

Projektdefinition ASAG2026

Ausgangslage

Im Rahmen eines realen Hochschulmoduls wurde ein Datensatz mit ca. 31'500 studentischen Kurzantworten erstellt.

Die Antworten wurden im Rahmen eines strukturierten LLM-basierten Bewertungsprozesses generiert und bewertet.

Der Datensatz bildet einen geschlossenen Grading-Loop ab:

- Fragegenerierung
- Studentische Antwort
- LLM-basierte Bewertung (Score + Feedback)

Ziel ist es, diesen Datensatz systematisch aufzubereiten und als Benchmark für Automatic Short Answer Grading (ASAG) nutzbar zu machen.

Ziel der Arbeit

Ziel der Bachelorarbeit ist:

- den Datensatz strukturell zu kuratieren,
- die Bewertungsdynamik im Grading-Loop zu analysieren,
- einen klar definierten Benchmark bereitzustellen,
- sowie eine stabile Version zu veröffentlichen.

Der Fokus liegt nicht auf umfangreicher Datenbereinigung, sondern auf systematischer Strukturierung, Metadaten-Anreicherung und Analyse des Bewertungsprozesses.

Projektansatz (Pipeline)

Die Pipeline ist als iterativer Prozess konzipiert, wobei Erkenntnisse aus der Benchmark-Phase zu strukturellen Anpassungen im Curation-Schritt führen können.

Base Dataset Construction

In einem ersten Schritt werden die relevanten Tabellen aus der bestehenden Datenbank zusammengeführt.

Ziel ist die Erstellung einer Basisversion des Datensatzes (.parquet).

Dataset Curation & Loop Analysis

In dieser Phase wird der Datensatz systematisch strukturiert und mit zusätzlichen Metadaten angereichert.

Anstatt Daten zu entfernen, werden strukturelle Eigenschaften, Modellkonfigurationen und Bewertungsparameter explizit gekennzeichnet (Tagging).

Im Mittelpunkt stehen zwei Split-Strategien:

- Question-Level Split (leakage-sicher)
- Answer-Level Split (zufällig)

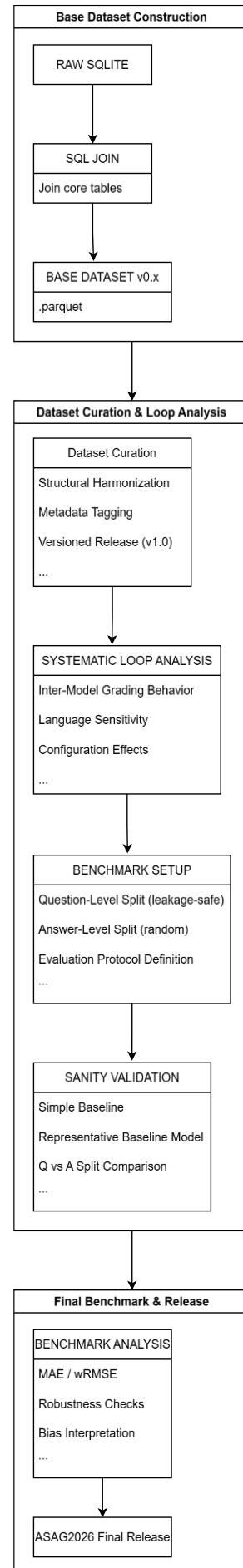
Das Evaluationsprotokoll wird transparent dokumentiert, um Reproduzierbarkeit sicherzustellen.

Zur Überprüfung der Benchmark-Struktur werden einfache Baseline-Modelle trainiert und evaluiert.

Final Benchmark & Release

Abschliessend werden die definierten Metriken angewendet und die Ergebnisse systematisch analysiert.

Die finale Version wird dokumentiert, versioniert und als stabile Grundlage für weitere Forschung bereitgestellt.



Erwartetes Ergebnis

Am Ende der Arbeit sollen folgende Ergebnisse vorliegen:

- Eine strukturierte und dokumentierte Version des Datensatzes
- Ein klar definiertes Benchmark-Protokoll (Question-Level und Answer-Level Split)
- Eine Analyse der Bewertungsdynamik im LLM-basierten Grading-Loop
- Reproduzierbare Baseline-Ergebnisse
- Dokumentation der Versionierung und Metadatenstruktur

Darüber hinaus soll auf Basis dieser Ergebnisse ein wissenschaftliches Paper für den NeurIPS 2026 Datasets & Benchmarks Track erstellt werden.

Das Paper orientiert sich an den offiziellen Anforderungen und Strukturen des NeurIPS Call for Datasets & Benchmarks.

Nach Abschluss der Konferenzversion wird die Bachelorarbeit aus dem finalisierten Paper abgeleitet und in das offizielle ZHAW-Format überführt.