

Shengwei (Jacob) Zhang

GitHub Project Portfolio

University of Pennsylvania

github.com/Bubu631

My portfolio presents research on reliable AI and decision-making, followed by applied machine learning and software projects. Explore the links for code, manuscripts, experiment configurations, recorded results, and verification instructions.

Research Projects

Base Model Upgrade Non Regression

[GitHub](#)

Studies how to certify model upgrades without degrading important data slices. Uses paired non-inferiority tests and retains the incumbent model when evidence is insufficient. Includes controlled simulations, public digits experiments, saved checkpoints, and validation tests.

Python, NumPy, SciPy, scikit-learn | Statistical testing and model release evaluation

[Read manuscript](#)

Edit Localization and Selection Bias

[GitHub](#)

Investigates how learning from selectively observed human corrections changes gradient bias and the training target. Combines mathematical analysis, importance weighting, Monte Carlo experiments, and fixed-checkpoint diagnostics using public translation post-edits and pretrained models.

Python, PyTorch, Hugging Face Transformers, NumPy, LaTeX, Make | Gradient analysis

[Read manuscript](#)

Budgeted Measurement for Wellbeing Prioritisation

[GitHub](#)

Studies which additional measurements are worth acquiring before prioritizing units for wellbeing review. Evaluates budget allocation across measurement channels using controlled simulations, public BRFSS replay, and an NHS staff survey illustration.

Python, pandas, SciPy, Matplotlib, openpyxl | JavaScript and Node.js for workbook checks

[Read manuscript](#)

The Audit Complexity of Group Relative Policy Updates

[GitHub](#)

Studies how many reference labels are needed to estimate a fixed group-relative policy update. Includes query-complexity theory, randomized estimators, controlled simulations, small-policy training, and frozen-model diagnostics on historical software-agent trajectories.

Python, NumPy, SciPy, PyTorch | Estimator implementation and independent checks

[Read manuscript](#)

Research Index

The research index links the four independent repositories and their release materials. It is a navigation hub for the projects above.

github.com/Bubu631/research

Repository coverage: 8 public project repositories, 1 public research index, and 2 private software repositories. Checked September 14, 2026.

Applied Machine Learning and Software

Public Applied Machine Learning Projects

Predicting the Spark

[GitHub](#)

A geospatial study of U.S. wildfire occurrence combining weather, historical fire records, and a vegetation fuel index. Includes data cleaning, raster processing, class imbalance handling, model comparisons, and diagnostic visualizations.

Python, scikit-learn, XGBoost, Rasterio, GeoPandas | Notebooks and analysis

Deep Learning Across Vision Language and Time Series

[GitHub](#)

A course project implementing image classification, extractive question answering, and time series forecasting. Compares CNNs and ResNet-50, fine-tunes DistilBERT on SQuAD, and evaluates recurrent models for stock price prediction.

Python, PyTorch, Hugging Face Transformers | Training and evaluation notebooks

NYC Property Sales Prediction

[GitHub](#)

A regression project using New York City property transactions. Cleans and encodes property features, compares linear regression with random forests, and examines prediction errors and feature importance through visualizations.

Python, pandas, scikit-learn, Matplotlib, Seaborn | Data analysis notebook

Heart Disease Classification

[GitHub](#)

A classification study of imbalanced public heart disease data. Compares logistic regression, random forests, and XGBoost with class weighting and recall-focused tuning to examine sensitivity and precision trade-offs.

Python, scikit-learn, XGBoost | Educational classification and evaluation study

Additional Software Projects

These two repositories are private and require repository access to view.

Course Notes Platform

[GitHub](#)

A web application for organizing and editing course materials, with administrator-only AI summaries and GitHub synchronization for Markdown notes. Includes client and server code, course content, and deployment configuration.

Private | JavaScript, Node.js, Express, Anthropic SDK

Hugging Face Model Explorer

[GitHub](#)

A collaborative CIS 5500 course project combining a Hugging Face metadata pipeline with model and dataset search, comparison, lineage exploration, and analytics. Includes normalized data tables, SQL queries, and a web interface.

Private fork of the team repository | React, Express, PostgreSQL, Python