Austin Silveria

austinsilveria@gmail.com

Educational Background

California Polytechnic State University, San Luis Obispo 09/2017-06/2021

- Computer Science | **3.84** GPA | Magna Cum Laude

Independent Research

Fused Sparse Tree Attention Triton Kernel 02/2024

- 70x faster than Flash-2 + tree attention mask at a 65k tree size, 130 line Triton kernel
- Only loads keys and values for the query block's unique tree ancestors and leaves

Sparsity Aware Inference with CPU Offloaded Language Models 11/2023-12/2023

- 15x faster than naive offloading, 733 lines of native PyTorch (no custom kernels needed)
- Exploits power law distribution and adjacent token similarity of parameter access

AI's Networked Memory Hierarchy Overhang 10/2022-04/2023

- Research blog post analyzing the paths to and implications of scalable edge inference
- Experimental results indicate that it may be possible to partition a pre-trained Transformer to accelerate streaming edge inference

Broadcasting Humans' Epistemic Affordances 05/2020-06/2021

- Initial paper introducing a CLI to embed personal search history, extract summaries/keywords, reduce dimensions with t-SNE
- Follow-on paper creating a mapbox web application to visualize, navigate, and organize hierarchical document clusters

Employment History

Software Development Engineer II 07/2022-08/2023

Amazon | Last Mile Jurisdiction Planning

- Led team's planning/delivery cycle of subsequent service/web app functionality to automate and optimize Amazon's last mile jurisdiction planning
- Mentored two part time Jr. Developers who are independently owning the design/delivery/iteration of a Director tracked project

Software Development Engineer I 08/2021-06/2022

Amazon | Last Mile Jurisdiction Planning

- Led a team of SDEs in designing, developing, and launching a new service and React web application for 50+ internal Amazon Logistics planning employees
- Led the migration of multiple software teams' data pipelines to consume from our newly launched planning system

Jr. Software Developer 04/2019-06/2021

Amazon | Last Mile Jurisdiction Planning

- Built full stack features for internal Mapbox/React/Redux web app