Augmenting, Not Replacing, Instructor Judgment: Generative AI for Assessment and Feedback in Programming Education

Christian Garmann Sørli[0009−0007−0544−1349],
Trond Aalberg[0000−0001−5593−0860], and
Michail N. Giannakos[0000−0002−8016−6208]

Norwegian University of Science and Technology, Trondheim, Norway
christian.g.sorli@ntnu.no
trond.aalberg@ntnu.no
michailg@ntnu.no

Abstract. Generative AI (GenAI) is increasingly being explored for assessment and feedback in higher education, yet the instructor perspective remains underrepresented in the literature. To address this gap, we investigate how instructors perceive the challenges and opportunities of using GenAI for assessment and feedback in programming education. We conducted an exploratory qualitative study based on eight semi-structured interviews with course coordinators and teaching assistants at two universities in Canada and Norway, analysed with a hybrid thematic analysis (challenges) and deductive coding (opportunities). We identified 11 challenges and 21 opportunities. All participants expressed concerns about the capabilities and reliability of current GenAI systems, particularly regarding hallucinations and factual errors, which put at risk a fair and trustworthy assessment. Participants also consistently raised concerns about loss of control and the potential transfer of assessment authority away from educators. In contrast, perceived opportunities were considerably more diverse, with only five identified by more than one participant. Overall, instructors predominantly proposed uses in which AI augments their work while they stay involved and retain responsibility for final decisions. We discuss the implications for the design and adoption of GenAI in programming assessment and feedback.

Keywords: Generative AI · Assessment and feedback · Programming education · Instructor perspectives · Higher education

1 Introduction

Generative AI (GenAI) tools have moved rapidly from research labs into classrooms and faculty offices. Driven largely by advances in large language models (LLMs), GenAI has become a prominent topic in education, yet there is little consensus on whether and under what conditions it should support teaching and learning. Early evidence on its use for assessment and feedback is promising but

consistently accompanied by documented limitations, including hallucinations, bias, and a lack of transparency [10].

In programming education, manual assessment is resource-intensive, and decades of automated tools, while able to grade consistently at scale, remain pedagogically insufficient [20]. In response, instructors have begun adopting GenAI for concrete assessment tasks, such as assisting with the grading of programming assignments [2] and generating formative feedback on student code, though evidence on effectiveness and student reception remains mixed.

Despite this growing adoption, the instructor perspective is underrepresented in the literature [1]. In assessment, GenAI raises questions not only about technical performance, but also about fairness, trust, transparency, and the distribution of decision-making authority [7]. As those ultimately responsible for assessment practices and outcomes, instructors are well positioned to identify where GenAI can help and where human judgment should remain central. Programming education, with its large cohorts, multidimensional assessment tasks, and limited teaching resources, makes both the value and the risks of automation particularly salient. We therefore ask:

RQ *Which challenges and opportunities do instructors perceive in using GenAI for assessment and feedback in programming education?*

We interviewed eight teaching staff at the University of Waterloo (Canada) and NTNU (Norway). The paper contributes an empirical account of how instructors perceive the role of GenAI in programming assessment and feedback, identifying both shared concerns and a diverse set of potential uses, many specific to programming. In particular, it examines where instructors see opportunities for GenAI to support assessment work while retaining human judgment and decision-making authority.

2 Background and Related Work

AI in education (AIED) builds on a long tradition of predictive analytics, adaptive learning, and intelligent tutoring systems, in which educators were rarely participants [1]. The release of ChatGPT in late 2022 shifted the field towards general-purpose generative models, producing a rapidly expanding but methodologically uneven literature [10, 23] in which instructors remain underrepresented as participants: Bond et al. [1] call for qualitative studies that include them and for international collaboration. Recent work has begun to address this gap but remains fragmented, converging across national contexts on instructors' scepticism towards AI-only feedback, and a preference for AI as support rather than replacement [9, 13, 14, 21], yet few qualitative studies focus on higher-education instructors' perspectives on GenAI for assessment and feedback in programming.

Programming education is a demanding assessment domain: large cohorts, iterative submissions, and multidimensional quality criteria (correctness, efficiency, readability, maintainability) make timely and fair human grading difficult [20]. Automated assessment tools long predate GenAI and now provide

richer feedback than Boolean correctness checks, but are often confined to pass/fail outcomes or output mismatches [20]. Human grading itself shows low inter-rater reliability in introductory programming [18], while LLMs can produce more readable and consistent feedback than human instructors [5]; yet effects on grading performance are mixed [15], reinforcing the case for human oversight.

