AI assisted Feedback

Findings from a second, formative pilot at UiT

July 2026

UNIwise

AI Assisted Feedback in WISEflow at UiT

Executive summary

This report presents findings from a second pilot of AI-assisted feedback in WISEflow at UiT – The Arctic University of Norway, conducted as part of the university's extended project "Nye vurderingsformer" (New Forms of Assessment). Building on a first pilot in autumn 2025, the second pilot deliberately shifted the emphasis to formative assessment: a written sub-module test in immunology (course MED-1501) used as structured preparation ahead of the final exam.

On 6 May 2026, 23 students submitted written answers to a formative immunology test in WISEflow. AI-assisted feedback was generated for every submission and returned to students four days later, together with the questions, their own answers and a short evaluation. Fourteen students completed the structured evaluation and two more submitted general comments.

The central lesson from the first pilot was that students valued the feedback but wanted it earlier, while it could still influence their learning. The second pilot was designed to meet this directly by using the tool in a formative setting and returning feedback during the teaching period, before the exam. On timing, 13 of 14 respondents agreed the feedback arrived at a useful point for learning, and 11 of 14 found it relevant to a large extent. Students were most positive about the concrete, forward-looking improvement suggestions.

From the assessor's perspective, Professor Tor Brynjar Stuge reported high satisfaction across quality, consistency and workflow. Using a single, refined formative prompt across all submissions, he rated the AI output as accurate, aligned with the marking guide, specific, actionable and appropriately toned, with hallucinations rare. Importantly, Tor Brynjar Stuge read and screened every AI-generated response before it was passed on to students, so no feedback reached a student without his review. He estimated roughly 5 minutes per submission with AI support against a 20-minute baseline, and would recommend continuing AI-assisted feedback in other courses and programmes at UiT.

Background and rationale

High-quality feedback is one of the most powerful drivers of learning, but it is resource-intensive - particularly in programmes with large cohorts and frequent assessment points. The feedback literature stresses that effective feedback clarifies standards, supports self-regulation, and helps learners close the gap between current and desired performance. It is also clear that timing matters: feedback is most useful when it can still shape what a learner does next.

The first UiT pilot (autumn 2025) tested whether an AI-assisted feedback module in WISEflow could produce qualitative, consistent and actionable feedback grounded in students' written responses and a marking guide. Students found the feedback clear, relevant and improvement-oriented, and a teacher trial confirmed that a single, well-specified prompt produced "basically correct" output across submissions. The clearest critique from that pilot was not about correctness but about timing: because of how the pilot was orchestrated, feedback reached students after the exam rather than before it. Students explicitly wanted the same feedback earlier, for example as a mid-course checkpoint.

The second pilot was designed to respond to that critique. Rather than testing AI feedback on work students had already finished with, it embedded the tool in a genuinely formative activity: a written sub-module test whose purpose is to help students identify what to study before the final exam.

Aim of the second pilot

The aim was to trial AI-assisted feedback in a formative setting, where its learning value is highest and the stakes are low, and to see whether the timing weakness identified in the first pilot could be resolved in practice. In addition to the original "quality and consistency" questions, the second pilot asked:

- Can AI-assisted feedback be delivered early enough in the learning cycle to genuinely support students' preparation for the exam (feed-forward)?
- Does a refined, explicitly formative prompt – including concrete next-step suggestions and an instruction to bound the level of detail – improve on the first pilot's output?
- How do students and Tor Brynjar Stuge experience AI-generated feedback when it is used as a low-stakes practice mechanism rather than a post-exam artefact?

Pilot context and design

Setting, participants and materials

The pilot was carried out in the immunology sub-module of MED-1501 (medicine) at UiT. A written formative sub-module test in immunology was held on 6 May 2026. Key elements included:

- 23 students submitted written answers to the immunology test in WISEflow on 6 May 2026.
- AI-assisted feedback was generated for every submission using the AI assistant in WISEflow.
- The marking guide (sensorveiledning) and assessment criteria were used as the reference standard for the feedback.
- Four days after the test, students received the AI-generated feedback together with the tasks, their own answers and a set of 10 evaluation questions.

This design is formative by construction: the test does not count towards the grade, and the feedback is delivered during the teaching period so that students can act on it before the exam.

