AI-GENERATED FORMATIVE FEEDBACK FOR ENGINEERING LAB REPORTS

Richard PYLE, Joel ROSS and Aydin NASSEHI
University of Bristol, United Kingdom

ABSTRACT
This project examined the use of AI-generated formative feedback in a first-year undergraduate engineering unit, aiming to develop students' AI literacy and critical thinking. Students who opted in received feedback on draft laboratory reports from a locally deployed large language model. The feedback was discussed in small groups during an in-person workshop. It is hoped that, alongside peer review, this offers complementary opportunities for a diverse cohort to engage critically with feedback. Of 830 enrolled students, 390 submitted a draft and 374 opted in for AI feedback. Pass rates on the subsequent summative report were 70% for those receiving AI feedback versus 57% for those who did not. This difference remains confounded by known correlations between engagement with formative activities and overall attainment.
A deliberate design feature was the inclusion of one intentional inaccuracy in each student's feedback. This was explained to students beforehand and intended to promote scrutiny of feedback over passive acceptance of AI output. The activity was supported by a short presentation on responsible AI use and ethical considerations, informing discussions about trust, reliability, and the role of AI in education.
Survey responses were limited (n = 32), but 71% of respondents found the feedback useful and 89% revised their work as a result. Students valued clarity, speed, and rubric-based structure but expressed mixed views on the introduced inaccuracies. Findings highlight the importance of transparency, opt-in participation, and careful pedagogic framing when integrating AI tools into learning and assessment.

Keywords: AI, large language models, formative feedback, laboratory reports, large cohorts

1 BACKGROUND

1.1 Useful Feedback
Feedback is a strong contributor to student development, but more is not always better. Literature repeatedly shows that feedback can strongly influence achievement, yet effects vary depending on what the feedback contains, how it is presented and how students use it [1], [2]. A common explanation is that for feedback to be effective it must be a comparison to an understood standard and enable action towards improvement. In practice, staff often report a "feedback gap", where time is spent producing comments, but many students do not engage with them in ways that lead to longer-term [3], [4]. This has led to greater emphasis on student feedback literacy: the capacity to interpret feedback, make judgements about quality, manage affect and act on advice [5]. These challenges are amplified in large cohorts, where providing detailed feedback is difficult to sustain at scale. The problem is further sharpened in contexts such as lab report writing where iterative writing is essential when learning the skill for the first time. Peer review is one established way to increase feedback opportunities and learn through evaluation of others' work [6]. However, despite the pedagogical benefits of peer review, students often report greater trust in staff feedback, perceiving it as more authoritative or reliable than peer comments [7]

1.2 Generative AI to Support Feedback
Generative AI (Gen-AI), and large language models (LLMs) in particular, are increasingly being trialled as tools to support formative feedback. Recent reviews of Gen-AI for automated feedback in higher education report potential benefits such as rapid turnaround, scalability and rubric-aligned commentary, which may help address staff workload pressures [8], [9]. Evidence from writing-focused studies suggests that LLMs can provide usable early-draft feedback, but that expert human feedback often has better accuracy, prioritisation and specificity [10]. A further concern is over-trust: LLM outputs can be

fluent but wrong, and students may accept them uncritically unless activities are designed to encourage checking and judgement [11], [12]. Emerging work on engagement with LLM-generated feedback also indicates that students' revision behaviours may be superficial without explicit guidance and opportunities for dialogue [13].

2 PROJECT OVERVIEW AND AIMS

This paper reports on a case study from a first-year engineering unit teaching experimental practice and lab report writing to a large first-year engineering undergraduate cohort (n = 830). Within this project students could opt in to receive AI-generated formative feedback on a draft laboratory report. This feedback was tightly structured around the marking rubric and produced by a locally deployed (i.e., run on a computer on campus) LLM. The feedback was reviewed by students in small groups during an in-person workshop immediately after giving peer feedback. This facilitated session created the opportunity for student-student and student-staff discussion about the different sources of feedback and how best to use them productively. It also exposed students to the real-world challenge of reviewing and using different sources of potentially conflicting feedback. A key design feature was the inclusion of one/two intentional inaccuracies in each student's feedback. This was disclosed in advance to students and aimed to position the AI output as something to evaluate rather than accept. The activity was also supported by a short presentation on responsible AI use and ethical considerations, informing discussions about trust, reliability, and the role of AI in education. It should be noted that most published examples of AI-generated feedback involve students using an LLM 'chat bot' to generate their own feedback [8]. In contrast, this project generated feedback with a staff controlled LLM system implemented by the authors – this provided consistency in feedback quality and placed the live session's emphasis on critically reviewing the provided feedback.

