AI assisted Feedback

Findings from a small-scale pilot at UIT

February 2026

UNIwise

AI Assisted Feedback in WISEflow at UiT

Findings from a SmallScale Pilot in Immunology (Dental/Medical Education)

Executive summary

A pilot at UiT - The Arctic University of Norway explored whether an AI assisted feedback module in WISEflow could produce qualitative, consistent, and actionable feedback for students, grounded in their written responses and a marking guide/assessment criteria. The pilot was run in an immunology teaching context with six volunteer secondyear dental students (ODO2008) who submitted written answers to five immunology tasks.

Student evaluation data (4/6 respondents) indicates that learners generally found the AI feedback clear, relevant, and able to highlight strengths and areas for improvement, with a strong theme that timing matters: students wanted feedback earlier in the learning cycle (e.g., as a midcourse checkpoint) to maximise usefulness.

From the teacher perspective, **Professor Tor Brynjar Stuge** described the output as “very helpful” and “basically correct” when a single, wellspecified prompt was applied consistently across all six students. He emphasised that the approach is particularly compelling for formative assessment at scale, where workload makes individualised feedback difficult.

Background and rationale

High quality feedback is widely recognised as a powerful driver of learning, but it is also resourceintensive, especially in programmes with large cohorts and frequent assessment points. Seminal research stresses that effective feedback clarifies standards, supports selfregulation, and helps learners close the gap between current and desired performance.

In parallel, interest in generative AI for feedback has grown quickly, with recent reviews suggesting potential to improve timeliness and personalisation while also raising questions about reliability and the continued need for human oversight.

Against this backdrop, UiT participated in a pilot under the initiative “Nye vurderingsformer” (New Assessment Forms), testing AI assisted feedback for written student work in immunology using the WISEflow environment.

Aim of the pilot

The stated goal was to trial AI assisted feedback on written student responses, initially in a formative setting (e.g., subtests during the semester). However, due to pilot orchestration the feedback was not released and provided to students prior to final exam, but after.

The pilot's practical focus aligned closely with a "quality + consistency" hypothesis:

Can an AI assistant generate feedback that is qualitatively useful to students?

Can it do so in a consistent manner across multiple student responses when grounded in the same tasks and marking guidance?

Pilot context and design

Setting, participants and materials

The pilot was carried out in autumn semester 2025 and feedback was generated I November/ December 2025 by module responsible Professor Tor Brynjar Stuge. Key elements included:

- Six volunteer students from ODO2008 (2nd year dentistry) submitting written responses to five immunology tasks.
- The used marking guide (sensorveiledning) and assessment criteria were (in this case) manually uploaded to support alignment between expected standards and the feedback generated.

Workflow in WISEflow

The pilot was conducted in WISEflow containing tasks, criteria/marking guidance, and student submissions, and tested via the AI assisted Feedback module, which interface provided:

1. an automatic "submission summary" of every student, and
2. an interactive "conversation" mode to prompt for targeted feedback on the student assignment, with both standard and user-initiated prompts were available.

A key prompt used to standardise the feedback across students were developed by the assessor (Tor Brynjar Stuge):

> "For each task, explain explicitly what was answered correctly and what was answered wrong, and what was left out that would have made the answer better."

This matters methodologically: rather than treating the AI output as a fixed artefact, the pilot operationalised feedback quality through structured prompting, enabling more comparable outputs across students.

Findings: Student evaluation

Student feedback was collected via a short evaluation in Canvas. By the extended deadline (late January 2026), 4 of 6 students had responded.

Overall usefulness and clarity

Across respondents, perceptions were broadly positive on core clarity dimensions:

- Usefulness: 3 of 4 respondents answered Yes to whether AI feedback was useful, while 1 answered No (explained later).
- Clarity on what was right/wrong: all 4 respondents answered Yes.
- Pointing out strengths and areas for improvement: all 4 respondents answered Yes.

