

Article

Prompting Better Feedback: A Study of Custom GPT for Formative Assessment in Undergraduate Physics

Ellie Mills, Arin Mizouri *  and Alex Peach

Department of Physics, Durham University, Durham DH1 3LE, UK; a.m.peach@durham.ac.uk (A.P.)

* Correspondence: arin.mizouri@durham.ac.uk

Abstract

This study explores the use of a custom generative AI (GenAI) tool, built using a prompt-engineered instance of ChatGPT, to provide formative feedback on first-year undergraduate physics lab reports. A preliminary survey of 110 students identified writing style as an area of low confidence and highlighted strong demand for more actionable, detailed feedback. Students expressed greater comfort with GenAI in formative contexts, particularly when used alongside human assessors. The tool was refined through iterative prompt engineering and supported by a curated knowledge base to ensure accuracy, clarity, and pedagogical alignment. A mixed-methods evaluation with 15 students found that the feedback was useful, actionable, and clearly written, with particular praise for the suggested improvements and rewritten exemplars. Some concerns were raised about occasional inaccuracies, but students valued the tool's structure, consistency, speed, and potential for interactive follow-up. These findings demonstrate that, when carefully designed and moderated, GenAI can serve as a valuable, scalable support tool within the broader formative assessment cycle for long-form scientific writing. The tool's flexibility, clarity, and responsiveness highlight its value as a supportive resource, especially as generative AI technologies continue to evolve in educational contexts.

Keywords: formative assessment; generative AI; large language models; physics education; GPT; feedback tools; lab reports



Academic Editor: Alison Voice

Received: 23 June 2025

Revised: 12 July 2025

Accepted: 16 July 2025

Published: 19 August 2025

Citation: Mills, E., Mizouri, A., & Peach, A. (2025). Prompting Better Feedback: A Study of Custom GPT for Formative Assessment in Undergraduate Physics. *Education Sciences*, 15(8), 1058. <https://doi.org/10.3390/educsci15081058>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Generative artificial intelligence (GenAI), especially large language models like ChatGPT, have advanced at a staggering pace in recent years. These widely-available tools can produce contextually relevant, structured output across a wide range of domains, quickly on demand, indicating the potential for transformative change across many institutions. While the conflict between AI and traditional forms of assessment occupies the foreground, we must also examine the enormous potential of GenAI as a positive, transformative technology for education (Cooper, 2023; El Fathi et al., 2025; Zambon et al., 2024). Previous work has already shown that generative AI models like ChatGPT can produce short-form physics essays graded at first-class level, raising foundational questions about the future of written assessment and authorship in science education (Yeadon et al., 2023). Meanwhile, the ingress of AI within the public domain underscores economic and institutional pressures that are rapidly driving the adoption of GenAI technologies, often ahead of clear evidence regarding their efficacy or ethical ramifications. Educational institutions, already facing tightening budgets and rising demands, may feel compelled to deploy these tools for teaching or assessment due to their novelty and cost-effectiveness (An et al., 2025; Dotan

et al., 2024; Ivanov et al., 2024). We respond to this pressure with caution, offering a timely contribution to the growing body of work that seeks to evaluate the real-world use of GenAI in education.

Our main goal in this work is to assess the potential for GenAI to be utilised effectively for the provision of feedback on student work, specifically with regard to scientific lab report writing for first-year undergraduate physicists. Written feedback plays a crucial role in developing students' scientific reasoning, communication skills, and conceptual understanding. However, as undergraduate physics cohorts grow in size, academic staff face increasing pressure to deliver timely, detailed, and high-quality feedback. In response to this challenge, there is growing interest in the potential for GenAI tools to support or augment the feedback process, particularly within the broader cycle of formative assessment.

Existing studies have explored the use of GenAI in assessing undergraduate physics examinations and conceptual problem responses. However, considerably less attention has been given to its suitability for evaluating extended, subject-specific scientific writing. Undergraduate physics lab reports represent a particularly demanding test case in this regard, as they require not only clarity of expression but also an understanding of experimental design, data interpretation, and scientific reasoning. These features make them a valuable use case in which to assess both the capabilities and the limitations of GenAI for longer-form assessment in physics education.

In this work, we investigate the use of a custom instance of ChatGPT (using the GPT-4o model) to provide formative feedback on first-year university physics lab reports. These reports were chosen as an initial focus due to their relative conceptual simplicity and the fact that students at this level often lack confidence in academic writing and so the greatest immediate benefit of feedback intervention may be seen. The tool was developed using a custom GPT from OpenAI, deployed via a user-facing, prompt-engineered interface aligned with course objectives and designed to generate subject-specific, structured feedback. This distinguishes the study from previous work, establishing a scalable foundation for future use in more advanced assessments. An exploratory survey of student perspectives on current feedback practices and AI in assessment was initially conducted to guide the design and relevance of the study. This was motivated in part by previous findings in UK physics education, which highlighted widespread use of GenAI tools for problem solving and conceptual support, but a general lack of trust in their role as evaluators (Zambon et al., 2024).

In this work we attempt to address the following key questions:

- How effectively can GenAI evaluate and provide feedback on longer-form physics content?
- What are the strengths and limitations of GenAI as a feedback provider, particularly in comparison to current assessor feedback?
- What are students' perceptions of GenAI's role in assessment, and how do these influence their evaluation of the tool?

Through this, we investigate how GenAI might support scalable, high quality feedback in large undergraduate courses while maintaining academic integrity and pedagogic value. Our findings also convey valuable insights regarding students' trust in AI-generated feedback, a key ethical and practical concern for future implementation.

1.1. Generative Artificial Intelligence (GenAI)

1.1.1. Why ChatGPT?

In late 2022, GPT-3.5 was released, alongside a public interface known as ChatGPT, which provided a conversational gateway to generative AI for the general public. Within 2 months ChatGPT surpassed 100 million users (Ebert & Louridas, 2023). The release of

GPT-4 in 2023 marked a significant leap in capabilities over GPT-3.5 (Ray, 2023). This new technology became accessible through the ChatGPT Plus subscription, offering users enhanced reasoning and compositional abilities.

In November 2023, OpenAI introduced a system called “GPTs”—custom versions of ChatGPT that allow users to define specific behaviours, upload documents, and assign persistent task-specific roles (OpenAI, 2023). These are sometimes referred to informally as “mini-GPTs” in user communities, though this is not an official designation. In this work, we utilise the custom GPT interface to develop a specialised feedback tool capable of evaluating undergraduate physics lab reports in alignment with institutional marking criteria and pedagogical goals.

While ChatGPT initially dominated public engagement with large language models, a growing number of alternative systems have since emerged, some surpassing it in specific capabilities. The choice of ChatGPT in this study is not intended to imply its superiority; rather, it reflects its accessibility, widespread adoption, and support for customisation. What is central to this work is not the specific model used, but the evaluative and prompting approach we take, which can be applied broadly to assess the efficacy of generative AI tools in educational contexts.

1.1.2. Benefits and Limitations

Generative AI offers a range of potential benefits when applied to formative assessment, particularly in high-enrolment subjects like physics (Yeadon & Hardy, 2024; Zambon et al., 2024). One of the most widely recognised benefits of GenAI is its accessibility and availability (Cooper, 2023). Unlike human assessors, who are constrained by institutional schedules and workload, GenAI can offer interaction and assistance at any time, providing continuous support to students (Ray, 2023). This level of accessibility significantly reduces barriers for learners who may not have regular access to feedback. This issue is particularly pronounced in higher education, where assessors may not have the capacity to engage directly with each student—especially where student numbers exceed the practical capacity for individualised human feedback. Accordingly, a lack of accessibility to individualised feedback undermines the wider process of formative assessment, of which individualised feedback is a core component.

Scalability and consistency are further strengths of GenAI (Yeadon et al., 2024; Zhang et al., 2025). Once configured, a GenAI system can apply the same evaluative criteria uniformly across all submissions, ensuring that each student receives feedback aligned with identical standards. This capability addresses a common student concern regarding the perceived subjectivity and variability of human marking. In addition to promoting consistency, GenAI tools are highly scalable and have already been widely adopted in other industries to manage large-scale, repetitive tasks efficiently (Zhang et al., 2025). This is particularly relevant in lower-level laboratory settings, where students often perform the same experiments year on year, and feedback tends to follow recurring patterns, either because students encounter regular pain points or because the expected outcomes are well-established. In theory, GenAI tools should be capable of processing and responding to large volumes of student work without compromising quality or accuracy. This makes them particularly well suited to higher education contexts, where expanding student cohorts place substantial strain on academic staff and limit opportunities for individualised feedback. By automating routine elements of the feedback process, GenAI can help address this bottleneck.

Timeliness is another key advantage of GenAI-based assessment. Unlike traditional formative and summative assessment processes which are strongly constrained by marking schedules and staff workload, GenAI can provide immediate responses during the learning

process (Hattie & Timperley, 2007; Wan & Chen, 2024). This accelerates feedback cycles and enables students to reflect on and apply feedback while their work is still recent and relevant. Real-time responsiveness is already a well-established feature in other applications of GenAI, such as customer service, where on-demand interaction enhances user engagement and satisfaction (Zhang et al., 2025). This timeliness of GenAI enables for the provision of automated, personalised feedback on written assignments, potentially conferring a drastic improvement to the speed and efficiency of the feedback component within formative assessment. This is perhaps more pertinent within STEM, where written assignments, such as scientific report writing, are substantially more prescriptive in nature compared with the humanities.

One of the key strengths of generative AI is its ability to be tailored to specific contexts through fine-tuning (Reynolds & McDonell, 2021; Schulhoff et al., 2024; Sortwell et al., 2024; Wei et al., 2022). Customisation is also possible through prompt engineering, which can, in some cases, offer even greater control over model behaviour and output, particularly for educational applications. These techniques allow developers to adapt a model's responses with task-specific goals, improving both its relevance and pedagogical value (Ray, 2023). While we focus on generating feedback for first-year student lab reports, the broader adaptability of GenAI positions our tool as a flexible test case. By modifying the underlying instructions and knowledge base, the same system could be reconfigured to evaluate other forms of student work such as, in the first instance, higher-level lab reports. This versatility makes GenAI particularly valuable in educational environments where assessment needs vary widely by subject, level, and institution.

Despite these advantages, there are notable limitations. Accuracy remains a critical concern, as GenAI models can produce plausible-sounding but factually incorrect outputs, commonly referred to as hallucinations (Blank, 2023; Cooper, 2023). These errors occur when the model generates content not grounded in its training data or the input it receives, resulting in outputs that lack factual basis. Because such responses generically carry an authoritative tone, users may be more susceptible to accepting hallucinations at face value without verification. This poses a substantial risk in educational settings, where incorrect feedback may be accepted uncritically and subsequently internalised (Ray, 2023).

