Anatomy of Assessment

Using AI to evaluate feedback and improve practice in a multi-marker module

Faculty: Medicine
Department: Faculty of Medicine Centre
Module name: Lab Pod 2
Programme name: BSc Medical Biosciences (BMB)
Level: UG
Approximate number of students: 155
Module type: Core

Case study summary

This case study used OpenAI GPT models to evaluate (not generate or mark) feedback on a formative scientific-report assessment in a Year 2 BSc Medical Biosciences module. Bespoke Python scripts extracted anonymised Turnitin comments, structured them in spreadsheets and submitted batches through the OpenAI API. The API was chosen over a chatbot to enable efficient, reproducible analysis. Using iteratively refined prompts, grounded in educational criteria and marker guidance, AI classified comments for strengths, improvement areas, actionability, tone and feedback type. Academic staff reviewed outputs and retained judgement. The project addressed variation in feedback quality across a ten-person marking team, time pressures and quality-assurance burden.

Overview and rationale

Providing feedback within a marking team of about 9–10 markers can pose challenges that impact the consistency, fairness, usefulness, and timeliness of student feedback.

1. Consistency of Feedback

When multiple markers are involved in marking an assessment, there is a risk that students receive feedback that varies in quality, detail, tone, and focus. Although grades may generally be consistent through moderation procedures, students receiving comparable scores might be provided with markedly different types and quantities of feedback, which can depend on the marker. Such variability can lead to perceptions of unfairness and negatively impact the student experience. Differences in feedback practices may also result in inconsistent messages about performance expectations and areas for improvement.

2. Time Pressures

Usually, assessment feedback is to be returned within strict institutional deadlines. Combined with high marking volumes, these time constraints can result in feedback that is brief, generic, or primarily focused on justifying grades rather than supporting learning and development. Consequently, opportunities to provide meaningful feedforward and actionable recommendations may be limited. As a result, students may be less likely to engage with the feedback or use it to improve future performance.

3. Quality Assurance

Ensuring that all markers provide feedback that is constructive and aligned with the assessment rubric and marking criteria can be challenging, particularly in large marking teams. Monitoring feedback quality across markers is often time-consuming and resource-intensive when undertaken manually. Furthermore, as marking progresses over time, there is potential for drift in marking standards and feedback quality. Achieving a shared understanding of effective feedback practices can also be difficult, as markers may hold differing views regarding the purpose, style, and content of feedback.

Dr Nitya Krishnan/ Emma Blyth
Senior Teaching Fellow/ Senior Instructional Designer

Anatomy of Assessment

Using AI to evaluate feedback and improve practice in a multi-marker module

Aims

The evaluation sought to investigate:

1. The extent to which feedback practices were consistent across markers.
2. Whether comments included key elements of effective feedback, such as identification of strengths, areas for improvement, feedforward guidance, and actionability.
3. The extent to which the tone of feedback was aligned across markers, including its constructiveness.
4. The types of feedback provided, including evaluative, descriptive, corrective, and feedforward-oriented feedback.
5. Ways to improve feedback practices to ensure they are constructive and beneficial for students, helping them improve subsequent linked assessments and contributing to the overall student learning experience.

AI tools are particularly effective for this type of evaluation, as they can analyse large volumes of qualitative text and systematically identify patterns across various indicators. This is especially useful when there are many feedback comments to review. Manually analysing such a large amount of data would be time-consuming and could lead to inconsistencies. One future goal for this analysis is to improve scalability and extend this approach to other modules within the programme. Using AI tools will support this implementation.

An analysis of marker feedback was conducted on a formative assessment in a Year 2 undergraduate core module, Lab Pod 2, on the BSc Medical Biosciences (BMB) programme, comparing feedback provided during the 2025 and 2026 academic years. Formative assessment for the Lab Pod 2 module's ICA includes a scientific report featuring the sections typical of a primary research article (Abstract, Introduction, Methods, Results, Discussion). This formative is linked to a summative assessment that measures students' skills in presenting the findings of their own research project in the format of a scientific report. In-text feedback is provided for each section, and comprehensive general feedback offers a summary evaluation of the report. Both sets of comments were analysed using AI-assisted thematic categorisation.