Workflow and the formative prompt

As in the first pilot, feedback quality was operationalised through structured prompting rather than treating the raw AI output as fixed. For this pilot Tor Brynjar Stuge refined the prompt into an explicitly formative instruction, applied consistently across all submissions, translated here from Norwegian:

> "For each task in the student's response: (1) give feedback on what is correct in the answer; (2) identify what is wrong or misunderstood, and explain why it is wrong; (3) point out what is missing, and how this would have improved the quality; (4) give concrete suggestions for how the student can improve their understanding or answer in future tasks. The feedback should be formative and focus on helping the student to learn and improve. Limit the level of detail in the feedback to what is specified in the marking guide and the assessment criteria."

Two features of this prompt address specific lessons from the first pilot. Point (4) makes next-step, feed-forward guidance an explicit requirement rather than a by-product. The final instruction – to bound the level of detail to the marking guide and criteria – directly targets the first pilot's finding that feedback could become "too specific" relative to what is realistic in an exam.

Findings: Student evaluation

Fourteen students answered the structured evaluation questions and two more submitted a general comment instead of answering directly. All but one response arrived within two weeks of the test; the last arrived after three weeks. The quantitative results below are based on the 14 structured responses.

Relative to the 23 students who took the test, 14 completed the structured questions and a further two responded with free-text comments – a response rate of roughly 61% for the structured evaluation, and about 70% for the evaluation overall. For a voluntary evaluation of a low-stakes formative test, returned four days afterwards, this is a healthy level of engagement that lends reasonable weight to the patterns below. It remains a self-selected sample, so the figures are best read as indicative rather than representative (see Limitations).

Evaluation question	Response distribution	Positive share
Overall, how useful was the feedback?	Very useful 6 · Somewhat useful 7 · Of little use 1	13 / 14
Was it clear what was right and wrong?	Yes 12 · No 2	12 / 14
How relevant was the feedback to your answer?	To a large extent 11 · Somewhat 3 · To a small extent 0	11 / 14
Did it increase motivation for further work?	To a large extent 6 · Somewhat 7 · To a small extent 1	13 / 14
Did it help you understand what to improve next time?	To a large extent 10 · Somewhat 4 · To a small extent 0	14 / 14
Was the feedback given at a useful time for learning?	Yes 13 · No 1	13 / 14

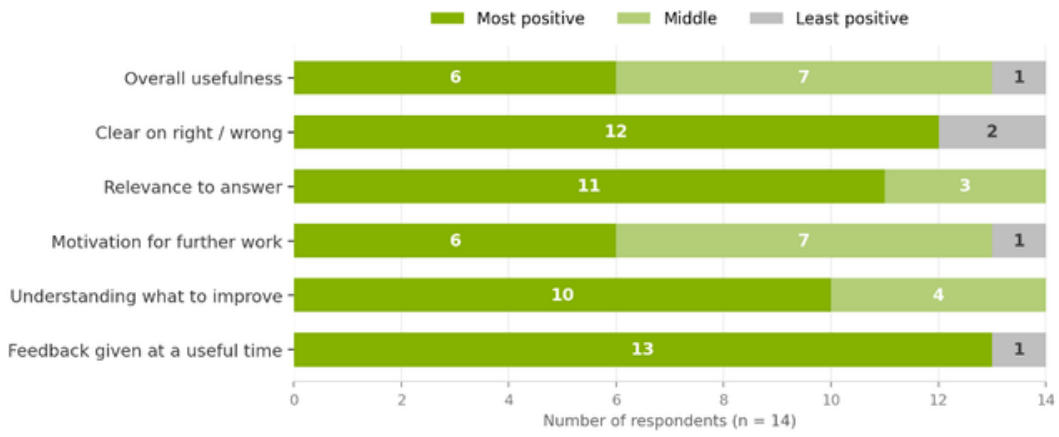

Figure 1. Student evaluation responses (n = 14). Each bar shows how many respondents chose the most positive, middle, or least positive option for that question. Questions with a Yes/No scale (clarity on right/wrong; timing) have no middle segment.