Another major challenge in employing GenAI for educational assessment is its lack of transparency and explainability (Blank, 2023; Dotan et al., 2024). GenAI models, such as ChatGPT, operate through complex neural architectures that are difficult to interpret. This opacity can impede efforts to understand how specific evaluative decisions are made, thus limiting the ability to assess the fairness and accuracy of feedback. Such concerns are particularly pertinent in high-stakes educational settings, where students rely on transparent and accountable grading processes to guide their academic progress. Without clear insight into how AI models arrive at their conclusions, their use in formal assessments risks undermining educational credibility (Ray, 2023).

Bias is an additional area of concern. AI models like ChatGPT are trained on vast, diverse datasets that may contain cultural, linguistic, and social biases. These can inadvertently be reflected in the feedback provided, potentially leading to differential treatment of students based on factors such as language or cultural references (Cooper, 2023; Dotan et al., 2024). In assessment context, this risks disadvantaging students from non-dominant backgrounds and exacerbating existing inequalities in educational outcomes (Zhang et al., 2025). To address this issue, it is crucial to ensure that training data is diversified, and mechanisms for detecting and mitigating bias are integrated into AI-assisted assessment tools. Although such concerns may be less pronounced in highly structured tasks like lab reports, they become more significant as AI tools are applied to a broader range of assessment formats. These risks are particularly relevant when student writing styles or

cultural cues deviate from the dominant norms embedded in the model's training data, raising important fairness concerns.

Finally, ethical and environmental concerns must be considered. The integration of GenAI into educational assessment raises a range of ethical concerns that extend beyond technical performance. A key issue is data protection, as in order to assess a piece of work, it must be processed by the GenAI tool. This prompts questions about consent, storage, and compliance with privacy regulations (Ray, 2023). While this is mitigated in this work by anonymisation of student data, it remains an important consideration for future applications. Additionally, the environmental impact of large-scale AI models is a growing concern, as these models require substantial computational resources, leading to high energy consumption and increased carbon emissions. This raises broader questions about the long-term sustainability of AI in education, particularly as institutions attempt to balance innovation with environmental responsibility (An et al., 2025; Dotan et al., 2024).

1.1.3. Past Works

GenAI has already been applied across a range of educational settings, as intelligent tutoring systems, personalised learning platforms, and adaptive assessment tools (Chassignol et al., 2018). Within physics education, most prior studies have focused on using GPT models to solve exam-style physics questions—evaluating the model's ability to generate correct answers rather than to assess student work (Sirnoorkar et al., 2024). These studies often highlight strengths in structured written responses and weaknesses in tasks involving mathematical computation, underscoring current limitations in relying on GenAI for assigning marks in subject-specific contexts, especially where deeper subject expertise is required. These findings reinforce the rationale for this study's focus on GenAI as a feedback generator rather than as a replacement for human assessors in grading.

Recent studies, however, have begun to assess GenAI's role in generating academic writing itself. For example, Yeadon et al. (2023) demonstrated that ChatGPT could produce short-form physics essays graded at first-class level, prompting concerns about the future of authorship and authenticity. Building on this, a follow-up study evaluated human and AI writing across a range of criteria, revealing challenges in benchmarking and interpreting quality in student-style prose (Yeadon et al., 2024). While useful, this body of work primarily investigates the performance of GenAI as a student, not as a marker or feedback provider.

More closely aligned with the current study, Wan and Chen (2024) investigated the use of GPT-3.5 to provide feedback on conceptual physics responses. Using few-shot learning and prompt engineering, they generated feedback that students rated as more useful than human-written alternatives. Importantly, approximately 70% of AI-generated comments required only minimal revision from instructors, suggesting strong potential for AI to reduce feedback workloads while maintaining pedagogical value.

The novelty of this work lies in our targeting long-form scientific writing in physics and by using a prompt-engineered, user-facing custom GPT model to generate formative feedback aligned with pedagogical goals. While prior studies demonstrate mixed success in applying GenAI to physics education, they collectively show that feedback-focused implementations may be more immediately valuable and appropriate than assessment-based ones. This is supported by recent work comparing GenAI—and human-authored physics essays, which found notable difficulties in defining and evaluating the quality of GenAI-generated writing, particularly when human assessors lacked consistent benchmarks (Yeadon et al., 2024).

To support this approach, the present study employed several distinct prompt engineering strategies. These techniques were used to shape the model's evaluative process, align feedback with domain-specific expectations, and scaffold clarity, consistency, and ped-

agogical relevance. Table 1 summarises the techniques implemented, their functions, and the rationale for their use in this educational context. These strategies draw on established prompting methods identified in recent surveys (Schulhoff et al., 2024; Wei et al., 2022).

Table 1. Prompt Engineering Techniques Used in the GenAI Feedback Tool.

Technique	Primary Function	Description	Purpose and Benefits in Educational Context
Chain of Thought (CoT)	Reasoning	Encourages step-by-step reasoning before answering.	Helps the model align with logical structures typical of lab report assessment.
Rephrase and Respond (RaR)	Tone Control	Requires the model to rephrase student input before evaluating.	Improves alignment with student writing and promotes clearer feedback phrasing.
System 2 Attention (S2A)	Deep Reflection	Simulates slow, deliberate cognitive processing.	Reduces superficial or overly generic responses by encouraging depth.
Few-shot Learning	Output Structuring	Incorporates worked examples into the prompt.	Establishes feedback style and tone consistency by modelling ideal responses.
Instruction-based Prompting	Behaviour Control	Uses explicit natural language directives for structure and constraints.	Ensures feedback format and scope are consistently followed.
Contextual Grounding	Content Anchoring	Conditions the model using domain-specific files (e.g., marking rubrics).	Anchors output in institutional expectations and scientific content norms.

1.2. Assessment and Feedback

While GenAI offers new possibilities for automating and scaling feedback, its effectiveness ultimately depends on aligning its use with established educational theory. A sound understanding of assessment principles, particularly those relating to formative feedback, is necessary to ensure that GenAI tools are implemented in ways that genuinely support student learning. While feedback is a critical mechanism within formative assessment, the process also includes the interpretation and use of that feedback to inform teaching and learning strategies. Figure 1 illustrates how a reliable and accurate GenAI tool might be integrated into existing student–teacher feedback loops, in an idealised case, complementing traditional feedback with scalable, iterative support within the cycle of formative assessment.

Scriven (1966) identifies formative evaluation as a method of measuring deficiencies in a curriculum during the process, with summative evaluation measuring the quality of what the process produced. These ideas were later adapted into formative assessments determining the student’s degree of mastery of a given task, as well as what can be changed to improve this mastery. More recently, formative assessment has been reframed as a continuous and dynamic process of gathering and using evidence to inform teaching and support student progress, rather than a single act of providing feedback (Sortwell et al., 2024). In contrast, summative assessments remained to evaluate the student’s total progress in the subject area, often through a grade or certification (Bloom, 1971). Later works, such as Sadler (1989)’s article on formative assessment, specifically highlight the importance of feedback as the “key element” for improving a student’s competence. Specifically, feedback loops are noted to be essential for satisfactory improvements, with students continually receiving feedback on their work and using that feedback to improve their performance. Within this system it is essential to have a teacher who can demonstrate a good example, and show how a bad example could be improved. With GenAI, students may now receive intermediate feedback and rewriting suggestions prior to or alongside teacher commentary,

potentially enhancing this loop. In recent years, formative assessment and feedback has been a main focus during discourse on educational assessment. While definitions vary, many choose to define feedback in terms of its effect: “Feedback is information about the gap between the actual level and the reference level of a system parameter which is used to alter the gap in some way” (Ramaprasad, 1983).

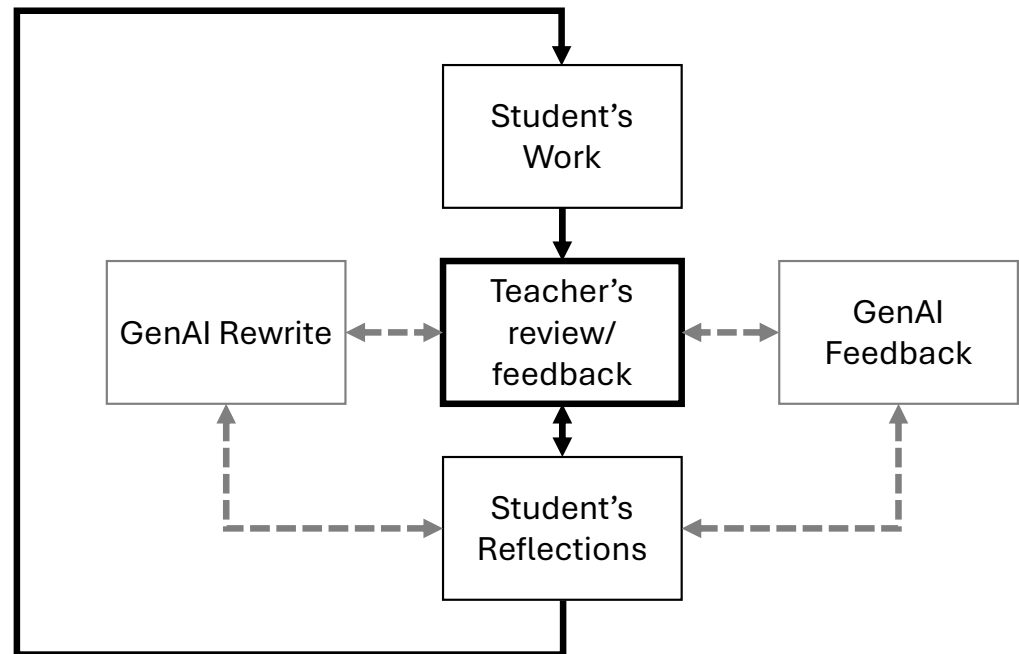


Figure 1. An idealised conceptual model of the student–teacher feedback loop with GenAI integration. Solid arrows represent the traditional feedback cycle, while dashed arrows indicate supplementary interactions enabled by the GenAI tool, including direct feedback on student work, support during reflection, and optional rewriting suggestions.

Many studies advocate for approaches to formative assessment such as Data Based Decision Making (DBDM) and assessment for learning (A4L) (Klenowski, 2009; Schildkamp & Kuiper, 2010). DBDM emphasises the systematic collection and analysis of data to inform teaching practices and guide the focus of feedback, while A4L promotes a continuous cycle of feedback through less formal methods such as classroom observation. Although both approaches offer valuable insights, their implementation is often constrained by the specific context in which assessment occurs. In higher education, and particularly in laboratory-based assessments such as lab reports, these models face significant challenges. The individual responsible for teaching (in this case, the laboratory supervisor) may not be the same person responsible for marking the report. This separation between teaching and assessment roles limits the opportunity for iterative feedback cycles and real-time instructional adjustments. This disconnection is especially problematic for feedback models reliant on sustained interaction and responsive dialogue, as it reduces the possibility of timely, context-aware intervention. In such contexts, GenAI may provide a continuous feedback touchpoint across disjointed teaching and assessment responsibilities. As such, the practical limitations of the university setting must be acknowledged when considering the applicability of these formative assessment frameworks.

