

Using Large Language Models to Generate Feedback for Educational Assessments: A Comparative Exploratory Study

Lucy Gitiria, Anna Nasyrova, Peter Tran, and Stephen G. Sireci

University of Massachusetts Amherst

DOI: <https://doi.org/10.51584/IJRIAS.2026.11080130>

Received: 02 September 2026; Accepted: 07 September 2026; Published: 17 September 2026

ABSTRACT

Feedback in online environments can be an essential learning tool. However, prior research indicates that poorly designed feedback, particularly when punitive or generic, can undermine student motivation and self-perception (Schartel, 2012). In educational assessment, high-quality feedback can provide learners with information about their performance and guidance for subsequent learning. Hattie and Timperley (2007) conceptualized effective feedback around three key questions: (a) Where am I going? (b) How am I doing? and (c) Where do I go next? From this perspective, useful feedback does more than indicate whether a response is correct; it helps learners understand the nature of their performance and identify appropriate next steps.

In digital assessment environments, feedback can be tailored to learners' responses and performance. Research suggests that such formative, adaptive feedback can support learning, particularly when it is specific, supportive, and prompts learners to reconsider their performance (Narciss et al., 2014; Shute, 2008). However, manually developing individualized feedback requires substantial time and resources (Leibold & Schwarz, 2015).

Multiple-choice assessments offer a useful context for response-specific feedback. Distractors may be designed to represent common errors or misconceptions, allowing feedback to address reasoning plausibly associated with a selected response (Glassey & Bälter, 2020). Response-contingent feedback can provide information beyond simple verification of whether an answer is correct or incorrect by addressing features of the learner's specific response (Shute, 2008). Feedback is also an important component of using multiple-choice assessment for formative learning purposes (Butler, 2018). However, selecting an incorrect response alone does not necessarily reveal the reasoning process that produced it. Consequently, explanations attributed to a particular distractor should be treated cautiously unless the underlying misconception has been independently established. This distinction matters especially when Large Language Models (LLMs) are asked to infer plausible errors associated with distractors.

Recent advances in large language models (LLMs) have created new opportunities to generate individualized assessment feedback. Emerging research has examined whether LLM-generated feedback reflects the characteristics of effective human feedback. For example, Dai et al. (2024) compared feedback generated by GPT models with human instructor feedback, evaluating its readability, inclusion of effective feedback components, and reliability. Their findings demonstrate the potential of automated feedback generation but also underscore the need to evaluate its mathematical accuracy and pedagogical appropriateness. Even plausible explanations may be incorrect, generic, overly verbose, or based on unsupported assumptions about why a learner selected a particular response.

The quality of LLM-generated feedback may also depend substantially on prompt design. Different prompting approaches can influence model responses and task performance. For example, providing in-context demonstrations can shape model performance in few-shot settings (Brown et al., 2020), while prompts that include exemplars with explicit intermediate reasoning steps can affect performance on reasoning tasks (Wei et al., 2022). In an educational context, prompts may therefore need to specify not only the content of the requested feedback but also its purpose, structure, level of detail, and intended audience. Before using LLM-generated feedback with learners, evaluate its characteristics against criteria grounded in established feedback research.

Recent work in educational measurement has emphasized assessment approaches that are responsive to individual learners and assessment purposes (Sireci et al., 2024). The present study was conducted within the Adult Skills Assessment Program (ASAP), which develops digitally delivered literacy and numeracy assessments for adult learners (Adult Skills Assessment Program [ASAP], n.d.). The investigation focused on multiple-choice numeracy items associated with the Curriculum for Adults Learning Math (CALM). Using these items, the study developed a structured prompting approach for generating response-specific, hint-based formative feedback, then compared feedback from three publicly accessible conversational artificial intelligence (AI) platforms on a separate holdout set of items. The comparison focused on five literature-informed characteristics: specificity, corrective guidance, personalization, constructiveness, and usability. The study evaluated the characteristics of the generated feedback rather than its effects on learner performance; therefore, it makes no claims about whether the feedback improves learning outcomes.

Research Questions

The study addressed the following research questions:

1. What prompting approach can be used to generate response-specific formative feedback for multiple-choice adult diagnostic assessment items?
2. How does the feedback generated by three publicly accessible conversational AI platforms using a standardized prompt compare with respect to specificity, corrective guidance, personalization, constructiveness, and usability?

