Erfahrungsbericht praktischer Umsetzung von Open Source LLMs für Unternehmen

Michael Tannenbaum

11. September 2023

Portfolio

Business-, Data-, IT-Consulting

ARTIFICIAL INTELLIGENCE & DATA INSIGHTS

IT-INFRASTRUCTURE & INFORMATION SECURITY

CUSTOM SOFTWARE SOLUTIONS & INTEGRATION

Problem Statement

Hohe Einstiegshürden für die Entwicklung von Foundation Models

1. Training von Foundation Models ist finanziell nur für sehr wenige Unternehmen realistisch oder sinnvoll
    1. Hohe Trainingskosten
    2. Hoher Aufwand, um mit Innovationen Schritt zu halten
2. Quasi-Monopol erlaubt OpenAI, den Markt zu diktieren

- Europäische Unternehmen sind abhängig von amerikanischen Konzernen, oder verdammt den GenAI-Hype zu verschlafen – **oder?**

Wie Open Source die GenAI-Revolution für Unternehmen rettet

1. Llama 2 (70B) bewegt sich in vielen Tests **zwischen GPT 3.5 und GPT 4** [1]
2. Open Source treibt **Innovationen** voran, die Training und Inferenz mit bezahlbarer Hardware erlauben (LoRA, Quantization)
3. Open Source LLMs sind **gut anpassbar**, auch als Basis für weitere Use Cases neben Chat
4. Das Fine-Tuning von Open Source LLMs benötigt oft vergleichsweise wenige Trainingsdaten → **günstiges Training**
5. Llama 2 (70B) ist deutlich **günstiger** zu betreiben als GPT 4 (ähnlich GPT 3.5)
6. Unternehmen behalten die **Kontrolle über ihre Daten** – Hosting on-prem oder in der Cloud möglich

1) https://www.promptengineering.org/how-does-llama-2-compare-to-gpt-and-other-ai-language-models/

Unser Setup

Fine Tuning
Sprachen, deutsche/europäische Spezifika, Branchenspezifika, ...

Hochperformante, effiziente Inferenz
ExLlama, GPTQ, …

Chat Interface

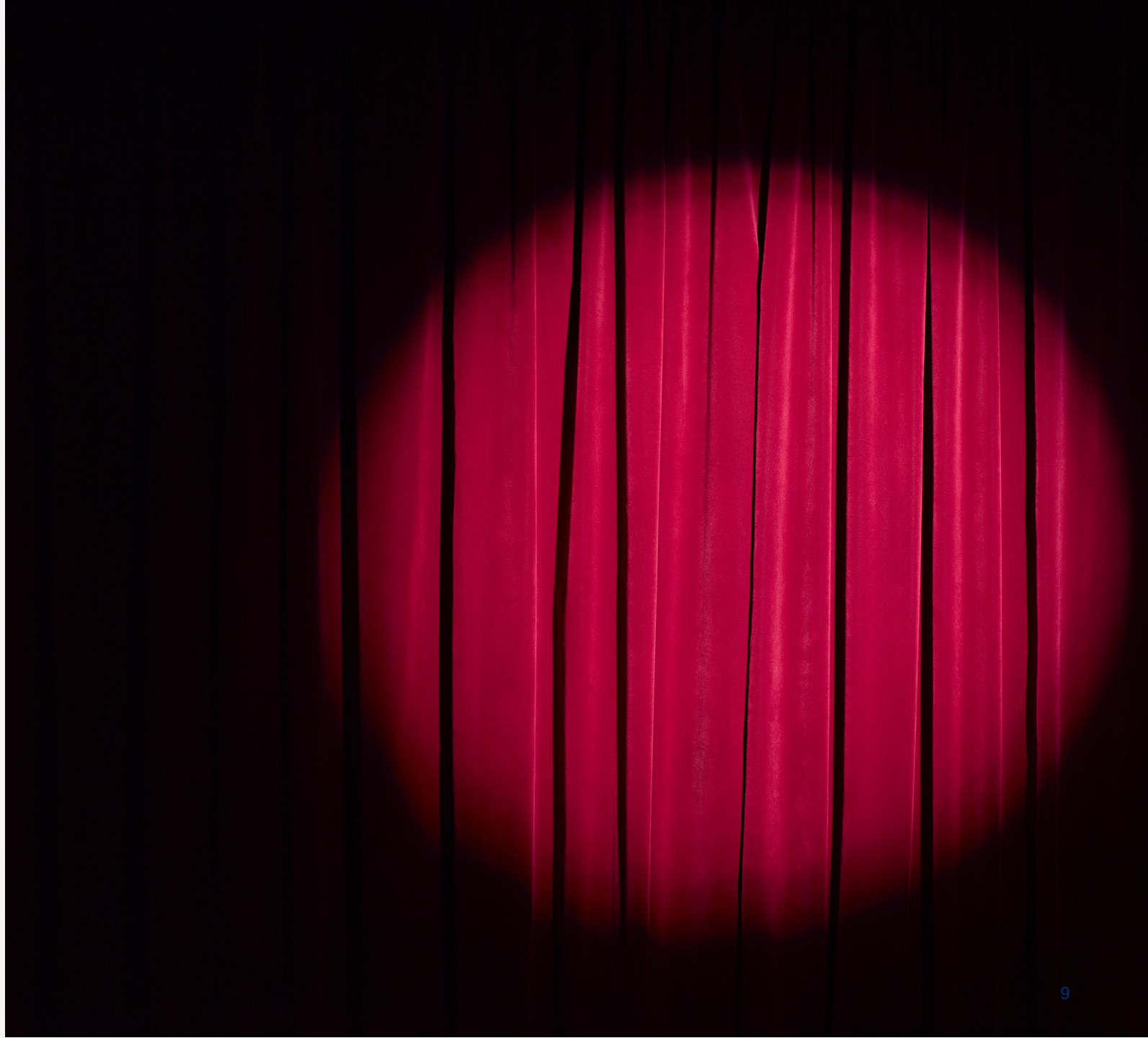
Training von LLMs ist einer der technisch anspruchsvollsten ML-Tasks

