Carolin Computer-Vision-Pries

Staff ML Engineer

carolin.pries@example.de | +49 9131 4485 2280 | Erlangen, Deutschland
carolin-pries.ai | linkedin.com/in/carolin-pries | github.com/cpries

Profil

Staff ML Engineer mit 11 Jahren Erfahrung bei Siemens Healthineers AI und Mercedes-Benz Tech Innovation. Verantwortlich fuer ML-Plattform-Strategie ueber 4 Squads und 28 Engineers mit messbarer Wirkung auf Trainings-Compute, p95-Latenz und Modell-Genauigkeit. Co-Autor bei NeurIPS 2023 und CVPR 2025 sowie Heisenberg-Stipendium-Empfaenger der DFG.

Berufserfahrung

Staff ML Engineer
07/2022 - heute

Siemens Healthineers AI — Erlangen, Deutschland

Staff Engineer mit Plattform-Verantwortung fuer 4 ML-Squads und 28 Engineers

- Konzeption und Einfuehrung einer einheitlichen ML-Plattform auf Kubeflow und Ray, 240 produktive Modelle ueber 14 Produktlinien
- Reduktion von Trainings-Compute um 38 Prozent (1,8 Mio. EUR p.a.) durch FlashAttention, Mixed-Precision und Speculative Decoding
- Eigentuemer der ML-RFC-Disziplin mit 84 Architecture Decision Records in 18 Monaten dokumentiert
- Direktes Mentoring von 6 Senior-ML-Engineers, 3 davon zu Tech-Leads befoerdert

Senior ML Engineer / Tech Lead
10/2018 - 06/2022

Mercedes-Benz Tech Innovation — Stuttgart + Berlin, Deutschland

Senior Engineer und Tech-Lead im In-Car-AI-Trainings-Team

- Training und Deployment von Whisper- und Conformer-Modellen auf 4.800 Stunden In-Car-Audio aus 28 Maerkten
- Reduktion von p95-Inferenz-Latenz von 280 ms auf 78 ms durch TensorRT und INT8-Quantisierung
- Aufbau einer DVC- und MLflow-Disziplin fuer 14 Datenversionen pro Produkt-Generation
- Co-Autor bei NeurIPS 2023 zu effizientem Speech-Pre-Training auf domaenenspezifischem Audio

Ausbildung

Dr. rer. nat. Maschinelles Lernen
10/2017 - 09/2021

Eberhard Karls Universitaet Tuebingen + MPI Intelligent Systems
Tuebingen, Deutschland

Maschinelles Lernen — 1,0 (summa cum laude)

M.Sc. Computer Science
10/2014 - 09/2017

ETH Zuerich — Zuerich, Schweiz

Computer Science — GPA: 1,1

Fähigkeiten

- Python (PyTorch, JAX, Triton)
- LLM-Trainings-Stack (DeepSpeed, Megatron-LM, FairScale)
- Ray Train/Serve & Kubeflow
- Triton Inference Server, TensorRT-LLM, ONNX
- Kubernetes, Helm, ArgoCD & Terraform
- AWS SageMaker, Vertex AI, Azure ML
- Vector-DBs (Weaviate, Qdrant, Milvus, Pgvector)
- Mentoring & Tech-Strategy

Zertifikate

AWS Certified Machine Learning Specialty
04/2025

NVIDIA Deep Learning Institute - Distributed Training mit Megatron-LM
11/2023

Heisenberg-Stipendium DFG
06/2022

ICML 2021 + NeurIPS 2023 + CVPR 2024 Co-Autor
03/2021

Projekte

siemens-mlplatform (intern, in Auszuegen Open Source)

06/2024 - heute

Standardisierte ML-Plattform fuer 14 Healthineers-Produktlinien, 240 produktive Modelle, 38 Prozent Compute-Einsparung

CVPR 2024 Submission: Vision-Transformer fuer Radiologie

10/2023 - 03/2024

ViT-Large-Fine-Tuning auf 2,4 Mio. CT-Scans, AUC 0,948 ueber Baseline 0,892, Top-3 Workshop-Award

Publikationen

Vision Transformer Variants for Radiology Workflow Optimization

06/2024

Domain-Adaptive Speech Pre-Training for In-Vehicle Voice Assistants

12/2023

Sprachen

Deutsch	Muttersprache
Englisch	C2
Franzoesisch	B2

Stärken

Plattform-Denken

Baue ML-Plattformen statt nur Modelle, mit dokumentierten Standards, Templates und Self-Service-Workflows

Architektonische Tiefe

Treffe robuste Trade-offs zwischen Modell-Qualitaet, Latenz, Speicher und GPU-Kosten mit RFC-Disziplin

Cross-funktionale Wirkung

Bin sichtbar in Research, Engineering und Produkt-Strategie auf C-Level mit klaren BWL-Argumenten