METHODOLOGY

Study Design

This study used a two-phase, exploratory, comparative design to assess publicly accessible conversational AI platforms' ability to generate formative feedback on multiple-choice adult numeracy assessment items. The first phase involved developing and refining a prompting strategy using a subset of assessment items. The second phase used a separate holdout set of items to compare feedback generated by three platforms using the same standardized prompt. Separating the prompt-development items from the holdout items reduced the possibility that the final comparisons would be based on items that had already influenced prompt development. The study evaluated the characteristics of AI-generated feedback; no learners participated, and no learner data were analyzed.

Assessment Context and Item Selection

Assessment items were drawn from the Adult Skills Assessment Program (ASAP), an initiative that develops digitally delivered assessments of literacy and numeracy skills for adult learners (ASAP, n.d.). The items used in this study were aligned with the Curriculum for Adults Learning Math (CALM) and focused on adult numeracy. We had access to and authorization to use the items for research purposes; however, the item bank is not publicly available.

We purposively selected 57 multiple-choice mathematics items covering fractions and probability from the items available to the research team. We selected these content areas because suitable items were available. Each item contained four response options: one keyed response and three distractors.

Five items were randomly selected from the 57-item pool for prompt development. After establishing the final prompting strategy, the remaining 52 items were used for the holdout comparison. Thus, none of the items used to develop or refine the prompting strategy were included in the final comparison of the three LLMs.

Development of Feedback Evaluation Criteria

The evaluation framework was informed by research on effective educational and formative feedback. Hattie and Timperley (2007) conceptualized effective feedback as information that helps learners understand their

goals, their current performance, and the actions needed for improvement. Research on formative and adaptive feedback further emphasizes the importance of feedback that responds to learners' specific errors rather than providing generic information (Narciss et al., 2014). Research on multiple-choice assessment supports the instructional value of response-specific feedback (Butler, 2018), while work on effective feedback emphasizes clarity, constructive communication, and guidance that facilitates improvement (Leibold & Schwarz, 2015; Schartel, 2012; Shute, 2008). Because judgments of feedback quality may involve subjective, higher-order considerations, explicit evaluation criteria can make the basis for such judgments more transparent (Glassey & Bälter, 2022).

Drawing on this literature, we evaluated five dimensions: specificity, corrective guidance, personalization, constructiveness, and usability. Table 1 summarizes the theoretical basis, operational meaning, and primary evaluation question associated with each criterion.

The five criteria guided the research team's qualitative judgments during the review. To make the distinctions underlying those judgments more transparent, the criteria are described using four interpretive levels: Low, Limited, Adequate, and High. Low indicates that feedback largely fails to exhibit the relevant characteristic; Limited indicates partial but insufficient evidence of the characteristic; Adequate indicates feedback that generally satisfies the criterion but has identifiable limitations; and High indicates clear and consistent evidence that the criterion has been met. These levels describe the qualitative distinctions used to interpret the outputs and were not treated as a numerical scoring scale.

Table 1. Conceptual and Operational Framework for Evaluating AI-Generated Feedback

Operational Criteria	Theoretical Emphasis	Operational Definition	Primary Evaluation Question
Specificity	Feedback should provide specific information about current performance and the nature of the performance gap (Hattie & Timperley, 2007; Shute, 2008)	Extent to which feedback identifies the mathematical or procedural issue associated with an answer option.	Does the feedback identify the mathematical or procedural issue associated with the response?
Corrective Guidance	Effective feedback should help learners determine what to do next (Hattie & Timperley, 2007; Schartel, 2012)	Extent to which feedback provides an appropriate direction, action, or next step for reconsidering or correcting the response.	Does the feedback provide a concrete and relevant direction for correction or further reflection?
Personalization	Feedback can be adapted to a learner's particular response, error, or relevant learning characteristics (Narciss et al., 2014; Glassey & Bälter, 2020)	Extent to which feedback is tailored to the answer option and its plausibly associated error.	Is the feedback meaningfully tailored to the selected option rather than being generic?
Constructiveness	Effective feedback should support and facilitate improvement without relying on punitive or discouraging language (Schartel, 2012; Shute, 2008; Leibold & Schwarz, 2015)	Extent to which feedback uses supportive, non-punitive language while facilitating recognition and correction of an error.	Is the feedback supportive while still clearly communicating what requires correction?
Usability	Feedback should be clear, appropriately detailed, and sufficiently concise to be usable	Extent to which feedback is understandable, concise,	Is the feedback clear, concise, appropriately detailed, and

	by the intended recipient (Shute, 2008; Leibold & Schwarz, 2015)	appropriately detailed, and suitable for adult learners.	understandable for an adult learner?
--	--	--	--------------------------------------

