Magdalena PyTorch-Brammer

Junior ML Engineer

magdalena.brammer@example.de +49 6221 4471 2208 Heidelberg, Deutschland magdalena-brammer.ai

linkedin.com/in/magdalena-brammer github.com/mbrammer

PROFIL

Junior ML Engineer mit M.Sc. Informatik (RWTH Aachen, Note 1,3) und 14 Monaten Werkstudent-Erfahrung bei Aleph Alpha und Mercedes-Benz Tech Innovation. Fokus auf PyTorch, Hugging Face und MLflow mit erster produktiver Erfahrung in deutschsprachigen LLM-Pipelines und Online-Serving via FastAPI auf AWS SageMaker.

BERUFSERFAHRUNG

Junior ML Engineer

10/2025 - heute

Aleph Alpha

Heidelberg, Deutschland

Junior-Engineer im LLM-Serving-Team an deutschsprachiger Inferenz-Infrastruktur

- Entwicklung von vLLM-basierten Inferenz-Endpunkten fuer Luminous-Pharia auf H100-GPUs, p95-Latenz von 1,8 s auf 0,6 s reduziert
- Implementierung eines Token-Streaming-Gateways in FastAPI, das 14 Konzern-Kunden ueber gRPC bedient
- Aufbau von 28 Regression-Tests fuer deutsche Prompt-Suiten in pytest, 4 stille Qualitaetsregressionen frueh erkannt
- Co-Mentor fuer 2 Praktikanten, woechentliche 1:1-Reviews und gemeinsame Hugging-Face-Workshops

Werkstudent ML Engineering

03/2024 - 09/2025

Mercedes-Benz Tech Innovation

Stuttgart, Deutschland

Werkstudent im In-Car-AI-Team an Sprachassistenten-Modellen

- Fine-Tuning von Whisper-Large-v3 auf 2.400 Stunden deutschem In-Car-Audio, WER von 14,8 auf 6,2 Prozent reduziert
- Aufbau einer Hugging-Face-Datasets-Pipeline auf S3 und Snowflake fuer 18 Datenquellen
- Konfiguration von Triton Inference Server fuer ONNX-Modelle, GPU-Auslastung von 38 auf 81 Prozent erhoeht
- Praesentation eines internen Posters auf der MBTI Tech-Konferenz 2025 vor 240 Teilnehmenden

AUSBILDUNG

M.Sc. Informatik

10/2023 - 09/2025

RWTH Aachen

Aachen, Deutschland

Informatik (Schwerpunkt Machine Learning)

1,3

B.Sc. Informatik

10/2020 - 09/2023

RWTH Aachen

Aachen, Deutschland

Informatik

GPA: 1,5

FÄHIGKEITEN

- Python (PyTorch, scikit-learn)
- Hugging Face Transformers & Datasets
- MLflow & Weights & Biases
- FastAPI & Docker
- Kubernetes & Helm
- PostgreSQL & Pgvector
- AWS SageMaker
- Airflow & Prefect

PROJEKTE

ragas-de (Open Source)

01/2026 - heute

Deutschsprachige Evaluations-Suite fuer RAG-Pipelines mit 480 GitHub-Stars und 2.100 PyPI-Downloads pro Monat

Masterarbeit: LoRA-Fine-Tuning fuer deutsche Rechtssprache

10/2024 - 08/2025

Adapter-Tuning auf 320.000 Urteilen aus openJur, BLEU +6,4 ueber Baseline, Note 1,2

ZERTIFIKATE

AWS Certified Machine Learning Specialty

02/2026

TensorFlow Developer Certificate

11/2025

Coursera Deep Learning Specialization (Andrew Ng / Stanford) 06/2025

SPRACHEN

Deutsch	Muttersprache
Englisch	C1

STÄRKEN

Statistische Sorgfalt

Bestehe auf sauberen Confidence Intervals und Power-Analysen, bevor ein Modell als besser deklariert wird

End-to-End-Denken

Sehe ML-Modelle als ein Glied in der Pipeline und arbeite Hand in Hand mit Data-Engineering und SRE

Verstaendliche Kommunikation

Erklaere komplexe ML-Konzepte ohne Jargon an Produkt- und Geschaeftsteams in regelmaessigen Brownbags