This project aimed to (1) explore whether LLMs could produce useful, rubric-based formative feedback at scale, and (2) use the feedback and live session to encourage development of students' critical evaluation and AI literacy. The remainder of the paper outlines the unit context, details the intervention and summarises student perceptions from a short post-activity survey.

3 INTERVENTION DESIGN

3.1 Unit Context

This project was situated within a University of Bristol undergraduate engineering unit called "Engineering Communication, Measurement and Data Analysis" (ECMDA). This unit is taken by almost all first-year engineering students, resulting in a cohort of 830 in 2024/25 when the intervention was first implemented. ECMDA covers the fundamentals of experimental practice, data analysis in Python and lab report writing. The teaching and learning of lab report writing skills (which this project is focused on augmenting) is primarily carried out in flat bed teaching space shown in Fig. 1a, timetabled for a sixth of the cohort at a time (~140 students), sat in groups of ~3 in the room.

As described fully in previous work [6] and summarised in Fig. 1b, ECMDA relies heavily on an iterative peer review process to scaffold students learning towards a final summative lab report submission. A series of formative lab reports are submitted by students throughout the year, each of which is accompanied by a facilitated live peer review session where students review their colleagues' reports in small groups. This same peer review session is conducted for students' draft submission of their final summative lab report. The currently considered intervention was conducted within this final live session: in the second hour of the session the students were given the AI-generated feedback (as described in Section 3.2) and supported to review in groups (as discussed in Section 3.3).

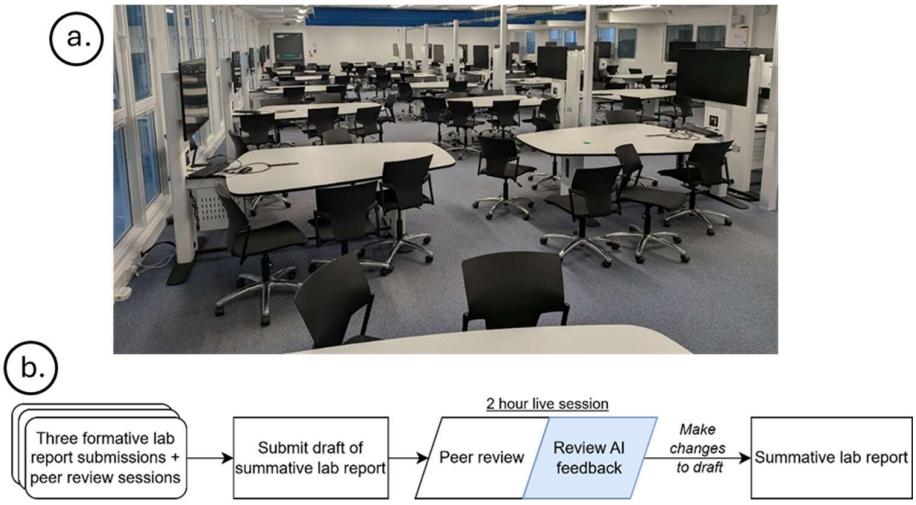

Figure 1. a) A photograph of the ECMDA teaching space used for peer and AI feedback review sessions (G.01, Ivy Gate, University of Bristol) and b) a diagram of the overall formative scaffold

3.2 Creating AI-Generated Feedback

We defined three design requirements for our AI-generated feedback:

Timely – Generation with a <24hr turnaround, for 400+ students (half the cohort as sessions were split across two weeks) with minimal human intervention.

Local compute – All computation should be local, using institution-controlled hardware, to ensure data privacy and, potentially, to reduce environmental footprint often associated with external services.

Structured – Feedback must be structured against the existing rubric and presented in a form that students can digest and critically review within the facilitated live session.