The original item-development records containing intended distractor rationales were unavailable for this analysis; therefore, statements generated by an LLM explaining why a learner may have selected a particular distractor were treated as plausible inferred explanations. The evaluation distinguished the feedback's mathematical correctness from the plausibility of the model's interpretation of the error.

Prompt Development

We randomly selected five of the 57 items to develop the prompting strategy. We developed prompts iteratively, starting with minimal instructions consistent with zero-shot prompting. Subsequent iterations introduced additional structure and examples consistent with few-shot prompting approaches (Brown et al., 2020). We also experimented with prompts requesting explicit stepwise mathematical explanations, informed by research showing that exemplars containing intermediate reasoning steps can influence model performance on mathematical and other reasoning tasks (Wei et al., 2022). These procedures were intended to structure the generated responses and should not be interpreted as providing access to the models' internal reasoning processes.

Over successive development cycles, the research team reviewed the feedback and refined the instructions to encourage error-specific explanations, corrective guidance, supportive language, response-specific feedback, and concise wording appropriate for adult learners. The resulting hybrid prompt combined expert-role framing, structured output instructions, error-specific guidance, and supportive language. Appendix 1 provides the final prompt.

The intended pedagogical approach was hint-based formative feedback. Specifically, for incorrect responses, feedback was intended to help learners reconsider their reasoning without immediately revealing the keyed answer. Consequently, premature disclosure of the keyed answer was considered an undesirable output characteristic during review.

Once established, we fixed the final prompt and made no further modifications for the holdout comparison.

Large Language Models and Generation Procedure

We evaluated three publicly accessible conversational AI platforms: OpenAI's ChatGPT-4o, Google's Gemini 2.5 Pro, and Microsoft's Copilot. We purposively selected these platforms because they were publicly accessible, general-purpose conversational AI systems available to the research team during the study period. The selection was intended to provide an exploratory comparison of accessible commercial systems rather than a comprehensive sample of available LLM platforms. We accessed the platforms in July and early August 2025 via publicly accessible conversational interfaces rather than API-based implementations. Consequently, generation parameters such as temperature, top-p, and random seed were not exposed to the researchers and could not be standardized across platforms. For the final comparison, we submitted the same standardized prompt and the same holdout items to each platform. We submitted each item in a new conversation session to minimize carryover effects from prior interactions and obtained one generation per item-platform combination.

Review and Evaluation of Generated Feedback

The team used their disciplinary expertise in educational measurement and the five literature-informed criteria to review the outputs collectively. We examined feedback at the level of individual answer options using the five literature-informed criteria described earlier. During the review, the team discussed discrepancies collectively, referring to the criterion definitions and relevant feedback literature, until it reached a consensus judgment.

RESULTS

Prompt Development Findings

The first phase of the study examined how different prompting approaches shaped the feedback generated for the five prompt-development items. Minimally structured prompts, consistent with zero-shot prompting, could yield brief or generic responses that merely identified an answer as incorrect without explaining the underlying mathematical issue or offering useful corrective guidance. For example, responses such as "Incorrect, try again" provided little information about why an option was incorrect or how a learner might reconsider the response.

As we introduced additional structure and examples, the feedback became more closely aligned with the characteristics the research team targeted. Few-shot prompts that included examples of the desired output yielded clearer, more structured responses, although some explanations remained general or failed to adequately address a plausible error associated with a particular answer option. Prompts requesting explicit stepwise mathematical explanations generated more detailed responses but also exhibited undesirable characteristics, including excessive detail and, in some cases, information that revealed the solution more directly than intended.

These observations informed successive revisions to the prompting strategy. We added instructions emphasizing option-specific explanations, plausible mathematical errors, corrective guidance, supportive language, and concise wording appropriate for adult learners. For example, we revised prompts to ask the model to address a possible calculation error for each incorrect option rather than simply stating that the option was incorrect. During development, this change elicited more targeted explanations, such as "This may result from adding only the whole numbers and miscalculating the fractions," instead of generic corrective statements.

