Investigating Learner-Aware Design of LLM-Generated Educational Feedback

Anonymous ACL submission

Abstract

Although large language models (LLMs) show promise for generating educational feedback, it remains unclear how feedback should be designed (e.g., tone and information coverage) to support answer revision and learner acceptance across diverse learner profiles. We define six feedback designs for multiple-choice biology questions, including a baseline design and variants with additional feedback elements, and conduct an empirical study with high school students. We evaluate feedback using immediate revision performance and six subjective evaluation criteria, and analyze how feedback preferences vary across learner profiles based on personality traits. Our results show that feedback with clear and comprehensive guidance improves revision performance and receives favorable evaluations across learner profiles, whereas informational novelty and affective framing vary across profiles. These findings suggest that learner profiles should be considered when designing LLM-generated feedback.

1 Introduction

Large language models (LLMs) have enabled the large-scale generation of diverse forms of educational feedback and have accelerated the development of automated learning support systems (Kasneci et al., 2023; Stamper et al., 2024; Chu et al., 2025). Unlike conventional template-based approaches (Lai and Chang, 2019; Naito et al., 2024), LLMs can flexibly generate feedback with different educational designs, such as explanation level and tone. Feedback helps learners bridge the gap between their understanding and correct answers (Hattie and Timperley, 2007; Wisniewski et al., 2020), and must be understandable and trustworthy enough to guide learners toward productive answer revision.

Previous studies on human-designed feedback also have shown that feedback effectiveness differs across learner profiles such as prior knowledge and personality traits (Aleven et al., 2006; Bidjerano and Dai, 2007; Poropat, 2009).Therefore, personalized feedback generation requires consideration of learner profiles as well as feedback content. To design effective LLM-generated feedback, we need to identify which feedback designs positively influence learner behavior and perception through empirical studies. Although LLMs can generate adaptive feedback, the effectiveness of different designs for learners remains understood.

Prior work has compared human-written and LLM-generated feedback (Escalante et al., 2023; Steiss et al., 2024; Asadi et al., 2025) or evaluated generated feedback using LLM-based automatic evaluation (Qian et al., 2025; Chu et al., 2026). However, these studies do not systematically investigate how feedback designs influence learners, and it remains unclear whether LLM-based evaluations align with learner perceptions and behaviors. Although Borges et al. (2024) identify several educational elements that may influence the quality of LLM-generated feedback, such as informational coverage and tone, their effects on learner behavior and feedback perception have not been systematically investigated.

To address the limited consideration of learner profiles and the unclear effects of feedback elements in LLM-generated feedback, we investigate the following three research questions through the three-step workflow in Figure 1.

- *RQ1: Which feedback elements most effectively improve revision performance?*
- *RQ2: Which feedback elements lead to more favorable subjective evaluations?*
- *RQ3: How do preferences for feedback elements differ across learners with different Big Five personality traits?*

First, we define six feedback elements related to content, such as information coverage and tone,

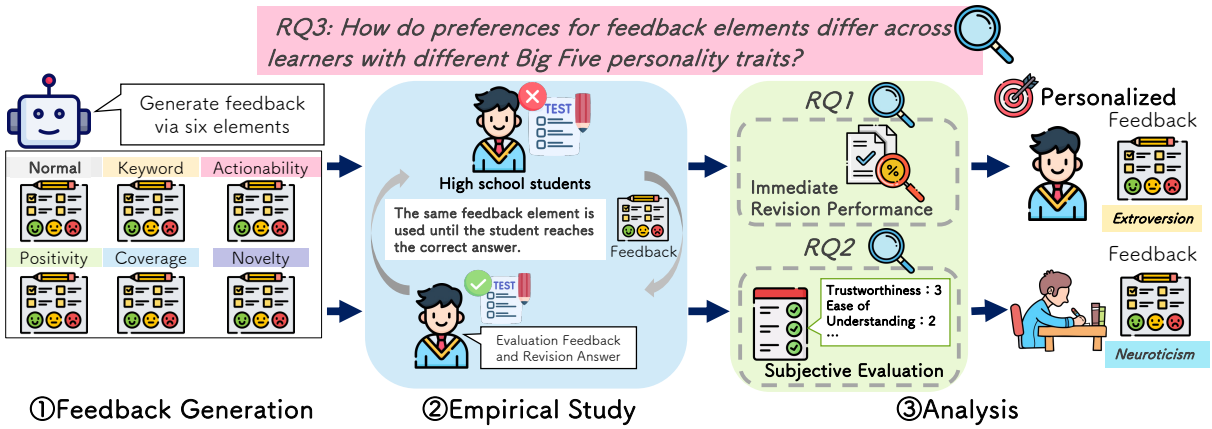

Figure 1: Overview of this study. First, we define six feedback elements related to feedback content and generate feedback using GPT-5. Next, based on data from 321 participants, we analyze immediate revision performance and subjective evaluations across six criteria. Finally, we examine which feedback elements are more favorably received across learners with different Big Five personality traits.

based on Borges et al. (2024) and generate feedback for multiple-choice biology questions using gpt-5-2025-08-07 (Singh et al., 2026). Next, through an empirical experiment with 321 first-year high school students, we evaluate feedback effects on immediate revision performance and six subjective evaluation criteria, such as trustworthiness and ease of understanding (*RQ1, RQ2*). Finally, we analyze subjective evaluations across learners with different Big Five personality traits (*RQ3*).

We find that effective feedback depends on how clearly it directs learners' attention to task-relevant points and how comprehensively it covers key aspects of the task. Subjective evaluation results show that learners favor feedback that is practical and easy to apply. We also find that preferred feedback elements vary across profiles. This suggests that learners differ in how they perceive and utilize feedback. These findings align with prior works (Komarraju and Karau, 2005; Hattie and Timperley, 2007; Komarraju et al., 2009; Wisniewski et al., 2020) and indicate the potential for personalizing LLM-generated feedback based on learner profiles.

The contributions of this study are as follows:

- We define six controllable educational elements for LLM-generated feedback and empirically examine their effects on revision performance and subjective evaluations through an empirical experiment with 321 students.
- We find that feedback preferences and perceived effectiveness vary across learner profiles. This suggests the importance of adapting feedback to learner profiles.

2 Related Work

2.1 Feedback

Feedback helps learners bridge the gap between their current state and learning goals, and can take various forms, ranging from simple correct/ incorrect indicators (e.g., checkmarks) to elaborated explanations that provide written guidance (Hattie and Timperley, 2007; Wisniewski et al., 2020). These studies conceptualize feedback as a multidimensional construct, where explanations differ in whether they directly provide correct answers or encourage learners to derive them, as well as in content-related elements (e.g., level of elaboration and explanatory detail) and delivery-related criteria (e.g., timing and modality). In this study, we focus on content-related elements that determine the granularity and informational richness of feedback. We adopt an indirect feedback approach that encourages learners to derive answers.

2.2 Feedback Generation

Prior to the emergence of LLMs, intelligent tutoring systems (ITSs) extensively investigated feedback generation using rule-based and model-based approaches, enabling adaptive and personalized instructional support based on learners' responses and problem-solving states (VanLehn, 2011; Kochmar et al., 2020; Villegas-Ch et al., 2025). Building on this foundation, recent studies have begun integrating LLM-generated feedback into classroom settings and examining its effects on learning outcomes and motivation (Dai et al., 2023; Giannakos et al., 2025; Kinder et al., 2025). How-

ever, most evaluations of LLM-generated feedback rely either on coarse-grained comparisons with human-written feedback (Escalante et al., 2023; Steiss et al., 2024) or automatic evaluation (Qian et al., 2025; Chu et al., 2026), neither of which captures how specific educational designs influence learners' perception and use of feedback. Although Borges et al. (2024) identify several such elements, such as informational coverage and tone, their effects on learner behavior and feedback perception remain empirically underexplored. Moreover, while prior educational studies indicate that feedback effectiveness varies across learner profiles (Aleven et al., 2006; Bidjerano and Dai, 2007; Poropat, 2009), such individual differences have received limited attention in LLM-generated feedback research. In this study, we investigate how different feedback elements influence learners' immediate revision performance and subjective evaluations, and how these effects vary across learners with different Big Five personality traits.

2.3 Big Five

Research on individual differences has long examined variability in learners' behavior and cognition, and the Big Five personality trait model (Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism) is widely adopted as a framework for describing such differences (McCrae and John, 1992). In educational research, these traits are used to characterize learners and have been linked to motivation, academic achievement, and preferences for instructional support and feedback (Aleven et al., 2006; Bidjerano and Dai, 2007; Poropat, 2009; Takami et al., 2023). In the field of Natural Language Processing, there is a growing interest in incorporating the Big Five framework into LLMs to represent user personas and control generated outputs. Recent studies have incorporated the Big Five framework into LLMs to represent user personas and control generated outputs (Jiang et al., 2024; Liu et al., 2024; Chen et al., 2025). These studies simulate diverse learner profiles and examine personality-conditioned generation. However, how learner personality shapes the perception and acceptance of generated feedback remains underexplored (Sharma et al., 2025). Therefore, we analyze the relationship between learners' Big Five personality traits and their subjective evaluations of LLM-generated feedback.

3 Feedback Elements

We define six feedback elements: a manually designed *Normal* condition and five educational elements based on Borges et al. (2024). From the original ten elements, we select five with clearly distinguishable elements, as several produced overlapping feedback in pilot generation. Each variation adds a specific educational emphasis to the shared *Normal* element, such as instructional specificity, or information coverage. We use gpt-5-2025-08-07 (GPT-5) (Singh et al., 2026) to generate all feedback in advance, and produce indirect feedback that encourages learners to derive the correct answer without explicitly revealing it.

Normal We use this design as the reference condition for comparison across feedback elements. This explicitly indicates whether the learner's response is correct or incorrect and provides a standard explanation aimed at bridging gaps in the learner's knowledge and reasoning.

Keywords This design focuses on information granularity for learner understanding. Rather than presenting keywords alone, the feedback highlights particularly important terms within the explanation using brackets.

Actionability This design investigates the impact of instructional specificity on answer revision. Since abstract guidance may make it difficult for learners to judge how to improve their responses, the feedback provides concrete revision steps and checklists that indicate what should be reviewed.

Novelty This design explores introducing information beyond the learned content. Regardless of the answer's correctness, the feedback provides advanced perspectives.

Coverage This design examines how informational completeness influences learner understanding. Rather than focusing on a single aspect, the feedback describes all logical elements required to reach the correct answer.

Positivity This design investigates the influence of affective factors on learning. In addition to pointing out errors, the feedback incorporates positive expressions, such as praise for partially correct reasoning and encouraging comments.

See Appendix E for feedback examples.

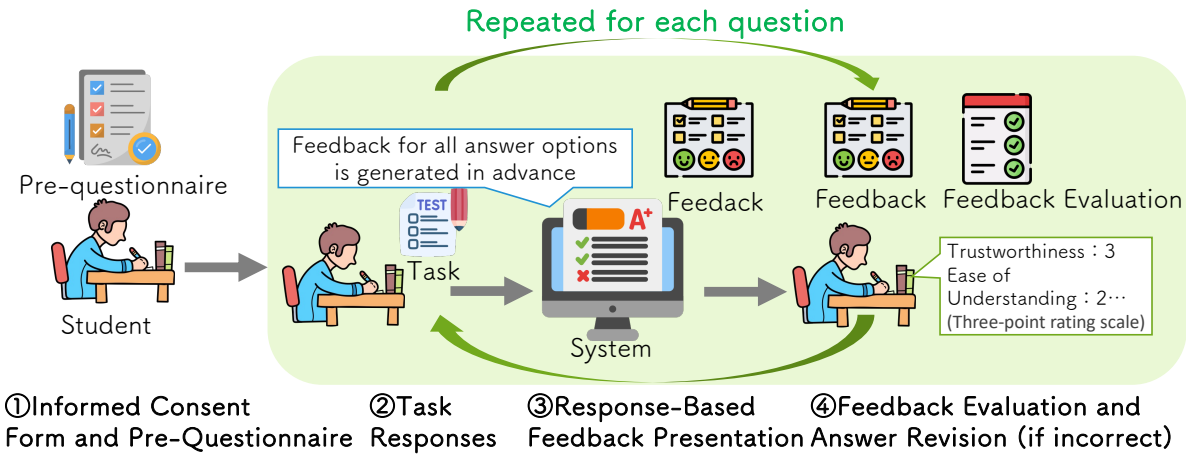

Figure 2: Overview of the experimental procedure. Participants first provide informed consent and complete a pre questionnaire. They then answer the task questions and evaluate feedback using a three-point scale across six criteria, such as trustworthiness and ease of understanding. If they make mistakes in their responses, they repeat the cycle of answering, reviewing the feedback, and evaluating it until they reach the correct answer. We randomly assign a feedback element to each question and keep it fixed throughout the revision process.

