From Proprietary to Open-Source: A Comparative Evaluation of LLMs for Automatic Feedback Generation

Elisabetta Mazzullo[1,*], Okan Bulut[2]

[1]Measurement, Evaluation, and Data Science, Faculty of Education, University of Alberta, Edmonton, AB T6G 2G5, Canada
[2]Centre for Research in Applied Measurement and Evaluation, Faculty of Education, University of Alberta, Edmonton, AB T6G 2G5, Canada

Abstract

This study contributes to theory-grounded evaluation of large language models (LLMs) by examining the effectiveness of four different models in generating automatic feedback for open-ended written responses, a complex task generally requiring human judgment. A state-of-the-art GPT model and the open-source Llama model were used, in their base form and in a fine-tuned version, to provide feedback on responses to a high-stakes situational judgment test. Model outputs were evaluated using a rubric grounded in literature recommendations for effective feedback and compared using automatic metrics. We analyzed model behaviors and error patterns before and after fine-tuning to understand how model configuration affects feedback quality. Results reveal systematic differences between GPT and Llama models, as well as between prompted and fine-tuned variants. Fine-tuning did not consistently outperform prompt engineering, suggesting that high-quality prompting may sufficiently approximate aspects of expert judgment. The strengths and limitations of our evaluation approach are discussed, and implications for research and practice are presented.

Keywords

LLMs, Automatic Feedback, Educational AI, Educational Technology, LLM evaluation

1. Introduction

Large language models (LLMs) are increasingly being adopted by professionals across various sectors, including education, to enhance the efficiency of routine tasks [1]. Among the more time-intensive responsibilities educators face is providing feedback on students' open-ended responses–a task for which artificial intelligence (AI) offers significant potential. Although feedback is widely acknowledged as a critical component of effective student learning, teachers often contend with substantial workloads and constrained resources [2], making it difficult to deliver individualized feedback to every student.

Several studies have explored the application of LLMs for automatic feedback generation (AFG), obtaining encouraging results. However, the vast majority of research to date uses proprietary LLMs, particularly OpenAI's ChatGPT, while overlooking open-source alternatives that might better suit schools' and educational companies' budgetary constraints and privacy requirements. Moreover, most studies used only one model and relied on careful prompting of LLMs, but did not explore the possibility of fine-tuning to align the models with user needs. Lastly, evaluating the feedback generated by LLMs is complex, and existing studies lack a unified measurement approach.

To address the aforementioned gaps, this study investigates the following research questions: (1) To what extent can LLMs generate high-quality feedback on short open-ended questions? (2) How does the performance of an open-source LLM compare to that of a state-of-the-art GPT model in generating such feedback? (3) Does fine-tuning enhance the ability of LLMs to produce effective feedback on short open-ended questions?

To evaluate the ability of both open-source and proprietary LLMs, both in their base and fine-tuned versions, to generate automatic feedback on short open-ended responses to complex items, we use

LLM Psychometrics: Applying Assessment Theories to Evaluate LLMs and LLM-Supported Learners, LAK 2026 Workshop
*Corresponding author.
mazzullo@ualberta.ca (E. Mazzullo); bulut@ualberta.ca (O. Bulut)
https://www.okanbulut.com/ (O. Bulut)
0009-0008-4847-9934 (E. Mazzullo); 0000-0001-5853-1267 (O. Bulut)
© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-11-15
CEUR-WS.org/Vol-4330/paper1.pdf

a rubric grounded in literature-based criteria for effective feedback. In addition, we compare model outputs using automated similarity metrics, both between models and against established evaluation guidelines.

2. Literature Review

Feedback improves students' performance, motivation [3], and learning strategies [4]. Effective feedback is personalized to a student's performance, delivered timely, and it not only points out mistakes, but also highlights strengths and offers actionable next steps for improvement [5, 6, 7]. To help students accept and process the feedback, these messages should talk to students directly [8], acknowledge their effort, and strike a balance between the positive and negative aspects of their performance [9].

Writing personalized, detailed, and thoughtful feedback is time-consuming for educators [10], and unfeasible in large-scale contexts. Existing systems for automating feedback provision struggle to incorporate expert knowledge and data-driven methods, thus failing to create feedback that is highly tailored to a task, to learners' individual responses, and to teacher preferences [11]. Thanks to their ability to understand context, follow instructions, and generate human-like text, LLMs offer a promising foundation for the next generation of AFG systems. LLMs may provide students with rapid and personalized feedback while keeping teachers in the loop, as they can guide model behavior by providing evaluation criteria and directions in the prompt to align outputs to their preferences and needs.