Current GenAI systems also have well-documented limitations that bear directly on assessment. Calibrated language models must hallucinate some arbitrary facts even with error-free training data [12], although this result concerns factual recall rather than reasoning about code. A related concern is automation bias: users may over-rely on AI, and explanations alone do not reliably prevent misplaced trust [22]. The EU AI Act classifies AI systems that directly affect students' evaluations as high-risk, imposing transparency obligations [6], and Corbin et al. [3] frame AI in assessment as a *wicked problem* that resists definitive formulation, placing irreducible responsibility on educators. These concerns motivate a hybrid intelligence framing, in which AI augments rather than replaces professional judgment by handling routine subtasks while the instructor retains decision-making authority [10, 11].

3 Method

3.1 Research Design

We adopted an exploratory qualitative design to map instructors' perspectives, experiences, and needs regarding the use of GenAI for assessment and feedback. The research was conducted in two phases: first at the University of Waterloo (Canada) from February to April 2025, then at NTNU (Norway) from May to June 2025.

3.2 Participants and Sampling

Participants were eligible if they had recent experience with assessment and feedback in a university-level programming course, with no specific exclusion criteria. We used purposive sampling across four subgroups, defined along two dimensions: national affiliation (Norway and Canada) and professional role (course coordinators responsible for the design and implementation of the course, and other teaching staff such as teaching and learning assistants). This ensured representation of course coordinators, whose perspectives were central despite their limited number, alongside other teaching staff across both contexts. Waterloo participants received a CAD 20 gift card, while NTNU participants were offered a NOK 250 gift card as compensation for their time. Throughout the paper, *instructors* refers collectively to teaching staff directly involved in assessment and feedback, including course coordinators, teaching assistants, and learning assistants. Table 1 summarises the participants.

3.3 Procedure

Participants received written information on the study's purpose, data storage, voluntary participation, and right to withdraw, and provided digital informed

Table 1. Overview of interview participants.

ID	University	Background	Role	Assessment process
P1	U. Waterloo	Faculty	Course coordinator (project, formerly CS1)	Project: manual + checklist; CS1: automated testing
P2	U. Waterloo	Faculty	Course coordinator (CS1, CS2, project)	Project: manual; CS1/CS2: automated testing
P3	U. Waterloo	PhD student	Teaching assistant (CS1, CS2, project)	Project: demo; CS1/CS2: automated testing
P4	U. Waterloo	PhD student	Teaching assistant (project)	Manual + rubric, functional testing
P5	NTNU	Faculty	Course coordinator (project, exam)	Manual + checklist; exam: 70% auto (MC), 30% manual
P6	NTNU	Faculty	Course coordinator (CS1, project)	Project: manual + rubric; CS1: demo + oral
P7	NTNU	Master student	Learning assistant (CS1)	Demo + oral feedback
P8	NTNU	Master student	Learning assistant (CS1)	Demo + oral feedback

consent before the interview. The Waterloo interviews were approved by the University of Waterloo Research Ethics Board (project #45231). In Norway, the study followed NTNU's data protection guidance and the NENT guidelines for research ethics in science and technology [19]. The first author, who conducted all interviews, had no relationship with the NTNU participants beyond working in the same department. Interviews lasted approximately one hour and were securely audio-recorded and transcribed. All but two (one per site, online) were conducted in person. Waterloo interviews were conducted in English and NTNU interviews in Norwegian, with Norwegian quotations translated by the first author. Identifiers were replaced with participant codes before transcription, and quotations were lightly edited for readability or anonymity. The semi-structured guide covered current practices and automation, opportunities with future and current AI, and concerns about AI (Appendix A). The interviews were framed around GenAI applications for assessment and feedback, although the interview guide used the broader term *AI*.

To encourage participants to consider possibilities beyond the limitations of current systems, the opportunities section introduced a speculative future scenario: participants imagined a future where AI is a common tool in teaching, with fewer limitations than today regarding availability, explainability, and privacy, and brainstormed how AI could support their assessment and feedback work in such a future. To anchor the discussion in participants' practice, they were encouraged to draw on their own teaching materials, inspired by a cognitive walkthrough. The scenario deliberately avoided naming specific AI systems or tools to reduce anchoring on existing products or capabilities and thus avoid steering participants' responses. The discussion then shifted from this future-

oriented scenario to current GenAI technology, asking participants to identify opportunities they considered feasible and actionable today, before concluding with broader concerns and challenges.

