YUXIN ZHU

yuxin.zhu@emory.edu | (404)451-0815

EDUCATION

Emory University, Laney Graduate School Atlanta, GA
PhD in Computer Science – Biomedical Informatics Track Aug.2024-Present
Major Courses: Introduction to Machine Learning, Introduction to Ethical Data Science and Informatics, Computational Methods for Biomedical Image Analysis

University College London (UCL) London, UK
Bachelor of Science in Statistics Sep.2020-Jun.2023
Major Courses: Statistical Inference, Medical Statistics, Computing for Practical Statistics, Optimisation Algorithms in Operational Research

PUBLICATIONS

⁎ **Zhu Y** et al. LLM-augmented therapy normalization and aspect-based sentiment analysis for treatment-resistant depression on reddit. [Under review by *npj mental health research*] [Preprint]

⁎ **Zhu Y**, Guo Y et al. Ensemble-Based Narrative Analysis for ADHD Diagnosis: An Integration of LLaMA3, RoBERTa and SVM Classifiers. [AMIA'25 Poster]

⁎ **Zhu Y**, Guo Y et al. 2025. Leveraging Large Language Models and Traditional Machine Learning Ensembles for ADHD Detection from Narrative Transcripts. In SciSoc LLM Workshop at KDD 2025. [Paper]

⁎ **Zhu Y** et al. Narratives After Naturalistic Movie Viewing in ADHD: A Large Language Model Analysis. [SOBP'25 Poster]

RESEARCH EXPERIENCE

Longitudinal Named Entity Recognition and Survival Analysis of Clinical and Social Impacts Among Opioid-Mentioning Reddit Users, *Project Lead* Jan.2026-Present

- Designed and conducted a longitudinal cohort study of 48,410 opioid-mentioning Reddit users, curating a pharmacologically structured lexicon of 500+ terms across 11 opioid groups and anchoring follow-up at each user's first opioid-related post (T0)
- Applied and extended a DeBERTa-large named entity recognition pipeline to identify clinical and social impact mentions in post-T0 narratives, then operationalized user trajectories through 30-day state sequences, first-event ordering, and Markov transition analyses
- Specified primary and sensitivity time-to-event models, including Cox proportional hazards, alternative event definitions, discrete-time hazards, post-count time scales, and IPW analyses; found that greater T0 keyword diversity predicted earlier clinical (HR 1.12, 95% CI 1.08-1.16) and social (HR 1.08, 95% CI 1.02-1.14) harms, while clinical impacts preceded social impacts in 62.2% of affected users

LLM-Augmented Therapy Normalization and Aspect-Based Sentiment Analysis for Treatment-Resistant Depression on Reddit, *Project Lead* June 2025 - Present

- This study, supervised by Professor Abeed Sarker at the Department of Biomedical Informatics, Emory University, is the first large-scale quantitative characterization of treatment-resistant depression (TRD) therapy discussions on Reddit, combining NLP-driven lexicon-based therapy normalization with aspect-based sentiment analysis to convert unstructured patient narratives into structured, therapy-level evaluations across 81 therapies and 16 therapeutic classes

