



Aissessor

VU Centre for Teaching & Learning + SOZ Procesregie



Problem

High workload of teachers
Overreliance on multiple-choice tests

Essays and open-ended questions support deeper learning and have higher validity

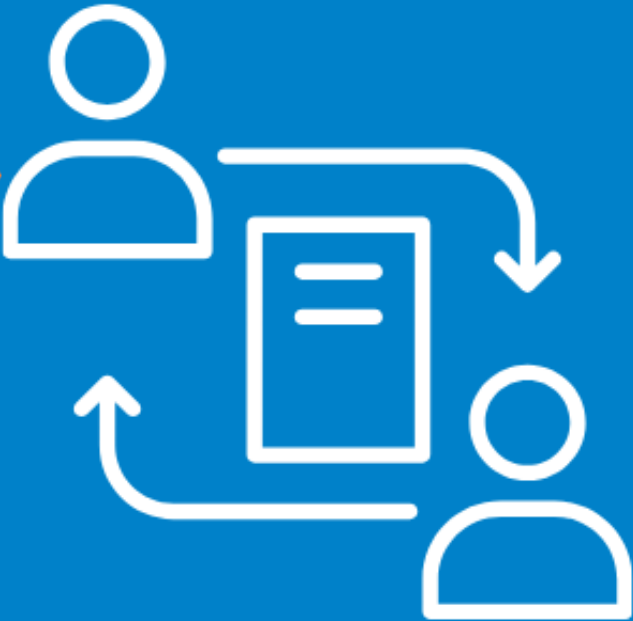
Feedbackloops are lacking

Make the use of essays, open ended exams and feedbackloops more scalable in frequency and for large cohorts!



General solution
Reduce workload with genAI

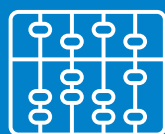
But how in a
responsible way?





High Risk (EU AI Act) →

Core Values



- Prevent injustice
- Keep teacher and students autonomous
- Keep education humane

Guiding Principles

- Be transparent
- Have human oversight
- Assure continuous evaluation and adjustment





Procedure

- 20-25 manual scores
- 5 to 10-runs of scoring with LLM (e.g. GPT-4o, GPT5)
- Select e.g. Mode value
- Inspect plots with correlation of manual vs LLM score

