# AI assisted Feedback

## Findings from a pilot in two master's courses at UiT

June - August 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). The pilot ran across two master's-level courses in supply chain and logistics: INE-3608 Manufacturing Logistics and INE- 3609 Supply Chain Management, coordinated jointly by Associate Professor Xu Sun and Professor Hao Yu (Professor of Industrial Engineering at UiT), who worked together throughout the pilot and evaluated it jointly.

Manufacturing Logistics had 7 students enrolled and Supply Chain Management had 5. AI-assisted feedback was generated in WISEflow for six Manufacturing Logistics submissions and four Supply Chain Management submissions. Two students – both from Supply Chain Management – completed the student evaluation survey; no Manufacturing Logistics students responded. Xu Sun and Hao Yu completed a joint, structured course-coordinator survey covering workflow, quality, consistency and future use.

The two students who did respond gave a positive but mixed picture. Both rated the feedback as clear, specific and relevant, and both agreed it would help them improve future work; one found the tone too soft and would have preferred a more "sterile", objective style even with a human reviewer in the loop, while the other actively preferred the combination of AI and human feedback over human-only feedback. Xu Sun and Hao Yu's own evaluation reflected the real difficulties they encountered during testing (discussed below): they disagreed that the AI's comments were reliably accurate and disagreed that hallucinations were rare in first-pass output. At the same time, they agreed the AI reduced grading time (roughly 30 versus 45 minutes per submission), strongly agreed they would use it again, and would recommend it to colleagues, provided rollout is gradual and paired with training.

Early in the pilot, Xu Sun and Hao Yu each ran into hallucinations in the AI's first-pass output – invented problems, missed real issues, and misread calculations. Based on their observations and testing data, UNIwise carried out a dedicated root cause analysis.

As summarised later in this report, these issues surfaced during testing and iterative prompting, not in the feedback ultimately released to students: both evaluators read and screened every AI-generated response before passing it on, corrected or discarded unreliable output, and released feedback only once they were confident in it.

The final feedback students received was therefore not affected by the hallucinations encountered along the way. Several of the findings from that analysis are already resolved or in progress in the new version of the WISEflow AI Assistant, due for release in October 2026, with the remaining recommendations under active consideration for future updates, as outlined at the end of that section.

## 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.

At master's level this challenge is sharper still. Feedback on specialised, technical coursework requires deep subject-matter expertise to give well, takes considerably longer to produce in depth than feedback on introductory material, and is inherently harder to make genuinely learning-oriented for the student – it is not enough to say what is right or wrong; the feedback has to meet the student at an advanced conceptual level and still point clearly to what to do next.

This pilot forms part of UiT's wider "Nye vurderingsformer" initiative, which is testing AI-assisted feedback across a range of course types and formats. This particular pilot focused on two small, master's-level courses with technical, quantitative and model-based assignments – inventory and lot-sizing calculations in Manufacturing Logistics, and mixed-integer network-design and optimisation modelling in Supply Chain Management – evaluated jointly by two coordinators working together. This combination of small cohorts, highly technical content and two evaluators comparing notes is well suited to understanding how AI-assisted feedback performs on structured, single-right-answer technical tasks rather than open, discursive answers.

## Aim of the pilot

The aim was to trial AI-assisted feedback on master's-level, technical coursework in two related supply-chain courses, with two evaluators working independently but comparing their experience directly. Alongside the established "quality and consistency" questions from earlier UiT pilots, this pilot specifically asked:

- How does AI-assisted feedback perform on closed-ended, quantitative and model-based assignments, where there is a single correct numeric answer or a specific modelling requirement to check, rather than open-ended discussion?
- Do students accept AI-assisted feedback on technical coursework, and what, if anything, do they want changed about its tone and style?

## Pilot context and design

### Setting, participants and materials

The pilot covered two master's courses at UiT:

INE-3608 Manufacturing Logistics (7 students enrolled), coordinated by Associate Professor Xu Sun, using an assignment covering EOQ assumptions, quantity-discount decisions, POQ and MRP calculations.