Other studies place greater focus on feedback itself. Hattie et al. classify feedback into four levels (Hattie & Timperley, 2007):

- Feedback about the task (FT)—related to correctness and specific knowledge;
- Feedback about the processing of the task (FP)—related to learning strategies;
- Feedback about self-regulation (FR)—such as prompting self-evaluation;

- Feedback about the self (FS)—relating to personal praise.

Through a study of empirical evidence from a range of literature, [Hattie and Timperley \(2007\)](#) concluded that FS is the least effective, FR and FP aid in deep processing and task mastery, and FT helps when task information is useful for improving strategy processing. Similarly, [Lipnevich and Smith \(2009\)](#) classify feedback into three categories: detailed feedback, grades, and praise. Based on empirical evidence collected for the study, it was found that detailed, specific feedback is the most advantageous approach to formative feedback. Grades were found to decrease the effect of this detailed feedback, possibly because it reduced a sense of self efficacy. Both studies emphasise the importance of a sense of autonomy for the student to improve the effectiveness of feedback. In this regard, the ability for students to interact directly with GenAI to receive tailored rewrites and clarification may increase perceived autonomy and feedback utility.

There is some focus on factors that can affect feedback efficacy. Timing of feedback return was one focus of several studies. [Hattie and Timperley \(2007\)](#) found that for FT, immediate error correction was often beneficial, resulting in faster rates of skill acquisition. However, for FP, immediate feedback disrupted the learning of automaticity. In an empirical study, [Fyfe et al. \(2021\)](#) found no single benefit to receiving feedback immediately, but recognised that in certain kinds of classes, there may be advantages for delayed feedback. The source of the feedback was also a focus. [Lipnevich and Smith \(2009\)](#) found no significant difference in responses based on whether feedback was provided by a human or a computer, however it was found that participants rated instructor feedback as more helpful than computer generated feedback. The integration of GenAI into this loop introduces new timing dynamics: feedback can be near-instant and available on demand, supporting immediate reflection and revision, especially within the formative assessment process, where timely feedback supports iterative improvement.

This study was conducted in two phases. In Phase 1, we carried out a preliminary, exploratory survey with undergraduate physics students at all levels to understand their experiences with written feedback and their attitudes toward the use of generative AI in assessment. These insights informed the design of a subject-specific GenAI feedback tool, developed using OpenAI's custom GPT interface and based on the GPT-4o model. In Phase 2, we implemented and evaluated this tool in the context of a large first-year undergraduate physics lab module. This two-phase design allowed us to align tool development with student priorities, and to evaluate its effectiveness and reception in authentic educational settings.

The paper proceeds as follows: Section 2 presents Phase 1 of the study, reporting on student experiences and perspectives from the initial survey. Section 3 describes Phase 2, detailing the development of the GenAI feedback tool and the evaluation of its implementation. Section 4 concludes with recommendations for future work.

2. Phase 1: Preliminary Survey—Results and Findings

2.1. Method

To evaluate the potential role of GenAI in physics assessment, this study first sought to identify limitations in current feedback and marking practices, as well as undergraduate student perceptions of GenAI. These insights provide essential context for assessing where GenAI tools might offer meaningful improvements. An preliminary, exploratory survey was designed to address three core objectives: (1) identify which aspects of lab reports students find most difficult, (2) evaluate student satisfaction with current feedback (A member of staff provides written annotations on their report and provides written general feedback on the report as a whole. This is provided to students two weeks following the submission of their report. Within three weeks of the submission of their report, the

same staff assessor provides verbal feedback to the student during a ten-minute interview), and (3) explore levels of trust with AI-generated feedback and marks. Questions were developed through iterative drafts aligned with these aims and informed by the structure and expectations of Year 1 Physics lab reports. The survey was hosted on Jisc Online Surveys, conducted anonymously, and approved by the Physics Department's Ethics Committee. Informed consent was obtained from all participants.

The preliminary survey contained 27 questions in five sections: demographics (discipline and academic year), lab report challenges, feedback experiences, GenAI familiarity, and attitudes toward GenAI in assessment. Each section targeted specific facets of broader themes. For example, student satisfaction with feedback was broken down into perceived usefulness, length, turnaround time, and clarity. A mix of five-point Likert-scale statements and open comment boxes enabled both quantitative and qualitative data collection. The survey was distributed to all undergraduates enrolled in physics modules, and responses were collected over a four-week period spanning the end of the first term and the start of the second term in the 2024/25 academic year. A total of 110 students responded (88 Physics, 22 Natural Sciences), covering all year groups (35 Year 1, 22 Year 2, 24 Year 3, and 29 Year 4). Our sample of 110 students represents approximately 12% of the total cohort ($N = 930$). With this sample size, estimates carry a margin of error of $\pm 9\%$ at a 95% confidence level. While sufficient for exploratory insights into perceptions of GenAI feedback, caution is warranted when generalising findings beyond this cohort. The free text responses were manually reviewed to identify the recurring themes.

2.2. Results

For the analysis of Likert scale questions, each answer has been assigned a numerical value (strongly disagree = 1, slightly disagree = 2, neutral = 3, slightly agree = 4, strongly agree = 5). Due to the ordinal nature of Likert scale data, the median was used when calculating average responses. Friedman tests were used to compare Likert responses across multiple related questions (e.g., perceptions of feedback usefulness, clarity, and timing), while Mann–Whitney U tests were used to assess group differences (e.g., between students who had or had not used GenAI). Statistical significance for both tests was defined as a $p < 0.05$.

Unless otherwise stated, results presented in figures include responses from all undergraduate year groups (Years 1 to 4).

2.2.1. Opinions on Lab Report Writing

Participants were first asked to rate their agreement with the statement *"I am confident in my ability to complete this aspect to a high standard"* on a Likert scale. The analysis for the results presented in Figure 2 was limited to Year 1 students, as they were the primary users targeted by the GenAI tool. As shown in Figure 2, students were most confident in the results (77%), method (71%), and data analysis (60%) sections of the report. Conversely, the error appendix (25%) and written style (34%) sections showed notably lower confidence. These aspects also have the highest levels of student disagreement (40% and 37%). Interestingly, students were notably divided regarding their confidence in the discussion section, which received relatively high levels of both agreement and disagreement.

When asked to comment on common lab report challenges, Year 1 students highlighted difficulties such as uncertainty about the relevance of specific content, struggling with academic language, and unclear or insufficient teaching guidance.

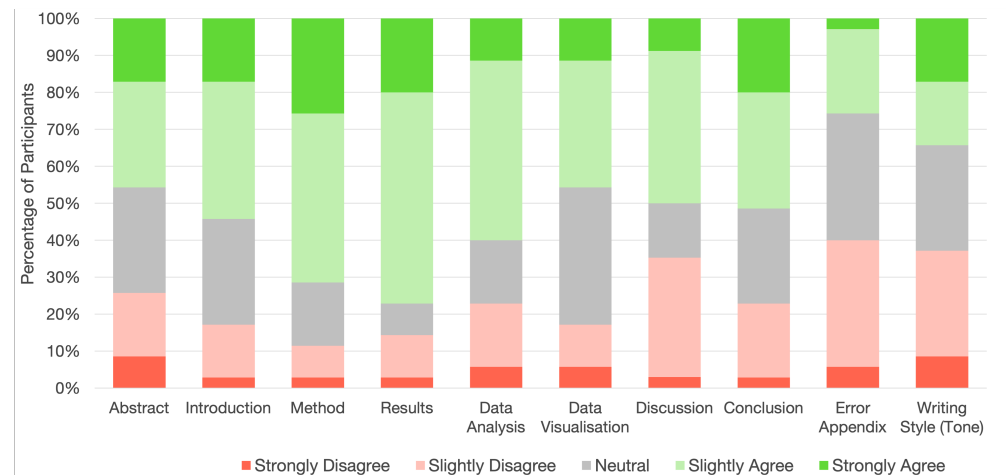


Figure 2. Likert scale responses from Year 1 students to Q5: “I am confident in my ability to complete this aspect to a high standard.” The chart shows confidence levels across different sections of the lab report, highlighting areas such as written style and error appendix where confidence was notably lower.

2.2.2. Opinions on Current Feedback

Participants rated the usefulness, length, timeliness, and clarity of their previous lab report feedback. As Year 1 students had not yet received lab report feedback at the time of survey distribution, their responses were excluded from the analysis presented in Figure 3). The highest agreement was for timeliness (69%), while length received the lowest (32%). Usefulness and clarity received intermediate agreement (50% and 54%, respectively). A Friedman test showed significant differences across these categories ($p = 1.38 \times 10^{-8}$), with post-hoc analyses identifying length (negative perception) and timeliness (positive perception) as major sources of variation.

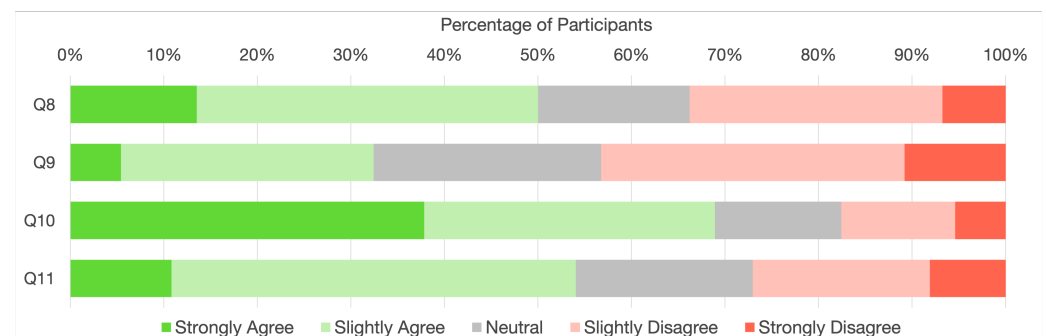


Figure 3. Likert scale responses to Q8–Q11, evaluating students’ perceptions of previous assessor feedback. Statements address feedback usefulness (Q8), length (Q9), timeliness (Q10), and clarity (Q11). Responses indicate relatively high satisfaction with turnaround time, but mixed views on clarity and adequacy of length. Year 1 students were excluded from this analysis, as they had not yet received lab report feedback at the time of the survey.

