W8

- Umstieg der Inferenz- und Entwicklungsumgebung von lokal auf RunPod.io, um genügend GPU-Ressourcen für größere Experimente bereitzustellen.
- Erste Versuche mit dem Mergen von Basis-Modellen mit spezialisierten Adaptern.
- Entwicklung einer robusten Merge- und Quantisierungs-Pipeline (HF → GGUF).
- Erweiterung des Trainingsdatensatzes um zusätzliche Beispiele und erste Experimente mit künstlicher Datenvervielfachung.
- Beginn der Implementierung einer serverseitigen FastAPI-Lösung für die Modellinferenz.