Kevin Lin

(469) 931-9689 | kevin.lin.cs1@gmail.com | klin.one | github.com/kevin314

Education

The University of Texas at Dallas May 2022
Bachelor of Science in Computer Science *GPA 3.72*

Experience

Software Engineer May 2022 – Present
Veryable Inc. *Dallas, TX*

- Designed and implemented a notification system to deliver and track over **10 million** daily notifications to users via email, SMS, and push, driving a **23%** increase in conversion
- Led the development of a resource-locking solution using Redis to ensure idempotent behavior across distributed payment systems, securing over **2,500** daily transactions by preventing duplicate payments
- Developed using TypeScript, Express, and PostgreSQL a gig work marketplace, implementing systems for managing job bidding, schedule conflicts, payment disputes, and more, handling over **66,000** monthly work contracts
- Implemented a data-centric service using CQRS, Prisma, and GraphQL to manage a read model updated via webhooks, enabling a **400ms** reduction in average response times
- Reduced development cycle time by **33%** through refactoring a monolith into independent microservices

Software Engineering Intern Oct 2021 – Jan 2022
Lvlup.ai *Remote*

- Developed full-stack features for a social networking platform using Dart/Flutter and Cloud Firestore, including a recommendation system that suggests user connections based on profile preferences
- Created and maintained detailed technical documentation for new features and existing APIs, expediting the onboarding of new team members

Open Source

vLLM | *Library for LLM inference and serving*

- Refactored params for updating metadata during speculative decoding *github.com/vllm-project/vllm/pull/8224*
- Fixed race condition with port assignment by using temporary socket *github.com/vllm-project/vllm/pull/8491*
- Added check to prevent running encoder/decoder models with CPU *github.com/vllm-project/vllm/pull/8355*

Projects

ChatTuring

- Developed and deployed an online chat application for a Turing game where users determine whether they are conversing with a real person or with a language model
- Fine-tuned the Meta-Llama-3-8B model and used in-context learning for generating more human-like responses
- Built a real-time chat room system using Phoenix LiveView and WebSockets, enabling low-latency messaging

Semantic Browser Search

- Developed a Chrome extension to enable semantic search on users' browser history, allowing for retrieval of pages based on contextual relevance
- Implemented an IndexedDB vector database to store text embeddings generated by a Sentence-BERT model

Alien-bot

- Implemented a multiplayer game for Discord where players strategize to identify an "alien" player through prompts provided by a chat bot, communicating via a shared text channel
- Designed a wrapper for Discord's API using WebSockets in Node.js for sending, receiving, and editing messages
- Registered custom commands within an event-driven architecture to facilitate real-time game management

Skills

Languages: TypeScript, JavaScript, Python, HTML, CSS, Java, Elixir, SQL, C++
Frameworks: React, Express, Jest, Phoenix, Apollo/GraphQL, Redis, RabbitMQ, PostgreSQL, MySQL, MongoDB
Infrastructure: Docker, Kubernetes, AWS (EC2, RDS), OAuth 2.0