Participants were asked whether they would benefit from more detailed feedback for each lab report section (Figure 4). There was broad agreement (61%) across all report sections, with Data Analysis (73%) and Discussion (84%) sections rated as most needing additional detail. When asked about their ability to effectively implement feedback, Year 1 students provided a median neutral response, while students in other years slightly agreed, suggesting the perceived inadequacy of current feedback rather than an inability to apply it effectively.

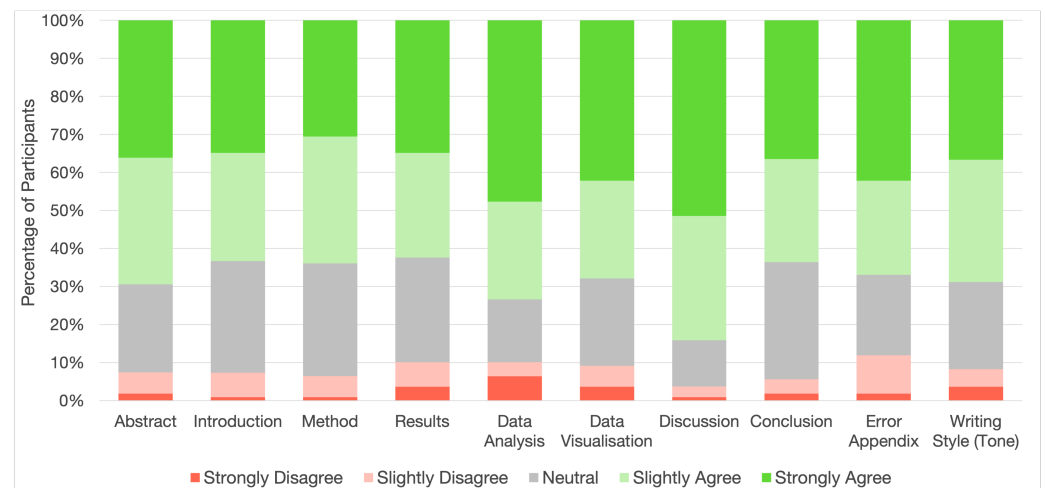


Figure 4. Likert scale responses to Q6: “I feel I would benefit from more detailed feedback in this aspect”. Students were asked to evaluate this across different lab report sections. The data highlight broad agreement, with particularly strong demand for more detail in the Data Analysis and Discussion sections. Responses include all undergraduate year groups (Years 1 to 4).

2.2.3. Student Familiarity with GenAI

The survey aimed to measure student familiarity with GenAI tools. Students selected from a list of popular GenAI systems and provided additional responses in an open-text box. Notably, 19% of respondents had never used GenAI, while 75% reported using the free version of ChatGPT, and 15% used the paid ChatGPT Plus service. These findings align with recent studies indicating that physics undergraduates primarily use GenAI for computational or clarifying tasks rather than evaluation or writing support (Zambon et al., 2024).

2.2.4. Student Trust in GenAI

Participants’ trust in GenAI was evaluated across five distinct contexts, with results summarised in Figure 5. Students responded to Likert scale statements addressing: the perceived accuracy of GenAI-generated content; its ability to provide relevant, constructive feedback; its capacity to assign a fair and accurate mark; and their willingness to implement AI-generated advice in future lab reports. A free-text box also invited participants to elaborate on any concerns about the use of GenAI in assessment.

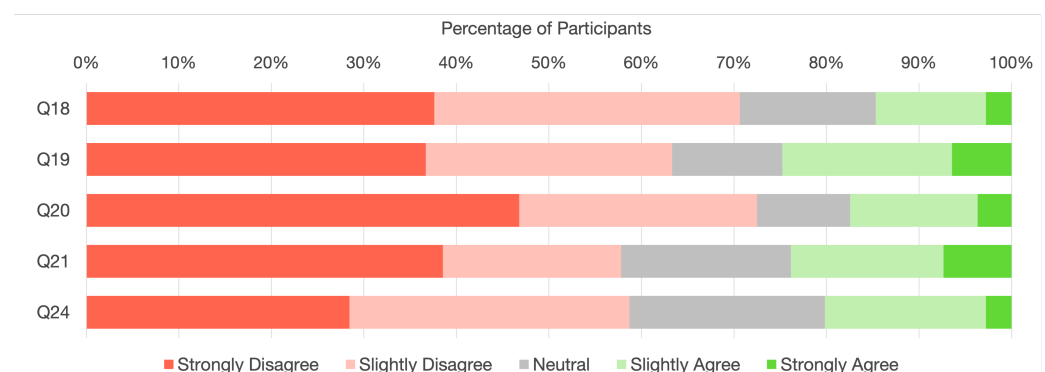


Figure 5. Likert scale responses to Q18–Q24, evaluating students’ trust in GenAI across different assessment-related contexts. Statements address: content accuracy (Q18), ability to provide relevant, constructive feedback (Q19), ability to assign accurate marks (Q20), trust in applying GenAI-generated feedback to summative work (Q21), and perceived fairness of GenAI-generated scores (Q24). Responses show generally low levels of trust, particularly in summative assessment. Responses include all undergraduate year groups (Years 1 to 4).

Across all five questions, the results indicate a general lack of trust in GenAI, with the majority of students expressing negative views. The lowest levels of trust were observed for GenAI's content accuracy (70% disagreement) and its ability to assign accurate marks (72% disagreement). These findings were reflected in free-text responses, where eight students raised concerns about inaccuracy in formative contexts and fourteen in summative contexts. For example, one student commented: "Very often ChatGPT gives the incorrect answers for technical questions and contradicts itself," while another stated: "The accuracy of GenAI is horrendous."

In contrast, perceptions of GenAI's ability to generate relevant, constructive feedback were slightly more favourable: 24% of students agreed with this statement, while 63% disagreed. A similar pattern emerged for willingness to implement such feedback in future reports (24% agreement, 58% disagreement). However, concerns about the quality of feedback were common, with students describing it as "[not] particularly deep," "[too] general", "less relevant", or "un-nuanced". Additional comments noted doubts about GenAI's ability to mark references and figures accurately.

Student perceptions of fairness were mixed. When asked whether a GenAI assessor would produce a fair lab report score, 28% strongly disagreed while only 2% strongly agreed. Neutral responses (21%) were more frequent here than for other statements, suggesting some uncertainty around what constitutes fairness or a limited understanding of GenAI's capabilities.

Importantly, prior experience with GenAI influenced trust. Among students who had never used GenAI, 90% disagreed that GenAI-generated content was accurate (Q18), compared to 66% of those who had used it. A similar trend was observed for trust in implementing GenAI-generated feedback (Q21). Mann–Whitney U tests confirmed that these differences were statistically significant: Q18 ($p = 0.0105$) and Q21 ($p = 0.0007$). These results suggest that greater familiarity with GenAI may lead to more positive perceptions of its academic utility.

2.2.5. GenAI in Relation to Assessors

Participants strongly agreed that GenAI-generated feedback on formative (87%) and summative (93%) reports should be moderated by human assessors, a sentiment also reflected in free-text responses (2 mentions for formative and 6 for summative moderation). The results are summarised in Figure 6.

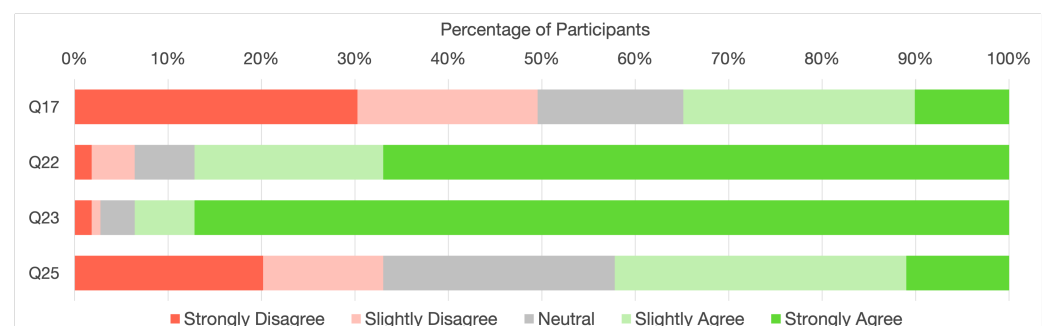


Figure 6. Likert scale responses to Q17, Q22–Q23, and Q25, evaluating student views on the role of human assessors in relation to GenAI. Statements address: comfort with assessors using GenAI to streamline feedback writing (Q17), the importance of human moderation for GenAI feedback in formative (Q22) and summative (Q23) contexts, and perceptions of whether GenAI would produce more consistently distributed lab report scores than human assessors (Q25). Responses show strong support for human moderation and mixed views on assessor use and score consistency. Responses include all undergraduate year groups (Years 1 to 4).

Opinions were divided regarding human assessors using GenAI to streamline feedback writing: 35% agreed while 49% disagreed. Prior experience with GenAI significantly influenced these views, with 71% of non-users disagreeing compared to 43% of users (Mann-Whitney U, $p = 0.0452$), suggesting familiarity increases openness to GenAI usage.

Perceptions regarding GenAI's consistency relative to human assessors were similarly mixed: 42% agreed that GenAI would produce more consistent scores, while 33% disagreed. Again, familiarity influenced responses significantly, with 52% of non-users disagreeing compared to only 28% of users (Mann-Whitney U, $p = 0.00342$).

These findings collectively suggest that students familiar with GenAI hold more positive perceptions about its potential role in feedback processes, highlighting familiarity as a key factor in acceptance.

2.2.6. Student Concerns Regarding GenAI Assessment

Analysis of free-text comments revealed concerns about using GenAI to mark formative lab reports, specifically the risk of inaccurate feedback negatively influencing summative assessments. This concern appeared in 13 comments, with examples including: "Strongly opposed, would have knock-on impact on summatives due to generic, inaccurate, or uncreative feedback", and "The AI might just be wrong, in which case it's going to be a problem when you try to put the feedback into your other assessments in the future".

Another prevalent concern (13 comments) was the perceived unethical or unfair nature of GenAI assessment, with students suggesting it could devalue their educational experience. Examples of strong opinions included: "what is even the point of paying for university to have work marked by AI", and "GenAI is not close to being an appropriate tool for this, and even if it was it would still be ethically dubious".

Respondents also suggested various appropriate uses for GenAI in assessment. Three students indicated they would be comfortable with GenAI providing supplementary feedback alongside existing human feedback. Two suggested restricting GenAI use to grammar or spelling checks. Three respondents proposed using GenAI as an initial template or to enhance the clarity of human feedback. The most frequently mentioned suggestion (5 comments) was using GenAI to moderate human assessments, reflecting student concerns about fairness and consistency highlighted elsewhere in the survey.