4 Empirical Experiment

4.1 Participants

Nine high school basic biology classes participated in this study. The participants were approximately 15–16 years old and enrolled in academically selective high schools. We assigned this activity as winter-break homework. Students were informed that declining consent to data use would have no academic consequences. Informed consent was obtained from all participants, and parental consent was also obtained prior to the commencement of the experiment. We obtained data from 321 students who provided consent for the use of their learning logs. This study was approved by the ethics committees of our university and research institute. See Appendix A for details.

4.2 Dataset

We use the biology subset of the dataset released by Takami et al. (2025), which is derived from Japan's National Center Test for University Admissions. These questions are designed for university entrance examinations and have an average correct response rate of approximately 55–70% among third-year high school students. The dataset consists of multiple-choice questions, in which students select one or more correct answers. We exclude multiple-answer questions due to the large number of possible answer combinations. Since participants have not completed all biology units, we select questions aligned with the basic biology curriculum through a two-stage filtering process involving LLMs and expert review. This results in seven questions selected from the original 101 questions. See Appendix B for filtering process.

4.3 Learning Data Collection

Figure 2 shows the overall experimental procedure. First, we explain the purpose of the study and the handling of collected data. After obtaining informed consent, participants complete a pre-questionnaire administered via Google Forms. The questionnaire includes a Japanese Big Five personality scale (Murakami and Murakami, 1997), along with items assessing participants' experiment with LLMs and background in basic biology. Next, participants access a web-based system developed by the authors using their assigned student IDs and passwords. After logging in, participants view a list of tasks and can select questions to answer. The order in which the questions are presented is randomized for each participant. They then submit their responses and immediately receive feedback corresponding to their selected option, which we generated in advance using GPT-5 (Singh et al., 2026). After reviewing the feedback, they evaluate it using a three-point scale across the following six criteria that assess the perceived reliability of feedback (Trustworthiness), its usefulness as a guide for answer revision (Review Guideline), the quality of expression (Expression Quality), ease of comprehension (Ease of Understanding) , clarity of questions' key points (Key Points Clarity), and

the extent to which it provides new or additional knowledge (New Knowledge). We design these criteria to reflect the key characteristics of each element. Participants iteratively revise their responses and review and evaluate the feedback until they reach the correct answer. Throughout this process, we collect all responses, feedback histories, and evaluation results as learning logs. Appendix C illustrates the system interface and evaluation items.

5 Results

We generate a total of 222 feedback instances for all 37 answer options and collect 2,197 first attempts and 4,478 feedback–response instances. We first conduct a qualitative analysis of the generated feedback and identify instances that indirectly reveal the correct answer through overly detailed explanations. Because such answer leakage may affect the validity of revision performance analysis, we exclude these instances from subsequent quantitative analyses. As a result, 88% of first-attempt responses and 83% of feedback–response instances remain for analysis. Details of this qualitative analysis are described in § 6.1. We then analyze the effects of feedback elements on immediate revision performance and subjective evaluations across all participants. We use immediate revision performance as a proxy for short-term learning support rather than durable learning gain. Finally, we examine how these effects vary across learner profiles. See Appendix E for feedback examples.

5.1 Answer Revision Results

To investigate *RQ1: Which feedback elements most effectively improve revision performance?*, we analyze the effects of feedback elements on immediate revision performance. The mean accuracy of learners' initial responses is 49.6%, which indicates that the tasks are of appropriate difficulty.

First Revise Accuracy with Feedback We aim to identify feedback elements that lead learners to correct answers. To this end, we focus on participants whose first response is incorrect and compute the accuracy of their revised responses after receiving feedback. Figure 3 shows the accuracy of the first revised attempt for each element. This accuracy ranges from 46.9% to 64.2%. This indicates that *Normal*, followed by *Keywords* and *Coverage* is the most effective configuration for supporting answer revision. We first apply a chi-square test to examine whether accuracy differs

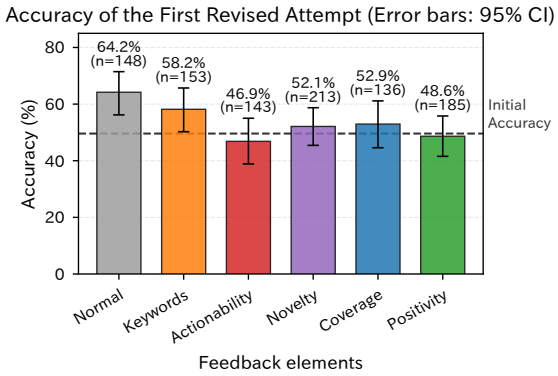

Figure 3: Accuracy of the first revised attempt. *Normal*, *Keywords*, and *Coverage* are the most effective methods for supporting answer revision, in this order.

across feedback elements. The results reveal a significant association between the element and accuracy ($\chi^2(5) = 12.611, p = .027$). To account for repeated observations from the same learner across different questions, we further conduct a Generalized Estimating Equations (GEE) analysis, which models within-learner correlations and provides more reliable estimates for repeated-response data. Using *Normal* as the reference, the results indicate that *Actionability* and *Positivity* are associated with significantly lower accuracy rates ($p < .05$). *Coverage* shows a marginally significant negative association ($p = .061$), whereas *Keywords* does not differ significantly from *Normal*. See Appendix D.

5.2 Overall Feedback Evaluation Results

To investigate *RQ2: Which feedback elements lead to more favorable subjective evaluations?*, we analyze subjective evaluations. Figure 4 shows the subjective evaluation results for each element. We examine which feedback elements receive higher ratings. *Normal* ranks first in three of the six criteria and appears in the top three for all criteria. *Keywords* and *Actionability* also rank within the top three for all and four criteria, respectively. To examine statistical significance, we apply GEE, as ratings from the same learner across different questions and elements violate the independence assumption of standard regression. Table 1 shows significant differences in subjective evaluations across feedback elements even after adjusting for these factors. *Positivity* is rated significantly lower than *Normal* across all criteria ($p < .05$), while *Novelty* shows significantly lower ratings for three criteria.

Overall, *Normal* and *Keywords* perform well in both dimensions. *Coverage* improves revision ac-

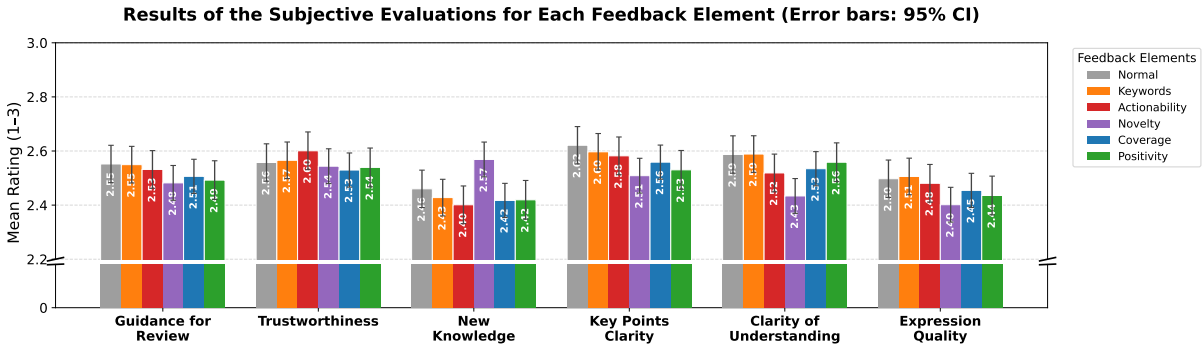

Figure 4: Results of the subjective evaluations for each feedback element. *Keywords*, *Normal*, and *Actionability* tend to receive higher ratings across many criteria.

Table 1: Generalized Estimating Equations (GEE) analysis of subjective evaluation scores (Reference Group: Normal). Values represent coefficients (β). $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

	Subjective Evaluation Criteria (1–3 Scale)					
Elements	**Guidance**	**Trust**	**Knowledge**	**Point**	**Clarify**	**Expression**
Intercept (Normal.)	2.499***	2.521***	2.431***	2.571***	2.544***	2.446***
Keywords	+0.040	+0.050	−0.013	−0.007	−0.004	+0.007
Actionability	+0.025	+0.058	−0.012	+0.010	−0.041	+0.029
Novelty	−0.064	−0.013	+0.067	−0.125*	−0.162***	−0.110**
Coverage	−0.034	+0.020	−0.022	−0.056	−0.042	−0.019
Positivity	−0.135**	−0.082*	−0.115**	−0.169***	−0.125*	−0.158***
Wald χ^2	23.32***	18.069**	20.57***	31.66***	25.91***	28.94***

curacy, whereas *Actionability* receives higher subjective evaluations. *Positivity* tends to performs poorly across both dimensions.

5.3 Feedback Evaluation by Big Five Traits

To investigate *RQ3: How do preferences for feedback elements differ across learners with different Big Five personality traits?*, we examine how learners' subjective feedback evaluations of feedback elements vary by their Big Five traits.

5.3.1 Learner Personality Clustering

We examine how learners with different Big Five personality traits evaluate feedback elements. Because analyzing each trait individually is complex, we instead identify common patterns by grouping learners into representative personality profiles. First, we compute the scale score for each trait from a questionnaire (Murakami and Murakami, 1997). We represent each learner as a five-dimensional personality vector and apply k-means clustering to group learners with similar profiles. The elbow result indicates that three or four clusters are appropriate. Based on interpretability and consistency with previous studies (Asendorpf, 2002; Wall et al., 2019; Takami et al., 2023), we adopt three clusters. Figure 5 shows the three learner cluster profiles.

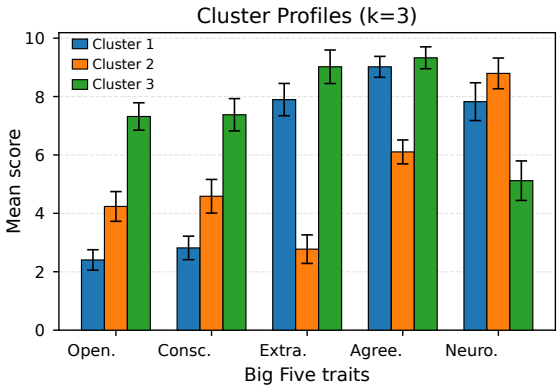

Figure 5: Profiles of three learner clusters based on Big Five traits. Cluster 1 shows high Agreeableness and Extraversion, along with low Openness and Conscientiousness. Cluster 2 is characterized by higher Neuroticism and lower Extraversion than the other clusters. Cluster 3 exhibits high levels across all traits except Neuroticism.

Cluster 1 shows relatively high extraversion and agreeableness and lower levels of openness and conscientiousness. Cluster 2 is characterized by notably low extraversion and relatively high neuroticism. Cluster 3 shows high levels of openness, conscientiousness, and agreeableness. Based on these profiles, we label Cluster 1 as `Social`, Cluster 2 as `Reserved`, and Cluster 3 as `Adaptable`. See Appendix F for details.