To meet these aims the following data processing steps were conducted for each lab report submitted by a student who opted to receive AI feedback:

Text from the lab report split into sections by an automated search for expected headings (possible as students used a provided template),

LLM used to iterative over the rubric, with each prompt:

Requesting feedback for one rubric criteria

Including only the relevant section of the report

Structuring responses to produce a) a *pass/fail* for the criteria, b) *reasoning* for this and c) supporting *evidence* from the report where relevant

One or two criteria which were originally assigned a pass chosen at random and the LLM prompted to provide evidence for the contrary. This was included, transparently to students, as encouraged inaccuracies (discussed further in Section 3.4),

Feedback against entire rubric formatted into a Word document and emailed to student during live session

The above process was implemented on a University of Bristol owned M3 Ultra Mac Studio and processing the full 33-criteria rubric for one lab report took ~60s. Example snippets of formatted feedback are provided in Fig. 2.

During development of this process the most popular publicly available LLMs at the time were qualitatively tested for feedback quality. The most effective LLM was found to be Microsoft's Phi4 [14]; which most consistently matched human assessment of the reports and adhered to output structure requirements (e.g., reasoning being concise and evidence being either blank or a quote from the report). It should be noted that at 14 billion parameters Phi4 is a relatively small LLM, especially when compared to commercial models such as OpenAI's GPT4, which is commonly reported as having over a trillion parameters.

Most LLMs tested were found to consistently match human judgement of reports for most rubric criteria. This is largely due to the narrow scope of most of the criteria, e.g.: *"the abstract describes the reports aim"* and *"the introduction presents motivation/background"*. Exceptions to this agreement were found for rubric criteria where relevant context/human knowledge is difficult to embed into the prompt, such

as *"there is no misidentification of sources of error"*. It is also challenging to moderate how lenient an LLM should be for criteria such as *"references are formatted correctly"* where humans tend to overlook minor issues such as misplaced punctuation.

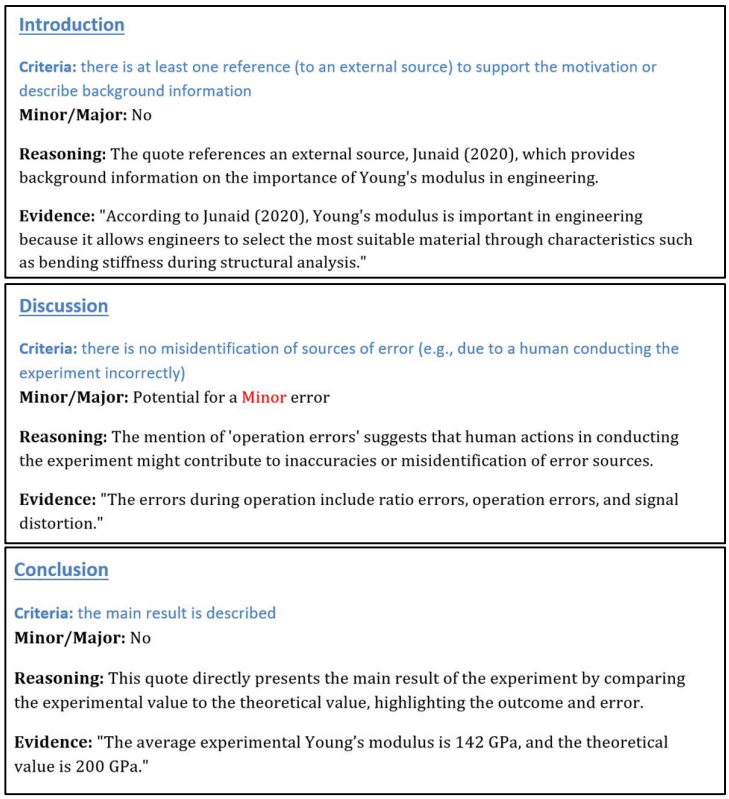

Figure 2. Three examples of AI-generated feedback provided to students as part of documents covering the 33 criteria in the rubric

3.3 Live Feedback Review Session

Placing review of the AI feedback in a live session and framing it against (different but also fallible) peer feedback was felt to be an important part of this intervention's success. Immediately after students reviewed peers' reports a short framing talk on responsible use of AI and its ethical and environmental issues was delivered. Students' AI-generated feedback was then emailed to them, and they were given ~30 minutes to critically review it in groups of ~3, with members of staff in the room actively encouraging and engaging in discussion of how best to interpret it.