2.3. Key Survey Insights

Survey findings revealed that Year 1 students had the lowest confidence in their written style when completing lab reports, identifying it as a priority area for support. This makes written style an appropriate focal point for the GenAI tool, particularly as it underpins clarity and effective communication throughout the lab report. Additionally, students expressed dissatisfaction with the length and clarity of assessor feedback, highlighting the need for more detailed and interpretable feedback—two key areas the GenAI tool should address in order to better support the goals of formative assessment. Finally, respondents raised significant concerns regarding the accuracy, fairness, and ethical implications of using GenAI in assessment. Thus, the most effective and acceptable implementation of GenAI appears to be as a complement to human feedback rather than a replacement. Concentrating on written style, frequently overlooked by assessors, positions the tool as a valuable supplementary resource that enhances feedback quality while maintaining academic integrity.

3. Phase 2: GenAI Tool Development and Evaluation

3.1. Tool Development

There are two principal methods for customising a GPT model using custom “GPTs”: an interactive method, where the user incrementally defines desired behaviours through messages, and a configuration-based method, which uses a structured instruction field and linked knowledge base. This study employed the latter, offering greater control and more consistent, replicable outcomes.

The text submitted to the instruction field is known as a prompt. These prompts were refined through prompt engineering, involving systematic adjustments to improve alignment between GPT outputs and task requirements (Schulhoff et al., 2024). Each iteration was evaluated for accuracy, consistency, and structural coherence, using both assessor feedback and objective correctness as benchmarks. Consistency was assessed by applying the same prompt to multiple lab reports and comparing the structure and tone of the resulting outputs.

The content of the knowledge base, including lecture materials, marking criteria, lab model reports, and departmental documentation, was designed to provide a structured, persistent context for the GPT, enhancing the prompt without overloading it. A detailed explanation of the knowledge base components is provided in below, with a full list of included documents available in Supplementary Material S3.

3.1.1. Instruction Box

The instruction box applied several refined prompting techniques to exert precise control over the GPT’s behaviour and output, with full text provided in Supplementary Material S2. The prompt was structured into three main sections: Context, Instructions, and Important Notes, each fulfilling a distinct role in directing the model.

The Context section defined the GPT’s role as a “lab report evaluator for a university physics department reviewing first-year student work”. It also set tone expectations (e.g., “detailed and constructive”) and established a critical constraint: the GPT must avoid commenting on scientific content or proposing any changes to it, to reduce hallucinations and preserve factual accuracy. Earlier versions included stylistic cues—e.g., to be “approachable yet authoritative”—but these were removed after testing showed that the role definition alone reliably shaped tone.

The Instructions section defined a sequential, multi-step process for interpreting, evaluating, and, where appropriate, rewriting the lab report. This process incorporated a modified combination of CoT, RaR, and S2A prompting techniques, instructing the GPT to consult knowledge base files at each reasoning step. These strategies, summarised in Table 1, were designed to support deeper evaluation, maintain academic tone, ensure alignment with marking criteria and encouraged reflective evaluation. A simplified version of this process is outlined below; the full prompt text is provided in Supplementary Material S2.

If the user submits a document:

1. Analyse the report structure using `AnalysisInstructions.pdf`:
 - Identify and label sections using heading and formatting cues.
 - Flag any deviations from the expected order.
 - Recognise *written style* as a distinct section.
2. Evaluate the report content using `MarkingCriteria.pdf`.
3. Generate feedback based on the formatting conventions in `FeedbackPresentation.pdf`.
4. Identify the weakest section and ask the user whether a rewrite is desired. If yes, proceed as follows using `RewritingGuidelines.pdf`.
5. End with an open invitation for student follow-up or clarification.

The rewrite guidelines were introduced to address a recurring issue observed in earlier tool iterations, where the GPT relocated content from other parts of the report when rewriting weak sections. Although intended to improve clarity, this behaviour risked distorting the structure and logic of the original submission. The finalised instructions effectively mitigated this by defining when and how existing content could be reused, ensuring that revised sections remained coherent and contextually grounded.

To ensure reliability and minimise hallucinations, the GPT was explicitly instructed to use only the information contained in the provided documents. Any task beyond this scope would trigger a fallback message: “Apologies, this request is outside the scope of this tool”. This constraint served as an important safeguard, ensuring that the tool remained aligned with course-specific content and avoided speculative or inaccurate output.

Additionally, to eliminate any risk of the model retaining or learning from submitted student work, both the *memory* and *data-sharing* features were disabled in the deployment of the custom GPT. The memory setting was switched off, preventing the model from retaining information between sessions. The *Improve the model for everyone* setting was also disabled, ensuring that user inputs were not stored or used to train future versions of the model. These precautions ensured that each session was isolated, preserving student privacy and preventing model drift during the study. (<https://help.openai.com/en/articles/8590148-memory-faq>, <https://help.openai.com/en/articles/5722486-how-your-data-is-used-to-improve-model-performance>, accessed on 14 July 2025).

The Important Notes section summarised additional constraints and reiterated critical behavioural boundaries. These included instructions to avoid commenting on graphs or figures due to limited visual understanding, and treat the specified PDFs as examples of high-quality student lab reports for benchmarking purposes.

3.1.2. Knowledge Base

The knowledge base was curated to support the instruction box by providing persistent access to context-specific resources. This ensured that prompts remained concise while the GPT could still deliver informed, domain-aligned feedback. Documents fell into five categories: Instruction documents (e.g., *AnalysisInstructions.pdf*) directly referenced in prompts, lecture transcripts and PowerPoints taken from Year 1 report writing lectures, model lab reports for benchmarking, assessment criteria and proforma, and departmental content to provide procedural and contextual clarity about the experiments referenced in student reports, as well as extra information on report writing ideals.

These knowledge files were integrated using an in-context learning paradigm. Each prompt included one or more model lab reports and sample feedback examples to illustrate the expected structure, tone, and phrasing. This gave the GPT a concrete sense of what successful outputs should resemble, while enabling generalisation to novel reports.

However, observed system constraints required several practical adaptations. The 8000-character instruction limit necessitated offloading details to the knowledge base, while the 20-document maximum forced selective curation. Additionally, the GPT’s inability to reliably parse PDF content, especially abstracts and figures, led to a workaround in which student reports were manually extracted and submitted as Word files. While this ensured compatibility, it also highlighted several robustness limitations in the current platform.

This modular structure enabled the GPT to deliver feedback that was accurate, consistent, and grounded in context, while minimising hallucinations and irrelevant output. By combining prompt engineering with carefully selected resources, the system maintained focus and reliability across a diverse range of student submissions.

3.1.3. Output Format

The GPT-generated feedback follows a structured format aligned with best practices in formative assessment. For each report section (including written style), it provides: (1) identified strengths, (2) identified weaknesses, (3) suggested improvements, and (4) a rewritten version of the weakest section. An example of the GenAI feedback and rewrite of a Year 1 abstract is shown in Figure 7. This format reflects research showing that detailed, constructive feedback—especially with modelled rewrites—is more pedagogically effective than general praise or grading. In line with this, the feedback system omits summative marks, as research shows that combining comments with grades can shift student focus from learning to performance (Hattie & Timperley, 2007; Lipnevich & Smith, 2009).

Abstract

Strengths:

- Provides a concise summary of the experiment and key findings.
- Includes quantitative results (e.g., $K\alpha 1$ and $K\beta 1$ wavelengths).

Weaknesses:

- Lacks a comparison to expected values.
- Could be more explicit in stating the physical conclusions.

Suggested Improvements:

- Briefly compare obtained wavelengths to known values from the literature.
- Emphasize the significance of findings in relation to the experiment's objectives.

(a)

Rewritten Abstract:

The energy spectrum of a material can be analysed using an X-ray tube and a Bragg spectrometer, applying Bragg's Law to determine the wavelengths of characteristic X-rays from measured intensity peaks. In this study, these techniques were used to investigate molybdenum's characteristic X-rays and the effects of emission current and voltage. The $K\alpha 1$ and $K\beta 1$ wavelengths were measured as $71.2 \pm 0.2 \text{ nm}$ and $63.3 \pm 0.2 \text{ nm}$, respectively. While variations in emission current and voltage did not alter these wavelengths, they significantly affected the measured X-ray intensity.

Improvements Made:

- Streamlined phrasing for better readability.
- Clarified the use of Bragg's Law.
- Removed unnecessary repetition of "X-rays."
- Strengthened the final conclusion for a more impactful summary.

(b)

Figure 7. (a) GenAI-generated formative feedback on a Year 1 lab report abstract, illustrating identified strengths, weaknesses, and targeted suggestions for improvement. (b) Rewritten abstract produced by the GenAI tool, accompanied by a list of improvements made. This showcases the tool's modelling function, offering students a concrete example of how to revise and improve scientific writing.

Each section begins with strengths, using quoted extracts where appropriate to exemplify effective writing. This supports Hattie and Timperley's model of task-level feedback (FT), reinforcing what was done well and providing clear performance benchmarks. The weaknesses section then identifies areas for improvement, again using examples from the student's work. This supports feedback strategies aimed at addressing the question, "How am I going?", and encourages reflective learning. According to Nicol and Macfarlane-Dick (2006), such feedback is key to helping students close the gap between current and desired performance. To reinforce this, suggested improvements offer specific and actionable guidance, directly linked to the weaknesses identified, to support revision. This feed-forward dimension enables students to understand what to improve and what to do (Hattie & Timperley, 2007).

To reinforce this, the GPT provides a rewritten version of the weakest section, accompanied by a list of the improvements made. This models the revision process, reflecting Sadler's emphasis on the value of showing, not just telling, students how to meet academic expectations. This modelling is particularly beneficial in scientific writing, where tone, structure and clarity often require explicit examples to master. These exemplars serve as a reference point, especially for students who are still developing confidence and fluency in academic writing.

A key advantage of this GPT-based system is its support for interactive engagement. After receiving feedback, students are invited to ask follow-up questions or request clarification. This interaction extends the feedback process into a two-way dialogue, supporting not just feed-back and feed-forward, but also the development of self-regulation (FR) (Hattie & Timperley, 2007). It positions GPT not merely as an assessor but as a viable tool within the formative assessment cycle, as envisaged in Figure 1. Importantly, it also enables consis-

tent, low-barrier access to feedback, which is often difficult to achieve in resource-limited academic settings.