3.4 Data Analysis

The first author coded all transcripts, with a different unit of analysis for challenges and for opportunities. Challenges were analysed with Fereday and Muir-Cochrane's six-step hybrid thematic analysis [8], combining inductive (data-driven) and deductive (theory-driven) coding. The challenge codebook used 17 codes from Bond et al.'s [1] meta-systematic review (e.g., shifting authority, bias, transparency, and staffing); as participants voiced them, we label the shifting-authority and capacity-of-current-AI codes *loss of control / shifting authority* and *AI capability and reliability*. Codes and themes were iteratively refined as new findings emerged from the data. The unit was the participant: each code was recorded once per participant who raised it, and n counts participants.

For opportunities, the unit was the proposed use. Each distinct use was recorded once under a descriptive name, with the participants who proposed it (each counting as one mention) and whether they saw it as feasible with current AI or as speculative. Opportunities were then classified deductively by Lubars and Tan's [17] degrees of AI involvement, judged against the assessment task each would serve: *human leads, AI assists* (machine in the loop) when the instructor does most of the work and AI offers suggestions or help; *AI leads, human assists* (human in the loop) when AI performs the task and the instructor gives input or confirms the result; and *full AI automation* when AI makes and carries out decisions without human involvement.

4 Results

Across the eight interviews, we identified 11 challenges and 21 opportunities: ten of the 11 challenges but only five of the 21 opportunities were mentioned by more than one participant. Eight of the a priori challenge codes were not observed in the interviews (infrastructure, variable data quality, common standards needed, students' lack of skills and awareness, accessibility, lack of real-world applications, lack of ethical consideration, and lack of professional development support), and two challenges emerged inductively: *students' acceptance of AI assessment* and *sustainability*. Table 2 summarises the observed challenges; Pn denotes participants and n the number mentioning each challenge.

4.1 Challenges

AI Capability and Reliability (n=8). All participants questioned the capabilities and reliability of current large language models, seen as a risk to fair and trustworthy assessment and an argument for human oversight: "my concern typically would be correctness, given how much we know that language models can hallucinate" (P4). This motivated a strong preference for human verification: "I think we'd still need to verify it because what if it makes a mistake" (P1).

Table 2. Identified challenges, with the number of participants (n) mentioning each and code origin (a priori or inductive).

Challenge	n	Origin
AI capability and reliability	8	A priori
Loss of control / shifting authority	8	A priori
Transparency	7	A priori
Bias	2	A priori
Curriculum development	2	A priori
Difficulty assessing effectiveness of AI	2	A priori
Privacy and security	2	A priori
Impact on staffing	2	A priori
Lack of teacher technical knowledge	2	A priori
Students' acceptance of AI assessment	2	Inductive
Sustainability	1	Inductive

Loss of Control / Shifting Authority (n=8). All participants expressed concern about loss of control if AI is given a central role in assessment, extending beyond technical control to the role humans should play in assessment in the future and to fear that assessment authority shifts away from educators. Both a Canadian (P1) and a Norwegian (P5) course coordinator raised the same reflection about what remains human if instructors use AI to create exams, students use AI to solve them, and AI evaluates the results: "What's the point of this all?" (P1); "What is left that is human in that?" (P5). P6 feared being reduced to a spectator: "I essentially just become a spectator ... I've just been directed or replaced." Participants stressed retaining the final say over grades: "I still am not in favor of having AI directly change the grade" (P4). P3 was similarly reluctant to delegate value-based judgments to AI even if a system could technically perform them: "Even if a system could do that, I wouldn't necessarily, as a professor, want to delegate that to something else".

Transparency (n=7). Participants identified lack of transparency and explainability as a central challenge, both regarding how assessment decisions are made and how models function: "I need to know the accuracy of the results. I need a number or something that I can point to" (P2); even when AI offers an explanation, "we don't really know, since the models are black box" (P8). One course coordinator (P1) also argued that openness towards students is required if AI is used in assessment, including how the system works and how quality assurance is carried out.