Students' written comments illuminate what "useful" meant in practice. One student highlighted how the feedback helped prioritise study effort, and another described the feedback as relevant and diagnostic:

"The feedback gives an overview of what you 'know enough' and what you should focus on next... helpful for knowing what to prioritise reading, which is often difficult for a student."

"To a large extent relevant and helped me understand what I did wrong or whether my answer was clear enough."

Relevance and level of detail: "helpful" vs "too specific"

A strong theme was that the feedback was relevant – sometimes even overly so. One student explicitly noted that the AI was good at capturing elements of their answer, but "in some cases too much":

"The feedback is good at catching things in the answer, and absolutely relevant... in some cases too much."

Two related concerns emerged:

Underspecified "what to improve"

ven when students appreciated being told that more detail was needed, they sometimes struggled to interpret exactly what the AI wanted

"When it gives feedback that it wants more details within a field, it is difficult to know specifically what it is looking for that the student is missing."

Perceived mismatch with realistic expectations

A student raised a legitimacy concern: AI may detect omissions or details beyond what is reasonable to expect in an exam context:

"AI can identify many more details than is realistic and can be expected of a student in an exam."

This is a classic feedback design tension. The feedback literature stresses that effective feedback should be specific and actionable, but not so complex or expansive that it becomes unusable.

Timing as the decisive factor

Perhaps the clearest practical insight from students was that when feedback is delivered shapes its value. The one respondent who answered “not useful” nevertheless acknowledged relevance, and explained why usefulness was low:

“Relevant but not very useful because these are topics I will not go through again. More useful if feedback came before the exam.”

Another student proposed a specific application pattern:

“Very nice to introduce something like this as a midcourse assessment... so you know what to work on more towards the exam.”

This aligns strongly with wellestablished formative principles: feedback is most powerful when it supports feed forward - guidance on what to do next.

Findings: Teacher/assessor evaluation

A qualitative interview with Tor Brynjar Stuge provides the teacher perspective on output quality, consistency, and feasibility.

Perceived quality and cinsistency across students

Tor Brynjar Stuge reported that the tool “worked very well”, particularly after he used a single prompt to make feedback explicitly address what was correct, incorrect, and missing. He applied the same prompt to all six students and found the output “basically correct”, with only minor issues.

This observation supports the pilot’s central hypothesis about consistent feedback when a standardised prompt is used.

The value of prompting vs default summaries

A nuanced finding was that the default "submission summary" automatically generated of each student submission as a quick deafult overview from the outset, could be too general and, in at least one case, misleading, where the teacher's own notes indicated weak answering but the summary suggested correctness. When prompted for explicit analysis, the system corrected this and recognised the answer was not correct.

In other words, the pilot suggests that how staff engage with the system (structured prompting) can materially affect perceived accuracy.

Feasibility at sscale and the "formative first" logic

The teacher's strongest argument for adoption was practical: formative written feedback is desirable, but workload is the barrier. He described the main benefit as enabling formative feedback on written tests taken during the semester, and stated he would trust the tool in nonsummative settings that support practice and learning.

He also framed a realistic scaling problem: if "everybody's going to want feedback" and a teacher has 100 students, even two minutes per student becomes substantial. This fits a wider pattern in the research: timeliness and teacher workload are key drivers behind interest in automated or AIassisted feedback, but oversight and taskfit remain critical.

Illustrative examples of AI feedbac

The pilot documentation included examples where the AI feedback was structured into: what was correct, errors/omissions, and what would improve the answer. For instance:

A highscoring answer received confirmation of correct conceptual distinctions and suggestions for deepening detail
(e.g., specifying involved Tcell types and time perspective)

A very lowscoring answer received a clear diagnosis of superficiality and missing core mechanisms, plus explicit statements of what would have made the answer better (e.g., describing recognition pathways and cell roles)

Notably, these structures map well to classic feedback models (task/process/next steps), suggesting why students experienced the feedback as clear and relevant even when they debated its level of detail.

