github.com/max-gartz
Maximilian Gartz
gartz.maximilian@gmail.com ✉

Maximilian Gartz

Software Engineer - Machine Learning, Berlin, Germany

SKILLS

- **Specializations:** Machine Learning, Software Engineering, MLOps & DevOps
- **Languages:** ***Python***, Java, Bash, SQL | German (native), English (professional proficiency)
- **Frameworks & Libraries:** FastAPI, PyTorch, TensorFlow, Huggingface Transformers & Datasets
- **Tools:** Git, Terraform, dbt, Docker, Kubernetes, Grafana, MLflow, Weights & Biases, Kubeflow, Flyte, Triton
- **Databases:** PostgreSQL, MySQL, BigQuery, Firestore, Prometheus
- **Clouds:** ***GCP***, Azure, AWS

WORK EXPERIENCE

Machine Learning Engineer @ML6, *Berlin, Germany* *06/2021 – present*

- Headed the development of a RAG based chatbot for a global leader in regulatory reporting, providing various solutions for financial institutions, regulators and tax authorities. The chatbot serves internal customer support by providing easier access to internal knowledge and documentation, but will be rolled out to clients directly to significantly reduce the load on the customer support team. The solution is based on open-source LLMs deployed on GCP's Vertex AI Endpoints using the vllm server and a hybrid (semantic and keyword) search on PostgreSQL using the pgvector extension.
- Lead the efforts of applying and evaluating LLMs for food recipe adaptations based on user queries for a US Retail giant, while implementing CI/CD pipelines across various streams of work. Deployed solutions utilizing the Azure Container Apps and Web App services.
- Driving the development of an internal standard solution for LLM and RAG projects, designed to seamlessly integrate various data sources, LLMs and embedding models through a centralized configuration. Deployments on GCP and AWS utilize various managed services, while on-prem deployments leverage Docker Compose or Kubernetes. This approach has significantly driven down development costs and enabled us to offer competitive pricing for RAG solutions.
- Served as Tech Lead on a demand forecasting project for a large retail company in Germany. Achieved a nearly 10% revenue increase by improving warehouse replenishment through the deployment of classical time series models. Lead the transition of POC stage solutions into production on GCP, leveraging Terraform, BigQuery, dbt, Cloud Run, and Vertex AI, while streamlining build and deployment processes.
- Acted as Lead Engineer on an MLOps project for a big Swiss insurance company. Implemented significant improvements in development workflows by establishing a standardized GitHub template with robust CI/CD pipelines for ML projects on GCP, adhering to stringent security protocols. Focused on end-to-end ML pipelines utilizing Vertex AI and facilitated extensive knowledge transfer to client's Data Scientists and Engineers.
- Took up an Engineer role in a POC aimed at detecting asbestos fibers in high-resolution SEM images using the Tensorflow Detection API and SAHI. Demonstrated the capability to automate fiber detection with human level accuracy, culminating in a labor reduction of up to 80%. Concurrently authored a well-received blog post with more than 15k views, elucidating the approach using the MMDetection framework and SAHI for small object detection in satellite imagery.
- Repeatedly mentored new recruits during their onboarding periods and first projects

Software Engineer @InsideM2M GmbH, *Osnabrueck, Germany* *05/2018 – 12/2020*

- Developed REST APIs for client-facing applications in IoT projects utilizing Java EE and built micro-services with Quarkus. Deployed solutions on Docker Swarm managed with portainer. Ensured code quality through unit testing with JUnit.
- Conducted comparison of time series databases InfluxDB and TimescaleDB to evaluate possible performance enhancement in handling transactional data from IoT devices over basic relational database solutions. Investigated downsampling and retention policies as well as visualizations with Grafana.

EDUCATION

Bachelor of Cognitive Science @University of Osnabrueck, *Osnabrueck, Germany* *10/2017 – 03/2021*

- Graduated with distinction and final grade 1.0 (German) equivalent to 4.0 GPA in USA
- Focused on Machine Learning, Bayesian statistics and Mathematics
- Bachelor's thesis: Bayesian Inference with Hamiltonian Normalizing Flows

CERTIFICATIONS

Google Cloud Certified Professional Machine Learning Engineer *since 10/2021*