Andreas Schaler

- Full-Stack, LLM Engineer -

andreas.schaler.cs@gmail.com
614 316 8982
https://www.linkedin.com/in/andreas-schaler-ab38b5201/
https://github.com/Andreas3333

Skills

Programming Languages - Python, C++, Javascript, Typescript, Bash, Django REST Framework, FastAPI, Flask, React
Cloud Services - Terraform, AWS CDK, AWS Cloudformation, EC2, ECS, ECR, EKS, Lambda, Event Bridge, SQS, SNS, S3, RDS
LLMs - Pytorch, Hugging Face tool chain, Tensorboard, KubeFlow, MLflow, Jupyter, Pandas, NumPy, Scikit Learn, KaTex, Safetensors, Bert models, domain specific supervised fine-tuning for NLP, GGUF, Apple CoreML, llama.cpp, vllm
Miscellaneous - Linux, Systemd, D-Bus, Docker, Podman, Kubernetes, Kustomize, Helm, Skaffold, Ansible, Packer, GItlab CI/CD, GitHub Actions, Pydantic, Open API Spec., RabbitMQ, Postgres SQL, Vite, UV, Poetry, MkDocs

Employment - 01/2023 - Present

Nimbis Services — DevOps Engineer

- Implemented multi stage container builds and flexible entrypoint service start up scripts to manage dependencies and build configuration supporting flexibility for running the service in development and production environments.
- Implemented parameterized container builds to provision model weights via container build arguments supporting LLM model artifact packaging and deployments impacting development and production.
- Implemented persistent container build caches in container files and persistent CI/CD runners which reduced build times to positively improve local and CI/CD build times.
- Extended Terraform to support containerized deployments of LLM services on AWS EC2 nodes.
- Implemented Library type Helm Charts used for packaging and managing service deployments including service dependencies enabling dynamic injection capabilities for configuring sub-charts and their services.
- Implemented DRF REST services using Domain Driven Design patterns and concepts improving clean internal and external interfaces supporting modularity, interface definitions, and code organization along with Open API Spec.
- Designed and implemented Event Driven cloud native solution architecture and application architecture for virus scanner solution which leveraged AWS Event Bridge, S3, ECS, and SQS monitoring a S3 bucket for malicious files.
- Co-Authored, built, packaged and deployed Vue 3 component library utilized in a SPA web application. Extended and contributed to multiple SPA web applications implemented in React utilizing the AWS Cloud Scape library.
- Implemented max concurrency capability enabling full saturation of remote compute resources managed through an asynchronous service manager.
- Implemented preemptive failure mechanisms for missing configuration prerequisites in pipeline job submissions improving successful job submission rates by 15%.
- Implemented Skaffold profiles and profile requirements to expand and support multiple configurations of services for local development and automated production deployment environments.
- Analyzed the security profile of Podman's rootless capabilities as a container manager and implemented solutions improving security, software artifact production, and deployment in a rootless container environment on RHEL.

Projects - 01/2023 - Present

LLM NER REST API — Named Entity Recognition (NER) classifier API

I along with two other colleagues completed system design, service implementations, and containerized model artifact deployment on AWS for this system. The NER REST API was part of the larger multimodal RAG on document system employing and training multiple open source LLMs for the use case. The NER REST API utilized a DeBerta model architecture for token classification. It was selected based on its benchmark performance results on the CoNLL03 NER dataset.

Personal Finances App — Web application for analyzing personal spending habits leveraging fine tuned base Bert

A traditional SPA and REST API project that allows users to upload their monthly transaction data and visualize their spending for the month. The transaction data is classified by a supervised fine tuned Bert model for multi class sentence classification. The classifications by the Bert model are then used as annotations on the data. With the data annotated visualizations are created allowing the user to assess their spending habits.

Extension and Adaptation of AWS RES — Fork and extension project

This system is an extension of the AWS RES service for the company. My contributions to this system involve reverse engineering of the solution to extend the SPA web application (utilizing the AWS Cloud Scape React component library), altering and extending Cloudformation, building, packaging and publishing of deployment artifacts, altering compute cluster bootstrapping mechanisms, and implementing hardened automated AMI builds for cluster node images and VDIs.

Education - 01/2020 - 12/2022

Kent State University, Bachelors — Computer Science (CS) Data Engineering Concentration

Certifications

CompTIA, Security + Certification — 03/2025