

CHRISTOPHER TON

christopher.ton@sjsu.edu — (669) 254-6967 — linkedin.com/in/chriztopherton — github.com/chriztopherton — US Citizen

Experience

Genentech

San Diego County, CA

Data Engineer

Apr 2024–Present

- Built and deployed a full stack LLM-powered chat assistant using Streamlit, owning end-to-end architecture including data ingestion, ETL, prompt engineering, model evaluation, inferencing, and deployment using Docker and Kubernetes.
- Designed and implemented scalable ETL pipelines to process and embed Confluence-based documentation into Azure AI Search, enabling semantic search through RAG and LangChain workflows.
- Developed serverless APIs using AWS Lambda, S3, Bedrock, and OpenSearch to support multiple Retrieval-Augmented Generation (RAG) applications, reducing latency and improving maintainability.
- Engineered performant data models in Snowflake using Talend, PySpark, and SQL across 7+ global sources; leveraged dbt to enforce testing and materializations via GitLab CI/CD pipelines.

Data Scientist Intern

Jun 2022–Apr 2024

- Reduced resolution times by 20% and drove improvements in operational efficiency by developing and monitoring predictive anomaly detection models using XGBoostRegressor, in partnership with C3.AI.
- Automated clinical reporting validation processes by building and deploying a sliding-window search module in R, reducing manual effort, streamlining cross-checks, and enhancing data accuracy and reliability.

Deloitte

Remote

Data Engineer via Brooksource

Feb 2022–May 2022

- Deployed data pipelines using Informatica TDM to automate sensitive data masking, ensuring compliance with data protection regulations and minimizing security risks.
- Developed Python-based solutions to model data distributions and population parameters, enabling risk assessment for data re-identification through quasi-identifier and pseudo-anonymization analysis.

Shell Recharge (formerly Volta Charging)

San Francisco Bay Area County, CA

Data Curator

Nov 2021–Mar 2022

- Automated and optimized data ingestion workflows by developing Python scripts to extract, transform, and curate datasets from over three vendor sources, achieving a 30% improvement in data quality and observability.

Data Marketing & Insights Intern

Jul 2021–Sep 2021

- Enhanced efficiency of EV network planning policies by conducting in-depth exploratory analysis with KMeans clustering to derive growth insights and integrated results into a reporting dashboard, enabling stakeholders to monitor charging performance and marketing initiatives in real time.

Publications

arxiv: A Comparison of LLM Finetuning Methods & Evaluation Metrics with Travel Chatbot Use Case

Implemented various fine-tuning techniques such as Supervised fine-tuning (SFT), Retrieval Augmented fine-tuning, and Reinforcement Learning with Human Feedback (RLHF).

Education

San Jose State University

Master of Science, Data Analytics

San Jose, CA

Graduated May 2024

University of California, Davis

Bachelor of Science, Statistics

Davis, CA

Graduated Jun 2020

Skills

Languages: Python, R, SQL

Data Engineering & Processing: PySpark, Pandas, NumPy, scikit-learn, SageMaker, LangChain

LLM & ML Frameworks: AWS SageMaker, AWS Bedrock, Dataiku DSS, PyTorch

Cloud Platforms: AWS (S3, Lambda, Redshift), GCP BigQuery, Azure AI Indexes

DevOps & Containerization: Docker, Git, CI/CD pipelines

Development Tools: VSCode, Jupyter Notebooks

Project Management & Documentation: Jira, Confluence