MAGGIE ZHANG

(510)-789-7615 | maz007@ucsd.edu | LinkedIn | Website | GitHub

EDUCATION

University of California, San Diego — La Jolla, CA

B.S. in Data Science, B.S. in Cognitive Science - Machine Learning — September 2023 – March 2027

- GPA: 3.8
- Minor in Linguistics

RESEARCH EXPERIENCE

Undergraduate Researcher — September 2024 – Present

UCSD Language Acquisition and Sound Recognition Lab — *La Jolla, CA*

- Process and clean **complex behavioral and speech datasets** using **Python (pandas)**, validating forced-speech aligner results against **human annotations** to ensure **data quality** for bilingual and child language research.
- Design and implement experimental workflows for the **L2Talk** study using **Qualtrics surveys** and **eye-tracking methods**, analyzing speech perception and production differences across bilingual and monolingual cohorts.
- Coordinate **on-site data collection** at **local preschools** by managing **IRB protocols**, designing demographic surveys, and administering consent forms to ensure compliance when working with **minor subjects**.
- Engineered a **Python** pipeline to evaluate OpenAI's **Whisper ASR** models against human transcription baselines, benchmarking performance on heavily accented and out-of-distribution speech.
- Calculated **Word Error Rates (WER)** using the *jiwer* library, implementing robust data cleaning protocols to normalize homophones, standardizing human typos, and isolating ASR hallucination/looping errors.
- Built **logistic mixed-effects models** to quantify the impact of syntactic context and accent familiarity on transcription accuracy, analyzing how Whisper's top-down language models differ from human bottom-up phonetic processing.

Research Assistant — January 2026 – Present

UCSD PURL (New Apostolic Reformation Study) — *La Jolla, CA*

- Collect and annotate video transcripts via **data mining**, transforming **unstructured text** into **structured datasets** for **supervised learning analysis** of modern-day ideological language.
- Develop and refine **codebooks** to ensure **high intercoder reliability** (target $\kappa > 0.80$), maintaining **data integrity** and resolving ambiguities across iterative coding cycles.

Student AI Engineer — July 2025 – September 2025

Swartz Center for Computational Neuroscience — *La Jolla, CA*

- Curated and validated large-scale EEG datasets using **Hierarchical Event Descriptors (HED)** within the **BIDS** standard, enabling **automated, scalable ML pipelines** for cross-study reproducibility.
- Expanded core **machine learning and signal-processing** capabilities in **EEGLAB** (100k+ users) by developing **production-ready Python/MATLAB modules**, automated tests, and Git version-controlled releases.
- Implemented and evaluated **machine learning classifiers** on high-dimensional EEG time-series data to support **BCI state decoding** and cognitive load estimation.

Bioinformatics Data Intern — January 2024 – March 2025

FDI Lab SciCrunch — *La Jolla, CA*

- Developed **automated Python** and **SQL-based data validation workflows** to improve accuracy, traceability, and **scalability** of large bioinformatics text extraction pipelines spanning **100,000+ scientific articles**.
- Improved **Python regex scripts** and SQL database storage to increase the accuracy and efficiency of text extraction for Research Resource Identifiers (RRIDs).
- Developed **Google Apps Script automation** to streamline data processing, optimize file management, and resolve data discrepancies across multiple information systems.

Bioinformatics Undergraduate Researcher Jan 2024 – Jun 2024, Sep 2024 – Mar 2025
UCSD FDI Lab *La Jolla, CA*

- Engineered automated **Python** pipelines to extract structured knowledge and entities from a corpus of **100,000+ scientific papers**, significantly improving parsing scalability.
- Designed and optimized **Python regex scripts** and **SQL** schemas to accurately identify, classify, and validate Research Resource Identifiers (RRIDs) within dense academic text.
- Built **Google Apps Script automation** to orchestrate large-scale document parsing workflows, resolve data discrepancies, and optimize information retrieval across multiple databases.

