Stefan AI-Safety-Aichinger

GENERATIVE AI ENGINEER

@ stefan.aichinger@example.de | +49 761 4421 3380 | Freiburg, Deutschland
stefan-aichinger.ai | linkedin.com/in/stefan-aichinger | github.com/saichinger

PROFIL

Generative AI Engineer mit 5 Jahren Erfahrung bei Black Forest Labs Freiburg und Zalando SE Berlin Data Science. Schwerpunkt SDXL- und FLUX-Pipelines, LoRA-Adapter und produktive Bild-Generations-APIs mit messbarer Wirkung auf Generations-Latenz, GPU-Kosten und A/B-Konversion. Co-Autor bei NeurIPS 2024 zu effizienten Diffusions-Modellen.

BERUFSERFAHRUNG

Generative AI Engineer
10/2023 - heute

Black Forest Labs
Freiburg, Deutschland

GenAI-Engineer im FLUX-Modell-Team an Diffusions-Inferenz und Adaptern

- Implementierung von Quantisierung und Speculative Sampling fuer FLUX.1, p95-Generations-Latenz von 4,8 s auf 1,4 s reduziert
- LoRA-Adapter-Training fuer 14 Marken-Stile auf 480.000 kuratierten Bildern mit DreamBooth- und Textual-Inversion-Pipelines
- Aufbau einer Hugging-Face-Datasets-Pipeline fuer 2,4 Mio. lizenzierte Bilder mit Wasserzeichen-Validierung
- Co-Autor bei NeurIPS 2024 zu effizienten Sampler-Algorithmen fuer Diffusions-Modelle

ML Engineer (Generative)
10/2020 - 09/2023

Zalando SE
Berlin, Deutschland

ML Engineer im Outfit-Visualisierungs-Team

- Fine-Tuning von SDXL-Modellen auf 2,4 Mio. Zalando-Produktbildern, FID 18,4 auf 8,2 reduziert
- A/B-Tests mit Statsig auf 18 Mio. Sessions, Klick-Rate +9,8 Prozentpunkte fuer AI-Visuals-Variante
- Aufbau einer Gradio-Demo-Plattform fuer 14 interne Stakeholder-Teams mit SSO und Audit-Trails
- Mentoring von 2 Junior-Engineers in Diffusions-Modell-Training und Modal-Labs-Serverless-Deployment

PROJEKTE

flux-lora-pack (Open Source)
01/2025 - heute

Kurierte LoRA-Adapter-Sammlung fuer FLUX.1 und SDXL, 4.800 Hugging-Face-Downloads pro Woche, 14 Marken-Adoptionen

Zalando AI-Visuals
06/2024 - 09/2024

Diffusions-basierte Outfit-Visualisierungen fuer 2,4 Mio. Produktbilder, Klick-Rate +9,8 Prozentpunkte

PUBLIKATIONEN

Efficient Sampler Schedules for Diffusion Models at Production Scale
12/2024

AUSBILDUNG

M.Sc. Computer Science
10/2018 - 09/2020

EPFL
Lausanne, Schweiz

Computer Science (Schwerpunkt ML)
1,2

B.Sc. Informatik
Albert-Ludwigs-Universitaet Freiburg
Informatik

10/2015 - 09/2018
Freiburg, Deutschland
GPA: 1,4

FÄHIGKEITEN

PyTorch + Diffusers + Transformers | Diffusion Models (SDXL, FLUX, Stable Diffusion 3) | LoRA, DreamBooth & ControlNet-Adapter | vLLM + Modal Labs + Together AI | Vector-DBs (Weaviate, Pgvector) | LangChain & LlamaIndex | FastAPI + Gradio + Streamlit | AWS SageMaker + Lambda + S3

ZERTIFIKATE

NVIDIA Deep Learning Institute - Generative AI mit Diffusion Models — 02/2026

Hugging Face Certified Diffusion Engineer — 11/2024

AWS Certified Machine Learning Specialty — 06/2023

SPRACHEN

Deutsch — Muttersprache
Englisch — C1

STÄRKEN

Produkt-Naehe
Sehe Generative AI als Bestandteil eines Konversions-Flows, nicht als Demo - mit klaren A/B-Tests und Guardrails

Latenz-Disziplin
Optimiere konsequent Sampler, Quantisierung und Caching, um Generations-Latenz fuer Endnutzer unter 2 s zu halten

Marken-Sicherheit
Implementiere Content-Filter, Wasserzeichen und LoRA-Whitelists fuer Marken-konforme Bild- und Text-Generation