Friedrich GenAI-Steinhuber

Senior NLP Engineer

friedrich.steinhuber@example.de
+49 221 4485 2270
Koeln, Deutschland
friedrich-steinhuber.ai
linkedin.com/in/friedrich-steinhuber
github.com/fsteinhuber

PROFIL

NLP / LLM Engineer mit 6 Jahren Erfahrung bei DeepL SE Koeln und Aleph Alpha Heidelberg. Schwerpunkt deutschsprachige LLMs, RAG-Pipelines und Inferenz-Optimierung mit messbarer Wirkung auf p95-Token-Latenz, MMLU-DE-Score und A/B-Konversion. Co-Autor bei ACL 2024 und Maintainer von rag-de-eval mit 2.140 GitHub-Stars.

BERUFSERFAHRUNG

Senior NLP Engineer

DeepL SE

Senior NLP-Engineer im Translation-Quality-Team an domaenenspezifischen LLM-Varianten

10/2022 - heute Koeln, Deutschland

- LoRA-Fine-Tuning von 13B- und 34B-Modellen auf 240 Mio. parallelen Saetzen aus 14 Rechts- und Medizin-Korpora
- Inferenz-Optimierung mit vLLM und Continuous Batching, p95-Token-Latenz von 1,1 s auf 0,28 s reduziert
- A/B-Tests auf 8,4 Mio. monatlichen Nutzern, Konversions-Rate +6,8 Prozentpunkte fuer DeepL Write
- Co-Autorschaft bei ACL 2024 zu domaenenadaptiver maschineller Uebersetzung fuer Fachsprachen

NLP Engineer

Aleph Alpha

NLP-Engineer im Luminous-Trainings-Team

10/2019 - 09/2022 Heidelberg, Deutschland

- Pre-Training und Fine-Tuning von Luminous-Modellen auf deutschem und englischem Common Crawl mit 1,4 Bio. Tokens
- Aufbau einer SFT- und DPO-Pipeline mit Hugging Face TRL fuer 14 interne Modell-Generationen
- Evaluierung mit MMLU-DE, HendrycksTest-DE und Helm-Lite ueber 240 Benchmark-Splits
- Mentoring von 3 NLP-Engineers in Inferenz-Optimierung und Tokenizer-Design

PROJEKTE

rag-de-eval (Open Source)

10/2024 - heute

Deutschsprachige Evaluations-Suite fuer RAG-Systeme, 2.140 GitHub-Stars und 18 Konzern-Adoptionen

DeepL Write LLM-Variant

06/2023 - 12/2023

Inferenz-Optimierung fuer Domain-Specific-Translation, p95-Token-Latenz von 1,1 s auf 0,28 s reduziert

AUSBILDUNG

M.Sc. Computational Linguistics

Universitaet des Saarlandes

10/2017 - 09/2019

Saarbruecken, Deutschland

Computational Linguistics • 1,1

B.Sc. Computerlinguistik

Universitaet Stuttgart

10/2014 - 09/2017

Stuttgart, Deutschland

Informatik + Linguistik • GPA: 1,3

FÄHIGKEITEN

Python (Hugging Face Transformers, PEFT, Datasets)

vLLM, TGI & TensorRT-LLM

LangChain & LlamaIndex

Vector-DBs (Weaviate, Qdrant, Pgvector)

DeepSpeed, Accelerate & LoRA-Finetuning

FastAPI, gRPC & Triton Inference Server

Ragas, DeepEval & Helm-Lite

AWS SageMaker & Vertex AI

ZERTIFIKATE

Hugging Face Certified NLP Engineer

03/2025

NVIDIA Deep Learning Institute - LLM Customization

09/2024

AWS Certified Machine Learning Specialty

06/2023

Coursera Stanford Deep Learning Specialization

11/2022

PUBLIKATIONEN

Domain-Adaptive Machine Translation for German Legal and Medical Corpora

08/2024

SPRACHEN

Deutsch

Englisch

Spanisch

STÄRKEN

Sprach-Sensibilitaet

Verstehe linguistische Feinheiten in deutscher Rechts-, Medizin- und Industrie-Fachsprache und sample Daten gezielt

Evaluations-Disziplin

Baue immer offline + online Evaluations-Pipelines mit Ragas, DeepEval und Helm-Lite vor jedem Modell-Rollout

Inferenz-Effizienz

Reduziere konsequent GPU-Kosten durch Speculative Decoding, Continuous Batching und Quantisierung INT4/INT8