INE-3609 Supply Chain Management (5 students enrolled), coordinated by Professor Hao Yu, using a facility-location assignment – a mixed-integer network-design and optimisation task marked on model formulation, data-file consistency and completeness.

AI-assisted feedback was generated in WISEflow for six of the Manufacturing Logistics submissions and four of the Supply Chain Management submissions, using the marking guide and assessment criteria for each course as the reference standard. Xu Sun and Hao Yu worked together throughout, comparing prompts and findings across the two courses, and completed a joint course-coordinator survey on the pilot. Two students – both from Supply Chain Management – subsequently completed the student evaluation survey.

**Workflow and prompting**

As in the earlier UiT pilot, feedback quality depended heavily on how the AI assistant was prompted rather than on the raw output being treated as fixed. Both evaluators iterated on their prompts: Xu Sun found that a vague, generic prompt produced unreliable output, while a structured prompt – stating the total number of questions, assessing correctness and completeness question-by-question, then summarising – produced markedly more complete and reliable feedback. Hao Yu likewise moved from a generic first attempt to specific, targeted prompts once he had identified a gap, for example asking directly: "For Problem A, are there any redundancy issues in the model file?" Both evaluators independently concluded that more structured, purpose-built prompting is what would most improve results, and both reported needing several rounds of interaction before a submission's feedback became reliable.

## Findings: Student evaluation

Only two students responded to the evaluation survey, both from Supply Chain Management; no Manufacturing Logistics students responded. With so few responses the results are illustrative rather than representative, but both students engaged substantively with the questions and their individual ratings are set out in full below.

| Dimension | Statement | Student 1 | Student 2 |
|---|---|---|---|
| Quality & relevance | The feedback was clear and easy to understand | Strongly agree | Agree |
| | The feedback was specific (concrete examples, quotes) | Strongly agree | Strongly agree |
| | The feedback was actionable (I knew what to do next) | Agree | Agree |
| | The feedback was relevant to my work | Agree | Agree |
| | The feedback was accurate about my work | Strongly agree | Agree |
| | The feedback aligned with the assessment criteria | Agree | Agree |
| | The tone was supportive and professional | Agree | Strongly agree |
| | The amount of feedback was appropriate | Strongly agree | Agree |
| | The feedback will help me improve future work | Agree | Agree |
| Fairness & consistency | The feedback felt fair | Agree | Agree |
| | The feedback felt objective (not subjective or biased) | Disagree | Strongly agree |
| | Consistent feedback across all students is important | Agree | Agree |
| | AI assistance improves consistency across students | Agree | Strongly agree |

| | | | |
|---|---|---|---|
| Trust & transparency | Comfortable receiving AI-assisted feedback | Agree | Strongly agree |
| | Clear how AI assistance was used in the process | Agree | Strongly agree |
| | Trust AI-assisted feedback when a human has reviewed it | N/A | Strongly agree |
| | Prefer AI-assisted + human feedback over human-only | Disagree | Strongly agree |
| Quantity & timeliness | The amount of feedback was appropriate | Agree | Agree |
| | The timeliness of feedback was satisfactory | Agree | Agree |
| | Can learn and act on the feedback received | Strongly agree | Strongly agree |
| Overall & future use | Overall satisfied with the AI-assisted feedback | Agree | Agree |
| | Would like to receive AI-assisted feedback again | Agree | Agree |
| | Would recommend AI-assisted feedback to other students | Agree | Agree |

Colour indicates strength of response, from strongly agree (green) through to strongly disagree (red); grey marks a not-applicable answer.

Both students rated the feedback highly on the core quality dimensions: both found it clear, specific and relevant to their work, both said it aligned with the assessment criteria, and both agreed it would help them improve future work. Asked what was most useful, one wrote:

> The most useful part of the AI-assisted feedback was its specific and objective comments. It clearly identified strengths and areas for improvement, provided actionable suggestions, and helped me understand how my work aligned with the assessment criteria.

The other singled out the same strength, more briefly:

> Very specific examples of what to improve.

