Edward Gan

Email : edgan8@gmail.com
Web : edgan8.github.io

Software engineer working at the intersection of ML platform development and research. I am especially interested in tools to empower people to work with and understand LLMs.

Experience

Scale AI New York, NY
Staff Software Engineer *Jan 2024 – Present*

- **LLM Data Quality Platform**: Led a small team in developing a platform for post-training + benchmarking models on our data products. Identified quality issues that drove major changes in Scale's data strategy. (Blog post).
- **Model Leaderboards**: Standardized eval execution infra, tracing, and results tracking across the research team. Provided technical mentorship to engineers to bring leaderboard refresh times down from days to hours and integrate leaderboard evals into the main user app.

Waymo New York, NY
Senior Software Engineer *Nov 2022 – Jan 2024*

- **Agent Prediction Model Evals**: Launched eval pipeline for the agent prediction model on multi-agent interactions. Collaborated with researchers to identify bugs in causal agent predictions.
- **Data Scaling**: Stabilized the training data distribution for agent prediction models to improve cache re-use and cut storage costs.

Databricks San Francisco, CA
Senior Software Engineer *Jun 2020 – Aug 2022*

- **Model Monitoring**: Led a group of 3 engineers in developing the metrics definition API and compute engine for the model monitoring product.
- **ML Product Usability**: Grew usage of ML Ops products by streamlining navigation, adding experiments search, and developing in-notebook data profiling.

Google Brain Mountain View, CA
Research Intern *Summer 2019*

- **TFX Platform**: Reduced ads training data validation runtime by 10% with streaming statistics.

Facebook Menlo Park, CA
Software Engineer *2013 – 2015*

- **Data Workflows**: Developed Python APIs, scheduling logic, and UX for ETL pipeline backfills.

Education

Stanford University Stanford, CA
PhD in Computer Science, advised by Peter Bailis *2015 – 2020*

- **Thesis**: Data summaries for scalable, high-cardinality analytics

Harvard University Cambridge, MA
A.B. Summa Cum Laude in Computer Science and Mathematics *2013*

Skills

Languages: Python, Java, C++, SQL. **Technologies**: Kubernetes, PyTorch, Spark.

Selected Publications

CoopStore: Optimizing Precomputed Summaries for Aggregation VLDB
***Edward Gan**, Peter Bailis, Moses Charikar* *2020*

– System for optimizing data summaries in high cardinality query engines.

Approximate Selection with Guarantees using Proxies VLDB
Daniel Kang, **Edward Gan***, Peter Bailis, Tatsunori Hashimoto, Matei Zaharia* *2020*

– Sample-efficient methods for calibrating models used for text/video retrieval.

CrossTrainer: Practical Domain Adaptation with Loss Reweighting DEEM
*Justin Chen, **Edward Gan**, Kexin Rong, Sahaana Suri, Peter Bailis* *2019*

– Automatic hyperparameter tuning for transfer learning.

DIFF: A Relational Interface for Large-Scale Data Explanation VLDB
*Firas Abuzaid, Peter Kraft, Sahaana Suri, **Edward Gan**, ..., Peter Bailis, Matei Zaharia* *2019*

– SQL operator for explaining differences between datasets.

Moment-Based Quantile Sketches for ... Aggregation Queries MLSys, VLDB
***Edward Gan**, Jialin Ding, Kai Sheng Tai, Vatsal Sharan, Peter Bailis* *2018*

– Memory-efficient algorithms for estimating quantiles in distributed systems.

Scalable Kernel Density Classification via Threshold-Based Pruning SIGMOD
***Edward Gan**, Peter Bailis* *2017*

– Efficient non-parametric outlier classification.

MacroBase: Prioritizing Attention in Fast Data SIGMOD
*P. Bailis, **E. Gan**, S. Madden, D. Narayanan, K. Rong, S. Suri* *2017*

– End-to-end system for explaining anomalies on multi-dimensional event log data.