W5 - KW21

Monday, 26 May 2025 18:41

@Tech Team

Deployment & Infrastruktur

- Deployment eines quantisierten Mistral-7B-GGUF-Modells erfolgreich auf bwCloud durchgeführt
- FastAPI-Server zur Inferenz eingerichtet und öffentlich erreichbar gemacht
- Hugging Face Cache auf Volume umgeleitet (60 GB) → Root-Disk entlastet
- Quota-Erweiterung auf bwCloud erhalten → Nutzung größerer Instanzen mit 8 vCPUs, 16 GB RAM und zusätzlichem Volume (60 GB) möglich

Training & Planung

- bwJupyter auf Trainingsfähigkeit explorativ geprüft: QLoRA-Setup mit A100 möglich
- Trainingsstrategie abgestimmt:
  - Training + Quantisierung (GGUF) auf **bwJupyter**
  - Deployment auf **bwCloud**

Modifikation der Matrixstruktur

- ursprüngliche Bewertungs-Matrix wurde erweitert und nach Rücksprache angepasst

@DataTeam

Labeling & Datensätze

- Finalisierung von Analysematrix
- Google Sheets Template für Labeling entworfen
- Datensätze nach Themenschwerpunkt aufgeteilt
  - Investments
  - Health
  - Naturkatastrophe
- Labeling von Datensätzen (Zial: ca. 200 bis EOW)