Third Normal Form) to maintain data integrity and support efficient query execution.
- Designed and optimized complex SQL queries, including featured reviews and nearby businesses based on user geo-locations, significantly improving query performance through indexing and reducing processing time for large datasets.
- Implemented a client-server architecture to facilitate seamless communication between the frontend and backend, ensuring efficient handling of user interactions and real-time data updates.

Patient Discharge Disposition Classification — *Dec. 2023*

AI for Health + Natural Language Processing — *GitHub · Report*

- Collaborated with a PhD student from Bioengineering to investigate the use of advanced NLP models for predicting patient discharge disposition (Home, Extended Care, Deceased) using MIMIC-IV emergency room admission notes.
- Utilized BERT, PubMedBERT, and GPT-3.5 to assess their efficacy in clinical data interpretation, contributing to enhanced predictive modeling for patient discharge outcomes.
- Achieved 78% accuracy with PubMedBERT, outperforming RoBERTa's 76%, demonstrating the advantages of domain-specific models in medical applications.
- Conducted in-context learning experiments with GPT-3.5, identifying a lower accuracy of 68%, and provided insights into the limitations of general-purpose models for healthcare data.
- Emphasized the importance of domain-adapted models like PubMedBERT for analyzing complex medical records and supporting clinical decision-making.
- Gained interdisciplinary experience by working closely with the Bioengineering department, fostering collaboration between computational methods and medical research.

Part of Speech Tagger with HMM — *Oct. 2023*

Natural Language Processing + Machine Learning — *GitHub · Report*

- Developed a scalable, hybrid neural network and Hidden Markov Model (HMM) for Part-of-Speech (POS) tagging, achieving a 96.06% F1 score on the Penn Treebank dataset.
- Leveraged large-scale datasets containing over 1.1 million words, incorporating advanced smoothing techniques and optimized inference algorithms to enhance tagging accuracy across diverse linguistic contexts.
- Engineered a robust, PyTorch-based method for handling out-of-vocabulary (OOV) words, improving unseen word classification accuracy to 83.89%, demonstrating significant gains in model generalization.
- Enhanced the Viterbi algorithm for faster state tracking using NumPy, reducing computational overhead by optimizing matrix operations and achieving real-time inference speeds.
- Conducted in-depth performance analysis across n-gram contexts, finding trigrams to provide an optimal trade-off between computational efficiency and predictive accuracy, guiding model selection for broader NLP tasks.

Skills

Programming	Python, R, SQL, LaTeX, C/C++ (intermediate), JAVA (intermediate)
Database	MySQL, Oracle, MongoDB, Neo4j
DevOps	AWS, Git, Docker
Back-end	Django, Node.js
Front-end	HTML, CSS, React
NLP / Machine Learning	Hugging Face, PyTorch, TensorFlow, BERT, GPT, BERT, Scikit-learn
Data Analysis	NumPy, Pandas, Matplotlib, SciPy, Seaborn
Languages	English, Mandarin
Research & Soft Skills	Data Analysis, Experimental Design, Critical Thinking, Problem Solving, Collaboration, Mentoring, Communication

Certificates

DeepLearning.AI Deep Learning Specialization

Deep Learning Skills *Aug 2023*

DeepLearning.AI Machine Learning Specialization

Machine Learning Skills *Aug 2023*

MITx MicroMasters Machine Learning

Machine Learning Skills *May 2023*

MITx MicroMasters Fundamentals of Statistics

Mathematical Skills *May 2023*

MITx MicroMasters Probability The Science of Uncertainty and Data

Mathematical Skills *May 2023*