Usefulness, clarity and relevance

Perceptions were broadly positive across the core dimensions. Thirteen of fourteen respondents found the feedback somewhat or very useful, all but two found it clear on what was right and wrong, and eleven of fourteen judged it relevant to a large extent. Every respondent said the feedback helped them understand what to improve next time (ten to a large extent, four somewhat).

When asked what was most helpful, students consistently pointed to the concrete, forward-looking suggestions and the explicit account of what was missing:

> The improvement suggestions! Especially that it comes with specific proposals such as "make a systematic list of..." and "use acronyms or memory techniques to remember."

> That it doesn't just say what is right and wrong, but also what was good, what is missing, and a new improvement suggestion. This gives the student easier insight into what needs to be corrected to achieve a better result.

> The most helpful part was the concrete suggestions about what I should learn more about and gain a deeper understanding of.

Timing: the critique from the first pilot, adressed

The most important change in this pilot was timing, and it is where the student response is clearest. Thirteen of fourteen respondents agreed the feedback was delivered at a useful time for learning. Where students elaborated, several still asked for even earlier and more frequent checkpoints – a request to do more of what the pilot had started, rather than a complaint that it came too late:

> Yes, but it would be good to have assessments earlier in the sub-module as well, to improve understanding along the way.

> Yes. I hadn't read enough on the topic [and this showed me where to focus].

This is a marked contrast with the first pilot, where the single respondent who found the feedback "not useful" attributed this precisely to its late arrival. Delivering feedback within the teaching period appears to have converted the same qualities that students already valued into feedback they could actually act on.

What students would improve

Students' suggestions were constructive and clustered around calibration, accuracy and trust. On calibration, several wanted a clearer sense of how much a gap mattered:

> The assessment of what is necessary for a correct answer should be more specific.

> It is a little difficult to know, when it lists both good and weak points per task, whether what was missing was very important or just a suggestion for further detail. A score per task might help.

On accuracy and syllabus alignment, some students reported what they perceived as occasional terminology or transcription issues – for example the tool not flagging a careless spelling slip in one place while, as they saw it, misreading a term in another, or questioning a term used in lectures. Several therefore asked that the feedback be checked against the syllabus and, ideally, reviewed by a person. As set out in the teacher evaluation below, Tor Brynjar Stuge screened every AI output before release and, in a follow-up interview, judged the AI to have been correct in the specific cases students raised:

> That a person goes through the feedback and double-checks that what the AI has written is correct and relevant.

> Make sure the feedback matches the syllabus and what we have learned.

Student views on AI-generated feedback

Attitudes to receiving AI-generated rather than human feedback were mostly pragmatic and positive, provided a human remains in the loop and the process is transparent. Several students valued the consistency and neutrality of the tool:

> I think AI can give more neutral and professional feedback and keep it fairly equal for all students – without the influence of being human: tired, hungry, a bad day, a good day.

> I have no problem with the feedback being AI-generated, as long as someone reviews it to check before it is sent out.

Others noted the limits of automated feedback – that it can feel impersonal, that a graded exam would still warrant human assessment, and that trust needs to be built through transparency about how the feedback is produced. One student captured the balance well, valuing tailored, frequent feedback that does not consume large amounts of teacher time, while cautioning against losing "the feeling of being seen," especially in large cohorts. Taken together, the responses support using the tool for formative, low-stakes feedback with human oversight, rather than as an unsupervised substitute for the teacher.

Findings: Tor Brynjar Stuge's evaluation as teacher and assessor

Professor Tor Brynjar Stuge completed a structured lecturer survey covering workflow, quality, criteria alignment, consistency, overall impact and support. He applied a single formative prompt across all 23 submissions and tested the tool for feedback rather than for grading. His ratings were strongly positive throughout, as summarised below.

