# Marcus Ma

mamarcus64.github.io
linkedin.com/in/marcusma
mamarcus64@gmail.com

## Education

### University of Southern California

### PhD in Computer Science (Advisor: Shrikanth Narayanan)

AUGUST 2024 – PRESENT

- Investigating the interaction of language, cognition, and artificial intelligence to create better representations of data-driven linguistic structures
- Analyzing data from a range of disciplines including LLM architectures, fMRI brain scans and EEG signals to understand the complex relationships between conscious thought and language

### Georgia Institute of Technology

### MS in Computer Science, Machine Learning Specialization (Advisor: Wei Xu)

AUGUST 2023 – MAY 2024

- 4.0 GPA; Relevant Coursework: Game AI, Advanced Natural Language Processing, Theory of ML
- Master's Thesis: leveraging novel NLP techniques and LLMs for cross-genre authorship attribution and verification; creating a new SOTA domain-adaptive attribution model for multi-genre documents

### BS in Computer Science, Intelligence and Theory Concentration

AUGUST 2020 – MAY 2023

- 4.0 GPA (graduated summa cum laude, Dean's List every semester)
- Relevant Coursework: Machine Learning, Natural Language Processing, Algorithms Honors, Data Structures, Advanced Linear Algebra, Number Theory, Computational Machine Learning, Computer Vision

## Selected Publications

- **Ma, M.**, Le, D., Kang, J., Dou, Y., Cadigan, J., Freitag, D., Ritter, A., Xu, W. CROSSNEWS : A Cross-Genre Authorship Verification and Attribution Benchmark. *Under Submission to AAAI 2025.*
- Hall, K., **Ma, M.**, Zhang, X., Swain, V., Kim, J. The Double-Empathy Problem & HCAI: Surfacing Misalignments Between LLMs and Neurodivergent Jobseekers. *Under Submission to CHI 2025.*
- **Ma, M.**, Kim, C., Hall, K., Kim, J. It Takes Two to Avoid Pregnancy: Addressing Conflicting Perceptions of Birth Control Responsibility in Romantic Relationships. *Proc. ACM Human-Computer Interaction*. 7, CSCW2, Article 282 (October 2023).
- **Ma, M.,** Glass, L. and Stanford, J. Introducing Bed Word: A New Automated Speech Recognition Tool for Linguistic Interview Transcription. *Linguistics Vanguard*, 2024.
- **Ma, M.** and Glass, L. (2022, Oct 13-15). BedWord: Computer-generated, human-edited TextGrids for faster sociolinguistic transcription. NWAV 50 Conference, Palo Alto, CA.
- Seeburger, D., Xu, N., Godwin, C., **Ma, M.,** Keilholz, S., Schumacher, E. (2022, Apr 24-27). Identifying the Neural Mechanisms of Zone State Performance using Time-varying Functional Connectivity Methods. CNS 2022 Conference, San Francisco, CA.

## Ongoing Research Projects

### PRECOG | USC SAIL

- Examining neural and biobehavioral signals for preconscious indicators of depression and suicidality
- Creating deep learning representations of eye tracking movement under specific affective and surprising stimuli

### SPAN | USC SAIL

- Visualizing language speech formation on a biological level, examining subject-level and longitudinal differences in speech patterns and production
- Creating a fully-automated pipeline for vocal tract area segmentation using a combination of hand-annotations and finetuned SAM-2 object segmentation models

### LASIDAD | USC SAIL

- Examining audio and textual contexts of older speakers to predict signs of mental impairment and/or dementia
- Comparing the effects of multilingualism as a deterrent against cognitive decline
- Investigating the role that societal, internal, and community exposomes have to influence markers of dementia and other cognitive impairments

## Teaching Experience

**Graduate Teaching Assistant for CS 4650: Natural Language Processing | Jan 2023 – May 2024**

- Assist in leading classes of 75+ students through NLP, covering topics from encoders to present-day LLMs
- Hold weekly office hours and grade 200+ assignments per semester

**Tutoring and Academic Services: One-on-one Tutor, PLUS Leader | Aug 2021 – Dec 2022**

- Tutored over 200 Georgia Tech students in intro-level math and computer science courses
- Prepared and taught ~30 weekly PLUS Sessions (~1200 student cumulative attendance) for Data Structures

## Previous Research Labs

**NLP-X Lab (Natural Language Processing) | Advisors: Wei Xu and Alan Ritter**

*Cross-Domain Authorship Analysis (Planned Submission, ACL 2024)* | January 2022 – August 2024

- Created the largest public cross-genre authorship dataset across hundreds of different news sites using a combination of hand-picked gold labels and thousands of silver-labelled authors determined from statistical models of social media network graphs and string linkage methods
- Leveraged this dataset to create a domain-adaptable authorship attribution model trained via contrastive learning on negative-mined paraphrases that reaches SOTA performance on various authorship tasks