- Designed and implemented a two-stage filtering pipeline to curate a TRD-specific Reddit corpus: collected 21,826 posts from 17,407 subscribers across 28 mental health subreddits (spanning general depression, treatment-focused, and professional/advice communities), applied keyword-based TRD detection incorporating alternative clinical labels (difficult-to-treat depression, treatment-refractory depression) to yield 5,059 TRD-relevant posts from 3,480 subscribers, then retained 3,839 posts containing at least one therapy mention from 2,700 unique subscribers
- Constructed a comprehensive therapy normalization lexicon grounded in established TRD treatment overviews and iteratively refined through observation of Reddit discourse; compiled brand names, abbreviations, and colloquial variants for 83 treatment entities, mapping each to a generic (nonproprietary) name aligned with International Nonproprietary Names and RxNorm normalized drug concepts
- Adapted the QMisSpell misspelling generator (trained on a phrase-level word embedding model from therapy-related social media posts) to automatically generate common misspellings, restricting candidate variants to tokens present in the Reddit TRD corpus; further augmented lexical coverage via LLaMA 3–70B constrained few-shot prompting (temperature 0.2, nucleus sampling top-p 0.9, max 220 tokens) with in-domain Reddit/X usage examples, yielding a final lexicon of 1,027 distinct variants across 81 therapies (median 12 variants per drug)
- Executed case-insensitive, full-token and multiword-expression matching using the lexicon on post titles and bodies, extracting 23,399 therapy mentions spanning 81 of 83 therapies after normalization; mapped all mentions to cohort metadata (post ID, subreddit, subscriber, timestamp) for multi-dimensional downstream analyses
- Trained and benchmarked three transformer architectures (BERT-base, RoBERTa-base, DeBERTa-v3-base) for aspect-based sentiment classification on the SMM4H 2023 English therapy-sentiment dataset (3,009 posts); replaced target therapy spans with a special placeholder token to force reliance on local linguistic context rather than surface-form drug names, and applied a 1,000-character context window centered on each mention for long texts
- Devised an LLM-based synthetic data augmentation strategy to address severe class imbalance (319 negative / 2,120 neutral / 570 positive in original training set): prompted LLaMA 3–70B-Instruct via few-shot examples to generate five synthetic posts per minority-class instance, preserving target therapy and sentiment while keeping co-mentioned therapies neutral, yielding 4,445 synthetic posts (2,850 positive; 1,595 negative) and an augmented training set of 7,454 instances
- Achieved micro-F1 of 0.800 (95% CI: 0.780–0.820) with the augmented DeBERTa-v3-base classifier, surpassing the top system reported in the SMM4H 2023 shared task overview (0.778); class-specific F1 scores: negative 0.467, neutral 0.874, positive 0.636; quantified uncertainty via nonparametric bootstrap (1,000 resamples)
- Applied the fine-tuned classifier to all 23,399 therapy mentions in the Reddit TRD corpus, generating mention-level sentiment predictions (negative/neutral/positive) mapped back to cohort metadata for analyses by therapy, subscriber, subreddit, and calendar year; performed manual spot-checks on prediction samples to assess validity
- Conducted inferential statistical analyses: Pearson chi-square test of independence on a 16 × 3 contingency table (therapeutic class × sentiment label) with Cramér's V effect size and standardized residuals; exact two-sided binomial tests per therapy for positive-vs.-negative asymmetry among non-neutral mentions; Benjamini–Hochberg FDR correction across all 76 therapy-level and 120 class-pair post hoc comparisons
- Identified statistically significant sentiment heterogeneity across therapy classes ($\chi^2(30, N = 23{,}399) = 686.07$, $p = 3.93 \times 10^{-125}$, Cramér's $V = 0.121$) and 20 therapies with significant positive–negative asymmetry after FDR correction; revealed that NMDA/rapid-acting therapies (ketamine, esketamine) exhibited comparatively favorable sentiment profiles while conventional SSRIs/SNRIs showed consistently higher negative proportions, consistent with TRD patients' prior nonresponse to standard antidepressants

Leveraging Large Language Models and Traditional Machine Learning Ensembles for ADHD Detection from Narrative Transcripts, *Project Lead* Sep.2024-May 2025

- Utilized 441 narrative transcripts from the Healthy Brain Network dataset—stratified into training (264), development (88), and test (89) sets—to evaluate classification models
- Engineered and iteratively refined a DSM-5-based prompt for LLaMA3-70B, optimizing for binary ADHD vs. nonADHD output and explanatory justifications
- Fine-tuned RoBERTa on interviewee-only text segments, applying a sliding-window (512-token) segmentation to accommodate long transcripts and hyperparameter tuning (learning rate, epochs) for sequence classification