The remaining eight challenges:

- **Bias (n=2)** in training data and models was raised by P6 and P8: "there's a lot of bias in them and we don't quite know what's happening" (P8).
- **Curriculum development (n=2)** was raised by two course coordinators (P1, P6), with direct implications for the validity of assessments built on

programming assignments: "a lot of professors are wondering what is the relevance of a course that emphasizes programming in assignments" (P1).
- **Difficulty assessing the effectiveness of AI (n=2)** was raised by P1 and P3, who questioned whether AI actually delivers efficiency gains: "I don't know that it actually will save time" (P3).
- **Privacy and security (n=2)** was raised only by NTNU participants: "you have problems with privacy and data security" (P6), while P5 saw university-approved solutions as a precondition for use.
- **Impact on staffing (n=2)** was raised by the two learning assistants (P7, P8): "you might not need learning assistants in the end because chatbots can do it better" (P7).
- **Students' acceptance of AI assessment (n=2)** emerged inductively: despite students' own likely AI use, participants expected resistance to AI grading: "students would not, even though they're probably using a lot of AI to generate their solutions, ... find it a terrible thing to use it for marking" (P1). P8 similarly distinguished feedback from grading: "feedback is one thing, but assessment with grades is something completely different".
- **Lack of teacher technical knowledge (n=2)** was raised by P5 and P6, whose limited hands-on experience made it hard to judge what AI could be trusted with: "I haven't tried it, so I simply don't know" (P5); "I haven't really used AI that much" (P6).
- **Sustainability (n=1)** emerged inductively as a concern about the resources required to train and run such models (P4): "there's climate concerns in terms of like how much you're using these models."

4.2 Opportunities

We identified 21 opportunities and coded each by degree of AI involvement (Section 3.4): eight human-led, nine AI-led, and four fully automated. Thus 17 of the 21 proposed uses keep the instructor involved, but the instructor leads the task only in the eight human-led ones. Because 16 opportunities were each proposed by a single participant, the counts indicate the breadth of proposed uses rather than their prevalence. Five were speculative (marked † in Table 3). Some proposals (compile and unit-test checks, SQL output comparison, speech-to-text) do not necessarily require GenAI; we kept them because participants proposed them as AI support. Proposals were unevenly spread (Table 3): Waterloo participants made 17 of 26 mentions and 10 of 11 human-led mentions (five by P3), while eight of nine NTNU mentions were AI-led or fully automated.

Human leads, AI assists (8 of 21):

- **Semantic search**: AI finds candidate matches in large codebases and documentation: "find me an instance of this design pattern" (P2).
- **Describing code**: AI explains what student code does: "you can give it a bit of code and say, talk about why this code is good" (P1).

Table 3. Identified opportunities, grouped by Lubars and Tan's degree of AI involvement, with proposer IDs and site (W = Waterloo, N = NTNU); a dagger (†) marks speculative opportunities.

Opportunity	IDs	Site
Human leads, AI assists (machine in the loop)		
Semantic search in codebases	P1, P2	W
Describing code	P1	W
Mechanical operations (compile, unit tests)	P2	W
Writing support for feedback consistency	P3, P4	W
Feedback guide	P3	W
Speech-to-text and note tools	P3	W
Error highlighting in student code	P3, P8	W, N
Smart version logging†	P3	W
AI leads, human assists (human in the loop)		
Evaluating test coverage	P2	W
Generating multiple-choice questions	P5	N
Summarisation of submissions	P8	N
Adapting assessment to individual needs†	P6	N
Analysing documentation for understanding†	P3	W
Insight into learning barriers†	P6	N
Code review	P1	W
Classifying feedback comments into a grade	P1	W
Evaluating software diagrams	P5	N
Full AI automation		
Evaluating user interfaces†	P1	W
Evaluating SQL queries	P5	N
Formative feedback on non-counting assignments	P4, P6	W, N
Explanation tool for students	P2, P7	W, N

- **Mechanical operations**: AI checks that code compiles and unit tests run: "I want things that are sort of just time sinks for somebody who's doing the grading that require very little skill" (P2).
- **Writing support for feedback consistency**: AI flags inconsistent wording within and across raters: "it might catch, like if your wording isn't consistent" (P3); "calibrate across different TAs, the language and make something that seems more coherent" (P4).
- **Feedback guide**: AI structures feedback the instructor writes: "it's basically telling me how to write it, but not writing it for me" (P3).
- **Speech-to-text and note tools**: AI takes notes during live oral assessments: "help me take notes on what I'm interested in while still focusing on a presentation that's happening live" (P3).
- **Error highlighting**: AI flags likely error locations in student code: "just point out where the error is. [...] that would save time as well" (P3).
- **Smart version logging**: AI maps claimed changes to actual changes between submissions: "automatically map what they say they've changed to the actual sections that have changed in the document" (P3).