Recent studies have focused on OpenAI's state-of-the-art ChatGPT and various prompting strategies. Results show that LLMs generate feedback almost as well as expert teachers [12, 13, 14]; this feedback can improve revision performance, increase motivation, and generate more positive emotional responses [15]. However, notable limitations remain. For instance, Jacobsen and Weber [12] found that one of the 20 automatically generated feedback messages evaluated in their study was of poor quality, and in [16] ChatGPT generated feedback containing misclassifications, redundancy, inconsistency, or inaccurate explanations in 48% of cases.

Existing studies focused almost exclusively on proprietary GPT models and relied heavily on prompt engineering for AFG. However, smaller open-source models could offer cost savings and higher privacy for researchers and educational organizations [17]. Also, fine-tuning could enable LLMs to generate feedback that is better aligned with instructors' pedagogical preferences, regardless of their prompt-engineering skills [17]. To date, little is known about whether open-source LLMs and lightweight fine-tuning can improve the quality of generated feedback.

To address these gaps, we used OpenAI's GPT-4o-mini and Meta's open-source Llama-3.2-11B to generate feedback for open-ended responses to Casper, a high-stakes situational judgment test (SJT). We prompted the models in their base versions; we also used a small dataset of human-written feedback and performed parameter-efficient fine-tuning (PEFT) to further train the models on the AFG task. The outputs generated by the four models were compared using a rubric and automatic metrics. Results were used to explore how different model configurations influence model behaviors in AFG, and to draw implications for future research and educational practice.

3. Methods

3.1. Data

To generate feedback, train the models, and evaluate performance, we used data from the Casper test [18], an SJT that measures skills related to social intelligence and professionalism. The dataset included test items, test-takers' responses, scores, and evaluation guidelines. All these elements were used to prompt and fine-tune the LLMs for AFG. Casper requires test-takers to type their responses to three open-ended questions about an ethical or professional challenge. Due to the open-ended nature of the questions, no single correct solution exists; responses can be very diverse, and evaluation guidelines

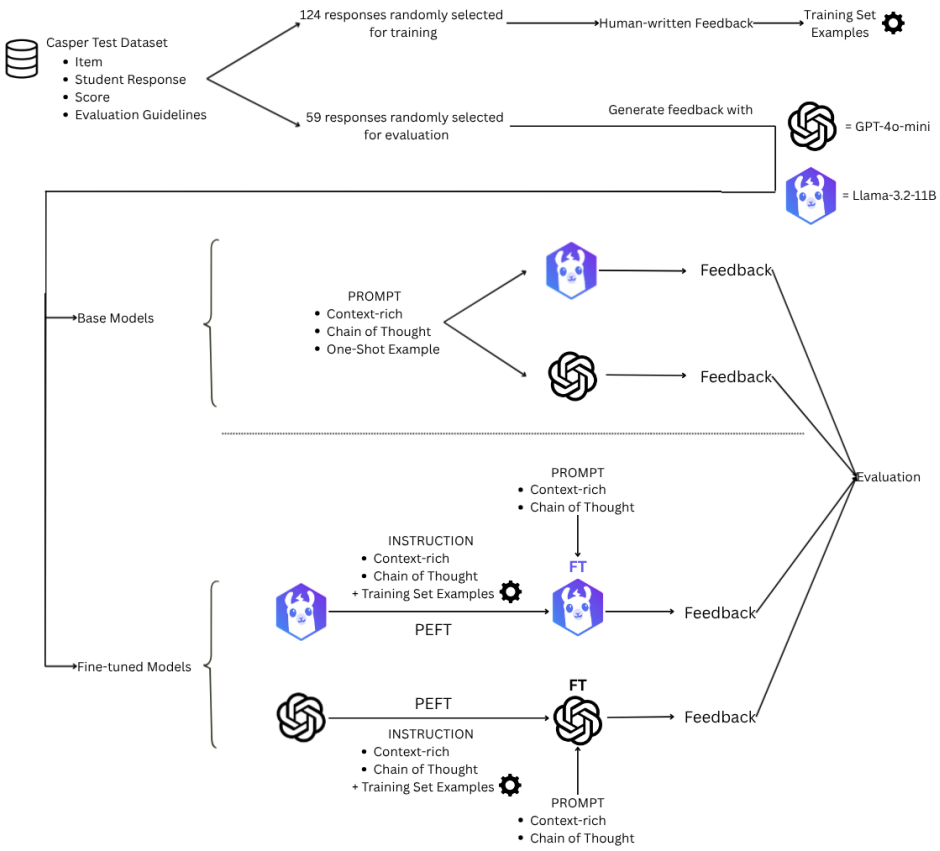


Figure 1: Overview of the Study Design

are not highly specific. These characteristics make existing AFG systems unsuitable for providing personalized feedback in this context, making this an ideal task to explore with LLMs.