PROFESSIONAL EXPERIENCE

Data Science Intern June 2026 – Present
AGCO Corporation *Duluth, GA*

- **Architect** an **AI/ML-driven Global Forecast Signal Dashboard** to pressure-test regional financial forecasts across global markets (EME, North America, South America, APA), transitioning standard financial reporting into **predictive, actionable intelligence**.
- **Modularize financial data pipelines** via **AI-assisted code structures**, enabling scalable, continuous integration of disparate inputs like retail sales, corporate financial performance, and external macroeconomic indicators.
- **Deploy time-series ML models** to uncover predictive financial trends, directly analyzing the variance between forecasted variables and historical actuals to quantify **forecast support, risk, and divergence**.
- **Translate complex ML outputs** into high-level strategic business insights, developing an end-to-end prototype designed to equip the **CFO and executive leadership** with data-backed risk assessments for corporate financial planning.

Data Science Intern June 2025 – August 2025
Cenergy Power *Aliso Viejo, CA*

- Built an **NLP sentiment analysis pipeline** to extract regulatory risk signals from unstructured municipal board meeting minutes, quantifying **Authority Having Jurisdiction sentiment** toward solar permitting.
- Integrated NLP models, **computer vision** (OpenCV, CNNs), and geospatial data into a **coordinated ML site-scoring system** that identified **5+ high-potential sites** and improved pipeline efficiency by **20%**.
- Designed actionable dashboards using **Tableau** and **Matplotlib** to synthesize **regulatory trends**, delivering data-driven insights to **non-technical stakeholders** to support permit decision-making.
- Engineered an **NLP sentiment analysis pipeline** in **Python** to parse unstructured municipal board meeting minutes, extracting and encoding structured regulatory risk signals and policy sentiment.
- Designed robust **Python workflows** for batch processing large-scale text and geospatial datasets, ensuring reproducible model inference and seamless data integration into production pipelines.
- Developed a **computer vision pipeline** (OpenCV, YOLO) to extract structured features from aerial and satellite imagery, translating raw visual patterns into quantitative variables (e.g., infrastructure density).

Mathematics Instructor December 2025 – Present
Mathnasium Learning Center - Balboa *San Diego, CA*

- Completed 100+ hours of instructor training in child development, metacognition, and instructional design, including use of manipulatives, reward systems, and adaptive redirection strategies.
- Delivered individualized mathematics instruction to K–12 students, fostering number sense, quantitative intuition, and confidence through developmentally appropriate teaching methods.
- Coordinate community-oriented educational events (e.g., Pi Day, Financial Literacy Month, summer programs), integrating interactive activities to cultivate a positive, high-engagement environment for STEM learning.

Related Technical Projects

Meeting Transcript QA via RAG Pipeline | *LangChain, Hugging Face, Vector Databases, Python* Spring 2026

- Engineered a Retrieval-Augmented Generation (RAG) system using LangChain to process and accurately query long-form, multi-domain meeting transcripts from the QMSum dataset.
- Integrated pretrained Hugging Face text encoders with vector databases for efficient neural memory storage and retrieval, establishing an evaluation pipeline to measure precision@k metrics.

Competitive Analysis of AI Infrastructure | *PyTorch, Python* Winter 2026

- Conducted a systematic competitive analysis of object detection paradigms (YOLO vs. Faster R-CNN) under strict hardware constraints (RTX 3050 Ti with 4GB VRAM).
- Mapped the Pareto frontier of accuracy vs. computational cost, identifying architectural bottlenecks to optimize performance for enterprise edge-compute platforms.

Multi-Agent LLMs for Causal Discovery | *Python, Gemini APIs, JSON* Winter 2026

- Architected autonomous multi-agent system with specialized roles (Generator, Judge, Decision) to integrate LLM domain expertise into trend justification and uncover subtle patterns missed by traditional statistical methods.
- Optimized performance, inference latency, and API costs via deterministic decoding and structured JSON outputs, improving model recall from 0.240 to 0.560 while maintaining a 0.520 F1-score for complex hypothesis exploration.

Understanding Hurricanes | *D3.js, JavaScript, HTML, CSS, pyproj* Fall 2025

- Developed an interactive data storytelling platform using D3.js to visualize 100+ years of longitudinal data, identifying intensity and frequency trends to communicate public impact.
- Engineered dynamic dashboards with filtering and geospatial simulations (pyproj) to translate complex data into actionable insights for stakeholders, including focused socio-economic impact case studies.

Hybrid Book Recommendation System | *Python, XGBoost, scikit-learn* Fall 2025

- Developed a hybrid recommendation engine for Goodreads book reviews, utilizing XGBoost and matrix factorization models to synthesize user behavior and item features, such as ratings and length.
- Optimized model parameters to achieve a highly accurate Root Mean Square Error (RMSE) of 0.8019 on large-scale validation datasets.