Formative assessment was selected because it provides extensive comments intended to improve the linked summative assessment. Therefore, improvements to feedback practice can enhance student engagement with feedback and have an immediate positive effect on their performance in the summative assessment. The purpose of the analysis was to identify examples of effective feedback practice and areas for enhancement, which subsequently informed revisions to marker guidance.

The choice of AI

This project did not use AI to mark student work or generate feedback. Instead, AI was used as a quality-enhancement tool to analyse feedback that had already been written by academic staff. The AI classified feedback comments according to predefined educational criteria. The findings were used to evaluate the consistency and quality of feedback, identify examples of effective practice and areas for improvement, and inform marker training and revisions to marker guidance.

OpenAI GPT models were used in two ways:

1. In the development of Python scripts, which were used in the workflow to extract anonymised feedback comments from Turnitin PDF reports, convert them into structured spreadsheets, and send them in batches for analysis via OpenAI's API.
2. The OpenAI API was used to classify the extracted feedback comments according to set criteria. This enabled large-scale analysis of feedback characteristics and patterns across the marking team.

OpenAI's API was chosen rather than a standard chat interface because it could be integrated directly into the Python workflow. The API route enabled us to automate the processing of thousands of feedback comments, with the python scripts extracting comments from Turnitin reports, sending them to the AI in

Dr Nitya Krishnan/ Emma Blyth
Senior Teaching Fellow/ Senior Instructional Designer

Using AI to evaluate feedback and improve practice in a multi-marker module

manageable batches, and automatically compiling the results into spreadsheets for analysis. This approach was considerably more efficient, consistent and reproducible than manually copying comments into a chat interface.

Implementation

The model's classification of the feedback comments was guided by predefined educational criteria and the existing marker guidance for the module. This ensured that the analysis reflected both evidence-based feedback practices and the expectations already established for markers.

The AI was provided with a structured classification prompt that defined each feedback category using explicit criteria and decision rules. The model was instructed to classify comments consistently according to predefined educational definitions and return the results in a structured format. The prompts were developed iteratively. Initial classifications were tested on small batches of feedback comments and manually reviewed to identify inconsistencies or ambiguities. Based on this review, the category definitions and decision rules were refined before the full dataset was processed. This iterative approach helped improve the consistency and reliability of the automated classifications.

The workflow was designed to be secure, transparent, consistent and reproducible. All data extraction and preprocessing took place locally using bespoke Python scripts, with only the extracted feedback comments submitted for AI classification via the OpenAI API. The category definitions and decision rules were developed iteratively and refined before processing the full dataset to improve the consistency and reliability of the classifications. Academic judgement remained central throughout, with AI supporting analysis.