Figure 1 summarizes the overall study design. To keep the workshop's focus on the theme, we will evaluate the outputs and compare the models used in our study. Further details about model selection, prompts, and training strategies can be found in the supplementary materials.

3.2. Feedback Generation

We selected OpenAI's GPT-4o-mini as the proprietary model for our study and Meta's Llama-3.2-11B as the open-source LLM. These models are built on a similar architecture and have a comparable size [19, 20]. From the Casper dataset, we randomly selected 124 responses for fine-tuning and 59 for evaluation. We ensured that, in each sample, responses and questions offered a broad representation of the skills assessed and the performance levels. We crafted a detailed, context-rich prompt that leverages the one-shot and chain-of-thought (CoT) prompting paradigms [21]. First, this instruction was used to prompt GPT and Llama in their base versions to generate feedback for each of the 59 responses held out for evaluation. Later, GPT and Llama were fine-tuned using human-written feedback on the 124 responses selected for training. The fine-tuned models were prompted to generate feedback for the same 59 responses.

3.3. LLM Evaluation and Comparison

Most automatic metrics for text evaluation are not suitable when tasks permit significant diversity, or cannot be computed without a reference text [22]. Thus, in the context of feedback on Casper responses and, often, in the broader context of open-ended questions, human evaluation remains the gold standard

Table 1
Percentage of Agreement Between Independent Raters Evaluating 59 Outputs for Each Model

	GPT-4o-mini	Llama-3.2-11B	Ft:GPT	Ft:Llama	Avg.
Linguistic Quality	100	84.7	96.6	98.3	94.9
Factuality	100	93.2	96.6	94.8	96.2
Personalization	93.2	96.6	98.3	87.9	94
Actionability	100	96.6	98.3	96.6	97.9
Affective tone	98.3	76.9	96.6	62	83.5
Second Persons	100	100	100	100	100

Note. Ft stands for fine-tuned.
Avg stands for average.

[23].

To evaluate the quality of feedback generated by our models and compare performance across LLMs, we developed a rubric and used automatic evaluation metrics. The rubric captures qualities that require human judgment, while the automatic metrics provide scalable, fine-grained indicators of textual similarity and structure that complement expert evaluation when human scoring is limited or unavailable.

3.3.1. Rubric

Based on literature recommendations for effective feedback, we identified six structure-related criteria: linguistic quality, factuality, personalization, actionability, affective tone, and use of the second person. Each feedback output was evaluated by two independent judges using these criteria, conceptualized into a scoring rubric.

Before starting the evaluation, the judges received rater training. Inter-rater reliability was generally high across all models and evaluation criteria, with percentage agreement ranging from 62% to 100% (Table 1). Percentage agreement was reported instead of Cohen's kappa because, in instances of very high inter-rater consistency, limited variability in ratings led to undefined Kappa values, a known limitation of the statistic under near-perfect agreement conditions [24]. Discrepancies were discussed and resolved prior to analysis.

3.3.2. Automatic Evaluation Metrics

After text preprocessing, the following automatic evaluation metrics were computed in Python to compare model outputs.

Polarity. Average polarity scores for each model were computed using lemmatized text processed with the TextBlob package. Polarity scores range from -1 to 1, with -1 denoting highly negative feedback, 0 indicating neutrality, and 1 representing highly positive feedback. In addition, we examined the correlation between polarity and item scores to assess whether the affective tone of the feedback varies in relation to test-taker performance.

Semantic Similarity. A similar message can be conveyed using different words and sentence structures. To analyze semantic similarity, we first created text embeddings using SentenceBERT [25]; then, we measured the similarity between these embeddings using the cosine metric. Cosine similarity ranges from -1 to 1: 1 indicates identical meaning, 0 indicates no similarity, and -1 indicates complete opposition. We computed the similarity between (1) feedback generated by each model, (2) automatic feedback and test-taker responses, and (3) automatic feedback and evaluation guidelines. These results inform us about (1) whether different models offer feedback with similar meaning, (2) whether the content of the feedback overlaps with that of the response, and (3) whether the feedback shares meaning with the evaluation criteria.

