Sihan(Michelle) Zhou

(757) 291-1113 | michelle.zhou.701@gmail.com | LinkedIn: Sihan (Michelle) Zhou

EDUCATION

College of William and Mary   **Williamsburg, VA**
Major: Computer Science+ Applied Mathematics   Expected Graduation 2023 Fall

- **GPA:** 3.97/4.0 | Dean's list (Fall+Spring 2020, 2021, 2022)
- **Relevant Coursework:** Adv Applied Machine Learning, Database, Software Development, Advanced Statistical Data Analysis, Algorithm, Data Structure, Advanced Linear Algebra, Probability, Econometrics

TECHNICAL SKILLS

- **Certificates:** Azure Data Scientist Associate (DP-100), AWS Certified Cloud Practitioner
- **Competition:** 1st place in Kaggle's CCI Machine Learning Disinformation Detection(Spring 2023)
- **Coding Languages:** Python(TensorFlow,Sklearn), SQL, R, Java, C/C++, Pytorch
- **Others:** Huggingface, LangChain, Axure, AWS, PowerBI, Tableau, Git, Drupal

RESEARCH & PROJECTS

DisinfoLab - Engineering team

- Build a robust synthetic media detector as an AI platform to analyze and flag artificially and digitally generated media
- Fully trained two models with high accuracy on the VideoSham dataset; use Meta's Segment Anything to improve model performance; integrate Weights & Biases MLOps into BA-TFD+ to track and manage training and testing

Geolab AI-NLP team

- Applied CoreNLP and fine-tuned large language models from HuggingFace with low-rank adaptation (LoRA) to achieve text summarization and named entity extraction functionalities for high-volume geoparsing articles

Triangulation and Desity Estimation based DBSCAN Clutering Algorithm

- Developed a novel clustering algorithm that solves adjacent problem for nonlinear data with noise and various density by combining Delaunay Triangulation with DBSCAN and KNN and kernel density estimation
- Presented at the lighting session of 2023 Symposium of Statistics & Data Science followed by poster session

WORK EXPERIENCE

AidData   **Williamsburg, VA**
Machine Learning Intern   *May 2021 – October 2022*

- Implemented the whole framework for performing few-shot learning on OECD's 5-digit encoding task, including from data processing, creating vector database for the training data pool, to running and testing the model.
- Deployed Google's language model Flan-ul2 on AWS Sagemaker and achieved model inference using EC2 G5.
- Improved the model accuracy from 88% to 97% by implementing dynamic retrieval of in-context examples; reduced annotation cost for sectors without training data by implementing fast vote-k selection of annotation.

Institute for Integrative Conservation   **Williamsburg, VA**
Database Management Intern   *Apr 2022 - March 2023*

- Designed the database architecture by setting up taxonomy, content type, blocks, fields and searching functions on Drupal; cleaned conservation data into standard form and import them into Drupal's database
- Researched content management systems such as Drupal and Hubzero, met with their sales, data analysts and developers to analyze the technical feasibility of each CMS for building a conservation solution tool website

Bytedance Ltd.   **Beijing, China**
Data Analyst Intern- Global Payment Department   *June 2021 - August 2021*

- Used SQL to retrieve bill reconciliation data between merchants and disbursement channels, queried doubtful transactions, sorted out currency and amount mismatched transactions, proposed to R&D team with tree diagrams of reconciliation logic and possible solutions
- Designed website prototypes, optimized product features based on customer response, draw cross function flowchart to manage the project requirements