3.2. Evaluation Methodology

To evaluate the effectiveness of the GenAI feedback tool, 15 participants were recruited (first-year undergraduate physics students). First, students completed an initial survey to gather information on prior use of GenAI and their initial attitudes towards its educational potential. Students then submitted their formative lab reports, which were processed through the GenAI tool to generate feedback. The process is summarised in Figure 8. This feedback was reviewed by an academic moderator for factual accuracy before being returned to the student, providing an ethical safeguard. The academic moderators were recruited from among experienced module assessors. The moderators were presented with the GenAI feedback for each student and asked to identify aspects that they deemed to be inaccurate or potentially misleading and, subsequently, to annotate comments addressing any problematic aspects identified. The annotated, moderated feedback was then distributed to the students. Following this, a second survey was distributed to collect student impressions of the feedback provided. Both surveys were hosted on Jisc Online Surveys and approved by the Physics Department Ethics Committee. Informed consent was obtained from all participants, and all responses were anonymised. Data collection occurred over a two-week period in March, beginning with the return of students' marked formative reports and ending at the close of the second term.

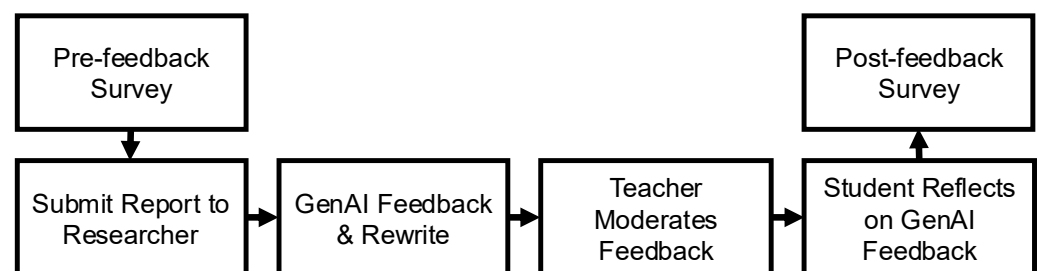


Figure 8. Overview of the evaluation study process. The diagram summarises the sequence from the initial pre-feedback survey, through report submission, GenAI feedback generation and moderation, to the final post-feedback survey used to assess the tool's impact.

The Post-Feedback Survey evaluated the effectiveness of the tool across five key categories: Usefulness, Accuracy, Clarity, Actionability, and Comparison to Assessor Feedback (i.e., feedback from the original module marker, not the GenAI moderator), with each question targeting a distinct aspect of these broader themes. For example, in the Clarity category, Q13 focused on the specificity of the feedback, while Q14 addressed the accessibility of its language. This structure allowed for a nuanced evaluation of the tool's perceived strengths and weaknesses from the student perspective.

A range of statistical techniques were used to analyse the survey data. All survey items used 5-point Likert scales, which are ordinal and seldom normally distributed with small samples (de Winter & Dodou, 2010). Accordingly we adopted non-parametric procedures throughout. Internal consistency of the multi-item themes was checked with Cronbach's α , as it is the standard index for Likert scales and remains reasonably stable even for samples as small as 10–20 (Bonett, 2002); values ≥ 0.70 are considered acceptable for exploratory work (Nunnally, 1978). The Friedman test is the rank-based analogue of a repeated-measures ANOVA and makes no distributional assumptions, making it suitable for within-participant comparisons of multiple Likert items when $n < 30$ (Conover, 1999). Where a between-group comparison was required we used the Mann–Whitney U test because it is robust to

non-normal data and small sample sizes (Field et al., 2012). Exact p -values were obtained for all tests, and we consider effect sizes (Kendall's W for Friedman, rank-biserial r for *Mann–Whitney*) to convey practical importance in a small- n context. Because only two participants fell in the negative-view subgroup, those comparisons are interpreted cautiously due to low power and potential Type II error.

3.3. Evaluation Results

3.3.1. Pre-Feedback Survey: Student Attitudes Toward GenAI

Participants were first asked about their prior experience with GenAI tools. Two reported never using GenAI, while the remaining thirteen had interacted with ChatGPT or other platforms. Of these experienced users, five described their experiences as strongly positive, six as slightly positive, and one each as neutral or negative. All participants agreed that GenAI could support their learning, and ten believed it could enhance the quality of their academic work, though four were neutral and one disagreed. Views on institutional integration were more mixed: four supported broader use of GenAI in academic tools, five were neutral, and two disagreed. Eleven participants felt comfortable using GenAI within university regulations, while four did not.

Opinions on GenAI-generated lab report feedback were similarly varied. Eight participants trusted GenAI to provide relevant, constructive comments, with five neutral and one disagreeing. When asked if they would trust such feedback enough to apply it to future summative reports, seven agreed, while four were neutral and four disagreed. Fewer participants (six) felt that GenAI feedback could be as helpful as human feedback, with six disagreeing, while five believed it was more objective. These findings provide important context for interpreting subsequent evaluations of the GenAI-generated feedback.

Free-text responses offered further insight. Three participants raised concerns about the accuracy of GenAI feedback, particularly the risk of undetected errors affecting future work. Another three emphasised the importance of moderation by a human assessor. Participants also highlighted strengths, including grammar correction, fluency, speed, and objectivity. However, they also noted limitations in GenAI's handling of visual content, discipline-specific feedback, and alignment with academic style expectations.

3.3.2. Post-Feedback Survey: Student Evaluation of GenAI Tool Usefulness

Participants responded positively to all three statements assessing the usefulness of the GenAI-generated feedback (Figure 9). Fourteen participants agreed that the feedback was valuable, and thirteen agreed it would lead to meaningful improvements in their writing and that it offered insights not provided by their human assessor. Each of the latter two items received one neutral and one disagreeing response.

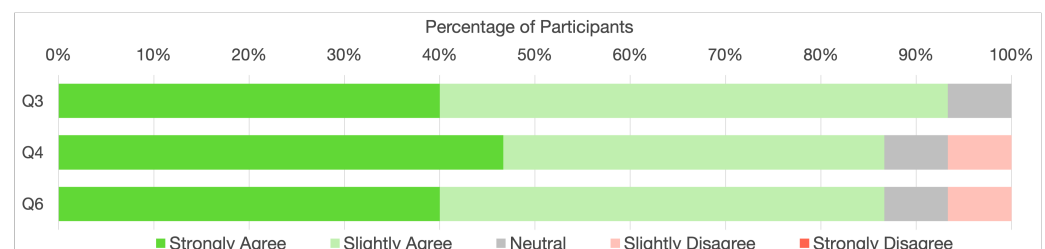


Figure 9. Likert scale responses to Q3, Q4, and Q6, evaluating the perceived usefulness of GenAI-generated feedback. Statements address: whether the feedback offered valuable insights (Q3), whether it could lead to meaningful improvements in student writing (Q4), and whether it provided new insights beyond assessor feedback (Q6). Responses indicate strong agreement across all items.

Free-text comments reinforced these findings. Several students praised the ‘improvements’ section, with one noting it suggested clear and actionable steps for development. Others appreciated the level of detail and comprehensiveness of the feedback. However, some limitations were identified. A few participants noted that the tool occasionally failed to recognise when information was appropriately located in other sections, and three reported contradictions between GenAI and human feedback, raising concerns about reliability and coherence.

Accuracy

Two questions were asked to participants regarding the accuracy of the feedback, with the results shown in Figure 10. Participants responded positively to the GenAI tool’s ability to identify strengths and weaknesses, with fourteen agreeing that it correctly identified these aspects and eleven agreeing that it aligned with the marking criteria. However, views were more mixed when it came to the presence of misleading comments.

In the free-text comments, three participants explicitly noted a “discrepancy between the assessor and the GenAI feedback”, suggesting occasional divergence in evaluative judgement. One participant observed that the GenAI occasionally recommended elaboration on content that was not directly relevant to the experiment, indicating that the tool may occasionally misjudge the appropriate level of detail or scope for a Year 1 report.

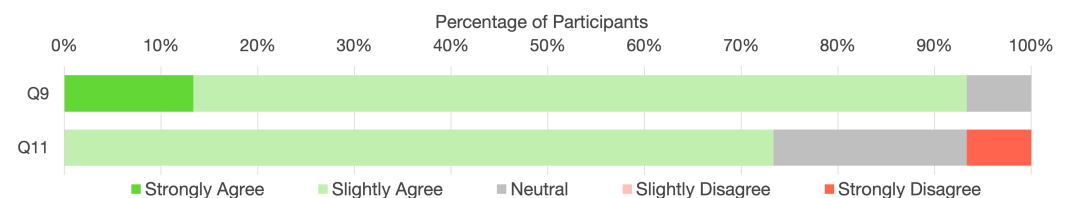


Figure 10. Likert scale responses to Q9 and Q11, assessing student perceptions of the accuracy of GenAI-generated feedback. Statements address: whether the feedback was aligned with the marking criteria (Q9), and whether it correctly identified strengths and weaknesses in student work (Q11). Responses suggest generally positive views on accuracy.

Clarity

Feedback on clarity revealed a contrast between language accessibility and comment specificity, as shown in Figure 11. All fifteen participants agreed the feedback avoided technical jargon or overly complex language. However, only nine felt the comments were clear and specific, while five disagreed. This suggests that although the language was broadly accessible, there was some variability in how precisely comments conveyed their intent. Supporting this, six students described the feedback as “clear and concise” in their free-text responses, and two commended the formatting, specifically the use of bullet points, as aiding readability and focus. However, three students felt the feedback was “a little vague in places” and indicated that it could benefit from greater specificity and elaboration. As echoed in earlier comments, these responses suggest that including more illustrative examples could improve both the clarity and applicability of the tool’s advice.

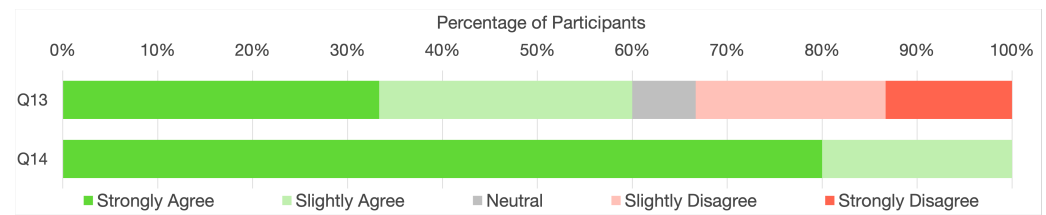


Figure 11. Likert scale responses to Q13 and Q14, evaluating the clarity of GenAI-generated feedback. Statements address: whether the feedback was clear and specific rather than vague (Q13), and whether it avoided technical jargon or unnecessarily complex language (Q14). Responses indicate that students found the language broadly accessible, with some variation in perceived specificity.