Table 2
Proportion of Feedback Meeting the Qualities of Effective Feedback Considered in the Rubric

		GPT-4o-mini	Llama-3.2-11B	Ft:GPT-4o-mini	Ft:Llama-3.2-11B*
Linguistic Quality	Correct	100	100	79.7	98.3**
	Spelling	0	0	15.3	0
	Syntax	0	0	3.2	0
	Semantic	0	0	1.7	0
Factuality	Pertinent	100	96.6	94.9	98.3
	Response	0	1.7	5.1	0
	Scenario	0	1.7	0	0
Personalization	Personalized	96.6	100	98.3	86.4
	Generic	3.4	0	1.7	11.9
Actionability	Actionable	89.8	89.8	86.4	86.4
	Not actionable	0	0	3.4	1.7
	9 N/A	5.1	1.7	10.2	5.1
	9 actionable	5.1	8.5	0	5.1
Affective tone	Balanced	1.7	30.5	30.5	33.9
	Suggestions	88.1	66.1	54.2	49.2
	Flaws	0	0	5.1	5.1
	9 not applicable	5.1	1.7	10.2	5.1
	9 suggestions	5.1	1.7	0	5.1
Second Person	Yes	100	52.5 + 39***	98.3	96.6
	No	0	8.5	1.7	1.7

Note. Ft: stands for fine-tuned * percentages do not add up to one because one output (1.7%) was the repetition of part of the prompt and could not be evaluated to fit into any of the rubric categories. ** one output (1.7%) was linguistically correct but repeated the same sentence twice. *** 39% of outputs provide second-person feedback, but only after analyzing the response in the third-person.

4. Results

4.1. Rubric-Based Evaluation

Results from the rubric evaluation are summarized in Table 2. Both GPT and Llama, in their base versions, consistently generate linguistically correct feedback. However, after fine-tuning, GPT's feedback begins to display linguistic errors. All models generally provide feedback that is pertinent and personalized. However, in a few cases, base Llama and the fine-tuned GPT generate feedback with factual mistakes; fine-tuned Llama presents the highest rate of feedback that does not reference specific aspects of a test-taker's response.

Our prompt was effective at directing the base GPT model to generate second-person feedback. After fine-tuning GPT, this does not happen in only one instance. On the other hand, prompting base Llama with the same instruction led to second-person feedback in just over half of the generations. 8.5% of outputs were in the third person, and in 39% of cases, the model analyzed the response in the third person before giving feedback in the second person. Fine-tuning successfully taught Llama to address the test-taker in the second person, with only one generation that failed to meet this criterion.

GPT and Llama, in their base versions, consistently offer actionable feedback. However, after fine-tuning, they fail to do so in two cases and one case, respectively. Interestingly, the two models display distinct patterns in how they direct test-takers toward improvement. The base GPT maintains a very supportive tone, often providing suggestions without explicitly pointing out flaws; the base Llama was better at highlighting both positive and negative aspects in its feedback messages. After fine-tuning, GPT learns to write more balanced feedback. In 5% of cases, the fine-tuned models generate feedback that only points out flaws.

Lastly, quite often, the foundational LLMs provided guidance for improving responses that received a perfect score. In the training data, feedback to responses with a score of 9 did not include actionable suggestions; after fine-tuning, GPT seems to learn more effectively than Llama that perfect responses