AI leads, human assists (9 of 21):

- **Evaluating test coverage**: AI reports gaps in test coverage for the instructor to verify: "Tell me what's missing in my tests [...] I would feel a heck of a lot more confident with somebody looking over its shoulder" (P2).
- **Generating multiple-choice questions**: AI drafts questions that the instructor refines: "you could imagine getting help to create questions" (P5).
- **Summarisation**: AI summarises extensive submissions, and the instructor checks the summary: "if there's a lot to look through and I just need a summary [...] then I think AI is good at that, and then I quickly check if it's correct" (P8).
- **Adapting assessment**: AI adapts materials to individual needs such as ADHD, with instructors deciding what to use: "say it was ADHD or other challenges. Can it be adapted? [...] AI can help us manage such things" (P6).
- **Analysing documentation**: AI judges evidence of understanding in student documentation for the instructor to assess: "you could get it to look at documentation ... does it seem like the student understands?" (P3).
- **Insight into learning barriers**: AI identifies where students get stuck, for instructors to interpret: "many struggle with programming, and maybe AI can help us find out what's actually getting stuck" (P6).
- **Code review**: AI reviews student code, and the instructor weighs its comments: "I think AI is pretty good at coding reviews in general" (P1).
- **Classifying feedback comments into a grade**: AI proposes a grade from the instructor's written feedback: "if feedback is about what you did correctly, what you can improve on and what you did wrong, implicitly it must be highly correlated with a grade" (P1).
- **Evaluating software diagrams**: AI evaluates diagrams against design rules for the instructor to assess: "As long as there is a correct solution, I think AI can contribute a lot. But when it is a matter of judgment, and there are many different designs, I struggle a bit to see how AI can help us [...] concepts like cohesion and coupling [...] make it possible to automate it to a larger degree" (P5).

Full AI automation (4 of 21):

- **Evaluating user interfaces**: AI marks user-interface code: "the biggest one would just be one that can mark user interface code somehow" (P1).
- **Evaluating SQL queries**: AI checks query output against the expected answer: "evaluating SQL and checking that the output matches what was asked for, I think that should be possible" (P5).
- **Formative feedback**: AI gives feedback on non-counting assignments: "on things that are not so serious if there are mistakes, like exercise feedback where you have room to fix, room to appeal" (P6).
- **Explanation tool for students**: AI explains concepts to students directly, without the instructor: "a very nice way to learn is to talk with such a chatbot to bounce ideas and concepts at any time of day" (P7).

5 Discussion

5.1 Challenges: Convergent Concerns

The three most prominent challenges recurred at both universities: all participants raised *AI capability and reliability* and *loss of control / shifting authority*, while seven of eight raised *transparency*. This pattern echoes the closest comparable studies: Ghimire et al. [9] found educators concerned about incorrect outputs and over-reliance, and Karunaratne and Lindblad [13] found concerns about fabricated information and maintaining human control over AI-supported assessment. Since calibrated models must hallucinate some facts even with error-free training data [12], the human verification participants emphasised is likely to remain necessary as models improve.

The loss-of-control concern extends beyond technological reliability to who holds authority and responsibility when assessment decisions involve competing considerations, which Corbin et al. [3] frame as a wicked problem in which every approach carries trade-offs and educators must weigh and take responsibility for them. Our participants echoed this by insisting on retaining the final say over grades, a stance consistent with Kleveland et al. [14], who found that teachers valued GenAI for saving time while firmly maintaining professional responsibility. Independently of participants' views, regulation points the same way: the EU AI Act classifies AI systems intended to evaluate learning outcomes as high-risk and requires effective human oversight [6], although it does not apply in Canada and reaches Norway only through the EEA Agreement.

Bias warrants particular caution in consequential assessment. Loukina et al. [16] show that automated scores can differ across groups and that different definitions of fairness may require different responses. Although raised by only two participants, this concern suggests a procedural safeguard: independent human review of contested AI-supported assessments, since an appeal evaluated by the same model may reproduce its specific errors or biases.

5.2 Opportunities: Collaboration, Not Replacement

A central finding is that few identified opportunities positioned AI as an autonomous assessor: only four of the 21 proposed uses were fully automated, while the other 17 kept the instructor involved, leading the task in eight and confirming AI output in nine. This aligns with the hybrid intelligence framing, where AI augments rather than replaces professional judgment [10, 11], and with proposals for hybrid LLM–instructor workflows [2]. It also echoes prior findings by Roe et al. [21], who report that staff prefer AI-assisted to AI-only marking, and Karunaratne and Lindblad [13], who frame opportunities as support, not replacement. Cukurova [4] argues that replacing human tasks with AI risks atrophy of critical competencies over time. Our participants' concerns echo this: the fear of becoming a spectator (P6) and the insistence on retaining the final say over grades are resistance to replacement.