The prompt-development process resulted in a hybrid prompting strategy that combined expert-role framing ("Consider yourself a numeracy expert"), structured output requirements, response-specific error explanations, corrective guidance, and supportive language. Compared with the earlier prompt formulations examined during development, the final prompt produced feedback that the research team judged to be more closely aligned with the five target characteristics. Once selected, the team fixed the final prompt and used it without further modification in the holdout comparison.

Comparative Findings Across Platforms

The final comparison included the holdout items not used during prompt development. Each item contained four answer options, and we reviewed feedback for each option separately. Table 2 summarizes the principal characteristics observed during the review across the five evaluation criteria.

Table 2 Qualitative Comparison of the Three AI platforms Across the Five Evaluation Criteria

Evaluation Criterion	ChatGPT	Gemini 2.5 Pro	Copilot
Specificity	High – Generally option-specific; some inferred error pathways were uncertain.	High – Detailed and option-specific; some inferred pathways did not clearly account for the distractor.	Adequate – Sometimes specific, but several explanations remained general.
Corrective Guidance	High – Generally provided concrete mathematical next steps.	High – Provided clear corrective guidance, though sometimes at excessive length.	Adequate – Guidance was present but sometimes nonspecific.
Personalization	High – Feedback was generally differentiated across answer options and related to features of the	High – Feedback was individualized across distractors, although some inferred explanations were	Adequate – Some feedback was option-specific, while other responses remained

	selected response.	uncertain.	generic.
Constructive ness	High – Generally supportive and non-punitive.	High – Frequently encouraging and supportive.	High – Generally constructive, though sometimes brief.
Usability	High – Generally clear and accessible, with occasional verbosity.	Adequate – Detailed and supportive, but often too wordy for efficient use.	Adequate – Generally clear but sometimes underdeveloped.

Note. Low, Limited, Adequate, and High are qualitative interpretive descriptors reflecting the research team's consensus judgments. They are not numerical scores.

Overall, the consensus review found that the three platforms differed in how consistently their feedback exhibited the five target characteristics. ChatGPT was characterized as High across all five criteria, though the team noted occasional concerns about verbosity and the plausibility of inferred error pathways. Gemini 2.5 Pro was also characterized as High on several criteria, particularly specificity, corrective guidance, personalization, and constructiveness, though lengthy responses sometimes reduced usability. Copilot was characterized as having a constructive tone but showed comparatively greater limitations in specificity, corrective guidance, personalization, and usability. Next, we present an illustrative cross-platform example highlighting these patterns.

Illustrative Cross-Platform Comparison

An example from the holdout set illustrates both the strengths and limitations observed in the generated feedback. Figure 1 presents an item requiring addition of two fractions, and Figure 2 presents selected feedback generated by the three platforms. The research team purposively selected the item because it exemplified a pattern noted during review of the holdout outputs. Across multiple outputs, the team observed that models could provide mathematically plausible explanations for relatively direct distractors while offering more uncertain or mathematically unsupported explanations for distractors with less evident error pathways. We present the example as illustrative rather than as an estimate of its prevalence across the holdout set.

Figure 1 Illustrative Item

Item: A delivery driver completed a route of $\frac{5}{8}$ miles in the morning and another $\frac{3}{10}$ miles in the evening. How far did the driver travel in total that day? A: $\frac{8}{18}$ B: $\frac{64}{80}$ C: $\frac{50}{40}$ D: $\frac{37}{40}$

Figure 2 Illustrative Feedback for Options B and C Generated by ChatGPT, Copilot, and Gemini 2.5 Pro for the Same Assessment Item

Option B: 64/80

ChatGPT

B: Incorrect — This result comes from converting both fractions to have a denominator of 80 but incorrectly adding

or simplifying afterward.

Copilot

B: Incorrect. While 64 and 80 might seem to reflect some conversion effort, this fraction doesn't simplify to a logical result of adding $5/8$ and $3/10$. Check the equivalent fractions carefully.

Gemini 2.5 Pro

B: Incorrect. This answer may have come from multiplying the numerator of the first fraction by the denominator of the second (5×10) and the numerator of the second by the denominator of the first (3×8) and then adding those results together, which is not the standard way to add fractions. The key is to adjust the fractions to have the same denominator first.

Hint: Remember that to add fractions, we need a "common" denominator.