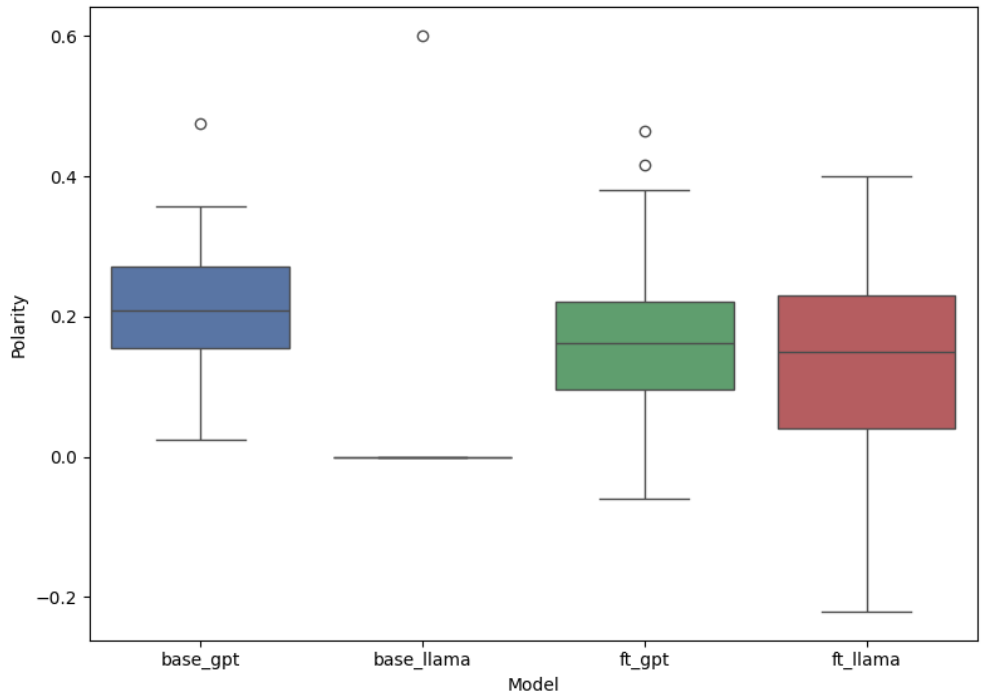


Figure 2: Boxplot of Polarity Scores of Feedback Generated by Each LLM

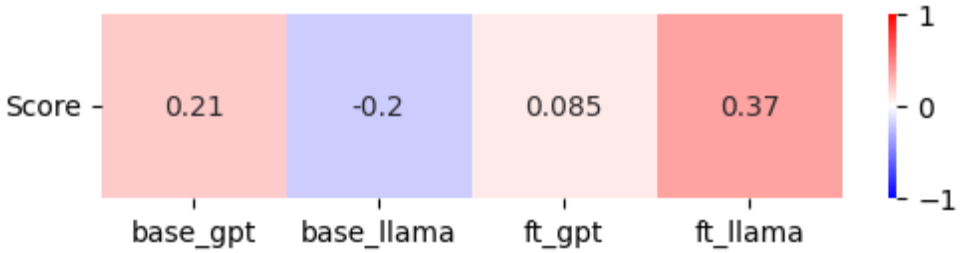


Figure 3: Heatmap of the Correlations between Response Scores and Polarity Scores of the Associated Feedback

do not require suggestions.

4.2. Automatic Metrics Evaluation

Polarity. Similar to what was observed from the rubric evaluation, on average, Llama maintains the most neutral tone in AFG, while base GPT tends to be slightly more positive. Fine-tuning increases the variability in the polarity of Llama-generated feedback, and the average polarity scores for the fine-tuned models become more similar. Only after fine-tuning, the models generate some feedback with negative polarity (Figure 2).

Correlations between response scores and the polarity of their feedback vary considerably between the four models (Figure 3). The base GPT tends to write slightly more positive feedback to responses that achieve higher scores, while the base Llama, surprisingly, shows an opposite tendency. Fine-tuning had a largely different effect on the two models: it attenuates GPT's tendency, and it inverts it drastically in Llama, as the correlation between polarity and scores becomes moderate and positive.

Semantic Similarity between Models' Outputs. The feedback messages written by the four models to the same response are, on average, moderately similar to each other in their content (Figure 4). The base Llama generates feedback semantically very different from all other models, but after fine-tuning, the feedback becomes moderately similar to that produced by the base and fine-tuned GPT models.

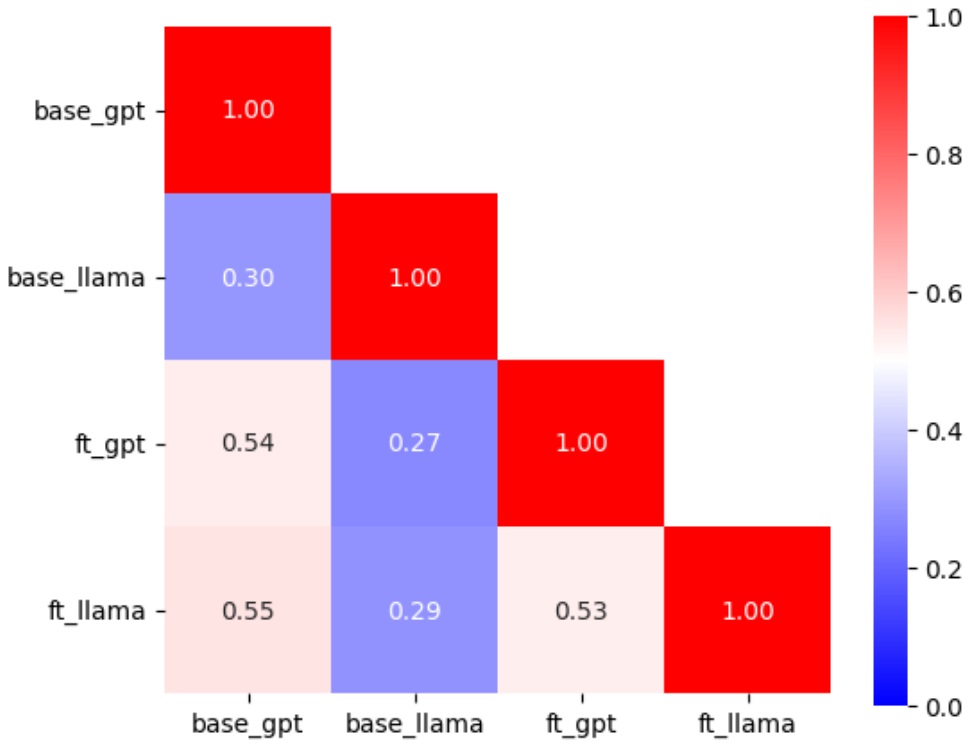


Figure 4: Heatmap of the Average Semantic Similarity between Feedback Generated by Each LLM