- Hohe **Hardwareanforderungen**: Llama-70b benötigt 120GB Grafikspeicher, die größten verfügbaren Grafikkarten besitzen 80GB
- Transformer sind eine der **kompliziertesten Architekturen** für neuronale Netze
- **Quantisierung** ist notwendig, aber **komplex**
- Einzelne Training Runs dauern oft Tage → **Trial & Error ist kein zielführender Arbeitsmodus**
- **Qualität der Trainingsdaten** ist essenziell, aber weder einfach herzustellen noch zu messen
- Komplizierte Evaluation
- Abwägung für die Nutzung interner Daten: **Training oder Embedding?**
- Llama-2 wird schnell outdated sein (s. Falcon-180b). Wie kann das Foundation Model ausgetauscht werden, ohne wieder von null anzufangen?

Hosting von LLMs ist einer der anspruchsvollsten MLOps-Tasks

- Viele Hardware-Optionen
  - Eine/wenige **sehr große GPUs**
  - Viele **kleinere GPUs**
  - Mixed mode **CPU/GPU**
  - **CPU only** per llama.cpp
- Alle Optionen sind **teuer** → Trial & Error ist kein akzeptabler Arbeitsmodus
- Machine Learning und MLOps müssen **Hand in Hand** gehen
  - Größere Hardware vs. Effizienteres Training & Inferenz
  - Schnellere Inferenz vs. Günstigere Hardware

Live Demo

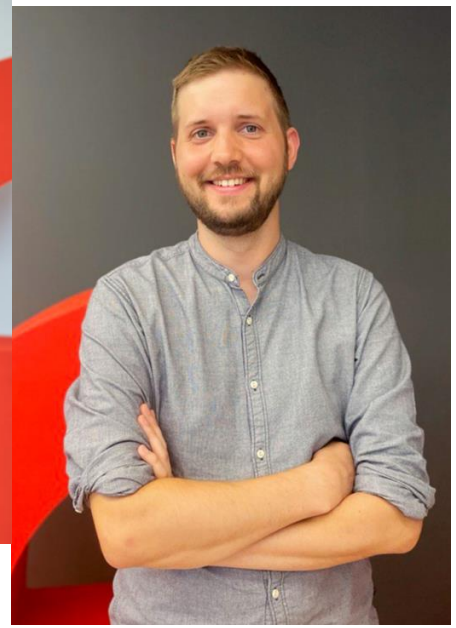


Wir freuen uns auf den weiteren Austausch.

Benjamin Schulte

Mitglied des Vorstands, COO

M +49 160 365 1375
E Benjamin.Schulte@comma-soft.com



Michael Tannenbaum

Lead Generative AI Products

M +49 160 9075 5963
E Michael.Tannenbaum@comma-soft.com