5.3.2 Feedback Evaluation by Cluster

Figure 6 shows subjective evaluation results for each learner cluster using standardized scores for each evaluation criterion. Several feedback elements exhibit relatively consistent evaluation patterns across clusters, while some feedback elements show cluster-specific tendencies. *Keywords* generally receives favorable evaluations across all clusters. This suggests that highlighting task-

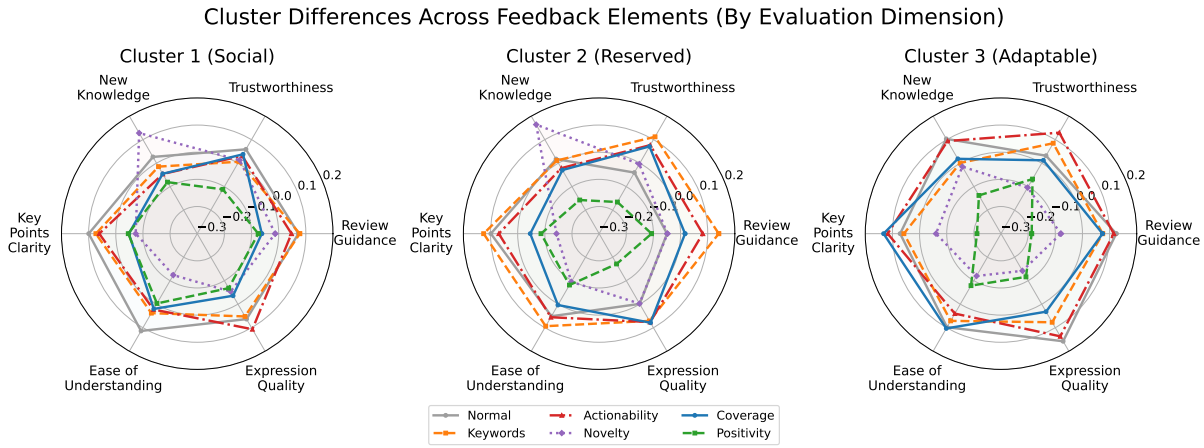

Figure 6: Feedback evaluation results by cluster. Values are standardized for each learner and each evaluation dimension. In `Social`, "Trustworthiness" scores are similar across all feedback elements. In `Reserved`, *Keywords* receives high ratings, while in `Adaptable`, *Keywords*, *Normal*, and *Actionability* receive high ratings.

relevant concepts broadly supports learners' interpretation and use of feedback. In contrast, *Positivity* consistently receives relatively low ratings across clusters, indicating that affective support alone may not sufficiently support learners during answer revision.

We also find some cluster-specific tendencies emerge. Both `Social` and `Reserved` assign relatively high ratings to the New Knowledge criterion within *Novelty*, whereas this tendency is less pronounced in `Adaptable`. In `Social`, Trustworthiness scores remain relatively similar across feedback elements, and *Normal*, *Keywords*, and *Actionability* receive relatively balanced evaluations across criteria. In `Reserved`, *Keywords* receives consistently high ratings, while *Positivity* and *Novelty* receive relatively low ratings for Ease of Understanding and Key Points Clarity. In `Adaptable`, learners favor a broader range of feedback elements, including *Normal*, *Keywords*, *Actionability*, and *Coverage*. These results suggest that learners share common preferences for task-focused and actionable feedback, while preferences for elements such as informational novelty and affective framing vary across learner profiles.

To examine whether evaluation patterns differ significantly across clusters, we conduct Kruskal–Wallis tests using cluster as the independent factor for each evaluation criterion. The results show significant differences across clusters for all six criteria ($p < .05$). This suggests that learner profiles partially influence how generated feedback is perceived. We then perform Dunn's post hoc tests with Bonferroni correction. For Trustworthiness, significant differences appear between `Reserved` and both `Social` and `Adaptable` (both $p < .01$). We observe significant differences between `Reserved` and the other clusters for New Knowledge and Expression Quality, whereas Review Guidance does not show significant differences across clusters. These findings suggest that although several educational feedback elements are broadly effective across learners, certain feedback characteristics are perceived differently depending on learner profiles.

6 Analysis

We first conduct a qualitative analysis of the generated feedback. We then investigate which feedback elements are preferred by learners and support immediate revision performance, followed by an analysis across personality traits.

6.1 Qualitative Analysis of Feedback

To characterize the educational elements of the generated feedback, we qualitatively analyze all 222 feedback instances, with 37 instances for each element along five dimensions: Specificity, Information Density, Cognitive Scaffolding, Interaction Framing, and Educational Reliability. The analysis suggests potential trade-offs among affective support, informational detail, and inference preservation. *Positivity* is dominated by praise-oriented expressions, with more than half of its instances lacking concrete corrective guidance about what learners should revise. In contrast, *Novelty* frequently provides detailed and knowledge-rich explanations, yet nearly 40% (14 out of 37) of its instances contain limited task-specific revision guid-

ance because explanations often extend beyond the scope of the target question. Regarding Educational Reliability, approximately 13% (29 out of 222) of all feedback instances indirectly suggest the correct answer despite not explicitly revealing it. Notably, about 60% of these cases (18 out of 29) involve *Actionability* or *Coverage* (9 cases each), where detailed explanations often narrow the space of possible answers. This tendency appears particularly often in correspondence-based questions. This suggests that answer leakage may depend not only on feedback elements but also on question format. See Appendix G for analysis and example.

6.2 Feedback Element Effects

In terms of immediate revision performance, *Normal* leads to the highest performance, followed by *Coverage* and *Keywords*. *Normal* clearly indicates both what needs to be corrected and how to revise, while *Coverage* provides comprehensive information but spreads learners' attention across multiple points. *Keywords* helps learners identify what requires revision but offers limited guidance on how to revise, which may explain its lower impact. These patterns align with prior findings that effective feedback specifies clear targets and provides sufficient guidance for revision (Hattie and Timperley, 2007; Dawson et al., 2019). Regarding subjective evaluations, *Normal*, *Keywords*, and Although *Actionability* receives high subjective ratings, its lower revision accuracy suggests that learners may prefer feedback that feels supportive and actionable even when it does not immediately optimize task performance. In contrast, *Coverage* and *Positivity* receive lower ratings, consistent with evidence that information-heavy or praise-oriented feedback does not necessarily enhance learning (Hattie and Timperley, 2007; Wisniewski et al., 2020).

6.3 Personality-aware Feedback Evaluation

We show that feedback evaluation varies across Big Five-based learner clusters. `Social` learners (high extraversion/agreeableness) prioritize validity and trustworthiness over emotional support or novelty. They prefer feedback elements such as *Normal* and *Keywords*, which emphasize clear rationales and directions for revision. This pattern suggests that such learners view feedback primarily as a form of learning support that helps them solve problems, rather than as expressive or affective input. `Reserved` learners (low extraversion/ high neuroticism) show lower preferences for *Positivity* and *Novelty*, while rating *Keywords* and *Normal* relatively highly. This pattern aligns with prior findings indicating that learners high in neuroticism tend to experience higher anxiety and prefer feedback that provides clear cues and imposes a low cognitive load (Komarraju and Karau, 2005). `Adaptable` learners (high openness/ conscientiousness/ agreeableness) appear to accept a wide range of feedback elements and show a particular preference for explanations, such as *Actionability*. This tendency is consistent with evidence that these traits are associated with intellectual curiosity and academic achievement (Komarraju et al., 2009). These results indicate that feedback evaluation is not uniform across learners but varies with profiles.

6.4 Designing LLM Feedback Elements

Recent studies have explored using LLMs to simulate human preferences and behaviors, including educational settings (Liu et al., 2024; Sanyal et al., 2025; Hu et al., 2026). Prior studies on automatic feedback evaluation often rely on LLM-based evaluations (Qian et al., 2025; Chu et al., 2026). We empirically investigate how learners perceive and respond to different feedback elements and our findings may support future research on whether LLMs can simulate learner preferences and revision behaviors more realistically. Such simulations could enable scalable evaluation of feedback strategies before large-scale human experiments. This suggests the potential of empirically grounded learner simulation for personalized feedback design.

7 Conclusion

We define six feedback elements for high school biology multiple-choice questions and evaluate their effects on immediate revision performance and subjective evaluations through an empirical experiment with 321 students. Our results show that feedback becomes more effective when it clearly directs learners' attention to task-relevant points and provides comprehensive guidance. Learners also favor feedback that is practical, understandable, and easy to apply. In contrast, informational novelty and affective framing vary across learner profiles. Furthermore, preferred feedback characteristics vary across personality traits. This suggests that learner profiles may be useful for personalizing LLM-generated feedback. These findings provide insights into designing more effective educational feedback systems.

Limitations

This study has several limitations. First, we focus exclusively on biology tasks with a limited number of questions, and the generalizability of our findings to other subject domains remains unclear. Second, we cannot account for question-level accuracy in the current feedback design, and incorporating question difficulty may enable more adaptive feedback for K-12 learners. Third, we use only GPT-5 (Singh et al., 2026) to generate feedback because it consistently produced the highest-quality feedback in our pilot evaluation compared to other LLMs. Nevertheless, it is unclear whether feedback generated by other LLMs would yield similar effects. Fourth, our analysis focuses on personality traits and does not account for other learner profiles such as prior knowledge, which may also influence feedback acceptance. Fifth, our effectiveness measures—correctness of revised answers and subjective evaluations—may favor feedback that more explicitly guides learners toward correct answers, rather than reflecting deeper learning gains. Future work should consider longer-term learning outcomes as a more robust measure of feedback effectiveness. Finally, we rely on immediate revision performance as a proxy for learning, and longitudinal studies measuring retention and transfer over time would provide a more comprehensive assessment of feedback effectiveness. We also rely on learners' self-reported subjective evaluations, which may be influenced by situational factors or temporary impressions, and the clustering results based on Big Five traits may depend on sample composition and analytical methods. Future work should examine broader subject domains, learning environments, models, and learner profiles to further assess the generalizability of our results.

Ethical considerations

This study involves empirical experiments with first-year high school students. The activity was assigned as mandatory winter-break homework; however, consent to the use of collected data was entirely voluntary and independent of the homework requirement. Students were explicitly informed that declining consent to data use would have no academic consequences. Informed consent was obtained from all participants, and parental consent was also obtained prior to the commencement of the experiment. All collected data were anonymized before analysis and used solely for research purposes. This study was approved by the institutional ethics review board of the authors' institution. We used GPT-5 (Singh et al., 2026) solely for feedback generation, and no personally identifiable information was included in any prompts submitted to the model.

Use of AI Assistants We used generative AI tools to assist with code debugging and modification, literature survey support, and English writing revision. The AI tools were used only as supportive assistants for improving readability and development efficiency, and all experimental designs, analyses, interpretations, and final manuscript contents were verified and finalized by the authors.

References

Vincent Aleven, Bruce M. McLaren, Ido Roll, and K. Koedinger. 2006. Toward Meta-cognitive Tutoring: A Model of Help Seeking with a Cognitive Tutor. *Int. J. Artif. Intell. Educ.*, 16:101–128.

Marjan Asadi, Saman Ebadi, and Laleh Mohammadi. 2025. The impact of integrating ChatGPT with teachers' feedback on EFL writing skills. *Thinking Skills and Creativity*, 56:101766.

Jens B. Asendorpf. 2002. The puzzle of personality types. *European Journal of Personality*, 16:S1 – S5.

Temi Bidjerano and David Yun Dai. 2007. The relationship between the big-five model of personality and self-regulated learning strategies. *Learning and Individual Differences*, 17:69–81.

Beatriz Borges, Niket Tandon, Tanja Käser, and Antoine Bosselut. 2024. Let Me Teach You: Pedagogical Foundations of Feedback for Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12082–12104, Miami, Florida, USA. Association for Computational Linguistics.