- Developed an SVM classifier using TF-IDF–weighted unigram through four-gram features (1,000-term vocabulary) augmented with engineered metrics (mean response length, response count, mean question length) for enhanced interpretability
- Implemented a majority-voting ensemble that combines LLaMA3, RoBERTa, and SVM predictions—boosting recall to 0.91 and achieving an F1 score of 0.71 (95% CI: 0.60–0.80)
- Conducted in-depth error analysis on false positives/negatives across models to identify failure modes (e.g., n-gram sparsity, discourse coherence) and inform future enhancements (e.g., weighted voting, cohesion metrics)

Research on Knowledge Fusion of Multi-source Heterogeneous Archives of Ancient Villages in the Context of Cultural Big Data, ***Research Assistant*** May 2023-Jun.2024

- This project, supervised by Professor Tianjiao Qi of Renmin University of China, is a Youth Program of the National Social Science Foundation of China, which integrates various types of archives of ancient villages and applies cultural big data and knowledge fusion technologies (data mining, information extraction, etc.) to explore the historical, cultural, and social information of ancient villages holistically and intensively, thus promoting cultural heritage preservation and studies • Completed 246 working hours as a research assistant, designing a sophisticated data collection strategy, extracting 11 village data variables from six different files containing 8,156 village lists, and performing data collection, analysis, and database building
- Developed an algorithm that adapted to the ever-changing search engine optimization strategy of Baidu Baike, which could accurately filter out irrelevant links and enforce the collected data reflected only authoritative Baidu Baike entries
- Designed an algorithm to respond to different content layouts to automatically process Baidu Baike's search results with inconsistent formats, headings, and nested sub-sections, incorporated regex-based pattern matching, heuristic analysis, and context-based extraction methods to transform unstructured textual data into practical information
- Employed a sophisticated RPC (Remote Procedure Call) technique to integrate a browser plugin with Python code to cope with Baidu Baike's network access limitations, thus maintaining consistent and unrestricted data retrieval

Spatial, Demographic, and Socioeconomic Factors Affecting the Distribution and Density of UK Pet Dog Population, *Researcher* Oct.2022-Aug.2023

- This dissertation is supervised by UCL's Professor Takoua Jendoubi, in collaboration with the UK's largest Dog Welfare Charity *Dog Trust*, aiming to explore factors associated with the density and distribution of the UK canine population by investigating the underexplored realm of quantitative analysis
- Exploited publicly available data sources to develop novel predictive models, using advanced machine learning techniques to creatively gather and analyze accessible data on a large scale, overcoming the shortcomings of conventional survey-based methods such as funding and location constraints, leading to increased research breadth, depth, and accuracy in this field
- Uncovered significant variations in pet dog proportions across UK postcode areas based on age, size, and cephalic categories, highlighting lifestyle and environmental factors
- Identified critical factors affecting the number of dogs per household through decision tree modeling, including accommodation type, age composition, and ethnic groups
- Utilized spatial modeling analysis to reveal intricate non-linear relationships between geography, demographics, socioeconomic factors, and dog ownership patterns
- Implemented the spatial Classification and Regression Trees (CART) model, exploring the relationship between brachycephalic dogs per household and various predictors, including urbanization, unemployment rate, median salary, and the percentage of the population over 65
- Produced interactive maps and visualizations using the tmap package to communicate research findings effectively, demonstrating the predicted distribution of brachycephalic dogs, comparing it with the actual distribution, and visualizing residuals to assess model accuracy
- Additionally explored the application of spatial random forest modeling, comparing and contrasting results with the spatial CART approach

Chinese Historical and Cultural Villages, *Research Assistant* Jul. 2022- Jan.2023

- This project, supervised by Professor Tianjiao Qi of Renmin University of China, showcases the geographic distribution of Chinese villages while examining the historical and cultural reasons for their emergence using QGIS and R
- Conducted the base analysis on a provincial scale, and undertook a nuanced comparison between Eastern and Western cultural influences within China, revealing spatial patterns rooted in historical and cultural contexts
- Performed a focused regional analysis on the critical ethnic areas in China to decipher potential correlations between higher ancient village counts in the east and ethnic settlement patterns