Option C: 50/40

ChatGPT

C: Incorrect — This error likely results from an incorrect conversion. The numerator is too large, and the fraction is greater than 1, which does not match the expected total. It may be from multiplying across rather than adding.

Copilot

C: Incorrect. This improper fraction has a numerator greater than the denominator, suggesting the values might have been multiplied rather than added. Remember, when adding fractions, we align denominators—not multiply across.

Gemini 2.5 Pro

C: Incorrect. This option is very close to a correct step, but it represents only the first part of the addition after finding a common denominator. It seems you correctly converted $5/8$ to $25/40$, but this is just one of the fractions, not the final sum.

Hint: You've done a great job with the first fraction! Now, what does the second fraction become with the new denominator, and what is their sum?

Note. For option A ($8/18$), all three platforms identified direct addition of the numerators and denominators as the likely error and recommended using a common denominator. For option D ($37/40$), all three correctly identified the keyed response and described the common-denominator procedure.

The explanations for options B and C revealed greater uncertainty in the models' inferred error pathways. For option B ($64/80$), ChatGPT attributed the response to an error after converting to a denominator of 80, but the explanation did not provide a clear calculation path leading to $64/80$. Copilot was more cautious, referring generally to "some conversion effort," although its recommendation to "check the equivalent fractions carefully" provided limited corrective specificity. Gemini offered a more detailed explanation, suggesting that the learner multiplied 5×10 and 3×8 and then added the products. However, these calculations produce 50 and 24, which sum to 74 rather than 64. Thus, greater explanatory detail did not necessarily correspond to a more mathematically plausible interpretation of the distractor.

A similar issue was evident for option C ($50/40$). Gemini suggested that the learner had correctly converted $5/8$ to $25/40$ but had stopped after completing only the first part of the addition. Because the selected response was $50/40$ rather than $25/40$, this explanation did not clearly account for the distractor. ChatGPT and Copilot also proposed possible explanations for the response, but the available item-development information did not allow verification of these interpretations against the intended distractor rationale or an actual learner's reasoning.

Overall, the example showed that feedback could exhibit desirable characteristics such as supportive language, response-specific wording, and corrective guidance while still providing an uncertain or mathematically unsupported explanation of the error associated with a distractor. Accordingly, we treated model-generated explanations of learner errors as plausible inferences rather than verified misconceptions.

DISCUSSION

We examined the development of a structured prompting approach for generating response-specific formative feedback and compared feedback from three publicly accessible conversational AI platforms. Overall, the findings suggest that the evaluated platforms could produce candidate feedback that exhibited several characteristics identified in the feedback literature as desirable, including specificity, corrective guidance, personalization, constructiveness, and usability. At the same time, the findings revealed important differences in how these characteristics were expressed and, more importantly, showed that feedback that appeared detailed, supportive, and response-specific was not necessarily based on a mathematically plausible interpretation of the selected distractor. The results therefore support a cautious view of LLM-generated assessment feedback: these systems may be useful for generating candidate feedback for expert review.

The prompt-development phase demonstrated the importance of specifying the desired form and purpose of feedback. Minimally structured prompts sometimes yielded generic responses that offered little explanation or corrective direction, whereas increasingly structured instructions elicited more option-specific and actionable feedback. However, prompts requesting extensive stepwise explanations could also produce excessive detail or disclose more information than the intended hint-based approach required. The final prompt therefore sought to balance mathematical explanation, corrective guidance, supportive language, and conciseness. This pattern is consistent with Brown et al. (2020), who demonstrated that model performance can vary across zero-shot and few-shot prompting conditions and that providing task demonstrations can influence model responses. More broadly, prompt engineering involves deliberately structuring instructions and contextual information to guide model behavior (Federiakin et al., 2024). The present findings extend these observations to assessment feedback by suggesting that prompt refinement involves not only eliciting more informative responses but also managing their specificity, structure, and level of detail.

The comparative review also noted differences in feedback characteristics across the three platforms. ChatGPT was characterized as High across the five target criteria in the team's qualitative consensus review, though some explanations of inferred error pathways were uncertain. Gemini 2.5 Pro frequently generated detailed, individualized, and constructive feedback, but excessive detail sometimes reduced usability and did not necessarily improve the mathematical plausibility of its explanations. Copilot generally produced constructive and understandable responses but showed greater limitations in specificity, corrective guidance, and personalization. These findings should not be interpreted as a formal ranking of model performance. The interpretive levels used in this study reflected consensus-based qualitative judgments rather than independent numerical ratings, and the study was not designed to establish statistically reliable differences among platforms.