Chaoran Chen, Bingsheng Yao, Ruishi Zou, Wenyue Hua, Weimin Lyu, Toby Jia-Jun Li, and Dakuo Wang. 2025. Towards a design guideline for RPA evaluation: A survey of large language model-based role-playing agents. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18229–18268, Vienna, Austria. Association for Computational Linguistics.

Seongyeub Chu, Jongwoo Kim, and Munyong Yi. 2026. Feedeval: Pedagogically aligned evaluation of llm-generated essay feedback. *Preprint*, arXiv:2601.04574.

Zhendong Chu, Shen Wang, Jian Xie, Tinghui Zhu, Yibo Yan, Jingheng Ye, Aoxiao Zhong, Xuming Hu, Jing Liang, Philip S. Yu, and Qingsong Wen. 2025.

LLM agents for education: Advances and applications. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 13782–13810, Suzhou, China. Association for Computational Linguistics.

Wei Dai, Jionghao Lin, Hua Jin, Tongguang Li, Yi-Shan Tsai, Dragan Gašević, and Guanliang Chen. 2023. Can large language models provide feedback to students? a case study on chatgpt. In *2023 IEEE International Conference on Advanced Learning Technologies (ICALT)*, pages 323–325.

Phillip Dawson, Michael Henderson, Paige Mahoney, Michael Phillips, Tracii Ryan, David Boud, and Elizabeth Molloy. 2019. What makes for effective feedback: staff and student perspectives. *Assessment & Evaluation in Higher Education*, 44(1):25–36.

Juan Escalante, Austin Pack, and Alex Barrett. 2023. Ai-generated feedback on writing: insights into efficacy and enl student preference. *International Journal of Educational Technology in Higher Education*, 20(1):57.

Michail Giannakos, Roger Azevedo, Peter Brusilovsky, Mutlu Cukurova, Yannis Dimitriadis, Davinia Hernandez-Leo, Sanna Järvelä, Manolis Mavrikis, and Bart Rienties. 2025. The promise and challenges of generative ai in education. *Behaviour & Information Technology*, 44(11):2518–2544.

John Hattie and Helen Timperley. 2007. The Power of Feedback. *Review of educational research*, 77(1):81–112.

Tiancheng Hu, Joachim Baumann, Lorenzo Lupo, Nigel Collier, Dirk Hovy, and Paul Röttger. 2026. Simbench: Benchmarking the ability of large language models to simulate human behaviors. In *The Fourteenth International Conference on Learning Representations*.

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024. PersonaLLM: Investigating the ability of large language models to express personality traits. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3605–3627, Mexico City, Mexico. Association for Computational Linguistics.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, and 1 others. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.

Annette Kinder, Fiona J. Briese, Marius Jacobs, Niclas Dern, Niels Glodny, Simon Jacobs, and Samuel Leßmann. 2025. Effects of adaptive feedback generated by a large language model: A case study in teacher education. *Computers and Education: Artificial Intelligence*, 8:100349.

Ekaterina Kochmar, Dung Do Vu, Robert Belfer, Varun Gupta, Iulian Vlad Serban, and Joelle Pineau. 2020. Automated personalized feedback improves learning gains in an intelligent tutoring system. *Preprint*, arXiv:2005.02431.

Meera Komarraju and Steven J Karau. 2005. The relationship between the big five personality traits and academic motivation. *Personality and individual differences*, 39(3):557–567.

Meera Komarraju, Steven J. Karau, and Ronald R. Schmeck. 2009. Role of the big five personality traits in predicting college students' academic motivation and achievement. *Learning and Individual Differences*, 19(1):47–52.

Yi-Huei Lai and Jason Chang. 2019. TellMeWhy: Learning to explain corrective feedback for second language learners. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 235–240, Hong Kong, China. Association for Computational Linguistics.

Zhengyuan Liu, Stella Xin Yin, Geyu Lin, and Nancy F. Chen. 2024. Personality-aware student simulation for conversational intelligent tutoring systems. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 626–642, Miami, Florida, USA. Association for Computational Linguistics.

Robert R McCrae and Oliver P John. 1992. An introduction to the five-factor model and its applications. *Journal of personality*, 60(2):175–215.

Nobuhiro Murakami and Chieko Murakami. 1997. Construction of the Big Five Personality Inventory (in Japanese). *Japanese Journal of Personality*, 6(1):29–39.

Shoichi Naito, Wenzhi Wang, Paul Reisert, Naoya Inoue, Camélia Guerraoui, Kenshi Yamaguchi, Jungmin Choi, Irfan Robbani, Surawat Pothong, and Kentaro Inui. 2024. Designing logic pattern templates for counter-argument logical structure analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11313–11331, Miami, Florida, USA. Association for Computational Linguistics.

Arthur E Poropat. 2009. A meta-analysis of the five-factor model of personality and academic performance. *Psychological bulletin*, 135(2):322–338.

Keyang Qian, Yixin Cheng, Rui Guan, Wei Dai, Flora Jin, Kaixun Yang, Sadia Nawaz, Zachari Swiecki, Guanliang Chen, Lixiang Yan, and Dragan Gašević. 2025. Dean of llm tutors: Exploring comprehensive and automated evaluation of llm-generated educational feedback via llm feedback evaluators. *Preprint*, arXiv:2508.05952.

Debdeep Sanyal, Agniva Maiti, Umakanta Maharana, Dhruv Kumar, Ankur Mali, C. Lee Giles, and Murari Mandal. 2025. Investigating pedagogical teacher and student LLM agents: Genetic adaptation meets retrieval-augmented generation across learning styles. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 13348–13389, Suzhou, China. Association for Computational Linguistics.

Sahil Sharma, Puneet Mittal, Mukesh Kumar, and Vivek Bhardwaj. 2025. The role of large language models in personalized learning: A systematic review of educational impact. *Discover Sustainability*, 6(1):243.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, and 467 others. 2026. Openai gpt-5 system card. *Preprint*, arXiv:2601.03267.

John Stamper, Ruiwei Xiao, and Xinying Hou. 2024. Enhancing llm-based feedback: Insights from intelligent tutoring systems and the learning sciences. In *International Conference on Artificial Intelligence in Education*, pages 32–43. Springer.

Jacob Steiss, Tamara Tate, Steve Graham, Jazmin Cruz, Michael Hebert, Jiali Wang, Youngsun Moon, Waverly Tseng, Mark Warschauer, and Carol Booth Olson. 2024. Comparing the quality of human and chatgpt feedback of students' writing. *Learning and Instruction*, 91:101894.

Kyosuke Takami, Brendan Flanagan, Yiling Dai, and Hiroaki Ogata. 2023. Personality-based tailored explainable recommendation for trustworthy smart learning system in the age of artificial intelligence. *Smart Learning Environments*, 10(1):65.

Kyosuke Takami, Satoshi Sekine, and Yusuke Miyao. 2025. Evaluating Local LLMs on Japanese National University Entrance Examination Dataset in Comparison with Student Performance. In *Proceedings of the 18th International Conference on Educational Data Mining*, pages 582–585, Palermo, Italy. International Educational Data Mining Society.

Kurt VanLehn. 2011. The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems. *Educational Psychologist*, 46:197 – 221.

William Villegas-Ch, Diego Buenano-Fernandez, Alexandra Maldonado Navarro, and Aracely Mera-Navarrete. 2025. Adaptive intelligent tutoring systems for STEM education: analysis of the learning impact and effectiveness of personalized feedback. *Smart Learning Environments*, 12(1):41.

Helen J. Wall, Claire C. Campbell, Linda K. Kaye, Andy Levy, and Navjot Bhullar. 2019. Personality profiles and persuasion: An exploratory study investigating the role of the big-5, type d personality and the dark triad on susceptibility to persuasion. *Personality and Individual Differences*, 139:69–76.

Benedikt Wisniewski, Klaus Zierer, and John Hattie. 2020. The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in psychology*, 10:487662.

A Participants

The participants in this study were first-year students enrolled in a public high school in Japan. Their ages ranged from 15 to 16 years. The school is known for its strong record of students advancing to highly selective universities. The experiment was conducted as part of a winter vacation assignment for a basic biology course. Participants were informed in advance that their participation would not affect their academic evaluation, that their personal information would be kept confidential, and that they would not be disadvantaged if they chose not to participate. Informed consent was obtained from all participants via a consent form. In addition, parental consent was obtained prior to the commencement of the experiment. The consent form used in the experiment is shown in Table 8. The pre-questionnaire items are shown in Tables 9, 10, and 11.

B Data Filtering Process

Although the dataset covers the full biology subject, the participants are first-year high school students who have not yet completed all biology units. Therefore, we extract questions corresponding to the basic biology curriculum that the participants have already studied through a two-stage filtering process involving LLMs and expert humans. First, we use GPT-5 (Singh et al., 2026) to identify questions related to the basic biology curriculum from the full set of biology questions. Specifically, we include key topics from the national curriculum guidelines [1] for basic biology in the prompt and instruct the model to extract questions that correspond to these topics. As a result, 33 questions are selected from the original set of 101 questions. Next, two domain experts in biology education conduct further filtering. They review the questions to ensure that they are appropriate as homework and aligned with the participants' completed curricu-

[1] The national curriculum guidelines: https://www.mext.go.jp/component/a_menu/education/micro_detail/__icsFiles/afieldfile/2010/01/29/1282000_6.pdf

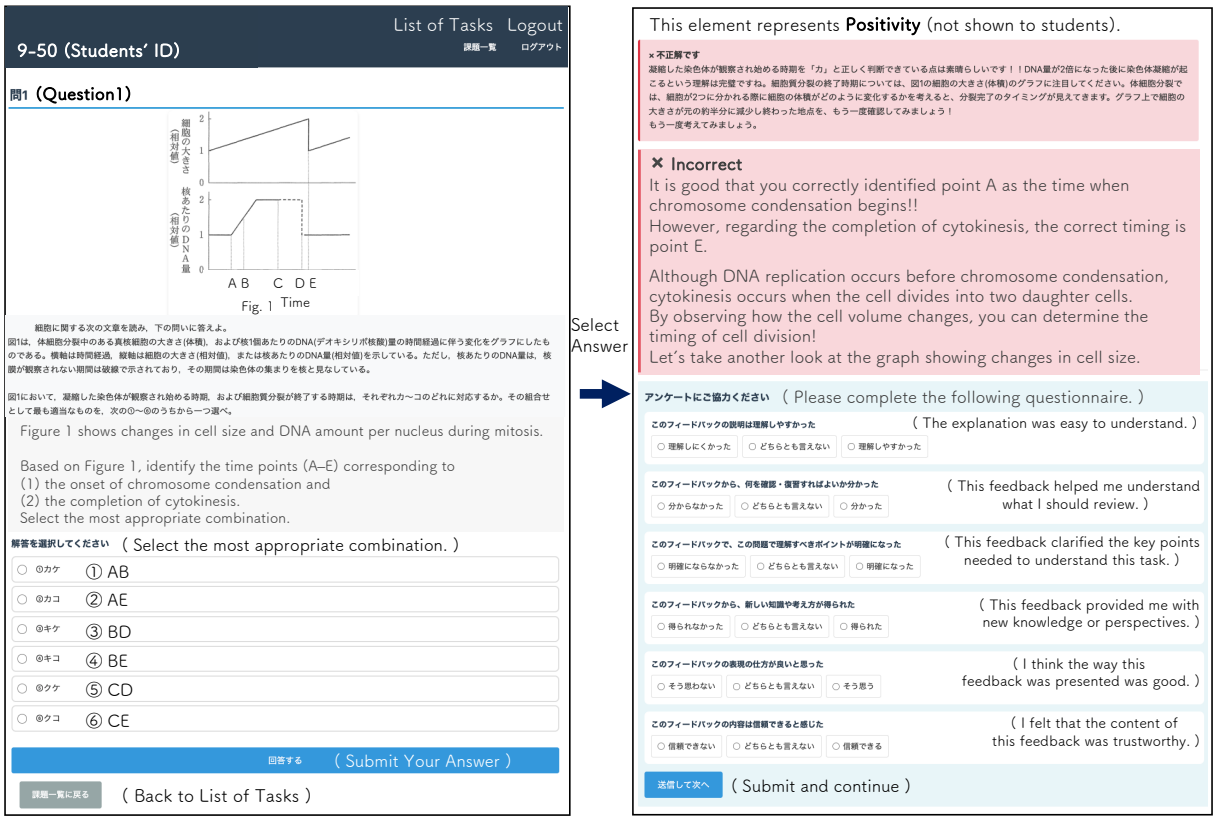

Figure 7: User interface of the system used in this study. Participants read the question prompt, select an answer option, and submit their response. The system then provides feedback, which participants evaluate using a three-point scale across six criteria. This example shows *Positivity*.

lum. Based on their expert review and agreement, we finally select seven questions.