The delegation levels also imply a risk gradient: human-led outputs are inspected within the instructor's own work, AI-led outputs are checked before use,

and fully automated outputs reach students unchecked. Risks also differ in kind within a level: checks on executable artefacts (compilation, unit tests, query output) mainly carry technical risk and are cheap to verify, whereas grades and inferences about understanding carry fairness and validity risks that a quick check may miss, and student-facing uses risk over-reliance [22]. This matters in programming, where executable artefacts make correctness mechanically checkable while design quality and readability still require judgment [20], a line P5 also drew for design tasks. Uses that evaluate learning outcomes, formative ones included, also fall under the EU AI Act's high-risk category [6]; policies and tool design should differentiate these profiles.

The delegation scale describes the instructor's role only; the student does not appear as an actor. Yet in the explanation tool the student is the active party, and several proposed instructor uses (code review, describing code, evaluating test coverage) are competencies students could be trained to perform themselves with AI support, a boundary also touched by the learning assistants' staffing concern (P7, P8). Whether assessment-related GenAI should be an instructor tool, a student tool, or both is an open design question; the interview guide's instructor framing is a limitation here.

The collaboration-not-replacement pattern in our data suggests prioritising bounded subtasks, such as error highlighting and feedback-consistency checks, with outputs that instructors can readily inspect and revise. Initial designs could explore these uses in low-stakes formative contexts, as P6 and P8 suggested, while preserving educator authority over grades and other consequential decisions. The practical value of this approach also depends on the effort required for verification: participants questioned whether AI would save time if checking its outputs were as burdensome as completing the task without AI. However, requiring human review does not by itself ensure meaningful oversight. Research on automation bias indicates that users may accept AI output without adequately evaluating it [22], and Cukurova [4] warns that human-in-the-loop correction can converge towards the AI's output rather than critical evaluation.

Only five of 21 were mentioned by more than one participant, underscoring the diversity of concrete tasks instructors envision. Rather than pointing to a single optimal integration, this diversity suggests that instructors' needs may vary across teaching settings, reinforcing the case for involving them directly in designing AI-supported assessment tools. Together, these findings suggest an asymmetry in our sample: challenges were generally framed as broad concerns about GenAI and converged across both sites, whereas opportunities were tied to specific tasks and teaching contexts, and their delegation levels varied by site and participant. With four participants per site, whether this pattern extends beyond our sample remains a question for further research, as does whether willingness to delegate depends on the kind of judgment and the stakes involved, not only on how much is automated.

5.3 Limitations

The sample, while appropriate for an exploratory study, is not representative of programming educators more broadly: all participants were men from Western universities, and several NTNU participants taught the same course, which may have produced overlapping perspectives. Recruitment may also have favoured participants already interested in or experienced with AI, and possibly more optimistic about it. By presenting availability, explainability, and privacy as limitations of current AI, the speculative scenario may have primed participants to raise these topics later, and questions 7, 8, and 10 (Appendix A) asked directly about human review and transparency, which may have prompted these concerns. Because participants had little hands-on experience with GenAI for assessment, many responses were anticipatory rather than based on experience. Translation of Norwegian quotations may also have affected nuance. The first author coded all data alone, so inter-rater reliability was not assessed. Some opportunities, such as the explanation tool for students and insight into learning barriers, lie on the boundary between assessment and broader learning support. GenAI tools, and instructors' experience with them, have continued to develop since data collection in the first half of 2025. Finally, the findings may not transfer to other educational levels, informal learning settings, or non-programming disciplines. Nevertheless, the goal was not comprehensive coverage but an exploratory mapping of the challenges and concrete uses salient to this group of instructors.

6 Conclusion

Through eight interviews in Canada and Norway, this paper investigated teaching staff's perceived challenges and opportunities for using GenAI in programming assessment and feedback.

The paper offers two primary insights. First, in these interviews, challenges at both sites clustered around shared concerns about GenAI, chiefly its capability and reliability, the potential loss of instructor control, and lack of transparency, whereas opportunities were more diverse and often tied to particular tasks and teaching contexts, suggesting that AI support may need to be tailored to particular courses, roles, and assessment practices. Second, willingness to delegate appeared to depend not only on AI's capabilities but on the kind of judgment: participants welcomed help with bounded, verifiable operations like compiling and testing code, but were more reluctant to hand over evaluative decisions involving interpretation, fairness, and accountability. Meaningful augmentation thus requires more than keeping a human in the loop: it calls for support whose outputs instructors can inspect and revise, with verification efficient enough to preserve potential time savings without reducing oversight to routine approval.

