Zezhou Huang

Now: Senior Researcher at **MSR AI Frontiers**, LLM agents & systems · Earlier: database systems (VLDB / SIGMOD, granted patents)

zh2408@columbia.edu | columbia.edu/~zh2408 | github.com/zachary62

Education

Columbia University — Ph.D. & M.S. in Computer Science, GPA 4.00 — Sep 2019 – May 2025

★ **Awards:** Google PhD Fellowship (2023) · Avanessian Fellowship (2023), **Advisor:** Prof. Eugene Wu

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

Advisors: Prof. AnHai Doan, Prof. Remzi Arpaci-Dusseau

Technical Skills

Languages: Python, C++, Java, SQL, Scala, Go, CUDA. **Data Systems:** Spark, Databricks, PostgreSQL, DuckDB, BigQuery, Snowflake, Kafka, Airflow. **AI / ML:** PyTorch, scikit-learn, vLLM, LangGraph; RAG, post-training (SFT / PPO / GRPO / RLHF), multi-agent orchestration, LLM evaluation & red-teaming, vector databases. **Infra:** Docker, Kubernetes, Ray, GCP, AWS, Azure.

Industry Experience

Senior Researcher — Microsoft Research AI Frontiers, Redmond, WA — May 2025 – Present

- **LLM agents, systems, & environment benchmarks.** Co-developed **Magentic Marketplace**: open-source agentic-market simulation covering search → match → negotiate → transact. Co-built **Red-Teaming a Network of Agents**, uncovering attack patterns (self-propagating worms, reputation manipulation, manufactured consensus, proxy chains).
- **Model training.** Reinforcement-learning post-training for LLM agent capabilities.

Research Intern — Microsoft Gray Systems Lab, Redmond, WA — May – Aug 2023

- Prototyped a GPU-accelerated SQL engine operating directly on lightweight-compressed (RLE / dictionary / bit-packed) data in CUDA; >**10**× faster and more cost-efficient than SQL Server and PowerBI on production workloads. Internship work led to the **VLDB 2026** paper and two granted **US patents** (12,277,123, 12,277,122).

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

- Built scalable data structures (Scala/JVM) for query optimization and view-coverage analysis in Spark, shipped in **Databricks Runtime 11.1**; designed incremental view maintenance over joins on **Delta Lake** tables (dynamic pruning, low-shuffle merge, deletion vectors) and contributed MV selection strategies to Databricks' **Enzyme**.

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

- Built ETL pipelines across three data sources for self-driving sensor data using **Python/Flask**, **Kafka**, **MongoDB**, **Docker/Kubernetes**, and **RESTful APIs**; applied **scikit-learn** for data-anomaly detection (>90% accuracy).

Research Experience

Graduate Research Assistant — Columbia University, New York, NY — Sep 2019 – May 2025

- **Wide-table query optimization & analytics.** Built **Reptile** (SIGMOD'22), **JoinBoost** (VLDB'23), and **Treant** (SIGMOD'24): a query layer deployable on PostgreSQL, DuckDB, and cloud OLAP engines (Redshift, BigQuery, Snowflake), grounded in probabilistic graphical models. Tree-model training across **100+ tables / TBs** of data, <**100 ms** interactive join dashboards, and aggregation-level explanations; up to 3 orders-of-magnitude faster than PostgreSQL on TPC-DS.
- **Private data search & AutoML augmentation.** Built **Saibot** (VLDB'23), **Kitana**, and **Fast-and-Private** (CIDR'24): a differentially private dataset-search platform with task-based query semantics and feature augmentation for AutoML.
- **LLM-driven data systems.** Built **Cocoon** (HILDA'24), **Data-Centric Text-to-SQL** (TRL@NeurIPS'24), **Transform Table to Database** (TaDA@VLDB'24), and entity-matching workflows for semantic profiling, text-to-SQL, and table-to-DB synthesis. The core LLM-workflow abstraction was later open-sourced as **PocketFlow** (10K+ ★).

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

- **Hierarchical Storage in WiscKey (WiscKeyHybrid).** In C++/Go, added a balancing layer between LSM tree and APIs to mediate range queries vs. random lookups on SSD; **+17.3%** throughput on a 100 GB database with 4-KB values.
- **Disguised Missing Value Detection (cloudFAHES).** Cloud data-cleaning system in **C++/Python** with **Docker**, detecting disguised missing values in tabular data via statistical pattern analysis; **0.82 F1** on real-world datasets.

Open Source & Outreach

- **PocketFlow** · ★10K+ : a 100-line, zero-dependency LLM framework using a graph-based core abstraction; supports multi-agent systems, workflows, and RAG without vendor lock-in.
- **PocketFlow-Tutorial-Codebase-Knowledge** · ★12K+ : an AI tool that analyzes any GitHub repository, identifies core abstractions, and auto-generates beginner-friendly tutorials with diagrams.
- **YouTube — @ZacharyLLM** · 35K+ subscribers : technical tutorials on LLM systems, agents, and AI engineering.
- **Writing & community.** PocketFlow Substack (2K+ subscribers) · Twitter/X @ZacharyHuang12 (5K+ followers).