Dimension	Statement	Rating
Workflow	Setup and onboarding were easy	Strongly agree
Workflow	Integrated well with my grading and feedback process	Strongly agree
Workflow	I had sufficient control to edit/curate outputs before releasing	Strongly agree
Workflow	The system was reliable (few errors, stable)	Agree
Workflow	The AI assistant was easy to use	Strongly agree
Quality	Output was accurate about student work	Strongly agree
Quality	Output aligned with the marking guide/criteria	Strongly agree
Quality	Produced specific, evidence-based feedback	Strongly agree
Quality	Produced actionable feedback	Strongly agree
Quality	Tone was appropriate and professional	Strongly agree
Quality	Hallucinations or misleading statements were rare	Strongly agree
Quality	Confident releasing feedback after my review	Strongly agree
Consistency	The AI improved feedback consistency across submissions	Strongly agree
Consistency	AI helped me avoid bias and remain objective	Strongly agree
Consistency	The amount of feedback per student was appropriate	Agree
Overall	Satisfied with the AI-assisted feedback	Strongly agree
Overall	Would use AI-assisted feedback again for similar assessments	Strongly agree
Overall	Would recommend AI-assisted feedback to colleagues	Strongly agree

Quality, consistency and criteria alignment

Tor Brynjar Stuge rated the output as accurate, aligned with the marking guide, specific, evidence-based and actionable, with an appropriate tone and rare hallucinations. He judged that the tool helped him assess students' achievement of the intended learning outcomes, captured key aspects of performance and supported his professional judgement – each "to a large extent." He also strongly agreed that the AI improved consistency across submissions and helped him avoid bias and remain objective, which is the central claim the pilots set out to test.

Human screening and the accuracy questions raised by students

A point Tor Brynjar Stuge stressed in a follow-up interview with UiT is that human oversight was built into the pilot from the outset: he read and screened every AI-generated response before it was passed on to students. No feedback was released without his review, and where necessary he was in a position to correct or withhold it.

This is directly relevant to the accuracy concerns voiced by a small number of students, who felt the AI had misread a term or been too rigid about terminology. Tor Brynjar Stuge did not accept that framing. On his expert review of the specific cases students raised, the AI's feedback was correct and it was the student's answer, not the AI, that was mistaken – in other words, what a student experienced as an AI error was usually a valid correction they had not yet recognised. This does not remove the value of calibrating detail or of grounding feedback in the marking guide, but it does reposition the "accuracy" theme: in this pilot the combination of a bounded prompt and consistent human screening kept genuine AI error rare, by Tor Brynjar Stuge's assessment.

Time and effort

Tor Brynjar Stuge estimated about 5 minutes per submission when using AI-assisted feedback, against a baseline of around 20 minutes without it. He noted explicitly that he tested the AI for feedback, not for grading, so the figures describe the effort of producing quality formative feedback – including his screening of every output – rather than assigning marks. Even with that caveat, the direction is clear: AI support makes individualised formative feedback markedly more feasible at scale, which was the main practical barrier identified in the first pilot.

Overall assessment and future use

Tor Brynjar Stuge was satisfied overall, would use AI-assisted feedback again for similar assessments, and would recommend it to colleagues. Asked what he had learned and what worked well, he answered that he "got new ideas for better prompts"; asked what he would do differently, he would "allocate more time to test and adjust a system prompt." His single most important request for future use was telling for the formative direction of this pilot: automatic feedback on students' responses to informal tests. Asked whether AI-assisted assessment should continue in other courses and programmes at UiT, his answer was an unqualified "Yes." He also strongly agreed that the dialogue with, and support from, the cross-disciplinary support team was satisfactory and useful.

Illustrative examples of AI feedback

Students' own accounts illustrate how the structured prompt translated into feedback that was diagnostic and actionable. In one detailed comment, a student described how the feedback pinpointed a genuine conceptual error rather than a surface issue:

> The absolute highlight was how it uncovered critical misunderstandings in my grasp of signalling pathways. It made very clear that I had confused the detection mechanism for viruses with the subsequent response. It was also precise in correcting terminology errors – for example that antibodies neutralise by binding to antigens, not receptors.
>
> Overall a very instructive evaluation that gives a concrete pointer to what the exam requires.

The same student also felt the tool was too rigid in one place, dismissing what they saw as a partially valid point. On Tor Brynjar Stuge's review, however, the AI's correction in cases like this was sound – a reminder that a confident, well-grounded correction can feel "rigid" to a student who has not yet accepted it, and that the human screening step is what confirms the call either way. These examples map onto established feedback models (what was correct, what was wrong or missing, and what to do next) and help explain why students rated the feedback as clear and relevant even where they debated its level of detail.