Actionability

Responses indicated that students generally found the GenAI feedback actionable, as seen in Figure 12. Most agreed that it provided clear, step-by-step guidance (twelve agreed, one was neutral, two disagreed), felt confident applying it to improve their writing (fourteen agreed, one disagreed), and believed it would benefit future scientific writing (fourteen agreed, one disagreed). Free-text responses echoed this sentiment, with three students highlighting how easy the feedback was to understand and apply. Several praised the improvement suggestions, and one specifically noted the value of included examples. As echoed in earlier comments, these responses suggest that including more illustrative examples could improve both the clarity and applicability of the tool's advice. One participant also expressed the need for human confirmation before applying the advice, indicating some hesitation in relying solely on GenAI-generated feedback.

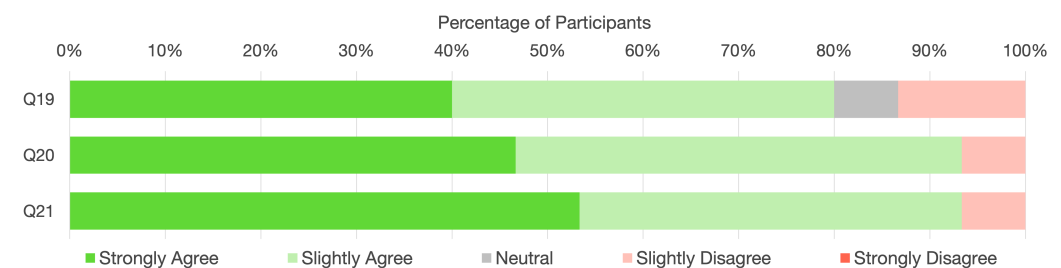


Figure 12. Likert scale responses to Q19–Q21, assessing the actionability of GenAI-generated feedback. Statements address: whether the feedback provided clear steps for improvement (Q19), whether students felt confident applying it to their writing (Q20), and whether it would be useful for future scientific assignments beyond the current report (Q21). Responses reflect strong agreement, indicating high perceived applicability.

Comparison of Categories

A comparison of the four thematic categories based on the summed Likert-scale responses averaged across the number of questions within each category reveals generally positive perceptions, with some variation in strength. Usefulness and actionability both achieved the highest average scores of 64, indicating that participants found the GenAI-generated feedback both valuable and implementable. Clarity followed closely with an average score of 62, suggesting that the feedback was generally well-articulated and accessible. Accuracy received the lowest average score at 51, reflecting a more cautious perception, likely influenced by noted discrepancies between GenAI and human assessor feedback. These findings suggest that while students appreciated the practical utility and comprehensibility of the feedback, concerns regarding its precision remain.

Comparison To Human Assessor

To evaluate how GenAI-generated feedback compared to traditional human feedback, participants were asked whether the GenAI feedback was equally or more useful, understandable, and actionable than what they had previously received from their assessors (i.e., the staff who originally marked their lab reports, not the GenAI moderators in this study). They also rated whether they preferred the length and level of detail provided by GenAI. The results are shown in Figure 13. Agreement was strongest for understandability and length, with twelve participants agreeing that the GenAI feedback was equally or more understandable than assessor feedback, and that they preferred its length. Ten participants also felt the GenAI feedback was more actionable, and nine preferred its level of detail. In contrast, only six agreed that it was equally or more useful overall. A Friedman test revealed a statistically significant difference across these responses ($p = 0.00176$), with usefulness receiving the lowest agreement. Free-text comments help contextualise this trend. Participants frequently praised the GenAI feedback for being more detailed, covering a broader range of areas, and offering more structure and consistency than human feedback. Several also emphasised its clarity and ease of implementation. Two participants noted the speed of return as a major advantage, with one remarking, “the longer I’m waiting on the feedback, [the longer] I’m not thinking about [it]”, underscoring the value of immediacy in promoting reflective learning.

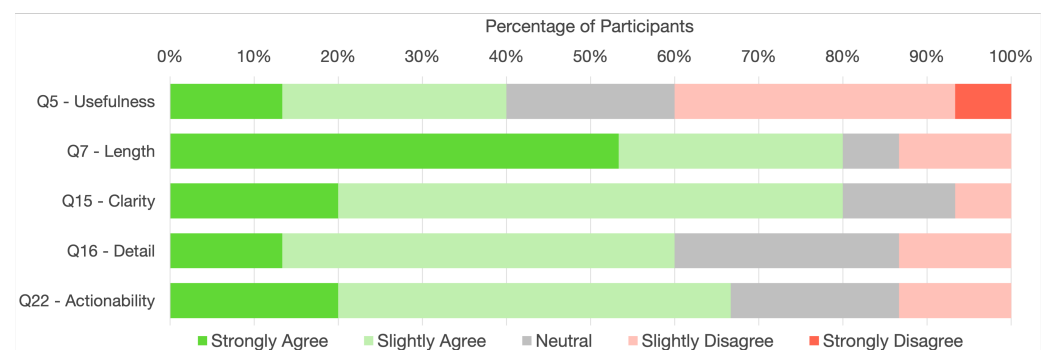


Figure 13. Likert scale responses to Q5, Q7, Q15, Q16, and Q22, comparing GenAI-generated feedback to human assessor feedback. Statements address: overall usefulness (Q5), preferred length (Q7), clarity and understandability (Q15), level of detail (Q16), and ease of implementation (Q22). Results suggest students generally found GenAI feedback comparable or preferable across several dimensions, particularly in clarity and length.

Rewrite

Participants were asked to evaluate the rewritten version of their work produced by GenAI, specifically in terms of style, clarity, and usefulness as a model. These results are shown in Figure 14. Agreement was generally high: eight participants felt the GenAI rewrite had a more appropriate style, eleven agreed it communicated ideas more clearly, nine said it helped them understand the writing style they should aim for, and ten indicated they would adopt some of the changes in their own writing. Free-text responses reinforced these results, with several participants describing the rewrite as helpful, particularly for improving clarity and scientific tone. However, some concerns were raised. Four students found the language overly polished or “inhuman”, which reduced its usefulness as a relatable model. Others noted that the rewrites were often too long for Year 1 lab reports, which are subject to strict page limits. This highlights a mismatch between GenAI output and actual assessment constraints.

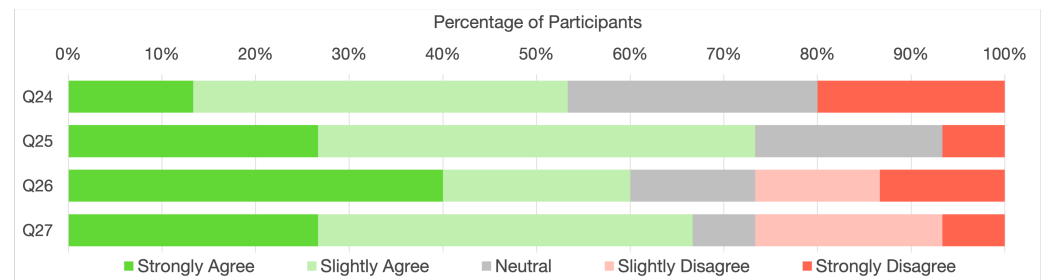


Figure 14. Likert scale responses to Q24–Q27, evaluating student perceptions of the GenAI-generated rewrite. Statements address: whether the GenAI rewrite demonstrated a more appropriate written style (Q24), whether it communicated points more clearly (Q25), whether it helped students understand the style they should aim for (Q26), and whether it would influence their future scientific writing (Q27). Responses indicate strong perceived value of the rewrite as a model for improvement.

3.3.3. Evaluation Results by Category: The Influence of Prior Attitudes on Feedback Evaluation

A comparative analysis was conducted to determine whether students' previous experience with GenAI or their general attitudes toward the technology influenced their evaluation of the feedback tool. A *Mann–Whitney U* test found no significant differences across the 28 survey items based on prior GenAI use, suggesting that familiarity alone did not affect perceptions. However, grouping students by overall sentiment, derived from a composite 'positivity' score based on Questions 7–10 of the pre-feedback survey (Cronbach's $\alpha = 0.84$), revealed meaningful differences. Of the fifteen participants, thirteen were classified as holding positive views and two as negative. Significant differences emerged for Q19 ('The GenAI feedback gave me clear steps I can take to improve my work', $p = 0.0456$) and Q24 ('I believe the GenAI rewrite has a more appropriate written style than my original work', $p = 0.0407$). Responses to Q20, Q21, and Q22 also trended toward significance, suggesting a broader pattern. These results indicate that while prior use of GenAI may not shape evaluations, students with more positive attitudes toward the technology tend to rate GenAI-generated feedback more favourably, particularly in terms of actionability and writing quality.

4. Conclusions

This study explored the potential of a GenAI tool, with prompt-engineered GPT model, as a support tool within the formative assessment of first-year undergraduate physics lab reports. With rising student numbers placing pressure on academic staff to deliver timely, consistent, and high-quality feedback, we examined whether a prompt-engineered GenAI tool could function as an effective supplement to traditional marking processes. Lab reports were selected due to their central role in developing core scientific communication and reasoning skills, as well as their high marking burden, particularly at Year 1 where feedback can have the greatest developmental impact.

Prior to developing the GenAI tool, a research survey was conducted to identify existing challenges in feedback and assessment, and to GenAI student perceptions of both current practices and the role of GenAI in education. The findings revealed that many students, especially at Year 1, lacked confidence in specific aspects of their lab reports writing, most notably in written style, the error appendix, and the discussion section. Across year groups, students expressed dissatisfaction with the clarity, usefulness, and length of current assessor feedback, although the timeliness of feedback return was generally viewed positively. Critically, students expressed a strong desire for more detailed actionable feedback, particularly in the discussion and abstract sections. This insight informed both the focus and design of the GenAI tool. Student attitudes towards GenAI in assessment

were cautiously optimistic: while sceptical of its use within summative assessment, they were more comfortable with its use within formative contexts, highlighting the need for moderation by a human assessor. These insights provided a clear rationale and a blueprint for the tool's development, ensuring it addressed pedagogical needs, student priorities, and contextual constraints within the assessment landscape.

The finalised GenAI system, built using a custom ChatGPT model and refined through prompt engineering and curated knowledge base, produced feedback that students generally rated as useful, actionable, and clearly written. Thematic analysis of free comment and survey responses indicated that the structured format, including sections for strengths, weaknesses, and suggested improvements, was particularly well-received. Students valued the practical guidance offered, especially in the form of rewritten exemplars, and noted its usefulness for both current and future scientific writing. However, concerns remained regarding the tool's occasional inaccuracies and misaligned comments. These findings underscore the need for continued oversight in any high-stakes application of such tools. Despite these limitations, key advantages of GenAI-based feedback systems were highlighted over traditional formative assessment, including greater consistency, broader coverage of feedback areas, increased detail, and faster turnaround. Students also valued the potential for interactive engagement, such as the opportunity to ask follow-up questions—a feature not typically offered in conventional feedback systems. This positioned the GenAI tool not merely as a static assessor, but as a responsive educational support tool, capable of facilitating reflection and promoting student autonomy. In this capacity, the GenAI tool augments the formative assessment cycle by providing a flexible feedback channel that promotes student engagement, supports instructional adjustment, and fosters reflective practice.

