Zezhou (Zachary) Huang

✉ zh2408@columbia.edu — 🌐 www.columbia.edu/ zh2408/ — github.com/zachary62

Education

Ph.D. in Computer Science, Columbia University, GPA: 4.00, Advisor: Prof. Eugene Wu — Sep. 2019 – Present

M.S. in Computer Science, Columbia University, GPA: 4.00 — Sep. 2019 – May 2021

B.S. in Computer Science, University of Wisconsin-Madison, GPA: 3.89 — May 2019

Industry Experience

Research Intern, Microsoft, Redmond, WA — May 2023 – Aug 2023

- Developed a prototype for database engines using novel hardware. Achieved $>$ 10× faster and more cost-efficient performance compared to SQL Server and PowerBI in production workloads.

Software Engineer Intern, Databricks, San Francisco, CA — May 2022 – Aug 2022

- Implemented data structures for query optimization and view coverage, delivered to Databricks Runtime 11.1
- Experimented with IVM over join using delta table with dynamic pruning, low shuffle merge, and deletion vectors
- Implemented MV strategies in Enzyme (advised by Prof. Yannis Papakonstantino)

Software Engineer Intern, Tusimple, San Diego, CA — May 2021 – Aug 2021

- Built ETL pipelines over three data sources for self-driving data

Research Experience

Graduate Research Assistant, Columbia University, New York City, NY — Sep. 2019 – Present

- **Wide-Table Optimization:** Developed a novel optimization layer on top of existing DBMSes for scalable, interactive and private analytics, based on the theories from probabilistic graph models. Enabling tree-based model training on 100+ tables and terabytes of data, interactive dashboard $<$ 100 ms, and data discovery with differential privacy.
- **Pipeline Automation by LLMs:** Built a project that uses LLMs to automatically crawl data and pipeline codes in enterprise data warehouses, automatically building new pipelines for data cleaning, transformation, and integration.

Undergraduate Research Assistant, University of Wisconsin-Madison, Madison, WI — Aug. 2018 – May 2019

- **Hierarchical Storage in WiscKey:** Optimized read/write performance of WiscKey and LevelDB on SSDs, improving performance by 17.3% on a 100-GB database by exploiting LSM tree structure and query balancing .

Publications

1. **Data Cleaning Using Large Language Models**
   Shuo Zhang, Zezhou Huang, Eugene Wu
   Under Review

2. **Data-Centric Text-to-SQL with Large Language Models**
   Zezhou Huang, Shuo Zhang, Kechen Liu, Eugene Wu
   TRL@NeurIPS 2024

3. **Transform Table to Database Using Large Language Models**
   Zezhou Huang, Jia Guo, Eugene Wu
   TaDA@VLDB 2024

4. **SET: Searching Effective Supervised Learning Augmentations in Large Tabular Data Repositories**
   Jiaxiang Liu, Zezhou Huang, Eugene Wu
   GUIDEAI@SIGMOD 2024

5. **Disambiguate Entity Matching through Relation Discovery with Large Language Models**
   Zezhou Huang
   GUIDEAI@SIGMOD 2024

6. **Cocoon: Semantic Table Profiling Using Large Language Models**
   Zezhou Huang, Eugene Wu
   HILDA@SIGMOD 2024

7. **Relationalizing Tables with Large Language Models: The Promise and Challenges**

Zezhou Huang, Eugene Wu
DBML@ICDE 2024

8. **The Fast and the Private: Task-based Dataset Search**
Zezhou Huang, Jiaxiang Liu, Haonan Wang, Eugene Wu
CIDR 2024

9. **Lightweight Materialization for Fast Dashboards Over Joins**
Zezhou Huang, Eugene Wu
SIGMOD 2024

10. **Data Ambiguity Strikes Back: How Documentation Improves GPT's Text-to-SQL**
Zezhou Huang, Pavan Kalyan Damalapati, Eugene Wu
TRL@NeurIPS 2023

11. **Saibot: A Differentially Private Data Search Platform**
Zezhou Huang, Jiaxiang Liu, Daniel Gbenga Alabi, Raul Castro Fernandez, Eugene Wu
VLDB 2023

12. **Kitana: Efficient Data Augmentation Search for AutoML**
Zezhou Huang, Pranav Subramaniam, Raul Castro Fernandez, Eugene Wu
Arxiv

13. **Random Forests over normalized data in CPU-GPU DBMSes**
Zezhou Huang, Pavan Kalyan Damalapati, Rathijit Sen, Eugene Wu
DaMoN@SIGMOD 2023

14. **JoinBoost: Grow Trees Over Normalized Data Using Only SQL**
Zezhou Huang, Rathijit Sen, Jiaxiang Liu, Eugene Wu
VLDB 2023

15. **Aggregation Consistency Errors in Semantic Layers and How to Avoid Them**
Zezhou Huang, Pavan Kalyan Damalapati, Eugene Wu
HILDA@SIGMO 2023

16. **Reptile: Aggregation-level Explanations for Hierarchical Data**
Zezhou Huang, Eugene Wu
SIGMOD 2022

17. **Calibration: A Simple Trick for Wide-table Delta Analytics**
Zezhou Huang, Eugene Wu
Arxiv

18. **Spatial and hedonic analysis of housing prices in Shanghai**
Zezhou Huang, Ruishan Chen, Di Xu, Wei Zhou
Habitat International 2017

Service

- TaDA@VLDB 2024 PC Member
- GUIDEAI@SIGMOD 2024 PC Member
- DEEM@SIGMOD 2024 PC Member
- DataPlat@ICDE 2024 PC Member
- DBML@ICDE 2024 PC Member
- TRL@NeurIPS 2023 PC Member
- DBML@ICDE 2023 PC Member

Awards

- Google PhD Fellowship 2023
- Avanessian Fellowship 2023