Publications

1. **GPU Acceleration of SQL Analytics on Compressed Data.**
   Zezhou Huang, Krystian Sakowski, Hans Lehnert, Wei Cui, Carlo Curino, et al. *VLDB 2026*.
2. **A Decade of Systems for Human Data Interaction.**
   Eugene Wu, Yiru Chen, Haneen Mohammed, Zezhou Huang. *Information Systems, Vol. 138 (2026)*.
3. **Magentic Marketplace: An Open-Source Environment for Studying Agentic Markets.**
   Gagan Bansal, Wenyue Hua, Zezhou Huang, Adam Fourney, Amanda Swearngin, et al. *arXiv 2025* · Code.
4. **Data Cleaning Using Large Language Models.**

Shuo Zhang, Zezhou Huang, Eugene Wu. *DAIS@ICDE 2025*.

5. **Data-Centric Text-to-SQL with Large Language Models.**
   Zezhou Huang, Shuo Zhang, Kechen Liu, Eugene Wu. *TRL@NeurIPS 2024*.
6. **Transform Table to Database Using Large Language Models.**
   Zezhou Huang, Jia Guo, Eugene Wu. *TaDA@VLDB 2024*.
7. **SET: Searching Effective Supervised Learning Augmentations in Large Tabular Data Repositories.**
   Jiaxiang Liu, Zezhou Huang, Eugene Wu. *GUIDEAI@SIGMOD 2024*.
8. **Disambiguate Entity Matching through Relation Discovery with Large Language Models.**
   Zezhou Huang. *GUIDEAI@SIGMOD 2024*.
9. **Cocoon: Semantic Table Profiling Using Large Language Models.**
   Zezhou Huang, Eugene Wu. *HILDA@SIGMOD 2024* · Code.
10. **Relationalizing Tables with Large Language Models: The Promise and Challenges.**
    Zezhou Huang, Eugene Wu. *DBML@ICDE 2024*.
11. **The Fast and the Private: Task-based Dataset Search.**
    Zezhou Huang, Jiaxiang Liu, Haonan Wang, Eugene Wu. *CIDR 2024*.
12. **Lightweight Materialization for Fast Dashboards Over Joins.**
    Zezhou Huang, Eugene Wu. *SIGMOD 2024*.
13. **Data Ambiguity Strikes Back: How Documentation Improves GPT's Text-to-SQL.**
    Zezhou Huang, Pavan Kalyan Damalapati, Eugene Wu. *TRL@NeurIPS 2023*.
14. **Saibot: A Differentially Private Data Search Platform.**
    Zezhou Huang, Jiaxiang Liu, Daniel Gbenga Alabi, Raul Castro Fernandez, Eugene Wu. *VLDB 2023* · Code.
15. **Kitana: Efficient Data Augmentation Search for AutoML.**
    Zezhou Huang, Pranav Subramaniam, Raul Castro Fernandez, Eugene Wu. *arXiv* · Code.
16. **Random Forests over Normalized Data in CPU-GPU DBMSes.**
    Zezhou Huang, Pavan Kalyan Damalapati, Rathijit Sen, Eugene Wu. *DaMoN@SIGMOD 2023*.
17. **JoinBoost: Grow Trees Over Normalized Data Using Only SQL.**
    Zezhou Huang, Rathijit Sen, Jiaxiang Liu, Eugene Wu. *VLDB 2023* · Code · Talk.
18. **Aggregation Consistency Errors in Semantic Layers and How to Avoid Them.**
    Zezhou Huang, Pavan Kalyan Damalapati, Eugene Wu. *HILDA@SIGMOD 2023*.
19. **Reptile: Aggregation-level Explanations for Hierarchical Data.**
    Zezhou Huang, Eugene Wu. *SIGMOD 2022* · Code · Video · News.
20. **Calibration: A Simple Trick for Wide-table Delta Analytics.**
    Zezhou Huang, Eugene Wu. *arXiv 2022*.
21. **Spatial and Hedonic Analysis of Housing Prices in Shanghai.**
    Zezhou Huang, Ruishan Chen, Di Xu, Wei Zhou. *Habitat International 2017*.

Patents

1. **System and Method for Performing Query Operations on Run-Length-Encoded Data.**
   Rathijit Sen, Zezhou Huang, Matteo Interlandi, Marius Dumitru, Krystian Sakowski, et al. *US Patent 12,277,123, granted Apr 2025*.
2. **System and Method for Accelerating Query Execution.**
   Rathijit Sen, Zezhou Huang, Matteo Interlandi, Marius Dumitru, Carlo Aldo Curino, et al. *US Patent 12,277,122, granted Apr 2025*.

Professional Service

Program Committee & Reviewing. SIGMOD'27 PC · TaDA@VLDB'26 PC · DEEM@SIGMOD'26 PC · ICLR'26 Reviewer · TRL@ACL'25 PC · TaDA@VLDB'25 PC · DEEM@SIGMOD'24,'25 PC · ICML'25 Reviewer · TRL@NeurIPS'23,'24 PC · TaDA@VLDB'24 PC · GUIDEAI@SIGMOD'24 PC · DataPlat@ICDE'24 PC · DBML@ICDE'23,'24 PC.