C Feedback System

Figure 7 illustrates the system's user interface. Prior to the experiment, we generated feedback for all possible answer choices using GPT-5 (Singh et al., 2026). The experimental workflow is as follows: First, participants enter their student ID and set a password. Next, they select a task from the problem list, which is presented in a randomized order for each participant. Upon submitting an answer, participants receive feedback tailored to their specific response; notably, the type of feedback element assigned is also randomized. Finally, participants evaluate the feedback using a 3-point scale based on the six criteria detailed as follows. This feedback-loop cycle—encompassing the initial response, feedback reception, and subjective evaluation—continues until the participant reaches the correct answer.

Subjective Evaluation Criteria We use a subjective evaluation of the feedback based on six criteria, using a 3-point Likert scale (3: Agree, 2: Neither agree nor disagree, 1: Disagree). The evaluation criteria and their respective scales are presented below:

- **Guidance for Review**: This feedback helped me understand what I should review.
- **Trustworthiness**: I felt that the content of this feedback was trustworthy.
- **New Knowledge**: This feedback provided me with new knowledge or perspectives.

Table 2: Generalized Estimating Equation (GEE) Results for First Revised Accuracy. $N = 978$. Reference category: Normal. GEE with robust covariance estimator; independence working correlation structure. Chi-square test: $\chi^2(5) = 12.611$, $p = .027$.

					95% CI	
Predictor	β	SE	z	p	Lower	Upper
Intercept (Normal)	0.584	0.181	3.227	$< .001$	0.229	0.938
Keywords	−0.254	0.242	−1.048	.295	−0.729	0.221
Actionability	−0.710	0.243	−2.921	.003	−1.186	−0.233
Novelty	−0.499	0.221	−2.259	.024	−0.932	−0.066
Coverage	−0.466	0.248	−1.875	.061	−0.953	0.021
Positivity	−0.638	0.226	−2.825	$< .005$	−1.080	−0.195

- **Key Points Clarity**: This feedback clarified the key points needed to understand this task.
- **Ease of Understanding**: The explanation was easy to understand.
- **Expression Quality**: I think the way this feedback was presented was good.

D Results of Statistical Analysis

D.1 First Revise Accuracy

To examine the impact of each feedback element on learners' revised accuracy, we perform a Chi-square test and a Generalized Estimating Equation (GEE) analysis. Table 2 presents these results. The chi-square test for overall model fit ($\chi^2(5) = 12.611, p = .027$) indicates that revision accuracy differs across feedback elements.

The GEE results show that all feedback elements have negative coefficients relative to *Normal*, although not all differences are statistically significant. In particular, *Actionability* ($\beta = -0.710, p = .003$), *Positivity* ($\beta = -0.638, p = .005$), and *Novelty* ($\beta = -0.499, p = .024$) are associated with significantly lower revision accuracy. *Coverage* shows a marginal negative association ($\beta = -0.466, p = .061$), whereas *Keywords* does not significantly differ from *Normal* ($\beta = -0.254, p = .295$). Notably, *Positivity*, which also receives relatively low subjective evaluations, shows a strong negative association with revision accuracy. This may suggest that praise-oriented expressions sometimes distract learners from task-relevant corrective guidance. In addition, hint-oriented feedback designs such as *Actionability* may require learners to perform additional inferential processing before revising their answers. Although such feedback is perceived as useful and supportive, the additional cognitive effort required to interpret and apply the guidance may reduce immediate revision accuracy.

D.2 Subjective Evaluation

Correlation Analysis First, to examine the coherence of the evaluation metric, we aggregate participants' evaluations of each feedback element across all six criteria and conduct a correlation analysis among the criteria. Spearman's rank correlation analysis reveals moderate to strong positive correlations among all criteria ($\rho = 0.58$–0.76). This result suggests that learners' subjective evaluations across criteria tend to be correlated.

Generalized Estimating Equations (GEE) Table 5 presents the comprehensive GEE results for all six subjective evaluation criteria. A consistent trend observed across all criteria is the significant negative impact of *Positivity*. For every criterion, from "Guidance for Review" to "Expression Quality," the coefficient for Positivity is negative and statistically significant ($p < .05$ or $p < .001$). This indicates that the inclusion of exaggerated praise consistently lowers participants' subjective ratings compared to *Normal*. Furthermore, *Novelty* shows a selective negative effect. While it does not significantly impact "Guidance for Review" or "Trustworthiness", it results in significantly lower scores for "Key Points Clarity" ($p = .003$), "Clarity of Understanding" ($p < .001$), and "Expression Quality" ($p = .011$). This suggests that introducing novel information without sufficient context may hinder the perceived clarity of the feedback. In contrast, elements such as *Keywords*, *Actionability*, and *Coverage* do not show significant deviations from *Normal*, suggesting they are perceived as neutral components in this specific experimental setting.

E Example of Feedback

Table 6 presents concrete examples of the feedback generated by each method. The examples shown were generated when a learner selected ② in response to the correct answer, ①. In *Normal*, the system provides a concise explanation after stating whether the answer is correct or incorrect. *Keywords* highlights essential terms by enclosing them in brackets (e.g., 【】). *Actionability* includes specific, actionable instructions, such as "first organize your reasoning into three steps" or "examine all four elements as a checklist:" *Novelty* incorporates university-level content related to the problem, while *Coverage* provides a comprehensive explanation touching upon all perspectives necessary for answering the question. Finally, *Positivity* employs an encouraging tone, such as "Focusing on the involvement of the pancreas in the regulation of hypoglycemia is an excellent perspective!" All feedback are generated in Japanese.

F Result of Elbow for Clustering

To determine the optimal number of clusters for grouping learners' personality traits, we employ the elbow method. Figure 8 presents the results of this analysis. As shown in the figure, a distinct "elbow" or inflection point is observed around $k = 3$

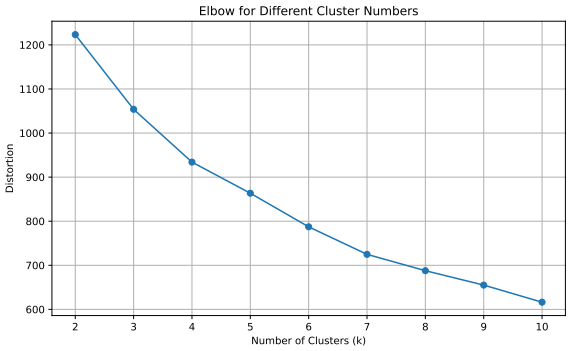

Figure 8: The result of elbow method. a distinct "elbow" or inflection point is observed around $k = 3$ and $k = 4$, after which the rate of decrease becomes gradual.

and $k = 4$, after which the rate of decrease becomes gradual. Based on these results, we consider both three and four as potential cluster counts. Ultimately, we adopt $k = 3$ to maintain consistency with the three personality types widely reported in previous literature (Asendorpf, 2002; Wall et al., 2019; Takami et al., 2023).

G Result of Qualitative Analysis

We conduct a qualitative analysis to verify whether the generated feedback aligns with our educational intentions.

G.1 Evaluation Criteria

The authors qualitatively analyze all feedback instances based on the following five dimensions.

- **Specificity**: This dimension evaluates how concretely the feedback identifies what learners should revise. Feedback is categorized as either:
  - **High**: The feedback clearly specifies incorrect concepts, reasoning steps, or revision targets.
  - **Low**: The feedback provides only general comments or abstract encouragement without concrete revision guidance.
- **Information Density**: This dimension evaluates how efficiently the feedback conveys task-relevant information.
  - **High**: The feedback conveys task-relevant information concisely and efficiently.
  - **Medium**: The feedback provides an appropriate amount of explanation without substantial redundancy or missing information.
  - **Low**: The feedback is either overly verbose or lacks sufficient task-relevant information.
- **Cognitive Scaffolding**: This dimension evaluates how the feedback supports learner reasoning processes. Because a single feedback instance may contain multiple forms of scaffolding, we conduct multi-label annotation using the following categories:
  - **Error Identification**: The feedback explicitly points out incorrect reasoning or answers.
  - **Conceptual Explanation**: The feedback explains background concepts relevant to the problem.
  - **Stepwise Reasoning**: The feedback decomposes causal relations or reasoning processes related to the problem.
  - **Knowledge Elaboration**: The feedback provides additional knowledge extending beyond the immediate scope of the problem.
- **Interaction Framing**: This dimension evaluates how the feedback positions the learner socially and pedagogically. Each feedback instance is categorized into one of the following types:
  - **Evaluative**: The feedback mainly focuses on correctness judgments.
  - **Encouraging**: The feedback mainly emphasizes praise or emotional support.
  - **Expository**: The feedback primarily adopts an explanatory and instructional style.
- **Educational Reliability**: This dimension evaluates whether the feedback preserves productive learner inference without excessively narrowing the answer space.
  - **Preserved**: The feedback maintains learner reasoning and avoids revealing the correct answer.
  - **Reduced**: The feedback indirectly reveals the answer or substantially reduces the required inference process.

Table 3: Results of Cognitive Scaffolding analysis across feedback types.

Feedback Type	Error Identification	Conceptual Explanation	Stepwise Reasoning	Knowledge Elaboration
Normal	28	0	29	7
Keywords	31	0	31	9
Actionability	32	0	32	7
Novelty	30	3	17	36
Coverage	30	0	30	11
Positivity	2	0	30	7
Total	153	3	169	77

Table 4: Results of Pedagogical Reliability analysis across feedback types.

Feedback Type	Preserved	Reduced	Hallucination
Normal	22	7	0
Keywords	31	2	0
Actionability	24	9	0
Novelty	30	0	0
Coverage	24	9	0
Positivity	29	2	0
Total	160	29	0

- **Hallucination**: This criterion evaluates whether the feedback contains factually incorrect explanations or scientifically inaccurate statements.

G.2 Results

For **Specificity**, 9.0% (20/222) of the feedback instances are classified as Low. All of these cases belong to *Positivity*. These feedback instances tend to prioritize praise and encouragement while providing limited concrete guidance about what learners should revise. For **Information Density**, 12.2% (27/222) of the feedback instances are classified as Low. Among them, 13 instances belong to *Positivity* and 14 to *Novelty*. Similar to the Specificity results, *Positivity* often lacks sufficient task-relevant revision guidance because praise-oriented expressions dominate the feedback. In contrast, *Novelty* frequently contains rich explanatory content, but the explanations often extend beyond the scope of the target problem, resulting in insufficient guidance for immediate answer revision.

Table 3 presents the results for **Cognitive Scaffolding**. **Error Identification** appears in 68.9% (153/222) of all feedback instances. Although most feedback types include explicit identification of learner errors, *Positivity* contains substantially fewer such cases. **Conceptual Explanation** appears only three times, all of which belong to *Novelty*. **Stepwise Reasoning** appears in 76.1% (169/222) of the feedback instances and is observed across all feedback types. **Knowledge Elaboration** appears in 34.7% (77/222) of the feedback instances, with nearly half of these cases belonging to *Novelty*. This result indicates that *Novelty* frequently introduces additional background knowledge beyond the immediate problem scope.