As mentioned above, every student's feedback included one or two intentional inaccuracies. These were included to encourage critical review of the feedback over passive acceptance. With the knowledge that both intentional and unintentional inaccuracies were likely to be present, the review task for students was to assign each feedback point with one of the following responses:

a) Agree
b) Disagree as the reasoning and evidence do not follow from the contents of their report (e.g., the LLM's evidence is "hallucinated")
c) Believe the LLM's reasoning and evidence to follow from their report but still disagree with the decision (e.g., because the evidence is not sufficient to demonstrate pass/fail on the criteria)

4 STUDENT SURVEY RESPONSES

Evaluation of intervention impact focused on student perceptions of usefulness and self-reported engagement with the feedback. A post-intervention survey was distributed to the full cohort (830 invited) and received 32 responses. Table 1 describes the questions asked and responses given.

Table 1. Post-intervention survey questions and responses (n = 32 respondents)

Question	Responses
Did you consider the assessment of your work by the AI tool to be useful?	Very useful - 4% Quite useful - 67% Not very useful - 22% Not useful at all - 7%
Did you make any changes to your report based on the AI feedback?	Lots of changes - 0% Some changes - 89% No changes at all - 11%
Has this AI feedback made you more/less likely to use AI to produce feedback for yourself in the future?	More likely to use AI for feedback - 33% No change - 48% Less likely to use AI for feedback - 19%
Since getting this AI feedback, do you think you will independently verify AI outputs…	More often - 22% As often as before - 71% Less often - 7%
Given your experience of both peer review and AI feedback, which of these enhanced your understanding and learning the most?	AI feedback - 7% The combination of AI and peer review - 52% Peer review - 41%
How many more opportunities for AI feedback would you like to see in this degree?	Lots more opportunities - 26% Some more opportunities - 56% Less opportunities - 15% AI should be avoided as a feedback tool - 3%
What were the key reasons / considerations for not opting for AI feedback? - A*sked only if student reported not opting in to AI feedback. No other questions asked to these students.*	*Free text, 4 respondents:* - Wanted staff feedback only (x 2) - Didn't expect to be useful (x1) - Strongly opposed to AI from ethical standpoint (x1)
What did you think about using AI for obtaining feedback on your report?	*Free text, 25 respondents, manual thematic analysis:* - Concerns about inaccuracies / hallucinations (x 14) - Helpful for spotting issues quickly (x 11) - Preference for staff feedback (x 3) - Helpful as a supplementary tool (x 3)
Please describe your experience of reviewing the AI output and evaluating hallucinations	*Free text, 23 respondents, manual thematic analysis:* - Inaccuracies / hallucinations were common (x10) - Experience general useful / positive (x7) - Confused by contradictions between criteria (x6) - Easy to spot inaccuracies / hallucinations (x4) - AI appeared to be scanning for key phrases (x3) - Interesting as an academic exercise (x1)

Despite the low response rate, the survey provides an indication that the intervention provided students with a useful supplementary source of feedback. A majority of respondents (71%) reported that the AI feedback was useful, and nearly all respondents (89%) reported making at least some changes to their reports based on the feedback. It is interesting to note that a majority of respondents (52%) reported the combination of AI and peer review feedback enhanced their learning more than either alone, and 82% of respondents wanted to see more opportunities for AI feedback in their degree.

The free text responses were manually coded for sentiment and theme. These contained requests for feedback from staff members but also comments on the usefulness of AI feedback for spotting issues quickly. There were negative responses towards the (intentionally included) hallucinations but also some that found them easy to spot and one response describing it as a 'useful academic exercise'. Pass rate on the subsequent summative report was 70% for those receiving AI feedback versus 57% for those who did not, though this difference remains confounded by known correlations between engagement with formative activities and overall attainment.