Discussion: what the pilot suggests

What worked

Clarity and relevance for students

The evaluation indicates strong agreement that the feedback clarified right/wrong elements and highlighted strengths and improvements.

Teacher confidence in formative use

The teacher reported high satisfaction with output quality and saw strong potential for scaling formative written feedback.

A "feedback literacy" affordance

Students' comments show engagement with how feedback could reshape assessment practices - e.g., supporting more open questions and midcourse checks. This echoes Nicol & MacfarlaneDick's emphasis on feedback that supports selfregulated learning and empowers students to act.

What needs care

Calibration of detail

Students raised a legitimate "expectation mismatch" concern: feedback can become too granular relative to assessment expectations. A design implication is to tune prompts (or product settings) to produce feedback that is sufficiently specific but also bounded, a point also consistent with formative feedback guidance in the literature

Timing and integration into the learning sequence

The strongest student critique was not about correctness but about usefulness when the feedback arrives after students have moved on. This argues for implementing AIassisted feedback primarily as feedforward during teaching periods (quizzes, checkpoints, drafts), not merely as a posthoc artefact.

Human oversight expectations

The teacher's comments reflect a pragmatic continuum: for small cohorts, quick review/editing is feasible; for large cohorts, the ambition shifts toward trustworthy automation—especially in formative contexts. Recent GenAI feedback reviews similarly conclude that human involvement remains important for alignment, accuracy, and appropriate use.

Limitations of the pilot evidence

This was an intentionally small pilot. Key limitations include:

- Small sample size (6 students; 4 evaluation responses).
- Volunteer participation, which may bias responses toward engaged or feedbackmotivated students.
- Outcomes measured were perceptions and usability, not learning gains or performance changes across time.

These are normal constraints for earlyphase pilots; they do not weaken the value of the qualitative insights, but they do bound the claims that can be made.

Conclusion and early success stories

Despite its limited scale, the UiT pilot produced several compelling success signals:

1

Students experienced the feedback as clear, relevant, and improvementoriented, with multiple comments describing how it supports prioritisation and motivation.

2

A studentdriven use case emerged organically: AI assisted feedback as a midcourse formative checkpoint, helping learners target study before exams.

3

The teacher reported that a single, welldesigned prompt produced “basically correct” feedback across all students, suggesting that consistency is achievable when the system is used in a standardised way.

4

The pilot demonstrated credible examples of feedback that distinguish correct elements, omissions, and next steps, which maps onto established feedback theory and likely explains the strong clarity ratings.

Overall, the pilot supports a practical, evidenceinformed conclusion: AIassisted feedback in WISEflow shows promise as a scalable feedback mechanism, provided that (i) feedback is timed to support learning and/or “what to do next”, (ii) prompts and output formats are calibrated to expectations, and (iii) adoption can begins in formative settings where learning benefit is high and risk is manageable.

External references used for this report

- Nicol, D. & MacfarlaneDick, D. (2006). Seven principles of good feedback practice and selfregulated learning. https://pureportal.strath.ac.uk/en/publications/formative-assessment-and-self-regulated-learning-a-model-and-seve/

- Hattie, J. & Timperley, H. (2007). The “feed up / feed back / feed forward” model of feedback. https://www.moedu-sail.org/wp-content/uploads/2016/03/feedback-model.pdf

- Shute, V. (2008). Focus on formative feedback (timely, specific, supportive). https://www.jstor.org/stable/40071124

- Lee, S.S. & Moore, R.L. (2024). Systematic review of GenAI for automated feedback in higher education. https://files.eric.ed.gov/fulltext/EJ1446868.pdf

Acknowledgement

UNIwise would like to thank UiT – The Arctic University of Norway for its participation and cooperation, and especially Professor Tor Brynjar Stuge and his students at ODO2008 (2nd year dentistry) for their structured feedback. Additionally we want to thank Torill Sommerlund for including this pilot into UiT’s project “Nye vurderingsformer” (New Assessment Forms).