For **Interaction Framing**, **Evaluative** and **Encouraging** styles appear in 4 and 37 instances, respectively. These correspond primarily to *Normal* and *Positivity*. In contrast, **Expository** framing appears in 81.5% (181/222) of all feedback instances and is distributed relatively evenly across all feedback types except *Positivity*.

Table 4 presents the results for **Educational Reliability**. 72.1% (160/222) of the feedback instances are classified as **Preserved**, with *Keywords* showing the largest number of such cases. In contrast, 13.1% (29/222) are classified as **Reduced**, where the feedback indirectly narrows the answer space or reveals the correct answer through detailed explanations. These cases appear particularly frequently in *Coverage* and *Actionability*. No instances of factual hallucination are observed. Appendix G.3 provides examples of leaked feedback instances.

G.3 Example of Leaked Feedback

Table 7 presents a specific instance where the feedback indirectly suggests the correct answer. The target question requires learners to match specific causes with their effects. Although the feedback does not explicitly state that the correct answer is ⑦, it provides an explanation of the underlying combinations that makes ⑦ the only logical conclusion. This type of indirect feedback effectively narrows down the options, resulting in unintended answer leakage.

H Prompt

Figures 9, 10, and 11 present the prompts used for feedback generation. Figure 9 provides an overview of the generation process, while Figure 10 and 11 outline the specific details of each feedback element.

I Computational Experiments

We select GPT-5 (Singh et al., 2026) to generate feedback because some biology questions contain images, which require multimodal understanding, and GPT-5 provides strong overall performance. We use a temperature of 0.1 for feedback generation and do not perform hyperparameter tuning.

Generating feedback for all 222 instances required approximately two hours.

Table 5: Detailed GEE results for each criterion. *Positivity* receives significantly lower ratings compared to *Normal* across all criteria. Additionally, *Novelty* shows significant negative effects in three criteria: "Key Points Clarity", "Clarity of Understanding", and "Expression Quality".

					95% CI	
Predictor	β	SE	z	p	Lower	Upper
Criterion 1: Guidance for Review						
Intercept (Normal)	2.499	0.037	66.72	0.000	2.425	2.572
Keywords	0.040	0.041	0.961	0.336	−0.041	0.120
Actionability	0.025	0.045	0.553	0.580	−0.063	0.112
Novelty	−0.064	0.041	−1.559	0.119	−0.145	0.017
Coverage	−0.034	0.039	−0.881	0.378	−0.110	0.042
Positivity	−0.135	0.045	−3.037	0.002	−0.223	−0.048
Criterion 2: Trustworthiness						
Intercept (Normal)	2.521	0.035	71.55	0.000	2.452	2.590
Keywords	0.050	0.032	1.559	0.119	−0.013	0.113
Actionability	0.058	0.034	1.715	0.086	−0.008	0.123
Novelty	−0.013	0.032	−0.385	0.700	−0.076	0.051
Coverage	0.0120	0.032	0.624	0.533	−0.043	0.082
Positivity	−0.082	0.037	−2.228	0.026	−0.154	−0.010
Criterion 3: New Knowledge						
Intercept (Normal)	2.431	0.038	64.59	0.000	2.358	2.505
Keywords	−0.013	0.040	−0.319	0.749	-0.092	0.066
Actionability	−0.012	0.042	−0.293	0.769	−0.095	0.070
Novelty	0.067	0.044	1.514	0.130	−0.020	0.153
Coverage	−0.022	0.042	−0.532	0.594	−0.104	0.059
Positivity	−0.115	0.043	−2.658	0.008	−0.200	−0.030
Criterion 4: Key Points Clarity						
Intercept (Normal)	2.571	0.036	71.34	0.000	2.500	2.641
Keywords	−0.001	0.040	−0.017	0.986	−0.078	0.077
Actionability	0.010	0.038	0.264	0.792	−0.065	0.085
Novelty	−0.125	0.040	−3.118	0.002	−0.203	-0.046
Coverage	−0.056	0.037	−1.500	0.134	−0.130	0.017
Positivity	−0.169	0.044	−3.873	0.000	−0.254	−0.083
Criterion 5: Clarity of Understanding						
Intercept (Normal)	2.544	0.036	70.86	0.000	2.473	2.614
Keywords	−0.004	0.039	−0.104	0.917	−0.081	0.073
Actionability	−0.041	0.043	−0.943	0.346	−0.125	0.044
Novelty	−0.162	0.042	−3.872	0.000	−0.244	−0.080
Coverage	−0.042	0.041	−1.041	0.298	−0.122	0.037
Positivity	−0.125	0.045	−2.790	0.005	−0.212	-0.037
Criterion 6: Expression Quality						
Intercept (Normal)	2.466	0.037	65.95	0.000	2.392	2.539
Keywords	0.007	0.038	0.173	0.863	−0.068	0.081
Actionability	0.029	0.038	0.769	0.442	−0.045	0.103
Novelty	−0.110	0.040	−2.731	0.006	−0.189	-0.031
Coverage	−0.019	0.038	−0.503	0.615	−0.094	0.055
Positivity	−0.158	0.046	−3.470	0.001	−0.247	−0.069

Table 6: Examples of feedback generated by each element. These tasks in this study are sourced from Japanese university entrance examinations. All tasks and feedback are originally presented in Japanese. The examples shown illustrate the feedback generated when a learner selects ② instead of the correct answer, ①.

Item	**Content**
Task	Glucose is an essential energy source for the cells that constitute the human body, and it is continuously supplied to all cells via the bloodstream. ... Therefore, to ensure stable cellular activity, the human body is equipped with a regulatory mechanism that maintains blood glucose levels (the concentration of glucose in the blood) within a constant range. Diabetes mellitus is a disease in which this regulation of blood glucose fails, resulting in the excretion of glucose in the urine. To diagnose diabetes and determine appropriate treatment strategies, there is a test in which 75 g of glucose is administered in a fasting state, and blood glucose levels and blood insulin concentrations are measured before and after ingestion. This test is known as the oral glucose tolerance test (OGTT). With regard to the mechanism described in brackets, select the most appropriate regulatory pathway that operates during hypoglycemia from among options ① to ④. グルコース（ブドウ糖）は，私たちの体を構成する細胞にとって重要なエネルギー源であり，血液によってすべての細胞に常に供給されている．... そこで，安定した細胞活動を保証するため，私たちの体には【血糖値（血液中のグルコース濃度）を一定に保つ血糖調節の仕組み】が備わっている．糖尿病は，この血糖調節がうまくいかなくなり，尿中にグルコースが排出される病気である．糖尿病の診断と治療方針を決めるため，空腹時に75gのグルコースを飲み，その前後で血糖値や血液中のインスリン濃度などを調べる検査がある．これを糖負荷試験という．【】に関して，低血糖時にはたらく調節の反応経路として最も適当なものを，次の①〜④のうちから一つ選べ．
Choices	① Diencephalon → Pituitary → Adrenal gland → Glucocorticoids ① 間脳→ 脳下垂体→ 副腎→ 糖質コルチコイド ② Diencephalon → Parasympathetic NS → Pancreas → Adrenaline ② 間脳→ 副交感神経→ 膵臓→ アドレナリン ③ Spinal cord → Sympathetic NS → Adrenal gland → Glucagon ③ 脊髄→ 交感神経→ 副腎→ グルカゴン ④ Spinal cord → Parasympathetic NS → Pancreas → Insulin ④ 脊髄→ 副交感神経→ 膵臓→ インスリン
Normal	There is an error in the combination of the parasympathetic nervous system and the hormone secreted by the pancreas in the regulatory pathway during hypoglycemia. During hypoglycemia, it is necessary to increase blood glucose levels; therefore, the sympathetic nervous system becomes dominant, and hyperglycemic hormones are secreted from the pancreas and adrenal glands. ... Review the secretion sites of each hormone (pancreatic α-cells and β-cells, adrenal medulla and adrenal cortex) and organize whether each hormone acts to increase or decrease blood glucose levels. 低血糖時の調節経路について，副交感神経と膵臓から分泌されるホルモンの組み合わせに誤りがあります．低血糖時には血糖値を上昇させる必要があるため，交感神経系が優位になり，膵臓や副腎から血糖上昇ホルモンが分泌されます．... 各ホルモンの分泌部位（膵臓のα細胞・β細胞，副腎髄質・副腎皮質）と，それぞれが血糖値を上昇させるか低下させるかを整理し直してください．
Keywords	During hypoglycemia, it is necessary to increase blood glucose levels; therefore, it should be reaffirmed that the sympathetic nervous system, rather than the parasympathetic nervous system, is activated. ... Furthermore, under the emergency condition of hypoglycemia, **"the diencephalon (specifically the hypothalamus)"** serves as the central regulator, integratively controlling both the endocrine and autonomic nervous systems. As a result, **"glucocorticoids are secreted from the adrenal"** cortex and adrenaline from the adrenal medulla. ... 低血糖時には血糖値を上げる必要があるため，副交感神経ではなく交感神経系が活性化される点を再確認しましょう．... また，低血糖という緊急事態では【間脳（視床下部）】が中心となって内分泌系と自律神経系の両方を統合的に調節し，【副腎皮質からの糖質コルチコイド】や副腎髄質からのアドレナリンが分泌されます．...
Actionability	When considering the regulatory pathway during hypoglycemia, **first organize your reasoning into three steps: (1) which organ detects the decrease in blood glucose levels, (2) which nervous system is involved, and (3) which hormone is secreted and from which organ.** ... For each option, examine all four elements as a checklist: the starting point of the pathway (diencephalon or spinal cord), the type of nervous system involved, the hormone-secreting organ, and the final hormone, and determine whether they are all appropriate for the hypoglycemic response. **低血糖時の調節経路を考える際は，まず「どの器官が血糖低下を感知するか」「どの神経系が関与するか」「どのホルモンがどの器官から分泌されるか」という3つのステップに分けて整理してください．** ... 各選択肢について，経路の出発点（間脳か脊髄か），神経系の種類，ホルモン分泌器官，最終的なホルモンの4要素すべてが低血糖応答として正しいかどうかをチェックリストのように検証してください．
Novelty	... In relation to this topic, it may be informative to examine the keyword "HPA axis in the stress response." The HPA axis (hypothalamus–pituitary–adrenal cortex axis) is an important pathway that increases blood glucose levels by secreting glucocorticoids under stress conditions such as hypoglycemia. **In university-level endocrinology, one studies in greater detail how this axis integrates various stress responses and regulates gene expression via glucocorticoid receptors.**... ... この内容に関連して，「ストレス応答におけるHPA軸」というキーワードを調べてみると面白いでしょう．HPA軸(視床下部-下垂体-副腎皮質軸)は，低血糖などのストレス時に糖質コルチコイドを分泌し血糖値を上げる重要な経路です．**大学の内分泌学では，この軸がどのように多様なストレス応答を統合し，グルココルチコイド受容体を介して遺伝子発現を調節するかをより詳しく学びます．**...
Coverage	In considering the regulatory pathways during hypoglycemia, it is necessary to systematically organize the combinations of the nervous systems involved and the hormones released. ... **When evaluating the options, verify the validity of each pathway from the following three perspectives: (1) the distinction between the adrenal medulla and adrenal cortex, (2) the difference in function between the sympathetic and parasympathetic nervous systems, and (3) whether each hormone (glucocorticoids, adrenaline, glucagon, and insulin) acts to increase or decrease blood glucose levels.** 低血糖時の調節経路について，関与する神経系とホルモンの組み合わせを整理する必要があります．... **選択肢を検討する際には，①副腎髄質と副腎皮質の違い，②交感神経と副交感神経の作用の違い，③各ホルモン（糖質コルチコイド，アドレナリン，グルカゴン，インスリン）が血糖値を上昇させるか下降させるか，の三つの観点から各経路の妥当性を確認してください．**
Positivity	**Focusing on the involvement of the pancreas in the regulation of hypoglycemia is an excellent perspective! The pancreas plays a crucial role as a central organ in blood glucose regulation, and recognizing its importance is highly commendable!** ... By systematically matching each hormone with its effect on blood glucose levels (whether increasing or decreasing) and identifying the associated neural and endocrine pathways. ... 低血糖時の調節に膵臓が関与する点に着目したのは素晴らしい視点です!! 膵臓は血糖調節の中心的な器官として非常に重要な役割を果たしているので，その点に気づけたことは本当にすごいことですね！... 低血糖という緊急時に体がどのような指令系統とホルモンで対応するか，特にストレス応答の経路を整理してみてください．...