Manual regrading

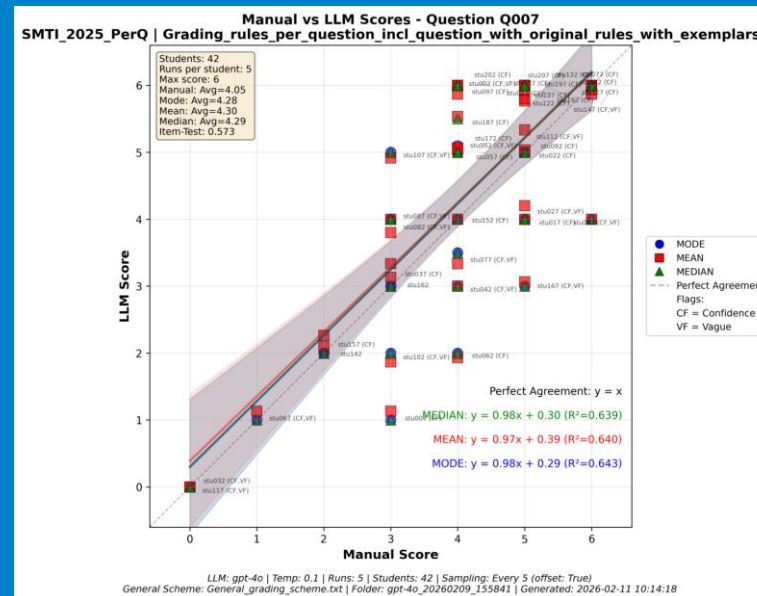
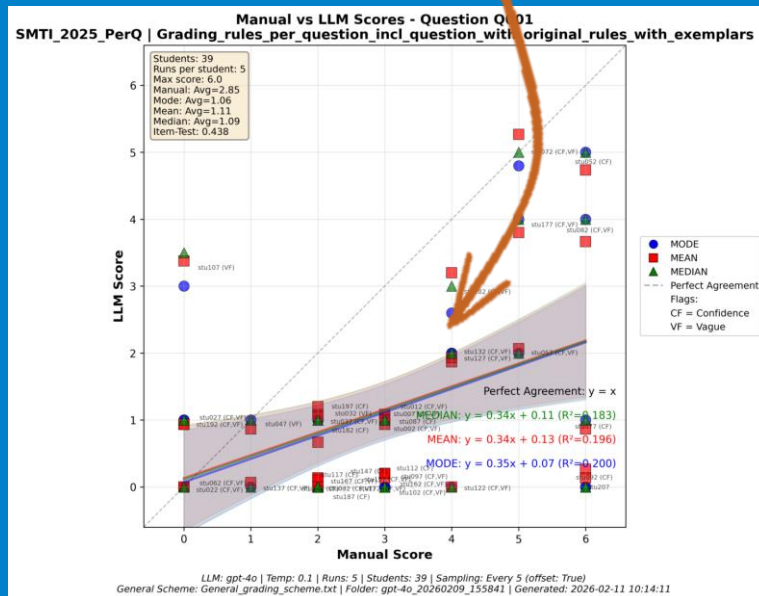
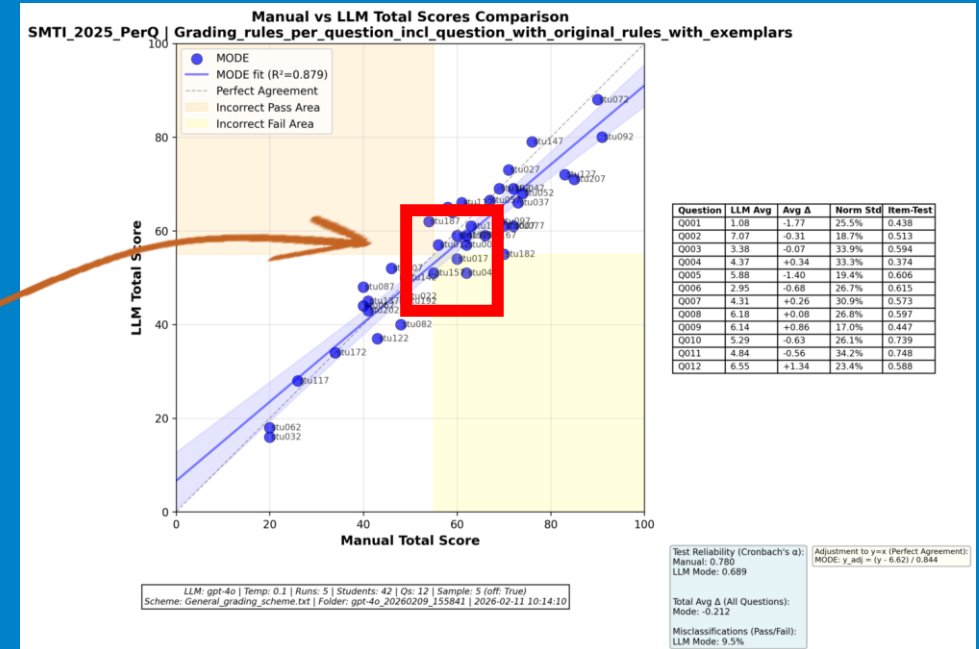
- Questions with odd behavior
- X-percentage of misclassifications around the pass-fail score

Conclusions

- High correlation at test level
- Some Regression to the mean
- Lower correlation at question level

X-percentage of misclassifications around the pass-fail score

Questions with odd behavior



Found 74 cases below 60% confidence

Detailed List Pivot Table (Student × Question)

Rows = Students, Columns = Questions. YES = Needs manual check

Student_ID	Q001	Q002	Q003	Q004	Q005	Q007	Q008	Q009	Q010	Q011	Q012	Total_Fla
stu002	YES	YES					YES		YES			
stu012	YES		YES					YES			YES	
stu022			YES				YES	YES	YES			
stu032								YES				
stu042	YES				YES	YES		YES		YES	YES	
stu052	YES		YES	YES	YES			YES	YES			
stu062								YES				
stu072	YES		YES									
stu082					YES			YES				
stu092		YES	YES		YES			YES			YES	



Follow-up

- If student want inspection → normal human-in-the-loop procedures are already in place (inzage, beroep, etc.)

Conclusions

- High correlation at test level
- Some Regression to the mean
- Lower correlation at question level
- AI-Act
 - Compliance with Core Values
 - Compliance with Guiding Principles



Further research

- Acceptance by students?
- Teacher training?
- Teacher simply clicks through?
- Too many appeals?
- Appeal appetite different for different groups?