Two threads warrant further work. First, the envisioned opportunities invite design-based research involving instructors and students, beginning with recurring opportunities such as feedback consistency and error highlighting in low-stakes formative contexts, and examining their effects on instructor work-

load, feedback quality, and student learning. Second, studies of actual assessment practice should examine how instructors verify AI outputs and whether the verifiable–evaluative distinction holds for other instructors. Ultimately, the design of AI-supported assessment should consider not merely how much work is automated, but which forms of judgment remain under educators' authority.

Acknowledgments. The authors thank Jian Zhao of the Cheriton School of Computer Science, University of Waterloo, for hosting the first author's research stay and facilitating the data collection at the University of Waterloo, and the ExcITEd centre (Excellent IT Education, NTNU's Centre for Excellence in Education) for facilitating the data collection at NTNU. This article has been partially funded by the Research Council of Norway (grant 357493) through the AI Centre for the Empowerment of Human Learning (AI LEARN).

References

1. Bond, M., Khosravi, H., De Laat, M., Bergdahl, N., Negrea, V., Oxley, E., Pham, P., Chong, S.W., Siemens, G.: A meta systematic review of artificial intelligence in higher education: A call for increased ethics, collaboration, and rigour. International Journal of Educational Technology in Higher Education **21**(1), 4 (2024). https://doi.org/10.1186/s41239-023-00436-z
2. Cisneros-González, J., Gordo-Herrera, N., Barcia-Santos, I., Sánchez-Soriano, J.: JorGPT: Instructor-aided grading of programming assignments with large language models (LLMs). Future Internet **17**(6), 265 (2025). https://doi.org/10.3390/fi17060265
3. Corbin, T., Bearman, M., Boud, D., Dawson, P.: The wicked problem of AI and assessment. Assessment & Evaluation in Higher Education **51**(4), 736–752 (2026). https://doi.org/10.1080/02602938.2025.2553340
4. Cukurova, M.: The interplay of learning, analytics and artificial intelligence in education: A vision for hybrid intelligence. British Journal of Educational Technology **56**(2), 469–488 (2025). https://doi.org/10.1111/bjet.13514
5. Dai, W., Tsai, Y.S., Lin, J., Aldino, A., Jin, H., Li, T., Gašević, D., Chen, G.: Assessing the proficiency of large language models in automatic feedback generation: An evaluation study. Computers and Education: Artificial Intelligence **7**, 100299 (2024). https://doi.org/10.1016/j.caeai.2024.100299
6. European Parliament and Council: Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence. Official Journal of the European Union, L 2024/1689 (2024), entered into force 1 August 2024. https://data.europa.eu/eli/reg/2024/1689/oj
7. Feng, T.H., Luxton-Reilly, A., Wünsche, B.C., Denny, P.: From automation to cognition: Redefining the roles of educators and generative AI in computing education. In: Proceedings of the 27th Australasian Computing Education Conference. pp. 164–171 (2025). https://doi.org/10.1145/3716640.3716658
8. Fereday, J., Muir-Cochrane, E.: Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development. International Journal of Qualitative Methods **5**(1), 80–92 (2006). https://doi.org/10.1177/160940690600500107
9. Ghimire, A., Pather, J., Edwards, J.: Generative AI in education: A study of educators' awareness, sentiments, and influencing factors. In:

