Konstantin RAG-LLM-Holzweissig

Senior ML Engineer

konstantin.holzweissig@example.de
+49 89 4471 6620
Muenchen, Deutschland
konstantin-holzweissig.ai
linkedin.com/in/konstantin-holzweissig
github.com/khlw

Profil

Senior ML Engineer mit 8 Jahren Erfahrung bei BMW Group AI Lab und Aleph Alpha. Tech-Lead eines 6-koepfigen Squads fuer LLM-Serving auf H100-Clustern mit messbarer Wirkung auf Modell-Latenz, GPU-Auslastung und A/B-Test-Konversion. Co-Autor bei NeurIPS 2024 und Maintainer von vllm-bench mit 3.200 GitHub-Stars.

Berufserfahrung

Senior ML Engineer
07/2022 - heute

BMW Group AI Lab — Muenchen, Deutschland

Tech-Lead eines 6-koepfigen Squads fuer GenAI-In-Car-Assistenten

- Architektur eines RAG-Stacks (vLLM, Qdrant, FastAPI) auf 32 H100-GPUs fuer 24 BMW-iX-Modellreihen, p95-Token-Latenz von 1,4 s auf 0,38 s
- A/B-Tests mit Eppo auf 4,8 Mio. monatlichen Nutzern, CTR-Lift +14,4 Prozentpunkte fuer Service-Buchungen
- Reduktion von Inferenz-Kosten um 38 Prozent (480.000 EUR p.a.) durch Speculative Decoding und Quantisierung INT8
- Tech-Lead fuer 6 Engineers, 2 Beforderungen ML-Engineer auf Senior in 14 Monaten dokumentiert

ML Engineer
10/2018 - 06/2022

Aleph Alpha — Heidelberg, Deutschland

ML Engineer im LLM-Trainings-Team fuer Luminous-Modelle

- Pre-Training und Fine-Tuning von 7B- bis 70B-Parameter-Modellen auf DGX-SuperPOD mit 384 A100-GPUs
- Implementierung von DeepSpeed Stage 3 und FlashAttention v2, Throughput +180 Prozent bei gleicher GPU-Auslastung
- Aufbau einer internen Evaluations-Suite mit MMLU-DE, HumanEval und Helm-Lite fuer 18 interne Modell-Versionen
- Co-Autor bei NeurIPS 2024 zu deutschsprachigen LLM-Evaluations-Methoden

Sprachen

Deutsch	Muttersprache
Englisch	C2
Mandarin	B1

Ausbildung

M.Sc. Computer Science
10/2015 - 09/2018

ETH Zuerich — Zuerich, Schweiz

Computer Science (Schwerpunkt Machine Learning)

1,1

B.Sc. Informatik
10/2012 - 09/2015

TU Berlin — Berlin, Deutschland

Informatik — GPA: 1,3

Fähigkeiten

Python (PyTorch, JAX, Triton CUDA), LLM Stack (Transformers, vLLM, DeepSpeed), Ray, Kubeflow & Argo Workflows, Triton Inference Server & TensorRT, Kubernetes, Helm & ArgoCD, AWS SageMaker, Vertex AI & Azure ML, Vector-DBs (Pinecone, Weaviate, Qdrant), Statsig, GrowthBook & Eppo A/B-Testing

Projekte

vllm-bench (Open Source)
01/2024 - heute

Benchmark-Toolkit fuer vLLM und TensorRT-LLM mit 3.200 GitHub-Stars und 18.400 monatlichen PyPI-Downloads

GenAI-Konversion-Lift bei BMW iX
01/2024 - 06/2024

RAG-basierter In-Car-Assistent fuer 24 Modelle, +14,4 Prozentpunkte CTR und 28 Mio. EUR jaehrlicher Service-Wert

Zertifikate

Google Cloud Professional Machine Learning Engineer
11/2024

AWS Certified Machine Learning Specialty
06/2024

NVIDIA Deep Learning Institute - LLM Customization
03/2023

Coursera Stanford Deep Learning Specialization
11/2022

Stärken

Architektonisches Urteil
Treffe robuste Trade-offs zwischen Modell-Qualitaet, Latenz und GPU-Kosten mit dokumentierten ADRs

Mentoring-Tiefe
Habe 5 ML-Engineers ueber 18 Monate von Junior auf Senior gefuehrt und 2 davon zu Tech-Leads befoerdert

Stakeholder-Klarheit
Uebersetze ML-Risiken in BWL-Sprache und liefere wirtschaftlich tragfaehige Roadmaps an C-Level