Yibo Sun

sunyibo2027@gmail.com | +1 (347) 324-6119 | github.com/SuperSheldonSun

EDUCATION

New York University | BA in Computer Science & Data Science — Sep. 2023 – May 2026

- **Major GPA:** 3.61/4.00
- **Coursework:** Probability & Statistics, Discrete Mathematics, Linear Algebra, Data Structures and Algorithms, Computer Systems Organization, Causal Inference, Data Management and Analysis, Deep Learning, Large Language Models, Natural Language Processing, Robot Intelligence
- **Awards & Honor:** 2025 Kringstein Family Research Scholar (Dean's Undergraduate Research Fund, May 2025)

INTERNSHIP

Data Scientist Intern | ByteDance – Remote — Jan. 2025 – Mar. 2025

- Conducted behavioral clustering on 10,000+ users using K-Means, with multi-dimensional user behavior data cleaned and normalized beforehand, which boosted campaign targeting accuracy by 15% and led to higher ROI in marketing spend
- Discovered high-retention user traits via 6-month user behavior time-series analysis; proposed UX improvements, which increased 7-day retention by +9% and 30-day retention by +6% in 50,000-user A/B testing
- Designed and monitored core metrics (watch time, retention, conversion) under AARRR to guide iteration; built dashboards for real-time DAU, completion and engagement insights, boosting decision efficiency

Data Analyst Intern | Education Examination Authority – Lianyungang, Jiangsu — Jul. 2024 – Sep. 2024

- Automated end-to-end exam data workflows, and tailored data pipelines—using Python; reduced manual processing errors by 40%, cut report generation time by 30%, and enhanced system scalability for peak-period large-volume data
- Created interactive dashboards with Tableau & Matplotlib to visualize score distributions and demographic trends, directly supporting education policy formulation and data-driven decision making

MAIN RESEARCH & PUBLICATIONS

Comprehensive Investigation into Bankruptcy Prediction Using Advanced Methods for Identifying Key Feature Impacts on Predictive Performance | Academic Research — Jan. 2025 – present

- Developed a hybrid Stacking Classifier integrated with Hyperband Optimization (HBO) and Improved Arithmetic Optimizer (IAO), which significantly enhanced the model's predictive accuracy and operational stability for bankruptcy assessment
- Used Recursive Feature Elimination (RFE) to identify high-impact predictors, showing feature competitiveness ($p=0.001$), operating risk ($p=0.012$) and credibility ($p=0.016$) as core bankruptcy prediction drivers

The Economic Aftermath of Wildfires on Urban Real Estate: A Synthetic Control Study of the Eaton Fire — Aug. 2025

NYU Center of Data Science | Independent Research

- Built a city-level monthly housing & climate panel (2000–2025) from Zillow, ACS, CAL FIRE, and NOAA; curated a 58-city donor pool for counterfactual construction.
- Estimated the Eaton Fire's (Jan 2025) impact on Altadena (ZHVI) with a time-decay Synthetic Control, achieving a 0.61% pre-treatment RMSPE; average treatment effect ~$32,125 and the gap peaked at ~$68,927 by July 2025
- Validated effects via placebo-in-space inference (post/pre RMSPE ratio $p=0.0508$, 10% level); automated ETL & geospatial workflows (Python) and produced reproducible econometric figures (matplotlib, seaborn, LaTeX).

A Hybrid Two-Stage Framework for the Change Detection | Academic Research — Aug. 2025

Shenzhen Research Institute of Big Data

- Developed an innovative two-stage change detection framework: its statistical module uses extreme value testing to identify candidate ROIs with FDR control (reducing false positives), paired with a lightweight U-Net integrated with a proprietary AttentionFusion module that leverages statistical maps as contextual priors for high-precision pixel-level refinement in ROIs
- Achieved significant computational efficiency (reducing complexity from O(WH) to O(N*wh)) and simultaneously strengthened the framework's robustness against complex noise patterns and seasonal fluctuations

An Emotion Text Classification Model Based on Llama3-8b Using Lora — Jul. 2024

2024 7th International Conference on Computer Information Science and Application Technology (CISAT)

- Conducted research on leveraging the Llama3-8B large language model for emotion text classification using advanced techniques such as LoRA fine-tuning and FlashAttention for improved training efficiency
- Contributed to the creation of a large-scale, six-class emotion dataset and performed systematic comparative evaluation against mainstream transformer baselines including BERT and RoBERTa, focusing on metrics like accuracy and F1-score
- Achieved 92.62% accuracy with the Llama3-8B model, demonstrating its strong effectiveness in NLP classification tasks

MAIN PROJECTS

Braille Recognition & Correction System | Github, Team leader — May. 2025

- Developed a Python-PyTorch transfer learning OCR model for automatic full-page Braille-to-English conversion (over 80% accuracy) and integrated lightweight LLMs to build an error-correction module; addressed manual transcription inefficiency and optimized output clarity for low-vision users and untrained writers
- Combined image augmentation techniques with the YOLO object detection algorithm for character segmentation, enhancing the model's robustness against noisy or blurred input and improving recognition stability in complex scenarios

StardewAI – Stardew Valley Game | Github, Team leader — Aug. 2025

- Built a Python-based, GPT-3.5-powered game assistant using Retrieval-Augmented Generation (RAG); implemented semantic retrieval across 400+ wiki pages with ChromaDB and SentenceTransformers for fast, accurate strategy matching
- Developed dynamic function-calling with LangChain and OpenAI agents, integrated multi-turn memory for context-aware Q&A; completed full-stack development (FastAPI, HTML/CSS/JS, Bootstrap, Jinja2) for practical tool deployment

Professor Rating Analysis | — Nov. 2024

- Processed over 500,000 RateMyProfessors entries with Python (Pandas, Scikit-learn) to identify rating bias factors and teaching quality predictors; used t-tests and Statsmodels regression models to quantitatively analyze how gender bias and teaching experience impact ratings, supporting educational evaluation fairness
- Constructed a classification model to predict professor satisfaction ratings, visualized results using Matplotlib and Seaborn, and achieved an AUROC (Area Under the ROC Curve) of 0.84, indicating excellent prediction accuracy

TECHNOLOGIES

- **Languages:** Chinese (native), English (Fluent)
- **Technical Skills: (1)** Languages: Python, C++, Java, C; **(2)** Frameworks: PyTorch, TensorFlow, FastAPI, LangChain; **(3)** ML/NLP: Scikit-learn, FlashAttention, LoRA, YOLO, LLMs (GPT, LLaMA); **(4)** Data Tools: Tableau, Power BI, Matplotlib, ChromaDB; **(5)** Others: Git, Bash, HTML/CSS/JS, SQL