Scalable Cloud Architecture for Retail ML | *Apache Spark, PySpark, Linux* Fall 2025

- Architected a scalable data pipeline using Apache Spark in a remote Linux cluster environment to process, flatten, and impute over 25 GB of complex, nested e-commerce datasets.
- Engineered an NLP feature extraction workflow using Word2Vec and PCA, and optimized a distributed machine learning model via rigorous hyperparameter tuning to maximize predictive performance.

Culinary Computing: Recipe Analysis & Prediction | *pandas, scikit-learn, Plotly, HTML* Spring 2025

- Engineered and validated predictive features using scikit-learn pipelines (GridSearchCV, Random Forest Regression) to forecast recipe cook time, followed by fairness and error analysis to evaluate model reliability.
- Conducted rigorous statistical analysis, including missingness assessment and permutation testing, interpreting p-values in context to identify drivers of prediction error and data inconsistency.

Char RNN Text Modeling | *PyTorch, NLTK, Transformers, pandas* Winter 2024

- Built GRU and LSTM character-level language models in PyTorch to automate stylistic text generation and classification, benchmarking against n-gram baselines.
- Conducted structured hyperparameter tuning to optimize model performance, establishing a foundation for AI-driven document review and text mining of unstructured data.

TECHNICAL SKILLS

Programming Languages: Python, R, SQL, JavaScript (ES6+), Bash/Shell, C++, Java, MATLAB, HTML/CSS, SAS
Natural Language Processing: Transformers (BERT, RoBERTa), LLMs, Prompt Engineering, TF-IDF, NLTK, VADER, jiwer, Text Classification, Sentiment Analysis
Machine Learning & AI: Deep Learning, Representation Learning, Recommender Systems, Causal Discovery
Computer Vision & Image Processing: OpenCV, YOLO, scikit-image, PIL, Image Processing
Time-Series & Signal Processing: Time-Series Analysis, Signal Processing
Frameworks & Libraries: PyTorch, TensorFlow, Hugging Face, LangChain, Scikit-learn, Pandas, NumPy, SciPy, Statsmodels, XGBoost, PySpark, Dask, BeautifulSoup, Django, Surprise
Data Engineering & Cloud: ETL Pipelines, PostgreSQL, Vector Databases, AWS (EC2, S3), Databricks, Hadoop
Developer Tools & MLOps: Git/GitHub, Linux/Unix, CI/CD, JSON, Regex, Google Apps Script
Statistics: A/B Testing, Hypothesis Testing, ANOVA, Bayesian Methods, Experimental Design, Logistic Mixed-Effects Models, Permutation Testing, Probability Distributions
Data Visualization: Tableau, Power BI, D3.js, Plotly, Matplotlib, Seaborn, Streamlit
Research Tools & Methods: Qualtrics, PsychoPy, Praat, EEGLAB, Eyetracker, IRB Protocols

RELEVANT COURSEWORK

Machine Learning & AI: Statistical NLP, Language Models as Cognitive Models, Advanced Machine Learning, Deep Learning for NLP, Neural Networks, Representation Learning, Recommender Systems & Web Mining, Probabilistic Models, Causality in ML/AI, Supervised Learning
Data Science & Engineering: Scalable Analytics (Distributed Computing), Data Science in Practice, Data Visualization, Signal Processing, Image Processing, Data Structures & Algorithms
Cognitive Science & Neuroscience: Cognitive Neuroscience, Systems Neuroscience, Brain-Computer Interfaces, Language Representation in the Brain, Learning, Memory, & Attention
Mathematics & Statistics: Applied Statistical Data Analysis, Statistical Methods, Probability, Linear Algebra, Vector Calculus, Statistical Research Methods
Linguistics: Information Theory, Phonetics, Language Development, Historical Linguistics, Language, Culture & Education

PROFESSIONAL MEMBERSHIPS

Women in Computing @ UCSD: Upperclassman mentor guiding freshmen in tech career pathways and enrollment.
Association for Computing Machinery (ACM) @ UCSD: Active member of the AI division.
Hackathons: Participant in DataHacks 2025 and 2026, UCSD's premier data science competition.
Triton Foodie Executive Board: Community Chair responsible for organizing large-scale student networking events.