**A split on tone and trust**

The two students diverged sharply on how they felt about receiving feedback generated by an AI. One was broadly enthusiastic about the combination of AI and human review:

> I was comfortable receiving AI-assisted feedback and felt that the role of AI in the feedback process was clearly explained. Knowing that the feedback had been reviewed by a human increased my confidence in its reliability and appropriateness. [...] Overall, I would prefer a combination of AI-assisted and human feedback rather than relying solely on human-generated feedback, as this approach has the potential to improve efficiency while retaining academic judgment and quality assurance.

The other was more sceptical, even knowing a human had reviewed the feedback, and specifically wanted a more neutral tone:

> Even with human supervision I would still prefer human-only feedback.

> When prompted with «it should be supportive» I feel like the AI might be a bit too nice, and less objective. A more robot-like answer is almost more preferable

This split matters more than the small sample size might suggest: it suggests that students are broadly comfortable with AI-assisted feedback provided a human remains in the loop and the process is transparent, but a minority will still prefer a human-only alternative regardless of review, and some experience an AI's attempt at a warm, encouraging tone as inauthentic rather than reassuring.

Both students agreed that consistency across students is important and that AI assistance can help achieve it; the disagreement was specifically about tone, not about accuracy or fairness.

**What students would improve**

Both students' suggestions were about calibration rather than substance. One wanted more personalised detail:

> I would like future versions to provide more personalized feedback, including more detailed examples and clearer suggestions tailored to my specific work. This would make the feedback even more actionable and useful for improvement.

The other wanted a plainer, more clinical register:

> A more «sterile» feedback if that makes sense, the AI trying to be nice and encouraging is a bit off-putting. I appreciate encouraging feedback from people, but it feels a bit off-putting from a machine, where you expect a purely objective and logical feedback.

## Findings: course coordinators' evaluation (Xu Sun and Hao Yu)

Xu Sun and Hao Yu completed a joint, structured course-coordinator survey covering workflow, quality, consistency and future use, drawing on their combined experience across the two courses. Their ratings were mixed, directly reflecting the hallucination issues discussed in the next section: strongly positive on workflow, tone and their intention to keep using the tool, but critical of accuracy and hallucination frequency in first-pass output.

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

Colour indicates strength of response, from strongly agree (green) through to strongly disagree (red).

The pattern is telling: Xu Sun and Hao Yu disagreed that the AI's comments were accurate and strongly disagreed that hallucinations were rare – a direct, first-hand echo of the issues later examined in UNIwise's root cause analysis – yet still agreed the tool integrated well, produced an appropriate tone and actionable feedback once curated, and strongly agreed they would use it again and recommend it. In their own words:

The AI-assistant is well integrated into the system, but the answers are not reliable in the initial interaction. In the initial feedback, there are many mistakes, which need several rounds of interaction to provide correct feedback.

And on quality specifically:

The quality of the initial answer could be extremely bad and not focus on the real problem, suffering significantly from hallucinations. However, after proper guidelines and conversation, the answer can be corrected and significantly improved, providing students with actionable feedback.

**Time and effort**

Hao Yu estimated around 30 minutes per submission using AI-assisted feedback, against a baseline of 45 minutes without it – a reduction of about a third – and agreed overall that AI reduced grading time. As the survey's own free-text answer notes: "AI can save time, but there will be a learning curve. This will be helpful with larger class." Much of the time in this pilot went into diagnosing and correcting the hallucinations discussed below – overhead that a first-time user on a small, technical assignment absorbs disproportionately, and that would be expected to shrink as prompting becomes more structured and the product changes described below take effect.

**Overall assessment**

Xu Sun and Hao Yu were satisfied overall, would use AI-assisted feedback again, and would recommend it to colleagues. Asked what they had learned, they answered:

I have learned the way of working together with an AI-assisted feedback system, say, a human-in-the-loop student assessment flow. Knowing the competence and limitations of the AI system and learning the best way to work together with it.

Asked what they would do differently, they said:

More structured prompts to guide the AI system and to make the assessment more efficient.