**Co-Well Lab (Human-Computer Interaction) | Advisor: Jennifer Gahee Kim**

*Personalized LLMs for Neurodivergent Jobseekers (Under Submission, CSCW 2024)* | Mar 2023 – Aug 2024

- Despite huge gains in conversational AI since the democratization of LLMs, chatbots are often unaccommodating of neurodivergent clients with personalized communication preferences
- Currently leading a pilot study on how to stylize LLMs via prompt engineering to adapt to different strengths and skills in the context of providing career advice to neurodivergent jobseekers

*Contraception Collaboration between Couples (Published in CSCW 2023)* | August 2021 – March 2023

- Investigated the expectations placed on women in relationships for taking birth control pills consistently, with findings indicating responsibility lies entirely on women and female-sex individuals
- Developed an app prototype designed to split the responsibility of taking pills more equitably between partners and interviewed nine couples to investigate the intersection of contraception responsibility and technology, findings published in CSCW 2023

**Energiesystemtechnik Jülich (Artificial Intelligence) | Advisor: Manuel Dahmen**

*RISE Scholar Recipient and Researcher* | July 2022 – October 2022

- Chosen as one of 320 scholarship recipients out of 1365 applicants to research in Germany at Forschungszentrum Jülich
- Investigated how to use McCormick convex-concave relaxations to improve verifiable robustness of neural networks, with hopes of later applying these findings to optimize resource consumption of green energy process systems
- McCormick relaxations were shown to be ~40% tighter than the alternative Interval Bound Propagation on several internal tests

**CoNTRoL Lab (Neuroscience) | Advisor: Eric Schumacher**

*Undergraduate Research Assistant* | August 2021 – May 2022

- Investigated the Quasi-Periodic Pattern relationship between different attention networks in the brain
- Wrote a script that parsed raw four-dimensional MRI scan data into detection of QPP's, with specific attention paid to the DMN and TPN brain network regions
- Ran data pipelining scripts, QPP detection algorithms, and MRI scan visualizations in Matlab and Bash on Georgia Tech's neuroscience Linux cluster

**Language and Politics in the New South (Sociolinguistics) | Advisor: Lelia Glass**

*Undergraduate Research Assistant (To Appear in Linguistics Vanguard 2023)* | January 2021 – April 2023

- Investigated the relationship between accent and political leanings, based on interviews from ~100 Georgia Tech undergrads; research findings featured in Wall Street Journal, NPR, Washington Post
- Interviewed, transcribed, and analyzed ~20 participants and visualized findings with R scripts
- Created an automated transcription tool, Bed Word, that leverages existing speech-to-text NLP models to make auto-generated TextGrid files, improving human transcription time by ~200%
- Integrated Bed Word with an existing linguistics website hub, which to date has been used to automatically transcribe 1000+ audio hours and 600+ transcriptions for the wider linguistics community

## Selected Work History

### Amazon Web Services

*Software Developer Intern: Optical Networking Team* | May 2023 – August 2023 |Full-Time Position Offered

- Spearheaded the transition of optical device networking communication from internal Amazon servers to the more scalable and maintained Native AWS environment, saving ~$10,000 of yearly server operation costs
- Created a DNS device hostname resolution script that was deployed to hundreds of existing optical devices in the AWS fleet, improving operational efficiency over previous method by ~40%
- Added a robust testing and logging system that monitors the hostname resolution script that automatically initiates rollback in case of failure

*Software Developer Intern: Postgres SQL Team* | May 2022 – August 2022 |Full-Time Position Offered

- Worked in Postgres source code to implement an optimization on the SQL Join command that used semi-join hash filtering to remove unnecessary table rows early in the execution tree
- Improved the speed of certain SQL join commands by up to 36% and integrated filter decisions for the merge join strategy into Postgres query planner, improving AWS Postgres' overall TPC-H database benchmark performance by ~5%
- Created an internal SQL test suite to verify improvement and deployed these tests to the entire AWS codebase, with future work looking to contribute semi-join filtering optimization to an upcoming Postgres community release

### Prognomiq

*Software Engineer Intern* | May 2021 – August 2021

- Created an automated medical records pipeline with MongoDB and AWS Lambda that merges existing field data with incoming forms and handles conflicts via database version control
- Improved query speeds by ~70% compared to previous method and saved about 15 hours of labor per week via a new data query interface
- Set up software and hardware tools for blood sample inventory management across sample freezers

## Programming Languages and Libraries

- **Expert:** Python, AWS, SQL, Pytorch, Huggingface, Pandas, Git, Bash
- **Proficient:** Java, Javascript, C++, C#, .NET, MongoDB, C, TensorFlow, Docker, Kubernetes, React, Flask
- **Familiar With:** HTML/CSS, PHP, Cypher, Go, Assembly, Ruby, Node.js, Swift