From a pedagogical perspective, the GenAI's feedback mechanism aligns closely with established principles of effective formative assessment. Its structured format reinforces student learning by identifying what was done well, diagnosing areas for improvement, and offering concrete, actionable suggestions. The inclusion of rewritten sections further enhances its educational value by modelling effective scientific writing, particularly for students still developing their academic voice. These features support the core elements of feed-back, feed-forward, and self-regulation, which are central to formative pedagogy. Importantly, the feedback quality stems not from the model's inherent capabilities, but from the intentional design of its instructional framework. This study demonstrated that the customisability and prompt design of the GenAI are critical to ensuring pedagogical value. Rather than being a fixed solution, GenAI represents a flexible platform, adaptable to the instructional aims, learner needs, and evolving educational contexts in which it is deployed.

In conclusion, this study demonstrates that generative AI can provide scalable, high-quality formative feedback in undergraduate physics education when designed with pedagogical intent and appropriate constraints. While the current version of the tool requires continued moderation and is not yet suitable for unmoderated deployment, it offers a strong foundation for future development. Its greatest value may lie in complementing rather than replacing human feedback—combining the clarity, speed, and consistency of GenAI with the subject expertise and contextual judgement of assessors. Future work should explore how this hybrid model could be integrated into institutional practice and optimised for accuracy and reliability. It is also essential to acknowledge that GenAI capabilities are evolving rapidly. Even within the timeframe of this study, OpenAI previewed its next-generation model, GPT-4.5, reflecting the pace of development in this space. As such, further investigations are likely to benefit from more advanced models and expanded functionality, raising new opportunities and challenges for educational practice.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/educsci15081058/s1>. Supplementary S1: Research Survey Questions; Supplementary S2: gAI Instruction Box; Supplementary S3: gAI Knowledge Base; Supplementary S4: Pre-Feedback Survey Questions; Supplementary S5: Post-Feedback Survey Questions.

Author Contributions: Conceptualization, A.M. and A.P.; methodology, A.M., A.P. and E.M.; software, E.M.; validation, E.M., A.M. and A.P.; formal analysis, E.M.; investigation, E.M.; resources, A.M., A.P. and E.M.; data curation, E.M.; writing—original draft preparation, E.M.; writing—review and editing, A.M. and A.P.; visualization, E.M.; supervision, A.M. and A.P.; project administration, A.M.; ethics, A.M., A.P. and E.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research doesn't receive external funding.

Institutional Review Board Statement: This research has been approved by The Physics Ethics Committee of Durham University. (approval code: PHY S-2024-2868-3530, approval date: 10 December 2024).

Informed Consent Statement: Informed consent was obtained from all participants involved in the study.

Data Availability Statement: The data supporting the findings of this study are available within the article and its supplementary materials. Survey instruments, GPT prompt text, and the knowledge base references used in the gAI tool are provided as downloadable supplementary files. Due to ethical considerations, individual student reports and free-text survey responses are not shared.

Conflicts of Interest: There are no known conflicts of interest associated with this study.

References

- An, Y., Yu, J. H., & James, S. (2025). Investigating the higher education institutions' guidelines and policies regarding the use of generative AI in teaching, learning, research, and administration. *International Journal of Educational Technology in Higher Education*, 22, 10. [CrossRef]
- Blank, I. A. (2023). What are large language models supposed to model? *Trends in Cognitive Sciences*, 27(11), 987–989. [CrossRef]
- Bloom, B. (1971). *Handbook on formative and summative evaluation of student learning*. McGraw-Hill.
- Bonett, D. G. (2002). Sample size requirements for testing and estimating coefficient alpha. *Journal of Educational and Behavioral Statistics*, 27(4), 335–340. [CrossRef]
- Chassignol, M., Khoroshavin, A., Klimova, A., & Bilyatdinova, A. (2018). Artificial Intelligence trends in education: A narrative overview. *Procedia Computer Science*, 136, 16–24. [CrossRef]
- Conover, W. J. (1999). *Practical nonparametric statistics* (3rd ed.). Wiley.
- Cooper, G. (2023). Examining science education in ChatGPT: An exploratory study of generative artificial intelligence. *Journal of Science Education and Technology*, 32(3), 444–452. [CrossRef]
- de Winter, J. F. C., & Dodou, D. (2010). Five-point likert items: t test versus Mann–Whitney–Wilcoxon. *Practical Assessment, Research & Evaluation*, 15(11), 1–16. [CrossRef]
- Dotan, R., Parker, L. S., & Radzilowicz, J. (2024, June 3–6). Responsible adoption of generative AI in higher education: Developing a “points to consider” approach based on faculty perspectives. The 2024 ACM Conference on Fairness, Accountability, and Transparency (pp. 2033–2046), Rio de Janeiro, Brazil [CrossRef]
- Ebert, C., & Louridas, P. (2023). Generative AI for software practitioners. *IEEE Software*, 40(4), 30–38. [CrossRef]
- El Fathi, T., Saad, A., Larhzil, H., Lamri, D., & Al Ibrahmi, E. M. (2025). Integrating generative AI into STEM education: Enhancing conceptual understanding, addressing misconceptions, and assessing student acceptance. *Disciplinary and Interdisciplinary Science Education Research*, 7(6), 6. [CrossRef]
- Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using r*. SAGE Publications.
- Fyfe, E. R., de Leeuw, J. R., Carvalho, P. F., Goldstone, R. L., Hourihan, K. L., Kerr, B., Lee, H., Motz, B. A., Nathan, M. J., Noelle, D. C., Pape, S. J., Ruprecht, C., Subban, P., Teasley, S. D., Thompson, C. A., Uz Zaman, T., Van Tassell, R., Yan, V. M., & Yip, D. (2021). ManyClasses 1: Assessing the generalizable effect of immediate feedback versus delayed feedback across many college classes. *Advances in Methods and Practices in Psychological Science*, 4(3), 1–24. [CrossRef]
- Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81–112. [CrossRef]

- Ivanov, S., Soliman, M., Tuomi, A., Alhamar Alkathiri, N., & Al-Alawi, A. N. (2024). Drivers of generative AI adoption in higher education through the lens of the theory of planned behaviour. *Technology in Society*, 77, 102521. [CrossRef]
- Klenowski, V. (2009). Assessment for learning revisited: An Asia-Pacific perspective. *Assessment in Education: Principles, Policy and Practice*, 16(3), 263–268. [CrossRef]
- Lipnevich, A. A., & Smith, J. K. (2009). Effects of differential feedback on students' examination performance. *Journal of Experimental Psychology: Applied*, 15(4), 319–333. [CrossRef]
- Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, 31(2), 199–218. [CrossRef]
- Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). McGraw-Hill.
- OpenAI. (2023, November). *Introducing gpts*. Available online: <https://openai.com/index/introducing-gpts> (accessed on 9 July 2025).
- Ramaprasad, A. (1983). On the definition of feedback. *Behavioral Science*, 28(1), 4–13. [CrossRef]
- Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, 3, 121–154. [CrossRef]
- Reynolds, L., & McDonnell, K. (2021). Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended abstracts of the 2021 chi conference on human factors in computing systems* (pp. 1–7). ACM. [CrossRef]
- Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, 18(2), 119–144. [CrossRef]
- Schildkamp, K., & Kuiper, W. (2010). Data-informed curriculum reform: Which data, what purposes, and promoting and hindering factors. *Teaching and Teacher Education*, 26(3), 482–496. [CrossRef]
- Schulhoff, S., Ilie, M., Balepur, N., Kahadze, K., Liu, A., Si, C., Li, Y., Gupta, A., Han, H., Schulhoff, S., Dulepet, P. S., Vidyadhara, S., Ki, D., Agrawal, S., Pham, C., Kroiz, G., Li, F., Tao, H., Srivastava, A., ... Resnik, P. (2024). The prompt report: A systematic survey of prompting techniques. *arXiv*, arXiv:2406.06608. [CrossRef]
- Scriven, M. (1966). *The methodology of evaluation* (No. 110). Purdue University.
- Sirnoorkar, A., Zollman, D., Laverty, J. T., Magana, A. J., Rebello, N. S., & Bryan, L. A. (2024). Student and AI responses to physics problems examined through the lenses of sensemaking and mechanistic reasoning. *Computers and Education: Artificial Intelligence*, 7, 100318. [CrossRef]
- Sortwell, A., Trimble, K., Ferraz, R., Geelan, D. R., Hine, G., Ramirez-Campillo, R., Carter-Thuiller, B., Gkintoni, E., & Xuan, Q. (2024). A systematic review of meta-analyses on the impact of formative assessment on K-12 students' learning: Toward sustainable quality education. *Sustainability*, 16(17), 7826. [CrossRef]
- Wan, T., & Chen, Z. (2024). Exploring generative AI assisted feedback writing for students' written responses to a physics conceptual question with prompt engineering and few-shot learning. *Physical Review Physics Education Research*, 20(1), 010152. [CrossRef]
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., & Zhou, D. (2022, November 28–December 9). *Chain-of-thought prompting elicits reasoning in large language models*. 36th International Conference on Neural Information Processing Systems (NeurIPS 2022) (pp. 1800–1813), New Orleans, LA, USA. Available online: https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf (accessed on 9 July 2025).
- Yeadon, W., Agra, E., Inyang, O.-O., Mackay, P., & Mizouri, A. (2024). Evaluating AI and human authorship quality in academic writing through physics essays. *European Journal of Physics*, 45(5), 055703. [CrossRef]
- Yeadon, W., & Hardy, T. (2024). The impact of AI in physics education: A comprehensive review from GCSE to university levels. *Physics Education*, 59(2), 025010. [CrossRef]
- Yeadon, W., Inyang, O.-O., Mizouri, A., Peach, A., & Testro, C. P. (2023). The death of the short-form physics essay in the coming AI revolution. *Physics Education*, 58(3), 035027. [CrossRef]
- Yeadon, W., Peach, A., & Testrow, C. (2024). A comparison of human, GPT-3.5, and GPT-4 performance in a university-level coding course. *Scientific Reports*, 14, 23285. [CrossRef] [PubMed]
- Zambon, C., Mizouri, A., & Stevenson, C. (2024). Navigating the gAI landscape: Insights from a physics education survey. *Enhancing Teaching and Learning in Higher Education*, 2, 16–38. [CrossRef]
- Zhang, Z., Zhang, J., Zhang, X., & Mai, W. (2025). A comprehensive overview of Generative AI (GAI): Technologies, applications, and challenges. *Neurocomputing*, 632, 129645. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.