Shuo Zhang

sz3177@columbia.edu (857) 350-2596 inverse0.github.io shuo-zhang-leo Inverse0

Education

Columbia University *Sept 2023 – Dec 2024*
MS in Computer Science; GPA: 3.95/4.00

Boston University *Sept 2020 – May 2023*
BA in Computer Science and Economics; GPA: 3.83/4.00

- Honors: Magna Cum Laude

Publications

Data-Centric Text-to-SQL with Large Langueg Models
Zezhou Huang, ***Shuo Zhang***, Kechen Liu, Eugene Wu
In Sumbission

Data Cleaning Using Large Language Models
Shuo Zhang, Zezhou Huang, Eugene Wu
In Sumbission

Policy Enforcement for IoT: Complexities and Emerging Solutions Sept 2024
Shuo Zhang, Luoyao Hao, Henning Schulzrinne
IEEE International Conference on Performance, Computing and Communications (***IPCCC***), 2024

Advancing IoT System Dependability: A Deep Dive into Management and Operation Plane Separation
Luoyao Hao, ***Shuo Zhang***, Henning Schulzrinne
In Sumbission

CAVE: Concurrency-Aware Graph Processing on SSDs May 2024
Tarikul Islam Papon, Taishan Chen, ***Shuo Zhang***, Manos Athanassoulis
Proceedings of the ACM on Management of Data (***SIGMOD***), 2024

Research Experience

Research Assistant *New York, NY*
Columbia University Database Group. Advisor: Prof. Eugene Wu & Prof. Kenneth Ross *May 2024 – present*

- Conducted research on data cleaning and Text-to-SQL conversion, contributing to the development of an advanced data-cleaning system that leverages Large Language Models (LLMs) to automate and enhance the data-cleaning process.
- Evaluated system performance by comparing accuracy, recall, and F1-scores with established data-cleaning systems like Holoclean and Raha on benchmark datasets, achieving F1-score improvements of up to 24x over baseline models.
- Demonstrated the critical importance of data cleaning in Text-to-SQL conversion, revealing that existing Text-to-SQL benchmark datasets contain up to 37% errors, even when relying on human-labeled ground truth.
- Conducted research on query optimization by leveraging a compiler, optimized for matrix multiplication tasks, to accelerate aggregation queries in DuckDB.

Research Assistant *New York, NY*
Columbia University Internet Real-Time Lab. Advisor: Prof. Henning Schulzrinne *Sept 2023 – present*

- Conducted research on IoT devices, systems, and applications, contributing to the development of a novel IoT system that enhances dependability and introduced the Identity-Independent Policy (IIP), a flexible and expressive policy specification.
- Developed a policy server prototype integrating IIP using Python to demonstrate system effectiveness. Evaluated expressiveness using real-world IoT IFTTT datasets, showing that up to 99% of policies are expressible under the IIP framework.
- Addressed key challenges in IoT policy enforcement, such as the steep learning curve of policy creation, and proposed

solutions leveraging Large Language Models (LLMs) to simplify and resolve these issues.

Research Assistant *Boston, MA*
Boston University DiSC Lab. Advisor: Prof. Manos Athanassoulis *June 2022 – May 2024*

- Conducted research on databases, data management, and storage systems. Contributed to the design and development of an innovative graph processing system that leverages SSD parallelism, significantly improving the efficiency of large-scale graph processing.
- Developed the system in C++, implementing a suite of parallelized graph traversal algorithms and designing analogous algorithms for baseline systems to enable comparative performance evaluations.
- Demonstrated the effectiveness of the system through rigorous testing on diverse real-world graph datasets across multiple SSD platforms, achieving up to 984x speedup compared to existing systems like GraphChi, Mosaic, and GridGraph.

Industry Experience

Research Intern *San Jose, CA*
VisionX LLC *June 2022 – Aug 2022*

- Developed an advanced object detection model utilizing PyTorch and a pre-trained YOLO framework, significantly enhancing the quality of a product designed for crop pest and disease identification. Achieved 98% accuracy and increased processing speed by 180%.
- Optimized the model based on real-world product applications, employing fine-tuning techniques to improve performance on blurry and low-resolution images.
- Constructed a Java-based Graphical User Interface (GUI) to enhance user experience.

Projects

Video Super-Resolution *2024*

- Developed a method that enhances both video resolution and smoothness by integrating advanced image super-resolution techniques with video interpolation strategies. Achieved a Peak Signal-to-Noise Ratio (PSNR) of 34.72 on a subset of the Vimeo-90K benchmark.

E-commerce Website *2024*

- Developed a shopping website, utilizing HTML for design and Flask for creating a RESTful API that interfaces with a MongoDB backend.
- Implemented SQL to develop essential e-commerce functionalities such as shopping carts and order tracking systems, mirroring those found on established shopping platforms.

Time Series Analysis *2023*

- Researched the potential of machine learning models for time series data forecasting.
- Experimented with various state-of-the-art machine learning models, comparing their performance with traditional methods like autoregressive integrated moving average (ARIMA) and advanced techniques such as Long Short-Term Memory (LSTM) networks.
- Evaluated on real-world GDP data and achieved an 85% improvement on Root Mean Squared Error (RMSE) compared with ARIMA prediction.

Data Cleaning and Integration *2022*

- Collaborated with Boston City Councilor Ruthzee to analyze small landlord data, identifying properties with the potential to join affordable housing programs.
- Applied machine learning techniques, including random forest, to classify owner-occupied and small landlord properties for integration into affordable housing databases.
- Cleaned and consolidated existing housing databases to identify eligible properties. Created machine learning models to cluster housing units and predict fair rental price ranges.

Technologies

Programming Languages: Python, Java, SQL, C++, HTML, R, OCaml

Technologies: PostgreSQL, MongoDB, MySQL, DuckDB, Flask, Node.js, MQTT, PyTorch, Stata