5 CONCLUSIONS

This paper demonstrates how Gen-AI can augment feedback opportunities for a large undergraduate cohort. In this pilot implementation, LLM-generated, rubric-aligned feedback was reviewed by students that opted-in to receive it, in a facilitated live session. The feedback was produced in under 48 hours (processing taking ~60 s per report on local hardware), was judged useful by most survey respondents (71%) and prompted revisions to work (89%).

Qualitative analysis of the AI feedback found good agreement with human judgement, with some limitations where rubric criteria required background/context to understand fully. An open dialogue in the live feedback review session created an opportunity for these limitations to be visible to students. To encourage scrutiny over passive acceptance of Gen-AI outputs one or two intentional inaccuracies were included in each student's feedback. This had mixed reception in the cohort but alongside framing of AI feedback against (also fallible) staff and peer review this process was positioned to build critical thinking and AI literacy.

Looking ahead, the ECMDA teaching team plan to further embed AI literacy and critical thinking in the unit by (i) exploring Vision–Language Models as a route to covering rubric criteria considering figures, tables, and document layout and (ii) running a workshop in which students are given live access to a locally run LLM to assist in critique of their report drafts.

REFERENCES

[1] Hattie J. and Timperley H. The power of feedback. *Rev. Educ. Res.*, vol. 77, no. 1, pp. 81–112, 2007.

[2] Shute V. J. Focus on formative feedback. *Rev. Educ. Res.*, vol. 78, no. 1, pp. 153–189, 2008.

[3] D. J. Nicol and D. Macfarlane-Dick, "Formative assessment and self-regulated learning: A model and seven principles of good feedback practice," *Stud. High. Educ.*, vol. 31, no. 2, pp. 199–218, 2006.

[4] Winstone N. E., Nash R. A., Parker M. and Rowntree J. Supporting learners' agentic engagement with feedback: A systematic review and a taxonomy of recipience processes. *Educ. Psychol.*, vol. 52, no. 1, pp. 17–37, 2017.

[5] Carless D. and Boud D. The development of student feedback literacy: enabling uptake of feedback. *Assess. Eval. High. Educ.*, vol. 43, no. 8, pp. 1315–1325, 2018.

[6] Selwyn R. K., Ross J. M. and Lancastle S. A. Perceptions and achievement of engineering undergraduates using peer review in group sessions to improve their report writing skills. *Eur. J. Eng. Educ.*, pp. 1–23, 2025.

[7] Ertmer P. A. et al. Using peer feedback to enhance the quality of student online postings: An exploratory study. *J. Comput. Commun.*, vol. 12, no. 2, pp. 412–433, 2007.

[8] Lee S. S. and Moore R. L. Harnessing Generative AI (GenAI) for Automated Feedback in Higher Education: A Systematic Review. *Online Learn.*, vol. 28, no. 3, pp. 82–106, 2024.

[9] Narreddy C., Joordens S. and Prompiengchai S. Harnessing Large Language Models for Scalable and Effective Formative Assessment in Higher Education: A Review. *Trends High. Educ.*, vol. 4, no. 4, p. 65, 2025.

[10] Steiss J. et al. Comparing the quality of human and ChatGPT feedback of students' writing. *Learn. Instr.*, vol. 91, p. 101894, 2024.

[11] Peláez-Sánchez I. C., Velarde-Camaqui D. and Glasserman-Morales L. D. The impact of large language models on higher education: exploring the connection between AI and Education 4.0. In *Frontiers in Education*, Frontiers Media SA, 2024, p. 1392091.

[12] García-López I. M. and Trujillo-Liñán L. Generative Artificial Intelligence in Education: Ethical Challenges, Regulatory Frameworks and Educational Quality in a Systematic Review of the Literature. In *Frontiers in Education*, Frontiers, 2025, p. 1565938.

[13] Zhan Y. and Yan Z. Students' engagement with ChatGPT feedback: Implications for student feedback literacy in the context of generative artificial intelligence. *Assess. Eval. High. Educ.*, pp. 1–14, 2025.

[14] Kamar E. Introducing Phi-4: Microsoft's Newest Small Language Model Specialising in Complex Reasoning. Dec. 12, 2024. [Online]. Available: https://techcommunity.microsoft.com/blog/azure-ai-foundry-blog/introducing-phi-4-microsoft's-newest-small-language-model-specialising-in-comple/4357090