Peter Zhao

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

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#, C++, Java, JavaScript, Clojure, Scala, Haskell, Solidity **Frameworks**: PyTorch, Tensorflow
- **Tools**: AWS, Azure, GCP, Ray, Terraform, Airflow, Elasticsearch, Kubernetes (Kubeflow, Argo, Istio, Cilium), Nomad, Docker, Spark, Kafka, Flink, Redis, .NET

Experience

OpenAI — San Francisco, CA

Member of Technical Staff, Fleet Infrastructure — *Oct 2021 - Present*

- Managed GPU scheduling and infrastructure for research and applied

Roblox — San Mateo, CA

Senior Machine Learning Infrastructure Engineer, ML Platform — *July 2021 - Oct 2021*

- Built Roblox's next-generation inference platform from first-principles on Kubernetes. Currently serving online text and content safety production traffic with over **800k RPS**. Owned the deployment control plane and inference custom resource definition
- Drove efforts with core Roblox compute infrastructure teams to scale all model inference for over **35 million concurrent users**, assisting with capacity estimation, hardware provisioning, and squeeze testing, defining appropriate SLO's for product surfaces.
- Lead and launched **AI APIs** to provide all creators and players on Roblox access to **text generation** and **translation models** in experience as a Studio API, building end-to-end serving and abuse reporting pipelines and supported infrastructure efforts in fine tuning an online safety critique model.
- Built the **Roblox ML Gateway**, a highly available entry point for models in a multi-tenant, distributed inference platform. Engineered features such as access control, usage attribution, intelligent routing/fair load balancing, and multi-cluster/fallback support for diverse inference backends.
- Integrated **vLLM** as the primary open source LLM inferencing engine at Roblox, driving adoption of relevant features such as **multimodal models** and **embedding** support.
- Architectured an improved model deployment API and optimized Roblox's CPU inference stack to utilize on-premise hardware, resulting in a cost reduction of **$20 million** annually.
- Spearheaded the design and implementation of **Frost**, Roblox's in-house feature store, establishing a robust machine learning data infrastructure. This increased data accessibility, improved feature engineering workflows, and lowered costs by **$7 million** annually from licensing and database costs. Currently serves **7000+ features** at **1m+ RPS**.

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.

Williams College Computer Science — Williamstown, MA

Research Intern, Professor Daniel Barowy — *June 2019 - Dec 2019*

- **SWELL**: Researched a self-repairing parser combinator library for fixing syntactic parsing errors by determining a possible fix using minimum edit distance.

Projects

- **Learned Filters**: A Pytorch implmentation of a Learned Bloom filter and a Sandwiched Learned Bloom filter using a deep neural network
- **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

Additional Experience & Achievements

- **Sigma Xi** Scientific Honor Society
- **2021 Ward Prize Recipient** for Best Computer Science Department Project
- **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