2024 IEEE Frontiers in Education Conference (FIE). pp. 1–9 (2024). https://doi.org/10.1109/FIE61694.2024.10892891
10. Giannakos, M., Azevedo, R., Brusilovsky, P., Cukurova, M., Dimitriadis, Y., Hernández-Leo, D., Järvelä, S., Mavrikis, M., Rienties, B.: The promise and challenges of generative AI in education. Behaviour & Information Technology **44**(11), 2518–2544 (2025). https://doi.org/10.1080/0144929X.2024.2394886
11. Järvelä, S., Zhao, G., Nguyen, A., Chen, H.: Hybrid intelligence: Human–AI coevolution and learning. British Journal of Educational Technology **56**(2), 455–468 (2025). https://doi.org/10.1111/bjet.13560
12. Kalai, A.T., Vempala, S.S.: Calibrated language models must hallucinate. In: Proceedings of the 56th Annual ACM Symposium on Theory of Computing. pp. 160–171 (2024). https://doi.org/10.1145/3618260.3649777
13. Karunaratne, T., Lindblad, E.: Imagining assessment futures through artificial intelligence in higher education teachers' perspectives. Discover Education **4**(1), 532 (2025). https://doi.org/10.1007/s44217-025-00987-5
14. Kleveland, J., Giannakos, M., Lindvig, Y.: Teacher–AI collaboration to support assessment and feedback: A case study in Norwegian secondary education. Evaluation Review (2026). https://doi.org/10.1177/0193841X261458416
15. Li, Y., Shan, Z., Raković, M., Guan, Q., Gašević, D., Chen, G.: When AI explains in natural language: Unveiling the impact of generative AI explanations on educators' grading and feedback practices. Education and Information Technologies **30**(17), 24931–24964 (2025). https://doi.org/10.1007/s10639-025-13741-z
16. Loukina, A., Madnani, N., Zechner, K.: The many dimensions of algorithmic fairness in educational applications. In: Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications. pp. 1–10 (2019). https://doi.org/10.18653/v1/W19-4401
17. Lubars, B., Tan, C.: Ask not what AI can do, but what AI should do: Towards a framework of task delegability. In: Advances in Neural Information Processing Systems 32. pp. 57–67 (2019). https://proceedings.neurips.cc/paper/2019/hash/d67d8ab4f4c10bf22aa353e27879133c-Abstract.html
18. Messer, M., Brown, N.C., Kölling, M., Shi, M.: How consistent are humans when grading programming assignments? ACM Transactions on Computing Education **25**(4), 1–37 (2025). https://doi.org/10.1145/3759256
19. Norwegian National Committee for Research Ethics in Science and Technology (NENT): Guidelines for Research Ethics in Science and Technology. The Norwegian National Research Ethics Committees (2024), third edition, ISBN 978-82-7682-118-5. https://www.forskningsetikk.no/en/guidelines/science-and-technology/guidelines-for-research-ethics-in-science-and-technology
20. Paiva, J.C., Leal, J.P., Figueira, Á.: Automated assessment in computer science education: A state-of-the-art review. ACM Transactions on Computing Education **22**(3), 1–40 (2022). https://doi.org/10.1145/3513140
21. Roe, J., Perkins, M., Ruelle, D.: Is GenAI the future of feedback? Understanding student and staff perspectives on AI in assessment. Intelligent Technologies in Education **1**(1) (2026). https://doi.org/10.53761/ITED/1.7
22. Romeo, G., Conti, D.: Exploring automation bias in human–AI collaboration: A review and implications for explainable AI. AI & Society **41**(1), 259–278 (2026). https://doi.org/10.1007/s00146-025-02422-7
23. Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., Gašević, D.: Practical and ethical challenges of large language models in education: A systematic scoping review. British Journal of Educational Technology **55**(1), 90–112 (2024). https://doi.org/10.1111/bjet.13370

A Interview Guide

The guide was originally developed in English, translated to Norwegian for the NTNU interviews, and used in the same form at both institutions.

Current Assessment Practices, Challenges, and Experience with Automation

1. Can you describe the current process for evaluating and providing feedback on coding assignments and exams?
2. What are the biggest challenges you face in assessing programming assignments and exams?
3. How do you balance speed vs. quality when giving feedback to students?
4. How do you ensure fairness and consistency in grading, especially across different instructors or TAs?
5. Have you used any automated grading, automated feedback, automated tests or other automated tools in the assessment process before? If so, what worked well and what didn't?

Possibilities with Future and Today's AI

6. Imagine a future where artificial intelligence has become a completely normal tool in education, where it:
   - is available to all instructors,
   - enables rapid development of services, systems, and tools to support assessment and feedback that previously would have required extensive time and resources,
   - has the ability to recognize patterns in students' work that previously required manual effort, and translate these into insights that strengthen assessment, academic follow-up, and teaching development,
   - can explain its assessments and decisions in a trustworthy way,
   - safeguards privacy and data security (it runs on systems your institution trusts, and there are no concerns about processing student data or sensitive information).

   Walk me through the teaching materials in your programming course, and let's brainstorm how AI could support you in assessment and feedback work in such a possible future. If you could design the ideal code assessment tool in this future, what would it look like and what features are essential to solve the challenges you mentioned?
7. What kind of tasks would you trust today's AI to automate in the assessment process without human review, if any?
8. What tasks would you trust today's AI to semi-automate with human review, if any?

Challenges and Concerns

9. What concerns and challenges do you identify about using AI for grading and feedback?

10. If AI did automate parts of the assessment process, what level of transparency and explainability would be necessary for you to trust it?