A particularly important observation concerned the distinction between desirable feedback characteristics and the validity of a model's inferred explanation for an incorrect response. The illustrative item showed that the platforms could sometimes identify a plausible error pathway. For example, associating the distractor 8/18 with directly adding the numerators and denominators was mathematically consistent with the response. In contrast, several explanations for the distractors 64/80 and 50/40 did not identify a clear mathematical pathway that produced the selected answer. In some cases, a more detailed explanation appeared pedagogically convincing, yet described calculations that did not actually account for the distractor. This finding matters because fluency, specificity, and supportive language can make an explanation seem credible even when the inferred reasoning is uncertain.

Accordingly, mathematical correctness and misconception inference should be treated as related yet distinct aspects of LLM-generated assessment feedback. A multiple-choice response provides evidence for the selected option but, by itself, does not establish the cognitive process that led to that selection. Without independently established distractor rationales or other evidence of learner reasoning, an LLM-generated account of why an option was selected remains an inference rather than a diagnosis. This distinction matters for personalization:

feedback may be personalized to an answer option without necessarily reflecting the learner's actual reasoning. The present findings therefore highlight the need for caution in claims that LLMs can automatically identify learner misconceptions from selected responses.

This distinction is consistent with broader work on formative feedback. Effective feedback should provide information about current performance and useful next steps (Hattie & Timperley, 2007), and formative feedback is generally most useful when it is specific, supportive, and responsive to learner performance (Shute, 2008). Research on adaptive feedback similarly emphasizes aligning feedback with relevant learner and task characteristics (Narciss et al., 2014). The present study suggests that LLMs can generate text with many of these desirable characteristics. Related research has also distinguished the content characteristics of AI-generated feedback from its reliability; for example, Dai et al. (2024) evaluated both effective feedback components and the correctness of assessment judgments when comparing GPT-generated and human feedback. The present results reinforce the importance of evaluating these dimensions separately.

These findings suggest a potentially useful but limited role for LLMs in developing assessment feedback. Rather than treating generated feedback as ready for direct delivery to learners, assessment developers could use LLMs to produce initial candidate explanations and hints for review by content and assessment experts. This human-in-the-loop approach would let developers leverage the speed and flexibility of LLM generation while retaining human responsibility for mathematical accuracy, appropriateness, and alignment with the intended interpretation of an assessment item (Bulut et al., 2024; Dai et al., 2024).

The results also suggest that expert review should go beyond checking whether feedback is grammatically clear or supportive in tone. Reviewers should verify that the proposed mathematical pathway can plausibly produce the selected distractor, that corrective guidance is specific and mathematically sound, and that the amount of information is appropriate for the intended feedback design (Dai et al., 2024). When known distractor rationales are available, LLM-generated explanations should be compared directly with those rationales rather than accepted merely because they sound plausible.

Limitations and Future Research

This study has limitations that should be considered when interpreting these findings. First, the evaluation was qualitative and consensus-based. Therefore, the Low, Limited, Adequate, and High levels provide transparent descriptors of the team's judgment. This approach was appropriate for this study, but future research could develop and validate a formal scoring rubric and use independent raters where feasible. Second, the study used publicly accessible conversational interfaces rather than fixed API implementations. Commercial LLM platforms are also updated frequently, which limits reproducibility over time. The output observed during the study may not be fully reproducible with current versions of the same platforms. Future studies seeking direct model comparisons should use fixed model versions where possible. Third, the assessment content was limited to multiple-choice adult numeracy items involving fractions and probability. The findings may not generalize to other mathematical domains, literacy assessments, open-ended responses, different age groups, or higher-stakes assessment contexts. In addition, no learners participated in the study. Future research could consider other assessment content and populations and evaluate whether expert-reviewed LLM-generated feedback benefits learners.

CONCLUSION

The evaluated conversational AI platforms varied in their capacity to generate candidate feedback that exhibited characteristics identified as desirable in the formative feedback literature. Structured prompting elicited more response-specific, corrective, and supportive feedback, but the results also showed that pedagogically appealing explanations may include uncertain or mathematically unsupported interpretations of distractor reasoning. These findings support further investigation of LLM-generated feedback under expert review. For assessment applications, the most defensible near-term role of LLMs may therefore be to generate candidate feedback within carefully reviewed human-in-the-loop workflows rather than as autonomous sources of learner-facing feedback.