Table 3
Average Semantic Similarity of the Automatic Feedback with the Evaluation Guidelines and the Response

Model	Evaluation Guidelines	Response
GPT-4o-mini	0.52	0.38
Llama-3.2-11B	0.20	0.16
Ft: GPT-4o-mini	0.50	0.38
Ft: Llama-3.2-11B	0.52	0.38

Note. Ft: stands for fine-tuned.

Semantic Similarity between Outputs and Evaluation Guidelines. Table 3 shows the average semantic similarity between the feedback and the evaluation guidelines for the item on which the model generates feedback. Llama, in its base version, produces feedback with the lowest content similarity to these guidelines. After fine-tuning, the model's feedback shows a moderate average semantic similarity to these guidelines, comparable to that of GPT models.

Semantic Similarity between Outputs and Test Responses. Similarly, all models except the base Llama generate outputs with low to moderate semantic similarity to the responses they provide feedback on (Table 3). Without fine-tuning, Llama shows extremely low semantic similarity with test-takers' responses.

5. Discussion

Overall, our evaluation suggests that both GPT and Llama typically generate feedback that meets desired qualities; however, sporadic but significant errors occur. We did not find an advantage in fine-tuning over carefully prompting LLMs for AFG. Insights from the rubric-based evaluation and the automatic metrics were complementary, capturing different dimensions of feedback quality. While the rubric reflected structure-related criteria aligned with qualities of effective feedback, automatic metrics offered additional information about semantic and content-related aspects of model outputs. These results evidenced meaningful differences between the GPT and Llama models in both the quality and style of

the feedback they produced, as well as in how their outputs changed after fine-tuning.

5.1. Patterns in LLM Performance

Both the rubric and automatic metrics suggested that the base version of GPT tends to offer the most positive feedback. It often encourages test-takers with statements such as "you are on the right track" or "keep up the good work"; however, it does so regardless of the actual quality of performance as indicated by the scores. GPT's encouraging tone may motivate test-takers, but its tendency to praise even weak responses threatens the clarity and usefulness of the feedback. In contrast, Llama maintains a more balanced tone, often analyzing the response in the third person before generating second-person feedback. This analytical style could foster deeper reflection, but it may also overwhelm some test-takers.

The affective tone of the message influences the effectiveness of feedback [26]. Therefore, when choosing to use LLMs in their out-of-the-box versions for AFG, educators should consider these tendencies in tone and style and create prompts that counteract them to further align outputs with their desired educational tone and purpose.

After fine-tuning, both GPT and Llama produced feedback that more closely matched the linguistic style and structural conventions of the training examples, indicating that PEFT can successfully shift model behavior toward a target feedback genre. Fine-tuning increased semantic convergence between the two models and notably improved alignment between Llama's outputs and the evaluation guidelines. This suggests that fine-tuning helped the model generate feedback that better adhered to evaluation criteria.

However, our results also revealed drawbacks to fine-tuning. Despite exposure to identical training data and instructions, GPT and Llama did not respond uniformly. GPT exhibited typos and grammatical errors, while Llama sometimes failed to tailor feedback to the content of the response. These deviations suggest that PEFT is effective in increasing style alignment, but may reduce content accuracy and relevance. Therefore, fine-tuning does not uniformly improve all dimensions of feedback quality and may even introduce new risks. The divergence between the two models also underscores that architecture-specific biases shape how different LLMs respond to fine-tuning.

5.2. Practical Implications

Similar to previous studies, we found that LLMs are "useful but fallible" [27, p. 2] in generating feedback on open-ended responses. Occasional linguistic errors and misinterpretations of test items or responses can undermine the pedagogical value of feedback, particularly for young learners or language learners. Therefore, at this stage, human oversight remains essential to ensure the pedagogical quality of automated feedback. Learners should also be encouraged and guided to engage critically with AI-generated feedback, while retaining access to human instructors who can clarify or contextualize the information provided by the model.

Although fine-tuning did not lead to a measurable improvement in feedback quality, effective prompting remains critical for achieving high-quality outputs. Crafting detailed and well-structured prompts requires a level of expertise that many educators may not yet possess. To promote equitable and responsible use of LLMs in assessment, educators need training and institutional support to develop prompt design skills and to integrate AI feedback tools in ways that enhance, rather than replace, human pedagogical judgment.