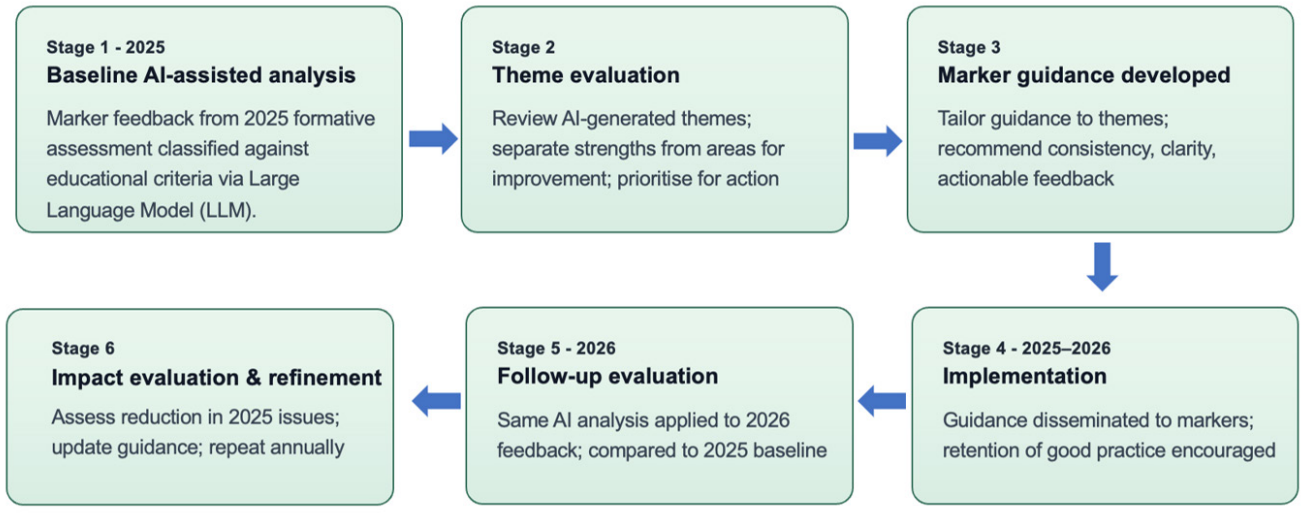


Figure 1. Project workflow

Compliance with the revised marking guidance was as expected. To ensure compliance, when the guidance was revised based on AI-supported analysis, a rationale was provided to clarify the reasons for these changes. This helped markers understand the purpose of the updates, ensuring their adherence to the new guidance.

The main risk was the potential for AI to misclassify qualitative feedback where comments contained nuanced language or could reasonably be interpreted in more than one way. Considerable time was therefore spent

defining categories that could be applied consistently and refining the prompts through iterative testing and manual review. This helped maximise the reliability of the classifications while recognising that human interpretation remained essential when drawing conclusions.

Developing the workflow required an initial investment of time to create the Python scripts and, in particular, to define and refine the feedback classification categories and decision rules. Once established, however, the workflow becomes highly reproducible, with subsequent analyses requiring relatively little preparation beyond extracting the Turnitin reports and running the existing scripts.

The calibration meeting was 60 minutes, which included presenting the guidance to markers and the findings of the evaluation study.

Currently, similar AI-supported evaluations of feedback provision are not performed within other BMB modules. However, discussions are ongoing with the Deputy Head of Quality (BMB) to determine the best approach for a programme-wide implementation.

2. Impact

The analysis of marker feedback on the formative assessment (2025) revealed several areas of strong practice within the module. Example: markers provided overarching comments that identified at least one area of strength (98%) and one area for improvement (99%). Furthermore, 93% of feedback comments included specific advice designed to help students improve future performance. This reflects a generally strong alignment with evidence-based principles of effective feedback. Research indicates that feedback is most effective when it not only highlights students' strengths and weaknesses but also offers specific guidance on how to improve and achieve learning goals (Hattie & Timperley, 2007; Carless & Boud, 2018). Good feedback should help students understand what steps to take next by providing practical suggestions that can be applied in future learning.

93% comments included actionable advice (2025)	**30% ⟶ 24%** hedged/indirect comments reduced after guidance update	**7% ⟶ 0%** person-focused (vs work-focused) comments eliminated	**98% / 99%** of comments named a strength / an area for improvement (2025)

Figure 2. Measured impact of the 2025 ⇢ 2026 cycle

One notable area for development identified in the 2025 analysis was the use of hedging language. Approximately 30% of comments contained hedged expressions such as "perhaps," "you can try," or "you could include." While hedging is often used as a politeness strategy to soften criticism, excessive indirectness can reduce the clarity of feedback. This issue may be particularly relevant for international students and multilingual learners, who may find implicit or indirect feedback more difficult to interpret. Research on formative feedback indicates that hedged or vague comments can lead to less understanding, lower confidence, and fewer successful revisions, as students—particularly international students—may be unsure about the specific actions they need to take (Ryoo, 2023).

In response to this finding, marker guidance was revised to encourage greater clarity and directness when communicating areas for development, while still maintaining a constructive tone. The rationale underpinning this guidance was shared with markers. Subsequent analysis of the 2026 formative feedback demonstrated a 5.8% reduction in the proportion of hedged comments. Although this reduction appears modest, it represents evidence of markers engaging with the revised guidance and beginning to modify established feedback

Dr Nitya Krishnan/ Emma Blyth
Senior Teaching Fellow/ Senior Instructional Designer

practices. Given that changes in assessment and feedback culture often occur incrementally, continued monitoring over subsequent years will be important to evaluate whether these improvements are sustained and further embedded within marking practice.

A second area identified in the 2025 analysis concerned the focus of feedback comments. Approximately 7% of comments were directed towards the student as a person rather than towards the work itself. Examples included comments that made judgments about effort or ability, rather than the quality of the submitted work. This distinction is significant because feedback that targets the individual rather than the task, as Hattie and Timperley (2007) point out, tends to be less effective for learning. Self-focused feedback offers limited guidance on how to improve performance compared to feedback aimed at the task, process, or self-regulation (Mandouit &Hattie, 2023).

