Yuyang Deng

626-413-9632 | ethand.2091@gmail.com | linkedin.com/in/ethan-yuyang-deng | github.com/ethand605

Professional Summary

Software Engineer specialized in Backend and Infrastructure, with accomplishments in high-growth, fast-paced startup environments. Also equipped with experience in Machine Learning and a strong Full Stack Development foundation. Skilled in driving product innovation and user engagement through scalable, data-driven solutions, and managing complex projects. Strong ability to collaborate across teams and rapidly learn and adapt to new technologies.

Technical Skills

Languages: Python, Go, Java, JavaScript, TypeScript, C++, C, HTML, CSS
Databases: PostgreSQL, Redis, MongoDB, Elasticsearch, BigQuery, DynamoDB, BigQuery, Firestore
Frameworks/Tools: Express, Node.js, FastAPI, Spring, Next.js, React, React Native, Angular, Firebase, Storybook
Cloud: Amazon Web Services(AWS), Google Cloud Platform(GCP), DigitalOcean, CoreWeave
DevOps: Docker, Kubernetes, Terraform, AWS(EKS, S3, ECR, EC2, VPC, ALB, RDS), Helm, DataDog, Grafana, Linux, CI/CD
Machine Learning: Pandas, PyTorch, Numpy, Transformers, LangChain, Scikit-learn, WandbB, vLLM, Axolotl, CUDA, Kserve

Work Experience

Kyte *February 2024 – Present*
Software Engineer *Remote, USA*

- Developing reliable and scalable backend infrastructure, enhancing observability, CI/CD, development experience, and security in collaboration with product, engineering, and operations teams
- Orchestrated an end-to-end enhancement of database performance by integrating Datadog APM on AWS RDS with EKS and Terraform, optimizing queries to achieve a 99% reduction in duration, alongside a 5% decrease in CPU usage on DB
- Implemented test monitoring with DataDog and Kubernetes and optimized integration test runtime by 20% in CI pipelines
- Streamlined RDS management processes and reduced costs by implementing workflows in Airflow with EKS orchestration
- Implemented API Gateway, coordinated with service owners to drive migration, enhancing observability and security

Flyx AI *December 2023 – February 2024*
Software Engineer *Remote, USA*

- Led the end-to-end integration of large language models(LLMs), encompassing backend development(APIs, prompting), infrastructure, model training, and model serving, resulting in scalable AI-driven features that enhance user experience
- Built recommendation system using Elasticsearch Vector DB, Go, PostgreSQL, contributing to 40% of user engagement
- Optimized model throughput on GPUs, **cutting infrastructure costs by $30k/month** with concurrency tuning and batching
- Developed A/B testing framework using Go, PostgreSQL, ClickHouse, refining model selection and improvement process
- Managed and optimized distributed serverless machine learning infrastructure using Kubernetes, ensuring high availability, fault tolerance, and scalability for **500k+ daily messages** with robust load balancing, autoscaling, and a failover system

Flyx AI *January 2023 – December 2023*
Software Engineer Intern *Remote, USA*

- Partnered with cross-functional teams to devise and execute technical strategies, resulting in **10% increase in user retention** and a **50% increase in session duration** for **30K+ customers**
- Spearheaded a fullstack natural language-driven data analysis and management feature using Next.js, Express, PostgreSQL, resulting in successful investor demos and early access to GPT models
- Designed and built ETL pipelines using Redis and PostgreSQL which resulted in a 7% annual savings in service costs
- Deployed and maintained ML infrastructure with Kubernetes, Knative, Grafana**,** achieving **scalability** and **99.5% uptime**
- Orchestrated an automated data pipeline using Python, Pandas, Kubernetes, BigQuery, and WandB, which performs extensive data wrangling and reduces recurring workload by 80%, efficiently processing a dataset with over 15M+ rows
- Achieved a 25% increase in user satisfaction by fine-tuning Large Language Model with curated data

SpaceFlare *January 2022 – June 2023*
Software Engineer *Irvine, CA*

- Led a 4-member team to prototype a clean energy platform for space rentals by building owners
- Spearheaded development of a full-stack application using Flask, AWS S3, PostgreSQL, and React/Redux

Projects

PeterPortal API | AWS(Lambda, API Gateway, RDS), Express.js, TypeScript

- Collaborated in a team of 7 to develop an API used across 3 platforms, averaging 2.5k users and 20k views per month
- Implemented an automation solution for data scraping, wrangling, and uploading of grades, reducing manual work by 80%
- Proposed and assisted in the development of a microservice architecture, improving system maintainability by 25%

Texera (UCI Research Project) | Scala, MongoDB, Angular, MySQL, Java

- Designed and implemented a comparison feature for workflows, streamlining user analysis by 15%
- Managed data workflows and provided real-time updates through WebSockets, improving data sync and user experience

Education

University of California Irvine *September 2023*
Bachelor of Science, Computer Science *GPA: **3.9***