Maggie Zhang

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

Education

University of California, San Diego — La Jolla, CA
B.S. Data Science & B.S. Cognitive Science (Spec. Machine Learning) — September 2023 - March 2027

- GPA: 3.8, Minor in Linguistics
- Coursework: Advanced Machine Learning, Statistical NLP, Deep Learning, Neural Networks, Causality, Scalable Analytics, Data Visualization, Data Science in Practice, Applied Statistical Data Analysis, Data Structures & Algorithms, Probability

Experience

Data Science Intern, Finance — June 2026 - Present
AGCO Corporation — Duluth, GA

- **Architect** an **AI/ML-driven Global Forecast Signal Dashboard** to pressure-test regional financial forecasts, translating complex ML outputs into actionable, data-backed risk assessments for the **CFO and executive leadership**.
- **Modularize financial data pipelines** via **AI-assisted code structures**, enabling scalable, continuous integration of disparate inputs like retail sales, corporate 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**.

Data Science Intern — June 2025 - August 2025
Cenergy Power — Aliso Viejo, CA

- Built an **NLP classification pipeline** using **TF-IDF** and **NLTK** to analyze municipal meeting minutes and historical permitting records, identifying sentiment and solar project approval features to directly inform a live permitting campaign.
- Engineered a **computer vision pipeline** (OpenCV, YOLO) to extract infrastructure features from aerial imagery, integrating **model inference** and **performance evaluation** for geospatial and text data into one reproducible pipeline.
- Designed **site-scoring algorithm** that improved selection efficiency by 20% and **identified 5+ high-potential sites**, collaborating with marketing and R&D teams to present model insights and evaluate location permitting trade-offs.

Data Curation Intern — January 2024 - March 2025
SciCrunch — La Jolla, CA

- Designed and optimized **Python** regex scripts to analyze unstructured, dense academic text, performing extensive **data wrangling** to accurately identify, classify, and extract Research Resource Identifiers (RRIDs).
- Engineered automated **SQL data pipelines** to continuously capture contextual metadata and track citation patterns across major publishers (such as Elsevier, Biorxiv, and Nature), optimizing database schemas for large-scale information retrieval.
- Conducted a long-term **metadata extraction study** to automate resource cataloging for RRIDs, building automation to orchestrate document parsing workflows and ensure **high data integrity** across biological databases.

Projects

Causal Discovery LLM Agent | *Python, OpenAI/Gemini APIs, JSON* — Spring 2026

- Developed a **multi-agent system** for causal discovery using BFS-style expansion, integrating an LLM-based **domain expert agent** to validate causal edges via real-world knowledge and statistical signals from data.
- Optimized system performance and inference latency through JSON-constrained outputs and caching to reduce LLM API overhead, implementing **confidence-gated decision workflow** with batch processing and transitive reduction.

RAG Meeting Transcript QA | *LangChain, Hugging Face, PyTorch* — Winter 2026

- Architected a **Retrieval-Augmented Generation (RAG) system** using LangChain and Python to process and accurately query long-form, multi-domain meeting transcripts, driving complex text generation and summarization tasks.
- Implemented pretrained **Hugging Face Transformers** with **vector databases**, enabling efficient semantic search and neural memory retrieval, evaluating with precision metrics to validate text summarization quality and model performance.

Competitive Analysis of AI Infrastructure (CV Benchmarking) | *PyTorch, Ultralytics* — Winter 2026

- Conducted a systematic comparison of object detection algorithms (YOLO vs. Faster R-CNN) under hardware constraints (RTX 3050 Ti, 4GB VRAM) to map the Pareto frontier of accuracy versus computational cost.
- Leveraged AI-assisted development workflows via Cursor and Claude to implement mixed-precision training, resolving architectural bottlenecks and maximizing performance on limited-compute infrastructure.

Technical Skills

Programming Languages: Python, Java, SQL, R, C++, Bash/Shell, JavaScript, MATLAB, HTML/CSS
Libraries: pandas, NumPy, scikit-learn, XGBoost, SciPy, Statsmodels, OpenCV, NLTK, Surprise, matplotlib, seaborn, Plotly
Frameworks: PyTorch, TensorFlow, Hugging Face, LangChain, PySpark, Dask, Django, D3.js
Tools & Infrastructure: Git/GitHub, Linux/Unix, CI/CD, AWS (EC2, S3), Vector Databases, REST APIs, Cursor, Claude
Machine Learning & AI Concepts: Generative AI (LLMs, Transformers), NLP, Supervised Learning (Classification & Regression), Random Forests, Gradient Boosting, Logistic Regression, Model Evaluation & Optimization