Asked whether AI-assisted feedback should continue in other courses and programmes at UiT, their answer was yes, with a caveat:

> Yes, I recommend it. It provides much more detailed and individualized feedback to support more effective learning outcomes without additional effort and time, and students will greatly benefit from it. However, the implementation must be in a gradual process, with training and experience sharing to help understand the limitations and the best way of human-AI collaboration.

## Hallucinations encountered – and how they were addressed

This was not something the pilot set out to test; it emerged as a consequence of testing. Early testing surfaced clear hallucinations in the AI's first-pass output on both courses. On Manufacturing Logistics, the assistant at one point accepted a factually incorrect statement about inventory-cost behaviour as correct, and had to be corrected by hand.

On Supply Chain Management, the assistant claimed a required data parameter was missing when it was in fact present, while separately missing a genuine model-redundancy issue the assignment's own rubric asked it to check for. Both evaluators found the assistant only produced reliable output after several rounds of guided correction, and Hao Yu additionally noted that a prompt and its answer were once lost mid-conversation.

Crucially, these issues surfaced during testing and iterative prompting – not in the feedback that reached students. Both evaluators read and corrected every AI-generated response before release, so no hallucinated content reached a student. Drawing on Xu Sun and Hao Yu's testing observations, UNIwise carried out a dedicated root cause analysis, using its knowledge of how the WISEflow AI Assistant is built (OCR-based text extraction, a compressed understanding of the exam and rubric, and a multi-turn "debate" step that justifies the grade already assigned).

The analysis traced the errors to six contributing causes, set out below together with the product team's current status on each. Notably, richer, fully-worked marking guidance made no measurable difference to the error rate for either evaluator, ruling out "a better rubric would fix it" as a stand-alone explanation.

| Cause | What we found | Status | UNIwise's response |
|---|---|---|---|
| 1. Gap in subject-matter reasoning | On both courses, the assistant invented problems or missed real ones on technical content, even with a fully-worked marking guide. | Under consideration | The plan is to first test the assistant's quality once it has access to the assignment text (see Cause 3), then assess whether an independent recalculation or re-derivation step is still needed on top of that. |
| 2. OCR/handwriting misreading | At least one flagged "calculation error" was actually a misread character from text extraction (e.g. × read as –), not a real student mistake. | In progress | The current pipeline's OCR has been confirmed to struggle with handwriting, consistent with this finding. The new pipeline is trialling a different OCR model, which has shown better results on handwritten text in initial testing; performance will continue to be monitored. |
| 3. No access to the assignment text | Without knowing the total number of questions, the assistant sometimes omitted unanswered questions from its completeness check. | Already built in | Giving the assistant access to the assignment text, alongside other supporting material, has been a design choice in the new pipeline from the outset. |
| 4. One-size-fits-all prompting | A generic prompt produced vague, unreliable output; a structured, question-by-question prompt worked far better – but had to be discovered by each marker. | Considered for future updates | The product team agrees a single generic approach under-serves at least one assignment type, and is exploring more specialised agents – for example one focused on quantitative aspects and another on qualitative aspects – with different instructions per assignment type. |
| 5. Risk of manufactured justification for a fixed grade | The assessor's grade is correctly and intentionally kept fixed and authoritative – the assistant does not and should not override it. The risk identified is narrower: when a submission does not cleanly support the given grade, the justification step can end up manufacturing or overstating supporting reasons for it. | Being tested | The grade-justification step has been redesigned into an agent loop that grounds its reasoning in actual quotes and examples from the assignment, rather than only a compressed summary – expected to reduce this risk while keeping the assessor's grade as the fixed anchor. |
| 6. Multi-turn conversation reliability | A prompt and its answer were once lost mid-conversation, undermining the correction process both evaluators relied on. | Resolved | The AI agents now have a longer context window and use an AI-driven conversation summariser that retains prior conversation history instead of discarding it. |

Status colour: green = resolved / already built in, blue = in progress, purple = being tested, orange = under consideration, grey = considered for future updates.