Consequently, the marker guidance was amended to explicitly emphasise that feedback should focus on the work submitted rather than on the student as a person. The impact of this change was evident in the analysis of the 2026 formative feedback, which found that 100% of comments were work-focused. This represents a substantial improvement and suggests that the combination of evidence-informed reflection and explicit guidance successfully influenced feedback practice.

Overall, the analysis demonstrates how systematic evaluation of assessment feedback can support feedback quality enhancement within modules. While existing feedback practices were largely effective, the findings identified specific opportunities to increase clarity and student-centredness. The resulting changes to marker guidance have already led to measurable improvements, particularly in reducing person-focused comments and decreasing the use of hedging language. Continued annual review will help determine whether these changes become fully embedded and continue to support student engagement with feedback.

This evaluation has supported reflective practice, encouraged consistency across the marking team, and contributed to ongoing improvements in feedback provision to align better with recognised principles of effective feedback provision to provide a better student learning experience.

3. Final reflection

Next time, additional aspects could be investigated, such as alignment of feedback comments with the rubric, pinpointing which components of the assessment receive the most and least feedback, evaluating the readability and accessibility of the comments, and exploring variations in feedback volume across different markers. These additions will provide a more comprehensive understanding of feedback practice within the module.

A key personal learning point from this process was discovering the extent to which feedback practices are aligned within this core teaching module and areas where there are variations. The variations have occurred even though markers are working from the same marker instructions, rubric and assessment criteria. This highlighted the importance of ongoing calibration, shared discussions about effective feedback, and regular review of feedback practices to ensure consistency and equity for students. The evaluation also demonstrated the potential of AI-assisted analysis to support quality assurance by identifying patterns and trends that may not be readily apparent through manual review alone. As a result, I have developed a greater appreciation of the need to evaluate the quality, tone, and developmental value of the feedback provided to students.

It is important to view AI as a tool to support, rather than replace, academic judgement. AI can efficiently analyse large volumes of feedback and identify patterns relating to consistency, tone, and feedback type. However, human interpretation remains essential to provide context and determine the educational significance of the findings. A useful starting point is to establish clear evaluation criteria based on recognised principles of effective feedback before undertaking the analysis. This ensures that AI is assessing feedback

Dr Nitya Krishnan/ Emma Blyth
Senior Teaching Fellow/ Senior Instructional Designer

Anatomy of Assessment

Using AI to evaluate feedback and improve practice in a multi-marker module

against meaningful pedagogical standards rather than simply generating descriptive outputs.
It is also important to anonymise student and marker information where appropriate and be transparent about how AI is being used within the evaluation process.

Embracing the evaluation findings with an open mind is also essential, as is clearly communicating the rationale and motivations behind them. All the markers need to be engaged and supportive of the process, which facilitates the effective downstream implementation of the findings.

Generative AI made it possible for us to develop the workflow by guiding us through the process of creating and refining Python scripts for each stage. The more important expertise was pedagogical rather than technical – understanding what constitutes effective feedback, defining meaningful classification categories, and critically evaluating the AI outputs. Human oversight remained essential throughout, both to refine the classification criteria and to recognise the limitations of AI, including the potential for misclassification of nuanced feedback comments.

References

- Carless, D., & Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. Assessment & Evaluation in Higher Education, 43(8), 1315-1325. https://doi.org/10.1080/02602938.2018.1463354
- Hattie, J., & Timperley, H. (2007). The power of feedback. Review of Educational Research, 77(1), 81-112. https://doi.org/10.3102/003465430298487
- Mandouit,L., & Hattie,J. (2023). Revisiting “The Power of Feedback” from the perspective of the learner. Learning and Instruction,84. https://doi.org/10.1016/j.learninstruc.2022.101718.
- Ryoo, S. M. (2023). Efficacy of hedges in formative feedback on L2 writing. International Journal on Studies in Education, 5(4), 550-567. https://doi.org/10.46328/ijonse.171

Dr Nitya Krishnan/ Emma Blyth
Senior Teaching Fellow/ Senior Instructional Designer