6. Limitations

The following limitations restrict the generalizability of these findings. First, the fine-tuning dataset was small and not developed by expert educators. Therefore, it is possible that (1) the training data was too small for the models to learn beyond surface-level characteristics of style, (2) the examples did

not provide a sufficiently good representation of the desired qualities, and (3) individual biases in how feedback was phrased and prioritized are present in the data.

Second, our rubric-based evaluation focused primarily on structure-related criteria and did not assess the completeness, focus, or pedagogical usefulness of the feedback. This limitation reflects our lack of access to test experts who could reliably judge these deeper qualitative dimensions. Although automatic metrics provided an initial, objective indication of how closely the model outputs aligned with the evaluation criteria, their interpretability is limited, and semantic similarity alone does not ensure that the feedback is accurate or educationally useful.

Therefore, future studies should draw on larger, expert-developed training datasets for fine-tuning. Moreover, incorporating evaluations from test experts and learners in authentic instructional settings is essential to determine whether the feedback aligns with intended criteria and supports students' understanding and performance.

7. Conclusion

This work highlights the importance of theory-informed approaches in evaluating LLM behavior in the context of AFG. Rubric-based evaluation revealed that LLMs can generate feedback that reflects many qualities associated with effective instructional support, and the open-source Llama model achieved these qualities at rates comparable to the state-of-the-art GPT model. Automatic metrics further showed that the two models exhibit distinct communication structures and stylistic tendencies. While PEFT helped models match the training examples, it did not yield additional measurable improvements in feedback quality based on the criteria used in our evaluation.

Significant differences were observed in how GPT and Llama behave when prompted to generate feedback. These insights can help educators anticipate the tendencies of different models and make informed decisions about which system better aligns with their instructional goals, or about how to adapt prompting strategies to override undesired behaviors. In addition, we demonstrate that rubric-based evaluations can be effectively complemented by automatic metrics, which provide useful information about content-level characteristics when expert, content-specific evaluation is unavailable.

As AI-driven feedback tools become increasingly integrated into instructional practice, responsible human–AI collaboration will require thoughtful model selection, human participation in prompt design, and the development of evaluation approaches to ensure that feedback aligns with expert pedagogical judgment and genuinely supports personalized learning and growth.

Acknowledgments

Thank you to Acuity Insights for sharing the dataset used in this study.

Declaration on Generative AI

During the preparation of this work, the author(s) used Chat-GPT-4 in order to: Grammar and spelling check. After using these tools/services, the authors reviewed and edited the content as needed and thus take full responsibility for the publication's content.

References

[1] J. A. Delello, W. Sung, K. Mokhtari, J. Hebert, A. Bronson, T. De Giuseppe, Ai in the classroom: Insights from educators on usage, challenges, and mental health, Education Sciences 15 (2025). URL: https://www.mdpi.com/2227-7102/15/2/113. doi:10.3390/educsci15020113.

[2] P. D. Jomuad, L. M. M. Antiquina, E. U. Cericos, J. A. Bacus, J. H. Vallejo, B. B. Dionio, J. S. Bazar, J. V. Cocolan, A. S. Clarin, Teachers' workload in relation to burnout and work performance,

International journal of educational policy research and review 8 (2021) 48–53. doi:https://doi.org/10.15739/IJEPRR.21.007.

[3] A. C. Koenka, E. M. Anderman, Personalized feedback as a strategy for improving motivation and performance among middle school students, Middle School Journal 50 (2019) 15–22. doi:https://doi.org/10.1080/00940771.2019.1674768.

[4] W. Matcha, D. Gašević, N. A. Uzir, J. Jovanović, A. Pardo, Analytics of learning strategies: Associations with academic performance and feedback, in: Proceedings of the 9th International Conference on Learning Analytics & Knowledge, LAK19, Association for Computing Machinery, New York, NY, USA, 2019, pp. 461–470. doi:https://doi.org/10.1145/3303772.3303787.

[5] J. Hattie, H. Timperley, The power of feedback, Review of Educational Research 77 (2007) 81–112. doi:https://doi.org/10.3102/003465430298487.

[6] R. D. Sadler, Beyond feedback: Developing student capability in complex appraisal, Assessment and Evaluation in Higher Education 35 (2010) 535–550. doi:https://doi.org/10.4324/9781315872322.

[7] Z. V. Zhang, K. Hyland, Student engagement with teacher and automated feedback on l2 writing, Assessing Writing 36 (2018) 90–102. doi:https://doi.org/10.1016/j.asw.2018.02.004.

[8] F. J. Prins, D. M. Sluijsmans, P. A. Kirschner, Feedback for general practitioners in training: Quality, styles, and preferences, Advances in Health Sciences Education 11 (2006) 289–303. doi:https://doi.org/10.1007/s10459-005-3250-z.