The Rise and Fall of Handicaps, *Research Assistant* May 2022- Jun.2024

- This project is supervised by Professor Paul X. McCarthy of the University of New South Wales, Sydney, using data science methodologies to analyze language changes over the last century concerning disability and rehabilitation and exploring how it reflects, lags, or anticipates broader social changes and attitudinal shifts in society
- Visualized language evolution (1900-2019) with Google Books Ngram data, highlighting shifts from older, pejorative terms (e.g., handicapped, retarded) to newer, respectful ones (e.g., identity-first, people-first language)
- Pioneered the creation of a time-series corpus encompassing a diverse range of over 2000 disability-related terms collected from Google Books using principles of sentiment analysis and principle component analysis (PCA)
- Implemented and adapted PCA methodology inspired by a paper entitled *The Rise and Fall of Rationality in Language* to derive insights into the dynamics of disability language over time
- Reviewed extensive literature on conceptual models of disability (moral/medical/social) to define research approaches

Ensemble Theory of Startups, *Research Assistant* May 2022

- This project explores what works and what doesn't for startup success using data analytics and machine learning techniques based on the Crunchbase database, which is supervised by Professor Paul X. McCarthy of the University of New

South Wales, Sydney

- Studied the conceptual framework of Ensemble Theory, a unique theory formulated by Professor Paul X. McCarthy, and outlined the trio of pivotal factors shaping the trajectory of startup ventures
- Researched and identified primary references delving into diverse perspectives for assessing startups
- Built insightful connections between distinct sources and their respective methodologies, contributing to a more holistic picture of startup assessment methodologies

PROFESSIONAL EXPERIENCE

Podcast Series: “AccessGranted: The Stories of UCL’s Most Inspiring Students”

Instructed by UCL's Professor Takoua Jendoubi Jun.2023-Aug.2024

AccessGranted is a bi-weekly podcast project sponsored by Student Success Fund and designed to raise awareness of minoritized communities, drive cultural and behavioral shifts, and foster a sense of community among UCL students

- Led a student team to facilitate effective collaboration through regular meetings, managing different parts of the project strategically, and ensuring smooth progress of interviews, recording, and editing
- Selected guests that aligned with the project's tone, conducted thorough research to identify customized episode topics and questions, designed a holistic interview framework, and composed scripts based on the interview content to define the episode structure to make the interviewees' "voices" accessible to the audience both informationally and emotionally
- Managed the post-production and editing of each episode, including audio quality, sound effects, and background music, reviewed the episode title and promotional material, and drove its launch on UCL's official social media and departmental channels
- Navigated through a labyrinth of meetings and qualitative analysis, amassing over 100 hours of team and interview interactions and meticulously processing more than 120,000 words of transcripts

Volunteer | Kith & Kids

Organization providing activities, opportunities, information, and support for individuals with learning disabilities or

autism. Oct.2022-Feb.2023

- Worked on a 1:1 or 2:1 ratio at the Kith & Kids Weekend Club, contributing over 40 hours during term-time
- Served as a mentor in the Employment & Life Skills Project, assisting individuals with mild/moderate learning disabilities and/or autism
- Shared personal experiences to help participants develop skills for employment, independent living, and community integration

Activating the High-quality Development of Cultural Tourism in Aba Prefecture with Qiang Embroidery

2022 ICH Big Data & Artificial Intelligence Innovation Competition Jul.2022-Sep.2022

• Collected data from four attractions in Aba Prefecture, Sichuan Province, encompassing 17,161 visitors, and extracted information on the source of visitors, age, gender, and travel mode

• Crawled the content and sentiment of over 1,000 reviews about attractions from Ctrip and Mafengwo to obtain tourists' comments regarding the tourism quality

• Investigated the online prominence of Qiang embroidery within the digital sphere, extracting insights from 200+ realtime search results on the "Qiang Embroidery" topic from social media platforms

ADDITIONAL INFORMATION

Languages: English (Proficient), Japanese (Basic), Korean (Basic)
Skills: R, Python, QGIS, MySQL, Stata, MATLAB, Excel, SPSS
Hobbies: Fiction Composition, Swimming, Yoga