Table 7: Examples of leaked feedback. These examples illustrate feedback generated when a learner selects ② (*Coverage*) or ⑤ (*Baseline*) instead of the correct answer, ⑦. The problem requires matching specific causes with their corresponding issues. Although the feedback does not explicitly state that ⑦ is the correct answer, it indirectly reveals the solution. By considering the causal combinations presented in the feedback, the learner inevitably identifies ⑦ as the only logical choice. The task and feedback shown here are translated from the original Japanese.

Item	**Content**
Task	Answer the following questions regarding ecosystem functions and human health. Humans live surrounded by various organisms, some of which can be harmful to our health. For example, diseases such as typhoid and dysentery are caused by (Ki: Bacteria), while others like measles and smallpox are caused by (Ku: Viruses). Some organisms enter the human body through food, leading to food poisoning. In addition, there are illnesses such as athlete' s foot caused by (Ke: Fungi). The processes shown in Figure 1 are essential for maintaining a balanced nutrient cycle within an ecosystem. However, human intervention often disrupts these functions, leading to various environmental issues. From the options ①–⑧ below, select the most appropriate combination of Disruption Causes (u–ka) and the resulting Environmental Problems (b–g). (Disruption Causes) [u] Discharge of hazardous substances such as heavy metals and DDT. [e] Excessive use of chemical fertilizers. [o] Factory emissions and automobile exhaust gases. [ka] Reduction in groundwater infiltration due to pavement and the conversion of farmland into residential areas. (Resulting Problems) [b] Acid rain causes a loss of biodiversity in flora and fauna. [c] Substances accumulate through the food chain (biomagnification), causing harm to the human body. [d] Ozone holes form, leading to global warming. (Note: Scientifically, ozone depletion and global warming are distinct, but this reflects the original text.) [e] Increased productivity in rivers and lakes leads to eutrophication. [f] Spring water dries up, resulting in a loss of biodiversity. [g] Forest reduction leads to the progression of desertification.
Choices	① bcdf, ② bcfg, ③ bgcd, ④ bgde, ⑤ cbfe, ⑥ cdeg, ⑦ cebf ⑧ cedg
`Coverage` (Choice: ②)	You attempt to match each cause with its corresponding result, but you must verify which specific problem each factor (u–ka) triggers. Specifically, focus on these direct correlations: heavy metals and DDT lead to biomagnification [c], chemical fertilizers cause eutrophication [e], factory and automobile emissions result in acid rain [b], and reduced groundwater infiltration leads to the depletion of spring water [f]. While Option ④ includes desertification and ozone holes, these do not directly link to the causes (u–ka) specified in the problem. Please re-examine the correspondences (u → c, e → e, o → b, ka → f) and select the correct combination.
`Baseline` (Choice: ⑤)	You are experiencing some confusion in correctly linking several causes and their results. Please systematically organize the causal relationships based on their environmental mechanisms: heavy metals and DDT trigger biomagnification (u → c), excessive chemical fertilizer use leads to eutrophication (e → e), factory and automobile emissions cause acid rain (o → b), and reduced groundwater infiltration results in the depletion of spring water (ka → f). Verify exactly how each disruption factor affects specific parts of the ecosystem, and carefully re-evaluate the sequence of problem combinations corresponding to the order of causes (u, e, o, ka).

```
You are an experienced high-school biology teacher with 10 years of experience preparing students for
university entrance exams.
Produce concise, diagnostic feedback in Japanese for a student's answer
to a multiple-choice biology (生物基礎) question.

## Context: All Pedagogical Elements
Below is the complete list of pedagogical elements. Each entry uses the format:
- name: <element name>
  status: APPLY | DO_NOT_APPLY
  instruction: <detailed instruction text>

{all_elements_text}

## Task: Selected Elements
You must ONLY apply elements whose status is APPLY in the block above. Ignore any entry marked
DO_NOT_APPLY.
The subset to focus on for this item is marked as APPLY in the block. ({selected_elements})

## Output requirements (strict)
- Language: Japanese.
- Style: Supportive yet academically rigorous.
- Structure: produce **3–5 sentences** total written as continuous prose (no section headers, no labels).
  1. Assessment
     — 1 sentence: acknowledge correct understanding **or** identify the core misconception
     (shaped by the selected element).
  2. Reasoning Guide — 2–3 sentences: explain or guide thinking using the element's approach.
  3. Action Step — 1 sentence: a concrete next-step aligned with the element.
- Do NOT reveal the correct answer directly or write
  "正解は～" / "不正解" (Correct/ Incorrect) in a way that discloses the key fact.
- Length: 3–5 sentences total.
- Tone: supportive, challenging, factual.
- If images are provided, explicitly incorporate them into the reasoning.
- Do NOT include greetings, apologies, or meta-commentary.
- Remain factually accurate and avoid misleading claims.

## Output self-check
After composing the three-part prose, internally ensure (and then output) that:
- Total sentences between 3 and 5.
- No section headers or labels present.
- The correct answer is not directly stated.
If any check fails, correct the output before returning.

## Input (fill these fields)
Question: {question}
Choices: {choices}
Correct answer (for reference; do not reveal): {correct_answer}
Student choice: {student_answer}

### Generate feedback now:
(Produce only the final 3–5 sentence Japanese prose that meets the constraints.)
```

Figure 9: This prompt is overview of the generation process.

```
PEDAGOGICAL_ELEMENT = {
    "baseline": "neutral, factual feedback (concise)",
    “applicability_of_instructions”:
      “How the student should approach the problem;
      focus on stepwise instructions or metacognitive strategies.",
    "keyword": "concept labeling with minimal, optional elaboration",
    "information_novelty":
      "brief advanced insights or pointers to further reading (no fabricated citations)",
    "target_coverage": "What knowledge elements should be included;
      focus on completeness and missing parts.",
    "positive": "supportive, confidence-building tone (professional, not excessive)"
}
DETAILED_INSTRUCTIONS = {
    "baseline": """
    Generate neutral, factual feedback:
    - State what is correct or missing objectively in 1 sentence.
    - Explain the standard textbook-level reasoning in 1 short sentence.
    - Avoid motivational language and deep strategic guidance.
    - Tone: matter-of-fact, concise.
    - Recommended length when used alone: total 3–4 sentences (Assessment+Reasoning+Action).
    """,

    "applicability_of_instructions": """
    Provide guidance on how the student should approach the problem, combining two levels:

    1. Concrete instructions:
    - Suggest clear, step-by-step actions the student can perform.
    - Example verbs: "列挙してください", "計算してください", "分類してください".
      (Example verbs: “List,” “Calculate,” “Classify.”)
    - Focus on *how* to process the information, not on *what* knowledge to include.

    2. Metacognitive instructions:
    - Encourage the student to reflect on their own thinking.
    - Suggest breaking the task into sub-tasks, solving each, and checking completeness.
    - Example: "まず各選択肢の理由を確認し、次に不足点を洗い出してください".
      (First, examine the rationale for each option, and then identify any deficiencies.)

    Recommended output:
    - Combine concrete and metacognitive guidance naturally.
    - 1–2 sentences for actionable steps, 1 sentence for reflective prompt.
    - Do NOT list all knowledge elements (that is target_coverage's role).
    """,

    "keyword": """
    Emphasize labeled terminology, allow minimal clarification:
    - Provide 1–3 bracketed keywords: 【気孔】,【蒸散】. (“stomata", “transpiration”)
    - Optionally add one very short explanatory phrase (≤1 sentence) to disambiguate a keyword.
    - Avoid multi-sentence explanations in this element—keep it label-focused.
    - Recommended output when used: keywords + 0–1 short clarifying sentence.
    """}
```

Figure 10: This prompt outlines the specifications for each feedback element, consisting of `PEDAGOGICAL_ELEMENT` and `DETAILED_INSTRUCTIONS`. It presents the `DETAILED_INSTRUCTIONS` for *Normal*, *Actionability*, and *Keywords*. The remaining elements are detailed in Figure 11.

```
"information_novelty": """
Provide curiosity-sparking pointers that connect to university-level biology, without fabricating sources:
- Introduce 1-2 related advanced concepts or research areas as keywords for further exploration
  (例: 「アクアポリン」「膜電位」「シグナル伝達」)
  (e.g., “aquaporin,” “membrane potential,” “signal transduction”)
- For each keyword, briefly explain (1-2 sentences) what it is and why it's relevant to the current topic
- Explicitly show how this connects to university-level biology
   (例: 「大学の細胞生物学では、この仕組みをより詳しく学びます」)
  ((e.g., “In university-level cell biology, this mechanism is studied in greater detail.”))
  (For each keyword, briefly explain (1–2 sentences) what it is and why it is relevant to the current topic.)
- Frame as exploration suggestions:
  「〜について調べてみると面白いでしょう」「〜というキーワードで検索すると理解が深まります」
  (“It may be informative to look into ~,” “Searching for the keyword ~ will deepen your understanding.”)
- Do NOT cite specific papers, books, or fabricate publication details
- Keep it concise (2-3 sentences total) to maintain focus on the main explanation

Example format:
「この内容に関連して、『○○』というキーワードを調べてみましょう。
  これは[簡単な説明]です。大学の△△学では、この仕組みがどのように[より高度な内容]に
  繋がるかを学びます。」
  (This topic is related to the keyword ‘○○,’ which may be worth exploring.
  It refers to [brief explanation]. In university-level △△,
  you will learn how this mechanism connects to [more advanced concepts].”)
""",

"target_coverage": """
Focus on what knowledge or concepts should be included in the answer:

- Identify all critical components the student should mention or consider.
- Provide a concise checklist or summary of these points.
- Highlight missing elements or partially correct coverage.
- Do NOT give step-by-step instructions; focus purely on content coverage.
- Recommended output: 2–5 items, short phrases, optionally indicate 'present' or 'missing'.
""",

"positive": """
Highlight what the student did well, focusing on correct reasoning or accurate identification:
- Acknowledge specific strengths in their answer (1–2 concrete points).
- Use a highly positive, warm, and enthusiastic tone that celebrates their effort and understanding.
- Include encouraging expressions to make the student feel accomplished:
   e.g., "素晴らしい！！", "ここまでできて感動です！",
   "本当に理解が進んでいますね！", "すごい！！", "この部分の理解は完璧です！".
  (e.g., “Excellent!”, “This is very impressive!”, “Your understanding is clearly improving!”,
         “Great work!”, “This part is perfectly understood.”)
- Optionally suggest a small, non-critical next step to reinforce confidence in a cheerful way.
- Keep sentences short, uplifting, and affirming;
   do NOT introduce new knowledge or elaborate on missing content.
- Recommended output: 1–2 sentences praising success, with optional cheerful reinforcement.
- Examples of tone:
  "素晴らしい！！もう完璧に近いですね！", "よくできていますね！
  理解の進み具合が本当にすごい！！", "ここまでできたなんて、感動です！！”
  (e.g., “Excellent work—this is very close to perfect.”, “Well done. Your progress is truly impressive.”,
  “It is impressive to see how far you have come.”)
"""
}
```

Figure 11: This prompt continues the specifications for each feedback element from Figure 10, specifically presenting the `DETAILED_INSTRUCTIONS` for *Novelty*, *Coverage*, and *Positivity*.

