W10

- Einführung der Q8-Quantisierungsmethode, um die Modellgröße bei hoher Qualität zu reduzieren.
- Test und Validierung der neuen Quantisierung mit längeren Prompts und realen Beispielen.
- Umbau des Servers: Migrationsplanung von RunPod auf BWCloud, um Kosten zu senken.
- Entwicklung eines automatisierten Build-Workflows mit CMake für die Modelltools.
- Optimierung der Logik-Komponenten, um alle relevanten Felder zuverlässig zu erkennen.