[9] J. Hill, K. Berlin, J. Choate, L. Cravens-Brown, L. McKendrick-Calder, S. Smith, Exploring the emotional responses of undergraduate students to assessment feedback: Implications for instructors, Teaching and Learning Inquiry 9 (2021) 294––316. doi:https://doi.org/10.20343/teachlearninqu.9.1.20.

[10] D. Boud, P. Dawson, What feedback literate teachers do: an empirically-derived competency framework, Assessment & Evaluation in Higher Education 48 (2023) 158–171. doi:https://doi.org/10.1080/02602938.2021.1910928.

[11] G. Deeva, D. Bogdanova, E. Serral, M. Snoeck, J. De Weerdt, A review of automated feedback systems for learners: Classification framework, challenges and opportunities, Computers & Education 162 (2021) 104094. doi:https://doi.org/10.1016/j.compedu.2020.104094.

[12] L. J. Jacobsen, K. E. Weber, The promises and pitfalls of chatgpt as a feedback provider in higher education: An exploratory study of prompt engineering and the quality of ai-driven feedback, 2023. doi:https://doi.org/10.31219/osf.io/cr257, preprint.

[13] J. Steiss, T. Tate, S. Graham, J. Cruz, M. Hebert, J. Wang, Y. Moon, W. Tseng, M. Warschauer, C. B. Olson, Comparing the quality of human and chatgpt feedback of students' writing, Learning and Instruction 91 (2024) 101894. doi:https://doi.org/10.1016/j.learninstruc.2024.101894.

[14] E. Fokides, E. Peristeraki, Comparing chatgpt's correction and feedback comments with that of educators in the context of primary students' short essays written in english and greek, Education and Information Technologies 30 (2025) 2577–2621.

[15] J. Meyer, T. Jansen, R. Schiller, L. W. Liebenow, M. Steinbach, A. Horbach, J. Fleckenstein, Using llms to bring evidence-based feedback into the classroom: Ai-generated feedback increases secondary students' text revision, motivation, and positive emotions, Computers and Education: Artificial Intelligence 6 (2024) 100199. doi:https://doi.org/10.1016/j.caeai.2023.100199.

[16] I. Azaiz, N. Kiesler, S. Strickroth, Feedback-generation for programming exercises with GPT-4, in: Proceedings of the 2024 on Innovation and Technology in Computer Science Education V. 1, ACM, Milan, Italy, 2024, pp. 31–37. doi:http://dx.doi.org/10.1145/3649217.3653594.

[17] D. Bergmann, What is instruction tuning, 2024. https://www.ibm.com/topics/instruction-tuning [Accessed: (2024.08.15)].

[18] Acuity Insights, Casper Technical Manual, Acuity Insights, 2024. URL: https://acuityinsights.com/resource/casper-technical-manual/, accessed: 2024-09-02.

[19] OpenAI, Gpt-4o mini: Advancing cost-efficient intelligence, 2024. URL: https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence.

[20] Meta AI, Llama 3.2, https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_2/,

n.d. Accessed: 2025-07-23.
[21] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa, Large language models are zero-shot reasoners, Advances in neural information processing systems 35 (2022) 22199–22213.
[22] A. Celikyilmaz, E. Clark, J. Gao, Evaluation of text generation: A survey, 2021. URL: https://arxiv.org/abs/2006.14799. arXiv:2006.14799.
[23] C. van der Lee, A. Gatt, E. van Miltenburg, S. Wubben, E. Krahmer, Best practices for the human evaluation of automatically generated text, in: K. van Deemter, C. Lin, H. Takamura (Eds.), Proceedings of the 12th International Conference on Natural Language Generation, Association for Computational Linguistics, Tokyo, Japan, 2019, pp. 355–368. doi:https://doi.org/10.18653/v1/W19-8643.
[24] A. R. Feinstein, D. V. Cicchetti, High agreement but low kappa: I. the problems of two paradoxes, Journal of clinical epidemiology 43 (1990) 543–549.
[25] N. Reimers, Sentence-bert: Sentence embeddings using siamese bert-networks, arXiv preprint arXiv:1908.10084 (2019).
[26] N. Storch, G. Wigglesworth, Learners' processing, uptake, and retention of corrective feedback on writing: Case studies, Studies in Second Language Acquisition 32 (2010) 303–334. doi:https://doi.org/10.1017/S0272263109990532.
[27] J. K. Matelsky, F. Parodi, T. Liu, R. D. Lange, K. P. Kording, A large language model-assisted education tool to provide feedback on open-ended responses, 2023. URL: https://arxiv.org/abs/2308.02439. arXiv:2308.02439.