Peter Zhao

pzhao1799@gmail.com | 703-889-0678 | github.com/pzhao1799

Experience

Roblox San Mateo, CA

Senior Machine Learning Engineer, ML Platform *July 2021 - Present*

- Leading **AI-as-a-Service** to provide all players and creators on Roblox access to generative models in any experience, supporting translation, text generation, speech to text/text to speech, and 3D generation models. Developed roadmap to forklift internal model API's for public access, scaled internal model servers for expected initial launch of **10k+** inferences per second, worked with Creator API team to expose the API's in a safe and versionable manner with Studio client libraries, and architectured a low-latency, multi-modal safety pipeline with abuse reporting, ensuring safety and civility for all generative APIs.
- Created the **ML Gateway**, a highly available API gateway for models in a multi-tenant, distributed inference platform. Engineered features such as rate limiting, usage attribution, load balancing, and fallback support for diverse inference backends such as **Nvidia Triton**, **vLLM**, and **OpenAI-compatible** cloud providers via HTTP and gRPC with streaming, along with a batch API leveraging **Ray** for distributed inference and an asynchronous/workflow API using **Temporal** for async/inference graph paradigms. The ML Gateway serves **28+ unique models for 22+ production use cases**, including the Roblox AI Assistant with **99.98% availability.** All inference at Roblox will eventually migrate to the ML Gateway.
- Built an improved model serving API and migrated Roblox's CPU inference stack to utilize on-premise hardware, resulting in a cost reduction of **$12 million** annually. This increased serving capacity over **10x**, reaching peak traffic of **600k+** inferences per second.
- Spearheaded the design and implementation of **Frost**, Roblox's in-house feature store on top of **Feast** and internal data infrastructure, increasing data accessibility, improving feature engineering workflows, and lowering costs by **$8 million** annually from licensing and database costs. Frost serves **100+ feature services at 500k+ feature requests per second under 30ms**, and ingests **300k+ features per second** with overall feature freshness under 1 day via batch and streaming pipelines.
- Delivered scalable **MLOps** on our AWS EKS hosted Kubeflow platform with **1500+ nodes**, supporting **600+ engineers** for distinct ML use cases (RecSys, Safety, Content Understanding, GenAI, Analytics, etc).

Software Engineer, Search & Discovery Platform

- Oversaw the infrastructure for search and recommendations, encompassing data pipelines, **Elasticsearch** database optimizations, and continuous integration/continuous deployment (CI/CD) processes.

De-Generate Remote

Blockchain Architect *Oct 2020 - June 2021*

- Engineered Solidity smart contracts on the Ethereum blockchain for a decentralized finance roboadvisor for crypto investments.
- Researched investment protocols and constructed strategy portfolios via staking, liquidity provision, and yield farming.
- Designed investment API and schemas built on top of DynamoDB. Built out testing framework for maintaining security and optimizing transaction fees using **Hardhat**.

Projects

- **Learned Filters**: A Pytorch implmentation of a Learned Bloom filter and a Sandwiched Learned Bloom filter using a deep neural network
- **RusTED**: A Rust implementation of the Zhang-Shasha Algorithm, which computes the edit distance between two trees
- **Firestone**: A Hearthstone clone developed in Clojure to focus on functional programming software design paradigms
- **Community Detection**: Implementation and testing of the Girvan-Newman Algorithm in Python for detecting communities in networks
- **C++uckoo**: A C++ implementation of a Cuckoo Filter using a variety of hashing techniques

Education

Williams College Williamstown, MA

B.A. Computer Science, B.A. Mathematics. *Sept 2017 - June 2021*

Relevant Coursework: Algorithm Analysis, Computer Architecture, Machine Learning, Natural Language Processing, Storage Systems, Programming Languages, Probability, Graph Theory, Combinatorial Optimization, Abstract Algebra, Algorithmic Game Theory

Study Abroad: Acquincum Institute for Technology in Budapest - Spring 2020

Technical Skills

- **Languages**: Python, Go, Rust, C#, C++, Java, JavaScript, TypeScript, Scala, Solidity
- **Frameworks**: PyTorch, Transformers, Ray, vLLM, Feast
- **Tools**: AWS (EC2, DynamoDb, S3, SNS/SQS, RDS, EFS, EKS, Fargate), Azure (OpenAI), Temporal, Terraform, Airflow, Elasticsearch, Kubernetes, Nomad, Docker, Kubeflow, Spark, Kafka, Flink, Redis, MySQL, GraphQL, Node.js, .NET, Arize

Additional Experience & Achievements

- **Sigma Xi** Scientific Honor Society
- **2021 Ward Prize Recipient** for Best Computer Science Department Project
- **Research Assistant** for Professor Daniel Barowy, working on SWELL
- **Teaching Assistant** for Theory of Computation (Spring '21), Algorithm Analysis (Fall '19, Fall '20), Linear Algebra (Fall '18)
- **Treasurer** of Williams College All Campus Entertainment: 2019-2021
- **President** of the Chinese American Student Association: 2018-2019
- Placed **4th** at Google Tech Challenge: Cambridge 2018
- Jack Kent Cooke Scholar and Questbridge Scholar