Discussion: what the pilot suggests

What worked

Timing solved in practice
Delivering feedback during the teaching period turned the first pilot's main weakness into a strength: 13 of 14 students agreed the feedback arrived at a useful time, and several asked for more of it, earlier.

Assessor connfidence and consistency
Tor Brynjar Stuge strongly agreed the output was accurate, aligned to the criteria and consistent across submissions, and screened every response before releasing it – with an estimated 5 versus 20 minutes per submission.

Feed-forward that students act on
The concrete next-step suggestions were the most-praised element, and every respondent said the feedback helped them understand what to improve. This is formative feedback working as intended.

A credible path to scale
A single refined prompt produced strong output across 23 submissions, supporting the case that quality formative feedback can be produced consistently and efficiently.

What needs care

Calibration of detail and importance
Some students found it hard to tell whether a noted gap was critical or merely a refinement. Prompts (or product settings) should keep feedback bounded, and could signal the relative importance of each point; a light-touch per-task indicator was requested by several students.

Human oversight and trust
Students accept AI-generated feedback most readily when a human checks it and the process is transparent. For low-stakes formative use this is straightforward; any move towards higher-stakes use would raise the bar for oversight and communication.

Perceived accuracy and the role of human screening
A few students perceived occasional terminology rigidity or misreadings; on review, Tor Brynjar Stuge judged the AI to be correct in the cases raised and the student mistaken. This is precisely the argument for the screening he applied – every output was checked before release – together with grounding the feedback firmly in the marking guide and course materials.

How the second pilot compares with the first

The two pilots are complementary. The first established that a single, well-specified prompt could produce clear, relevant and broadly correct feedback, but delivered it after the exam, which limited its usefulness. The second kept those quality gains, added an explicitly formative prompt with next-step guidance and a detail-bounding instruction, and – crucially – delivered the feedback while students could still act on it. The result is that the dimension students were most critical of in the first pilot, timing, is the dimension they are now most positive about. Tor Brynjar Stuge's strongest forward-looking request – automatic feedback on informal tests – points to the natural next step: making this formative use routine and even earlier in the module.

Limitations of the pilot evidence

This remains an early-phase pilot and its claims should be read accordingly:

1. Small, single-course sample (23 submissions; 14 structured evaluation responses plus 2 general comments)
2. Self-selected engagement and self-reported perceptions, which may skew towards motivated students
3. Outcomes measured were perceptions, usability and workflow – not learning gains or exam performance
4. Tor Brynjar Stuge's evaluation reflects a single, experienced assessor using a carefully refined prompt, who also screened every AI output before release

These are normal constraints for a pilot of this kind. They bound the strength of the conclusions but do not diminish the value of the qualitative and workflow insights, which are consistent across students and assessor and align with the wider feedback literature.

Conclusion and early success stories

Within its limited scale, the second pilot produced several clear success signals:

- The timing weakness from the first pilot was resolved: 13 of 14 students agreed the feedback arrived at a useful point for learning, and every respondent said it helped them understand what to improve.
- Students valued the feedback most for its concrete, forward-looking suggestions – exactly the feed-forward the refined prompt was designed to produce.
- Tor Brynjar Stuge reported strong, consistent quality across 23 submissions from a single prompt, screened every AI output before release, and had high confidence in the feedback – with an estimated four-fold reduction in time per submission.
- Both students and Tor Brynjar Stuge endorsed continued, expanded use in a formative, human-supervised setting, and Tor Brynjar Stuge recommended extending AI-assisted assessment to other courses and programmes at UiT.

Overall, the second pilot strengthens the practical, evidence-informed case from the first: AI-assisted feedback in WISEflow shows real promise as a scalable feedback mechanism when feedback is timed to support learning, prompts and output are calibrated to expectations, and adoption begins in formative settings where the learning benefit is high and the risk is manageable. On the evidence of this pilot, that formative promise is now being realised.

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 continued participation and cooperation, and especially Professor Tor Brynjar Stuge and his students in MED-1501 for their thoughtful and structured feedback. We also thank the project team behind “Nye vurderingsformer” (New Forms of Assessment) for including this second pilot in the extended project.