REFERENCES

1. Adult Skills Assessment Program. (n.d.). What is ASAP? University of Massachusetts Amherst. <https://websites.umass.edu/asap/what-is-the-asap>
2. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., . . . Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.
3. Bulut, O., Beiting-Parrish, M., Casabianca, J. M., Slater, S. C., Jiao, H., Song, D., Ormerod, C. M., Fabiyi, D. G., Ivan, R., Walsh, C., Rios, O., Wilson, J., Yildirim-Erbasli, S. N., Wongvorachan, T., Liu, J. X., Tan, B., & Morilova, P. (2024). The rise of artificial intelligence in educational measurement: Opportunities and ethical challenges. *Chinese/English Journal of Educational Measurement and Evaluation*, 5(3), Article 3. <https://www.ce-jeme.org/journal/vol5/iss3/3/?utm>
4. Butler, A. C. (2018). Multiple-choice testing in education: Are the best practices for assessment also good for learning? *Journal of Applied Research in Memory and Cognition*, 7(3), 323–331.
5. Dai, W., Tsai, Y. S., Lin, J., Aldino, A., Jin, H., Li, T., Gašević, D., & Chen, G. (2024). Assessing the proficiency of large language models in automatic feedback generation: An evaluation study. *Computers and Education: Artificial Intelligence*, 7, Article 100299. <https://doi.org/10.1016/j.caeai.2024.100299>
6. Federiakin, D., Molerov, D., Zlatkin-Troitschanskaia, O., & Maur, A. (2024). Prompt engineering as a new 21st-century skill. *Frontiers in Education*, 9, Article 1366434.
7. Glassey, R., & Bälter, O. (2020). Put the students to work: Generating questions with constructive feedback. In *2020 IEEE Frontiers in Education Conference (FIE)* (pp. 1–8). IEEE.
8. Glassey, R., & Bälter, O. (2022). Quality agreement on learner-sourced multiple-choice questions. In *Proceedings of the Workshop on Learnersourcing: Student-Generated Content @ Scale 2022* (pp. 45–49). CEUR-WS.
9. Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81–112. <https://doi.org/10.3102/003465430298487>
10. Leibold, N., & Schwarz, L. M. (2015). The art of giving online feedback. *The Journal of Effective Teaching*, 15(1), 34–46. <https://files.eric.ed.gov/fulltext/EJ1060438.pdf>
11. Narciss, S., Sosnovsky, S., Schnaubert, L., Andrès, E., Eichelmann, A., Gogvadze, G., & Melis, E. (2014). Exploring feedback and student characteristics relevant for personalizing feedback strategies. *Computers & Education*, 71, 56–76. [10.1016/j.compedu.2013.09.011](https://doi.org/10.1016/j.compedu.2013.09.011)
12. Schartel, S. A. (2012). Giving feedback—An integral part of education. *Best Practice & Research Clinical Anaesthesiology*, 26(1), 77–87. [10.1016/j.bpa.2012.02.003](https://doi.org/10.1016/j.bpa.2012.02.003)
13. Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, 78(1), 153–189. DOI: 10.3102/0034654307313795
14. Sireci, S. G., Suárez-Álvarez, J., Zenisky, A. L., & Oliveri, M. E. (2024). Evolving educational testing to meet students' needs: Design-in-real-time assessment. *Educational Measurement: Issues and Practice*, 43(4), 112–118.
15. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35, 24824–24837.

APPENDIX

Appendix 1. Final Standardized Prompt

The following prompt was used for all three LLMs in the holdout comparison:

“Consider yourself a numeracy expert. Each question includes multiple answer options labeled A:, B:, etc. Your task is to evaluate each option using precise mathematical reasoning. Use a positive tone and include a brief, supportive hint that encourages reflection.

For incorrect answers, explain one likely error without revealing the correct answer. Each explanation should:

Target a specific error or action

Contrast what was done vs. what should be done

Be easy for an adult learner to understand and offer a fix

Be supportive and individualized per option

Format responses as follows, with no additional commentary:

A: [Correct or incorrect] explanation why

B: ... and so on.”