Taken together, the causes most directly tied to the pipeline itself – assignment-text access, OCR quality, and conversation reliability – are resolved or in progress in the version due in October 2026, and the redesigned grade-justification step is currently being tested. The remaining recommendations – an independent recalculation step and more specialised, assignment-type-specific agents – are still under consideration for future updates rather than committed to this release. Together with the standing recommendation that a human reviews every AI-generated response before release, these changes target the specific mechanisms behind the hallucinations observed here rather than only the symptom.

## Discussion: What the Pilot Suggests

### What worked

**Human review caught every issue.**

Despite hallucinations in first-pass output, no unreliable feedback reached a student: both evaluators screened every response before release.

**Iterative prompting worked, eventually.**
Both evaluators independently found that a structured, specific prompt produced substantially better output than a generic one – a finding that will inform a built-in, differentiated prompt strategy.

**A real, if modest, time saving**

Even with the overhead of correcting hallucinations, AI assistance still reduced grading time by roughly a third on a small, technical cohort.

**Sustained intent to continue**
Despite the rockier experience, both evaluators would use the tool again and recommend it to colleagues, and one student clearly preferred the AI-plus-human combination to human-only feedback.

**What needs care**

**First-pass reliability on technical, model-based tasks**
Both evaluators needed several rounds of correction before output was usable – the central finding of the root cause analysis and the main focus of the product changes described above.

**Tone on sensitive or informal feedback**
One student found the AI's attempt at supportive language "off-putting" and preferred a more neutral register; this is a calibration question distinct from accuracy.

**Rollout pacing**
Both evaluators explicitly asked for a gradual rollout with training and shared experience, rather than immediate wide deployment, so markers understand the tool's current limitations before relying on it for a full cohort.

## Limitations of the pilot evidence

- Very small, single-institution sample: six Manufacturing Logistics and four Supply Chain Management submissions were tested, and only two students – both from one of the two courses – completed the evaluation survey.
- The course-coordinator survey reflects the joint, combined experience of two evaluators working together rather than two fully independent data points, and several optional fields were left unanswered.
- Outcomes measured were perceptions, usability and workflow – not learning gains or exam performance.
- The pilot deliberately tested closed-ended, quantitative and model-based assignments; the findings should not be read as describing how AI-assisted feedback performs on open, essay-type coursework.

## Conclusion

Within its small scale, this pilot adds a distinct and useful data point to UiT's evaluation of AI-assisted feedback: on closed-ended, quantitative and model-based master's assignments, the AI assistant's first-pass output required substantial correction before it was reliable, and both evaluators encountered real hallucinations before reaching usable feedback. The pilot's most important contribution is not the ratings themselves but the testing data it generated, which fed directly into UNIwise's root cause analysis: a structured account of why the errors occurred, mapped to specific product changes, several of which are already under way. With human review catching every issue before release, a still-positive time saving, and both evaluators willing to continue and recommend the tool, the pilot supports continuing AI-assisted feedback on technical assignments at UiT – provided it is rolled out gradually, with training, and alongside the product changes the root cause analysis has already prompted.

## External references informing this report


- Nicol, D. & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: a model and seven principles of good feedback practice.
- Hattie, J. & Timperley, H. (2007). The power of feedback: the feed-up / feed-back / feed-forward model.
- Shute, V. (2008). Focus on formative feedback (timely, specific, supportive).
- Lee, S. S. & Moore, R. L. (2024). Systematic review of generative AI for automated feedback in higher education.
- UNIwise (2026). Root Cause Analysis: Hallucinations and Errors in AI-Assisted Feedback – based on testing by Xu Sun and Hao Yu (internal report).


## Acknowledgement


UNIwise would like to thank UiT – The Arctic University of Norway for its continued participation and cooperation, and especially Associate Professor Xu Sun and Professor Hao Yu and their students in INE-3608 Manufacturing Logistics and INE-3609 Supply Chain Management for their thoughtful testing and structured feedback, including their willingness to document and investigate the challenges they encountered. We also thank the project team behind "Nye vurderingsformer" (New Assessment Forms) for including this pilot in the extended project.