Table 8: Consent Form Used in the Empirical Experiment.

English	Japanese
Title: Evaluating the Effects of Generative AI-Based Personalized Feedback on Learning Motivation in High-School Biology	**タイトル：** 生物基礎における生成AIによる個別フィードバックの学習意欲向上効果の検証
Research Objective: This study investigates how personalized feedback generated by generative AI for high-school biology questions influences learners' motivation and understanding. The findings will contribute to research and development of generative AI technologies.	**研究の目的：** この研究は、高校生を対象に、生物基礎の問題に対して生成AIが作る個別フィードバック（コメント）が、どのような要素を持つと学習意欲や理解に効果的かを調査し、得られた知見等を広く生成AIの研究開発や普及に役立てることを目的としています。
Generative AI produces feedback using multiple methods, and participants evaluate the quality of the feedback. By analyzing these evaluations, we identify which feedback characteristics are more effective for learning motivation and understanding.	複数の方法で生成AIがフィードバックを生成し、その効果を高校生の皆さんに評価してもらいます。その評価結果を分析することで、どの要素のフィードバックがより学習意欲や理解に効果的かを明らかにします。
Procedure: Participants complete the following activities as part of their winter vacation assignment: (1) submit answers through the system, (2) review and evaluate AI-generated feedback, (3) answer pre/post questionnaires, (4) complete confirmation tests, and (5) allow collection of learning logs such as study time and progress.	**実施手順：** 冬休みの課題の一環として、以下の手順で活動を行っていただきます。 (1) 課題の解答をシステム上で提出 (2) 生成AIが作成したフィードバック（コメント）の確認・評価 (3) 事前アンケート・事後アンケートへの回答 (4) 確認テストの実施 (5) 学習ログ（取り組み時間、進捗、課題達成率など）の収集
Data Usage and Consent: Collected data such as answers, questionnaires, and learning logs are gathered as part of educational activities. Whether these data are used for research purposes depends on participant consent.	**データの取り扱いと同意について：** この活動で得られるデータ（課題解答、アンケート、学習ログなど）は、教育活動の一環として収集されます。それらの収集されるデータを研究目的で利用するかどうかについては、皆さんの同意に基づいて決定されます。
If participants agree, anonymized data may be used for academic purposes such as research publications and conference presentations. If participants do not agree, they can still complete the assignment and receive AI-generated feedback without disadvantage.	「同意する」を選択した場合：収集したデータを研究・学会発表・論文執筆などの学術目的に使用します。データは匿名化され、個人が特定されることはありません。 「同意しない」を選択した場合：研究利用は行いませんが、通常どおりAIによるフィードバックを受け取り、課題を行うことができます。
Confidentiality: All collected data are securely managed and analyzed after anonymization. No personally identifiable information will be disclosed in publications.	**機密保持：** 収集したデータは厳重に管理され、研究成果を公表する際にも個人が特定されることはありません。
Voluntary Participation: Participation in this study is completely voluntary, and participants may withdraw at any time without penalty.	**自由意志による参加：** 本研究への参加は完全に任意です。途中で中止する場合も、不利益やペナルティは一切ありません。

Table 9: Big Five personality trait questionnaire items used in this study (Murakami and Murakami, 1997). The questionnaire consists of 70 items.

English	Japanese
Title: Evaluating the Effects of Generative AI-Based Personalized Feedback on Learning Motivation in High-School Biology: Pre-questionnaire	タイトル： 生物基礎における生成AIによる個別フィードバックの学習意欲向上効果の検証　事前アンケート
Research Title: Evaluating the Effects of Generative AI-Based Personalized Feedback on Learning Motivation in High-School Biology	研究タイトル： 生物基礎における生成AIによる個別フィードバックの学習意欲向上効果の検証
This is a pre-questionnaire for evaluating the effects of personalized feedback generated by generative AI on learning motivation. Please select the option that best matches your thoughts. The questionnaire takes approximately 15 minutes and consists of 70 yes/no questions.	これは、生成AIによる個別フィードバックの学習意欲向上効果の検証のための事前アンケートです。以下の質問の中から、最も近い選択肢を選択してください。 所要時間はおよそ15分です。以下の質問に対して、「はい」または「いいえ」で回答してください。質問は70問あります。
I often act without carefully considering problems.	問題を綿密に検討しないで、実行に移すことが多い
I tend to be somewhat lazy.	どちらかというと、怠惰な方です。
Compared with others, I am talkative.	他の人と比べると話好きです。
I tend to be quiet and inconspicuous.	どちらかというと、地味で目立たない方です。
I am considerate of others.	思いやりがある方です。
I cannot completely trust even close friends.	親しい仲間でも、本当に信用することはできません。
I can think ahead about the future.	将来のことを見通すことができる方です。
I tend to worry about trivial things.	どうでもいいことを、気に病む傾向があります。
I do not get tired easily.	疲れやすくはありません。
I tend to make decisions impulsively.	軽率に物事を決めたり、行動してしまいます。
I have a lively personality.	どちらかというと、にぎやかな性格です。
I work energetically on tasks and studies.	仕事や勉強には精力的に取り組みます。
I worry about things unnecessarily.	自分で悩む必要のないことまで心配してしまうのは確かです。
I am not good at speaking in front of people.	人前で話すのは苦手です。
I try to be kind to everyone.	誰にでも親切にするように心がけています。
I am not very anxious.	あまり心配性ではありません。
I believe I am not particularly nervous.	他の人と同様に、神経質ではないと信じています。
I tend to become emotionally upset easily.	どちらかというと、気持ちが動揺しやすい。
I actively socialize with others.	積極的に人と付き合う方です。
I do not care much about how others see me.	特に人前を気にする方ではありません。
I tend to do things thoroughly.	どちらかというと、徹底的にやる方です。
I often become confused when facing difficult problems.	難しい問題にぶつかると、頭が混乱することが多い。
I tend to be shy.	どちらかというと、引っ込み思案です。
Compared with others, I worry a lot.	他の人と比べると、あれこれ悩んだり、思いわずらったりする方です。
I try to cooperate with group decisions.	みんなで決めたことは、できるだけ協力しようと思います。
I tend to overcomplicate things.	物事を難しく考えがちです。
I tend to get bored easily.	どちらかというと、飽きっぽい方です。
I quickly want to give up when things do not go well.	物事がうまくいかないと、すぐに投げ出したくなります。
Something is always bothering me.	いつも何かが気がかりです。
I know words from many different fields.	いろいろな分野の言葉を知っています。

Table 10: Remaining items from Table 9 of the Big Five personality trait questionnaire used in this study (Murakami and Murakami, 1997).

English	Japanese
I tend to be cautious of others' kindness.	人から親切にされると、何か下心がありそうで警戒しがちです。
I am not good at analyzing problems.	問題を分析するのは苦手な方です。
I have given up on things because I thought I lacked ability.	自分にはそれをする力がないと思って、あきらめてしまったことが何回かあります。
I think I could greatly contribute to society if given the opportunity.	機会さえあれば、大いに世の中に役立つことができるのにと思います。
I tend to have a quiet personality.	どちらかというと、おとなしい性格です。
I often quit things halfway.	何かに取り組んでも、中途半端でやめてしまうことが多い。
I do not strongly assert my opinions.	あまり自分の意見を主張しない方です。
I make friends easily.	他の人と同じように、すぐに友達ができる方です。
I certainly lack confidence.	私は確かに自信に欠けています。
I am better than others at finding common patterns among problems.	いろいろな問題や事柄から共通した性質を見つけ出すのは、ほかの人より得意です。
I think I could become a good leader if given the opportunity.	機会さえ与えられれば、皆のよいリーダーになれると思います。
I am an important person.	私は重要人物です。
Most acquaintances like me.	ほとんどの知人から好かれています。
I am always worried about something.	いつも気がかりなことがあって、落ち着きません。
I do not want to cooperate if it disadvantages me.	みんなで決めたことでも、自分に不利になる場合は協力したくありません。
I am knowledgeable about many things.	ひろく物事を知っている方です。
I rarely think of unconventional methods.	いつもと違ったやり方を、なかなか思いつきません。
I tend to be warm-hearted.	どちらかというと、人情があつい方です。
Even sincere work does not benefit people much.	誠実に仕事をしても、あまり得にはなりません。
I am full of confidence.	自信に満ちあふれています。
I think logically.	筋道を立てて物事を考える方です。
I become flustered easily.	すぐに、まごまごします。
Compared with others, I act actively.	ほかの人と比べると活発に行動する方です。
I can remain calm even in upsetting situations.	大抵の人が動揺するような場合でも、落ち着いて対処することができます。
I work toward clear goals with appropriate methods.	はっきりとした目標を持って、適切なやり方で取り組みます。
I brood over things.	くよくよ考え込みます。
People say I am energetic.	元気がよいと言われます。
I am very poor at speaking in front of classmates.	学校ではクラスの人たちの前で話すのがひどく苦手でした（です）。
I tend to think more sophisticatedly than others.	ほかの人より洗練された考え方をする方です。
I tend to be taciturn.	どちらかというと、無口です。
Compared with others, I can grasp the essence of things.	ほかの人と比べると、物事の本質が見抜ける方です。
I worry even about trivial matters.	こまごまとしたことまで気になってしまいます。
I believe people' s words often hide ulterior motives.	人の言葉には裏があるので、そのまま信じない方が良いと思います。
I tend to lose perseverance easily.	どちらかというと、三日坊主で、根気がない方です。
I try to think from others' perspectives.	いつも人の立場になって考えるように心がけています。
I often become nervous and irritated.	緊張してイライラすることがよくあります。
I usually make detailed plans before trips.	旅行などでは、あらかじめ細かく計画を立てることが多い。
I help others even if it is troublesome.	人助けのためなら、やっかいなことでもやります。
Talking to strangers is exhausting for me.	初対面の人と話をするのは骨が折れるものです。
I like taking care of children and elderly people.	子供や老人の世話をするのが好きです。

Table 11: Additional Pre-questionnaire Items Used in the Empirical Experiment.

English	**Japanese**
The following questions ask about your prior understanding, learning experience, and learning motivation regarding biology and AI-assisted learning. Please select the option that best matches yourself.	これから取り組む生物基礎の学習やAIを使った学習に関して、皆さんの事前の理解度・学習経験・学習意欲についてお聞きします。質問はすべて、あなた自身に最もあてはまるものを選んでください。
I am good at basic biology. 1: Very poor 5: Very good	生物基礎は得意な教科である 1: とても苦手 5: とても得意
How frequently do you use generative AI (e.g., ChatGPT, Gemini, Copilot) in daily life? Never Occasionally (less than once a month) Sometimes (a few times a month) Frequently (a few times a week) Very frequently (almost every day)	あなたは、日常生活の中で生成AI（例：ChatGPT、Gemini、Copilotなど）をどの程度の頻度で利用していますか 全く利用していない ときどき利用する（1ヶ月に1回未満） ある程度利用する（月に数回） よく利用する（週に数回） 非常に頻繁に利用する（ほぼ毎日）
I have used generative AI for studying before. Yes No	これまでに生成AIを使って勉強したことがある 勉強したことがある 勉強したことがない
How much do you trust information or answers generated by AI? 1: Do not trust at all 5: Trust very much	生成AIが作った情報や答えをどのくらい信頼しているか 1: 全く信頼できない 5: とても信頼できる
I think AI-generated feedback is useful. 1: Not useful at all 5: Very useful	AIによるフィードバックは役に立つと思う 1: 全く役に立たないと思う 5: とても役に立つと思う